Title
Computational approaches to macromolecular and supramolecular processes in biology

Author
Han, Karen F.

Publication Date
1996

Peer reviewed|Thesis/dissertation

COMPUTATIONAL APPROACHES TO MACROMOLECULAR AND SUPRAMOLECULAR PROCESSES IN BIOLOGY:
FROM PROTEIN FOLDING TO CHROMOSOME FOLDING

by

KAREN F HAN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOPHYSICS

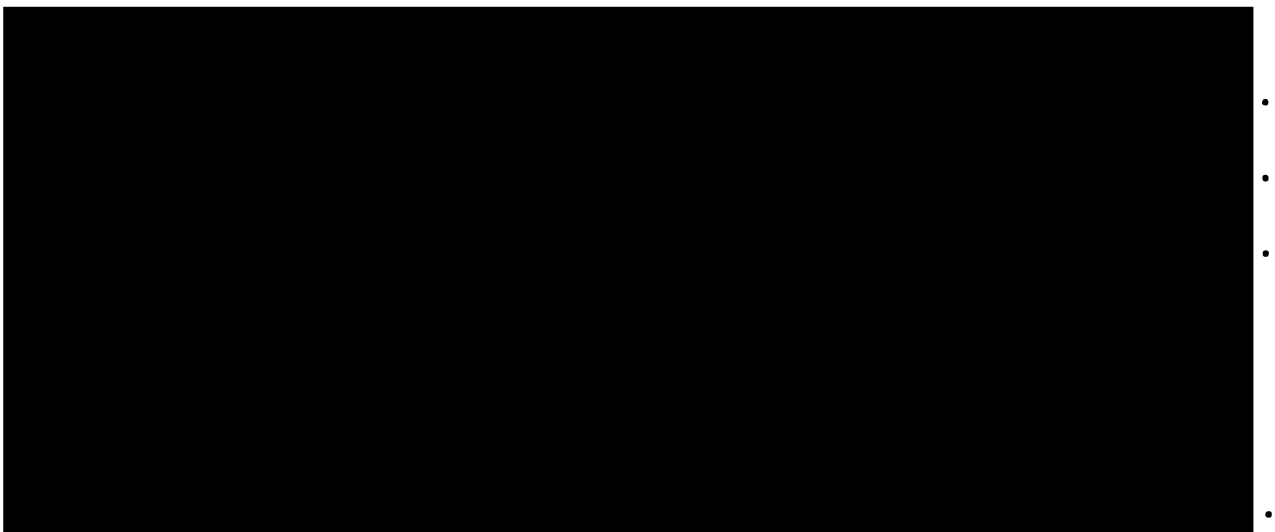
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco



Date **University Librarian**

Degree Conferred: . . . . . .

To my Mother, Sharon (1941-1995),

who has given me life in this world,

feelings to love,

courage to live,

and strength to move on.

Mother, I will always remember you and make you proud!

Preface

This thesis is dedicated to the loving memory of my Mother, Sharon, who died of an aneurysm on November 25, 1995.

My Mother was a traditional Chinese woman who devoted her life to my Father, my Brother, my grandparents, and myself. She ran the family business so that finances are taken care of. She was our family doctor, although she was never formally trained. Most importantly, she was our family's emotional buffer during the most stressful times. She tried to keep all possible worries away from us so we could each focus on our individual goals and careers.

Although my Mother was not a scientist, and she never fully understood my interest in science, she always told me that I should pursue my passions while I have the opportunity. She said that opportunity is only there if you are ready to receive it. I took the opportunities offered to me at UCSF to learn many different aspects of science and problem solving. These are skills that I will take with me as I explore the curiosities and interests I have in many fields. I am lucky to have enjoyed my experience in graduate school studying exactly what I wanted to learn about. This a luxury that is invaluable.

I have now returned to medical school where the daily confrontation with pain and suffering of patients keeps reminding me that life is not only about enjoyment and that pain is inevitable. Hopefully at the end of my training, I will be able to ease some of this inevitable pain.

Acknowledgments

I am indebted to many who made this thesis possible. I would like to thank my Mother and Father for the love, support and encouragement they have given me over the years. Thanks also to Grandma, who has always encouraged me to do my best at what I enjoy most. To Joe, who has been a great brother to argue with and care about. To Sander Gubbens, a wonderfully supportive boyfriend and colleague; to Maria Longuemare, my best friend in graduate school; to The Lunch Club (Richard Wagner, Ted Mau, Andy Shiau and Sarah Gillmor), Julie Ransom, Rob Cerpa, Glenn Gobble, Stephen Rader, Jennifer Fung and many others, who made my life in graduate school a colorful and fun experience.

To Prof. David Agard and Prof. John Sedat, who have been most supportive advisors, who provided me with an excellent scientific environment to work in, as well as giving me the freedom to explore and experience many different fields in science. To Prof. David Baker for his incredible energy and inspiration, and for teaching me the love for science. To my colleagues of the Agard and Sedat laboratories for their scientific resources, challenge and support. To the members of my oral-examination and thesis committees: Professors Robert Fletterick, Tack Kuntz, Joe Gray, William Morgan and Ken Downing, for their encouragement, time, and valuable inputs.

This thesis was completed under support of the Howard Hughes Medical Institute Predoctoral Fellowship award (1991-1996).

Abstract

Computational Approaches to Macromolecular and Supramolecular Processes In Biology: From Protein Folding To Chromosome Folding.

by

Karen F. Han

The biological function of macromolecules and their supramolecular assemblies are intimately related to their structures. It is therefore important to develop an understanding of the 'rules' by which these structures are formed. This thesis presents two areas of technical developments in the fields of protein and chromosome folding that will add to our understanding of the architectural organization of these structures.

At the level of protein folding, a new approach has been developed to explore the relationship between amino acid profiles in multiple sequence alignments and their three-dimensional (3D) structures. The resulting classification of these variation patterns, revealed sequence rules for new structural motifs. The global feature of the mapping between protein sequence and structure showed that 44% of all sequences map to a single structure type, whereas 28- and 8% of the remaining sequences map to only two and three distinctive structure types respectively, accounting for approximately 80% of all sequences. These results have two important implications: 1) tertiary structure prediction must allow degenerate mapping rather than single-state secondary

structure predictions, and 2) some of the newly identified structural motifs represent folding initiation sites which aid in tertiary structure prediction.

At the level of chromosome folding, electron microscope tomography was used to investigate the assembly of 30 nm chromatin fibers into higher order structures. Due to limited spatial resolution, it was not possible to precisely and continuously trace the trajectories of the highly compact chromatin fibers in the 3D reconstructions, and was therefore important to improve our imaging and restoration techniques. A thorough analysis of the nature of image formation for thick biological specimens in transmission electron microscopy has led to corrections of resolution degrading factors. Contrary to common assumptions, thick biological specimens still exhibit a significant phase contrast component Furthermore, the amplitude contrast component relates to the specimen mass density by a logarithmic relationship, whereas the phase contrast component follows a linear relationship. The optimal strategy to image thick specimens is using a high primary voltage (>200keV) with zero-loss energy filtering and exit wavefront reconstruction. The improved contrast and resolution will aid in the interpretation of the 3D reconstruction.

Table of Contents

List of Tables

List of Figures and Illustrations

Part I: Optimal strategies to image thick biological specimens in transmission electron microscopy

Summary:

An accurate three-dimensional (3D) tomographic reconstruction requires the proper interpretation of image intensities as they relate to the projected specimen mass density. A detailed understanding of the nature of image formation is required to properly restore images. In thick biological specimens, the large multiple inelastic scattering component contribute to image distortion, making image intensities difficult to interpret.

Chapter 1 provides an overview for this section of the thesis. Chapters 2 and 3 address the nature of image formation for thick biological specimens using both through focus series exit wavefront reconstruction and electron energy-loss filtering techniques. It was demonstrated that the coherent (interpretable) component is contributed only by the elastically scattered electrons. The importance of using higher accelerating primary voltage is emphasized to maximize the proportion of elastically scattered electrons in thick biological specimens. Chapter 3 is the first quantitative study of image contrast improvement using an imaging filter as compared to CCD data collection alone for thick specimens. The total number of contributing coherently scattered electrons were tabulated as a comparison.

Since electron tomography also suffers from beam damage, it is necessary to develop approaches to restore images using as few focus levels as possible to minimize electron dose. Chapter 4 presents image restoration approaches for thick specimens using a range of 4 to 8 focus levels, combining the Schiske filter with the corrected low-resolution component. The restored images showed enhanced contrast compared to the in-focus and 1 μm underfocus images. The fractional root-mean-square deviation of the 4-focus

level restored images compared to the 40-focus level images is 5 %, compared to compared to over 25% for the unrestored image.

From these studies, the importance of correcting for image aberrations in TEM is demonstrated. In particular, the assumption that the image intensity directly reflect specimen projection mass density is incorrect. This has direct consequences to in the 3D tomographic reconstructions, where the accuracy of the 3D model relies on the proper interpretation of the many tilt projection views of the specimen. It is concluded that a combination of TEM operating high primary voltages (an a small objective aperture) coupled with an electron energy-loss imaging filter is necessary for imaging thick biological specimens. In addition, image restoration is needed to extract the amplitude and phase components of the image wave giving rise to the absorptive and linear components which can then be related to the true mass density of the specimen. It is recommended that restoration be done routinely for 3D EM tomography.

The appendix A provides a short summary of procedures to determine the nature of image formation of thick specimens in general, and recommendations of specific filters to be used in the restorations.

Chapter 1

Overveiw of Image Restoration of Thick Biological Specimens For Transmission Electron Microscope Tomography

An accurate tomographic reconstruction requires proper transformation of the recorded image intensities to projected specimen mass densities. For thick biological specimens, images are not only aberrated due to the objective lens aberrations, but also due to the large proportion of inelastic multiple scattering. Using a combination of energy filtering and through focus series restoration, these aberrations can be removed. The corrected images show enhanced contrast and resolution.

1. Introduction

Three-dimensional (3D) transmission electron microscope tomography is a powerful technique to study the higher order structural organization of supramolecular assemblies and cellular organelles. Tomography involves the computational 3D reconstruction of an object from many tilt-projection views (Turner 1981). Recent studies have clearly demonstrated the power of this technique by elucidating details of large amorphous cell substructures at 5 nm resolution, not achievable by means of light optical microscopies (Belmont et al. 1987; Fung et al. 1994; Horowitz et al. 1994; Ladinsky et al. 1994; Moritz et al. 1995). In recent years, the advent of digital image acquisition and computer-controlled electron microscopy has made tomography a more practical technique for routine use (Koster et al. 1993).

The accuracy of a 3D reconstruction relies on the correct interpretation of the collected image intensities. Because in transmission electron microscopy (TEM) the depth of focus is large with respect to the achievable lateral

resolution, the detected image intensities are typically assumed to be a close approximation of the projected specimen mass density. In reality, there are two sources of aberrations that 'distort' the images: electron microscope lens aberrations and electron-specimen interactions. The former causes aberrations independent of the nature of the specimen, whereas, the latter is very much specimen-dependent. Because of the natural dimensions of the typical cellular organelles studied by tomography, the specimens are necessarily thick (0.5-1 μm), and the primary source of aberration is at the level of electron-specimen interactions. It is therefore critical to correctly understand the electron specimen interactions and their importance on the image formation process in order to properly relate the detected image intensities to specimen mass density.

2. Sources of aberration in TEM micrographs of thick biological specimens

2.1 Objective lens aberration:

A brief review of simple linear imaging theory in transmission electron microscope is presented in this section. Spherical aberration and defocus levels cause a systematic aberration in the images as described by the well-known lens aberration function:

$$\exp[i\chi(\Delta f,\kappa)] = \exp[i\frac{\pi\lambda\kappa^2}{2}(Cs\frac{\lambda^2\kappa^2}{2} - \Delta f)] \qquad (1)$$

where λ is the wavelength, Δf is the defocus level, C_S is the spherical aberration, and κ is a vector in the reciprocal image plane. The resulting aberrated wave $\hat{\psi}_{sc}^{ab}(\Delta f,\kappa)$ in reciprocal space is simply the unaberrated scattered wave multiplied by the wave aberration function:

$$\hat{\psi}_{sc}^{ab}(\Delta f,\kappa) = \hat{\psi}_{sc}(\kappa)\exp[i\chi(\Delta f,\kappa)] \qquad (2)$$

The image formed by the aberrated scattered wave is simply its amplitude. The amplitude and phase components of the unaberrated specimen exit wave

can be restored by solving for the unaberrated wave through a systematic perturbation of focus levels.

$$\hat{I}_{\det ected}(\Delta f,\kappa)=\delta(\kappa)+2\hat{\psi}_{amp}(\kappa)\cos[\chi(\Delta f,\kappa)]-2\hat{\psi}_{phs}(\kappa)\sin[\chi(\Delta f,\kappa)] \tag{3}$$

To solve for the amplitude and phase components in Equation (3), at least two images taken at different focus levels are required. The exit wave is then recovered by combining the amplitude and phase components of the scattered wave.

The most commonly used restoring filter is that derived by Schiske and other authors (Hawkes 1980; Saxton 1978; Schiske 1968):

$$\hat{\psi}_e(\kappa)=\sum_{\Delta f_n=1}^{N}\hat{I}(\kappa,\Delta f_n)\cdot r(\kappa,\Delta f_n)$$

$$r(\kappa,\Delta f_n)=\exp[i\chi(\Delta f_n,\kappa)]\frac{\{N-\sum_{\Delta f_m=1}^{N}\exp[2i[\chi(\Delta f_m,\kappa)-\chi(\Delta f_n,\kappa)]\}}{\{N^2-|\sum_{\Delta f_m=1}^{N}\exp[2i[\chi(\Delta f_m,\kappa)]]|^2\}} \tag{4}$$

In A restoration using many (>30) focus levels at constant defocus interval, Equation (4) reduces to:

$$\hat{\psi}_e=\exp(i\pi C_s\lambda^3k^4)\frac{1}{N}\sum_{\Delta f_n=1}^{N}\hat{\mathbf{I}}(\kappa,\Delta f_n)\exp(-i\pi\lambda\kappa^2\Delta f_n) \tag{5}$$

where N is the total number of through focus images. In the field of high resolution electron microscopy, authors have used Equation (5) coupled with *a priori* knowledge and maximum likelihood approaches to restore the exit wave at resolutions close to $(0.2\ nm)^{-1}$ (Coene et al. 1992; van Dyck et al. 1993).

2.2 Aberrations contributed by electron-specimen interactions

Electron-specimen interactions include single elastic and inelastic, and multiple elastic and inelastic scattering. For thick biological specimens, the

multiple inelastic scattering component dominates and results in a blurring of the images due to the chromatic aberration of the microscope's objective lens. Using electron energy-loss spectroscopy (EELS) and electron spectroscopic imaging (ESI), one can separately investigate the contribution of the various scattering mechanisms to the image formation (Colliex et al. 1989; Han et al. 1995c; Langmore et al. 1992; Reimer et al. 1991). Figure 1 plots examples of EELS spectra for thick biological specimens, showing that majority of the imaging electrons are indeed multiply and inelastically scattered.

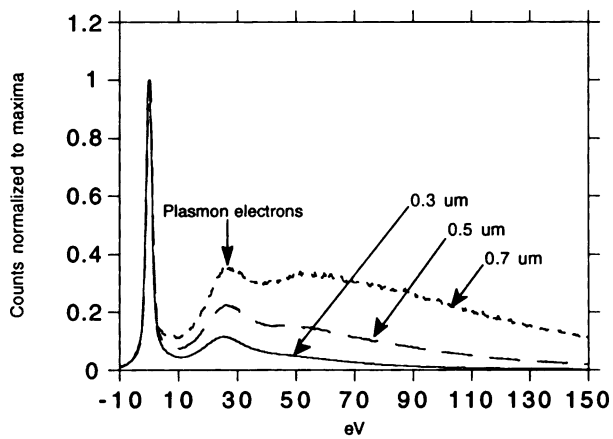


Figure 1. Electron energy-loss spectra of 0.3, 0.5 and 0.7 um thick specimens at 200 keV normalized by by peak elastic counts (reprinted (Han et al.,1996a) with permission from J. Microscopy).

For those specimens where the elastically scattered electrons account for only a very small proportion of the imaging electrons, it was found that employing ESI to image only those electrons in a energy-loss window centered on the most probable energy-loss, results in the best image contrast (or signal to noise ratio) (Colliex et al. 1989). For tomography, however, more important than image contrast is the need for a direct relationship between the image intensities and the projected mass densities.

Studies have shown that the 3D power spectrum of a through focus series (taken at equal intervals as in the exit wavefront reconstruction) can be used to

evaluate the proportion of coherent electrons contributing to imaging (Han et al. 1995). By combining this technique with ESI, it was shown that for thick biological specimens, only the elastically (zero energy-loss) scattered electrons exhibit the linear imaging behavior through focus, which is characteristic of the coherent image component (Han et al. 1996a). Table I shows the relative contribution of the coherent, incoherent, and background image components based on the analysis.

Table I. Relative amount of the contributing components to image formation of thick biological specimens. (reprinted (Han et al.,1996a) with permission from J. Microscopy).

Thickness	Energy filter experiment	Elastic electrons %	Parabola (coherent) %	Central (partially (in) coherent) %	Back-ground (incoherent & noise) %
0.5 μm	Unfiltered	21.4	10.1	33.5	56.6
	Zero-loss		13.1	23.3	63.6
	Plasmon		3.0	13.7	83.3
	30eV-130eV		0.0	16.2	83.8

The results imply that for thick biological specimens, it is essential to image at intermediate to high accelerating voltages in order to maximize the signal from the elastically scattered electrons. In addition, although contrast may be optimized by imaging at the most probable energy-loss, only the restored zero-loss filtered image can be readily interpreted and related to specimen mass density.

3. Restoration of thick amorphous specimens

The fundamental assumption in the restorations of equations (3) and (4) is that the entire scattered wave is effected by the wave aberration function (Eq. 1) in the same way. This is in generally true for thin specimens where most electrons are single elastically scattered. As mentioned in section 2.2, for thick

amorphous specimens, where most electrons are (multiply) inelastically scattered, it is expected that a sizable portion of the scattered wave does not propagate through focus in the expected manner as described by equation (2) above. However, because only the single elastically scattered electrons follows the linear imaging properties of the microscope, it is possible to recover the exit wave using standard restoration filters. Although equation (4) is the result of an incomplete description of the images detected for thick specimens, since all other components such as multiple inelastic scattering add incoherently, the coherent image component will be enhanced as more through focus images are used to restore the exit wavefront by a factor of $\sqrt{N}$, where N is the total number of images used in the restoration. A better separation between the parabolic and incoherent components will reduce the width of the parabola in ζ (proportional to sinc(Z), Z is the full range of focus levels).

With increasing specimen tilt angle θ (equivalent to an increasing effective specimen thickness following $1/\cos\theta$), the average image intensity decreases logarithmically (data not shown). This is in agreement with the logarithmic decrease of the elastically scattered electrons as a function of specimen thickness (Han et al. 1995). Both of these observations are consistent with the absorption model for thick specimens. Therefore, the interpretation of the restored amplitude component of the exit wave should be interpreted as logarithmically proportional to the specimen mass density, and the phase component as linearly proportional. Figure 2 compares an image restored using equation (7) with the unrestored in-focus image, and demonstrates the enhancement in resolution. Furthermore, restoration using a through focus series of zero-loss filtered images shows an additional improvement in resolution and contrast (Han et al. 1995a).

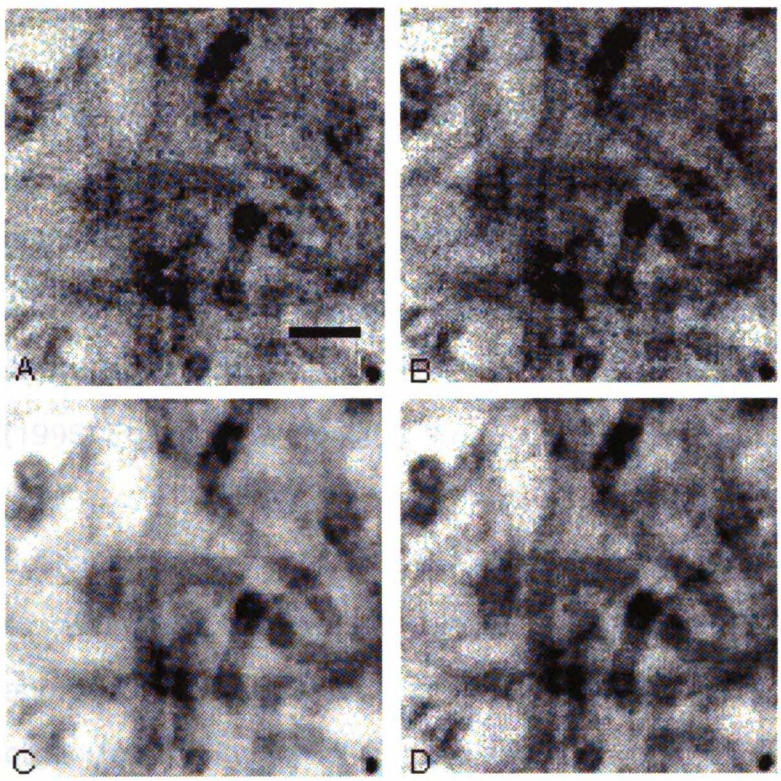


Figure 2. Zero-loss filtered restored (A) and in focus (B), unfiltered restored (C) and in-focus (D) images of a 0.5 μm specimen. Scale bar: 60 nm. (C,D reprinted (Han et al.,1995b) with permission from J. Microscopy).

4. Conclusion

Extensive analysis of the image formation for thick biological specimens has shown that it is valid to restore images using the restoring filters derived from linear imaging theory. The restored amplitude and phase images show respectively a logarithmic and linear relationship to the projected specimen mass density. For best results in the imaging of thick biological specimens, exit wavefront restoration should be used in combination with zero-loss filtering and operation at intermediate to high primary voltages. The improved image contrast and the proper interpretation of the image intensities of the aberration corrected image will result in an enhanced resolution in 3D tomographic reconstructions of thick biological specimens.

5. Acknowledgments

The authors thank M. Braunfeld and M. Moritz for providing the specimens. K.F.H. is supported by the Howard Hughes Medical Institute Predoctoral Fellowship in the Biological Sciences. This work is supported by grants from the National Institutes of Health (GM 31627 for D.A.A.; GM25101 for J.W.S.) and by Howard Hughes Medical Institute. This chapter is an approved reprint of the material as it appears in K.F.Han, A.J.Gubbens, J.W.Sedat and D.A.Agard (1995) Focus on Microscopy 95.

6. References

1. Belmont, A.S., Sedat, J.W. & Agard, D.A. (1987). A three-dimensional approach to mitotic chromosome structure: Evidence for a complex hierarchical organization. *Journal of Cell Biology,* **105**, 77-92.

2. Coene, W., Janssen, G., Op de Beeck, M. & Van Dyck, D. (1992). Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. *Physical Review Letters,* **69**, 3743-3746.

3. Colliex, C., Mory, C., Olins, A., Olins, D. & Tence, M. (1989). Energy filtered STEM imaging of thick biological sections. *Journal of Microscopy,* **153**, 1-21.

4. Fung, J.C., Agard, D.A. & Sedat, J.W. (1994). Three-dimensional reconstruction of the synaptonemal complex from high-pressure frozen maize meiocytes using IVEM tomography. *Proc. 53rd Ann. Microscopy Society of America,* 14-15.

5. Han, K.F., Sedat, J.W. & Agard, D.A. (1995). Mechanism of image formation for thick biological specimens: *exit wavefront reconstruction and electron energy-loss spectroscopic imaging. J. Microscopy,* **178:2**, 107-19.

6. Han, K.F., Gubbens, A.J., Sedat, J.W. & Agard, D.A. (1996a). Optimal strategies for imaging thick biological specimens: *exit wavefront reconstruction and energy filtering. J. Microscopy,* (in press).

7. Han, K., Sedat, J. & Agard, D. (1996b). Practical Image restoration of thick biological specimens using multiple focus levels in transmission electron microscopy. *J. Microscopy,* (Submitted).

8. Hawkes, P.W. (1980). Image processing based on the linear theory of image formation. Computer Processing of Eletron Microscope Images. Berlin, Pringer-Verlag.

9. Horowitz, R.A., Agard, D.A., Sedat, J.W. & 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.

10. Koster, A., Braunfeld, M., Fung, J., Abbey, C., Han, K., Liu, W., Chen, H., Sedat, J. & Agard, D. (1993).Towards automated 3D imaging of large biological structures using intermediate voltage electron microscopy. *MSA Bulletin,* **23**, 176-88.

11. Ladinsky, M.S., Kremer, J.R., Furcinitti, P.S., McIntosh, J.R. & Howell, K.E. (1994). HVEM tomography of the trans-Golgi network: structural insights and identification of a lace-like vesicle coat. *J Cell Biol,* **127**, 29-38.

12. Langmore, J.P. & Smith, M.F. (1992). Quantitative energy-filtered electron microscopy of biological molecules in Ice. *Ultramicroscopy,* **46**, 349-73.

13. Moritz, M., Braunfeld, M., Fung, J., Alberts, B., Sedat, J. & Agard, D. (1995). Three-dimensional structural characterization of centrosomes from early drosophila embryos. *J. Cell Biol.,* **130**,1149-59.

14. Reimer, L., Rennekamp, R., Fromm, I. & Langenfeld, M. (1991). Contrast in the electron spectroscopic imaging mode of a TEM. IV. Thick specimens imaged by the most-probable energy loss. *Journal of Microscopy,* **162**, 3-14.

15. Saxton, W.O. (1978). Computer Techniques for Image Processing in Electron Microscopy. New York, Academic Press.

16. Schiske, P. (1968). Zur Frage der Bildrekonstruktion durch Fokusreihen. 4th European Conference on Electron Microscopy. 145.

17. Turner, J.N. (1981). Introduction to Stereo Imaging. Three-dimensional Ultrastructure in Biology, Methods in Enzymology. New York, Academic Press.

18. van Dyck, D., Op de Beeck, M. & Coene, W. (1993). A new approach to object wave function reconstruction in electron microscopy. *Optik,* **93**, 103-7.

Chapter 2

Mechanism of Image Formation for Thick Biological Specimens: Exit wavefront reconstruction and electron energy-loss spectroscopic imaging

With increasing frequency, cellular organelles and nuclear structures are being investigated at high resolution using electron microscopic tomography of thick sections (0.3-1.0 μm). In order to accurately reconstruct the structures in three dimensions from the observed image intensities, it is essential to understand the relationship between the image intensity and the specimen mass density. The imaging of thick specimens is complicated by the large fraction of multiple scattering which gives rise to incoherent and partially coherent image components. Here we investigate the mechanism of image formation for thick biological specimens at 200 and 300 keV in order to resolve the coherent scattering component from the incoherent (multiple scattering) components.

Two techniques were used: Electron Energy-Loss Spectroscopic Imaging (ESI), and exit wavefront reconstruction using a through focus series. Although it is commonly assumed that image formation of thick specimens is dominated by amplitude (absorption) contrast, we have found that for conventionally stained biological specimens phase contrast contributes significantly, and that at resolutions better than ~10nm, superposed phase contrast dominates. It is shown that the decrease in coherent scattering with specimen thickness is directly related to the increase in multiple scattering. It is further shown that exit wavefront reconstruction can exclude the microscope aberrations as well as the multiple scattering component from the image formation. Since most of the inelastic scattering with these thick specimens is

actually multiple inelastic scattering, it is demonstrated that exit wavefront reconstruction can act as a partial energy filter. By virtue of excluding the multiple scattering, the 'restored' images display enhanced contrast and resolution.

These findings have direct implications for the three dimensional reconstruction of thick biological specimens, where a simple direct relationship between image intensity and mass density was assumed, and the aberrations were left uncorrected.

1. Introduction:

The structures of cellular organelles and nuclear components are most appropriately studied at high resolutions using transmission electron microscopy (TEM). For analytical studies and three dimensional reconstructions, it is essential to properly relate the observed image intensities to the specimen mass densities. This necessitates an accurate understanding of the mechanism of image formation for these specimens in the transmission electron microscope. Images taken by TEM are not always direct representations of the specimen mass density. In fact, there are two sources of aberrations that effect the image formation: microscope lens aberrations, and electron-specimen interactions. Unlike the microscope lens aberrations, the electron-specimen interactions contribute differently to the image aberrations for imaging thick and thin specimens.

The practical resolution obtainable for biological samples is typically limited by a combination of sample preparation limitations and by specimen damage resulting from beam exposure. For the study of large macromolecular assemblies such as proteasomes and ribosomes, the typical achievable resolution is limited to 1.0-1.5 nm (Frank et al. 1992). By contrast, for

supramolecular structures such as cellular organelles and chromosome the inability to do signal averaging and non-optimal staining generally limit the resolution to 4.0-10.0 nm.

Our efforts have focused on the study of organelles and chromosomes. In order not to limit the achievable resolution, much effort has been put into the preservation of the three-dimensional structures during sample preparation (Belmont et al. 1989; Belmont et al. 1987; Dahl et al. 1989; Hohenberg et al. 1994) . Because these structures are large, complex and not symmetric, a three dimensional analysis is often required to obtain useful information. While two major approaches to three dimensional analysis have been used: serial thin section reconstruction and tomographic reconstruction (Turner 1981), we have concentrated on the latter because it provides the unique ability to study internal structures with nearly isotropic resolution. In tomography, a large set of typically ~120 tilted views with +/- 75^{o} tilt range of the same specimen are collected and subsequently processed into a three dimensional reconstruction. Examining larger structures intact necessitates the use of thick specimens (0.3 -1.0 μm) and generally higher acceleration voltages. Compared with thin specimens, thick specimens have larger multiple and inelastic scattering components that cause aberrations in the images which effect the three dimensional reconstruction.

The mechanism of image formation for thick biological specimens has been previously investigated by a number of authors (Colliex et al. 1989; Langmore et al. 1992; Reimer et al. 1991). It is generally assumed that the main contrast mechanism for these specimens is so called amplitude or absorption contrast. It is also understood that as a result of the chromatic aberration of the TEM objective lens, the large fraction of inelastic scattering causes

considerable image blurring. Therefore, much attention has been focused on reducing the effects of the chromatic aberration by increasing the accelerating voltage, and energy filtering (Bazett-Jones 1992; Olins et al. 1989). By allowing only a small range of electron energies to contribute to the image, blurring can be significantly reduced. In cases where only a very small percentage of scattered electrons are zero-loss, it was found that imaging at the most probable energy-loss provided optimal image contrast and signal-to-noise ratio (Colliex et al. 1989).

In material science much effort has been devoted to resolution extension. In recent years, successful phase retrieval of the specimen exit surface wavefront by holographic approaches has been achieved either by the physical detection of an interference pattern (Lichte 1986) or by the construction of the Ewald sphere using a through focus series (Lichte 1986; Taniguchi et al. 1991; Van Dyck et al. 1990). By reconstructing the exit surface wave immediately after the specimen, the aberrations from the objective lens are in principle removed. In practice, these approaches have improved the correction of microscope aberrations, and have so extracted information to high resolutions (Coene et al. 1992; de Ruijter 1992; Van Dyck et al. 1993).

In this work we use a number of techniques to empirically analyze factors responsible for image degradation for thick biological specimens, and explore the possibility of separating these components from the coherent component. By construction of the Ewald sphere from a through focus series, the degree of coherent image transfer for thick biological specimens can be analyzed (see theory). Restoration of the electron exit wave front right after the specimen using the Ewald sphere construction and subsequent back transformation is shown to exclude most of the incoherent component from the images.

Parallel studies using electron energy loss spectroscopy (EELS) and electron energy-loss spectroscopic imaging (ESI) show a direct proportionality between the degree of incoherent transfer and the fraction of inelastic scattering. It is shown that the source of the incoherent component is largely due to multiple scattering. Since most inelastically scattered electrons are also multiply scattered in thick specimens, the exit wavefront reconstruction therefore acts in part as an energy filter. Restoration of the images by extracting the component on the Ewald sphere and back transformation of the 'filtered' coherent component can thus lead to enhanced contrast and resolution. More importantly, the restoration shows that in addition to the low-resolution amplitude contrast, there is a significant amount of phase contrast for thick specimens and that phase contrast actually dominates at moderate and high resolutions. It is proposed that the intensities in the restored amplitude image are due to absorption (inelastic and multiple scattering) and therefore exhibit a logarithmic relationship with mass-density. The intensities in the restored phase image are a direct measure of the projected mass densities. These results will greatly effect the three-dimensional reconstruction of thick biological specimens from multiple images.

2. Materials and Methods:

(a) Thick Biological Specimens:

The specimens used in the experiments described below are isolated nuclei (HeLa) or whole-mount tissues (maize and lily anthers) embedded in epon and stained with uranyl-acetate and lead-citrate (Fung et al. 1994). The regions of interest in the embedded specimens are the chromosome and synaptonemal complex structures in the nucleus. Specimens were cut to 0.3, 0.5, 0.7 and 1

μm thickness. Some of the thickness-dependent experiments were actually performed by tilting the specimens to increase their effective thickness.

(b) Microscopy:

Aside from the EELS and ESI experiments, all micrographs were digitally recorded at 300 keV on a Philips EM430 TEM. This microscope, which has been previously described (Koster et al. 1992a), is equipped with a LaB_6 filament, a C400 computer interface to control all lenses, and stage positioning, and a prototype Gatan large-format slow-scan CCD camera (1024x1024 19μm pixels) all controlled by a MicroVax III processor. The images were taken at a TEM magnification of 30,100x, and were subsequently binned twice corresponding to a final specimen pixel size of 1.34 nm. The focus levels were calibrated using beam-tilt induced image shifts as described (Koster et al. 1992b). For optimal beam coherence, an effective spot-size of 47 nm was used. In addition, the optimal beam divergence (intensity setting) was chosen for imaging to further minimize the effect of the coherence envelope. As evident in figure 6, the contrast transfer rings of the through focus series are limited by the pixel resolution (2.68 nm^{-1}).

(c) Through Focus Series:

Through focus series consisting of 30 images were recorded from 17.3 μm under focus (weaker objective lens current) to 17.3 μm over focus with a focus step size of 1.152 μm. To minimize specimen alterations and shrinkage (Braunfeld et al. 1994) during data collection, the specimens were stabilized by pre-irradiating with approximately 1000 e^-/nm^2. The individual images were aligned using a combination of cross-correlation and fudicial gold markers. Image processing and visualization were done on a DEC VAX-9000 and a

Silicon Graphics Iris workstation using Priism, the image visualization software developed in our laboratory (Chen et al. 1994).

(d) Electron Energy-loss Spectroscopy and Imaging:

These experiments were done with a Gatan Imaging Filter (GIF) (Gubbens et al. 1993; Krivanek et al. 1992) mounted on a JEOL 2010, a Philips CM12 and a Philips CM20 TEM. The images were acquired with the built in slow scan camera using the DigitalMicrograph software. The energy dispersion window used for spectroscopy was 0.5 eV per CCD pixel, a 15eV energy window was used for energy filtered imaging. The images were transferred to the DEC VAX-9000 and the Silicon Graphics Iris workstation for processing and analysis.

3. Theory:

Holography allows the reconstruction of the electron exit wave front immediately following the specimen surface. The advantage of image restoration using holographic techniques is that microscope aberrations are automatically excluded. In addition, it naturally functions as an energy filter and a coherence filter in that the incoherent component does not interfere constructively. Thus, theoretically holography is the optimal technique for imaging as demonstrated by a number of authors (de Ruijter et al. 1993; Gribelyuk et al. 1993; Joy et al. 1993; Lichte 1993; McCartney et al. 1994). The draw back of holography through the physical detection of an interference pattern is the requirement for a reference wave (and so a specimen edge), and its low detection signal.

D. Van Dyck and coworkers (Coene et al. 1992; Van Dyck et al. 1990; Van Dyck et al. 1993) showed that the exit wave front can also be restored using a through focus series. Using the theory of wave propagation, they demonstrated that the exit wave front is simply the component which falls on the

Ewald Sphere. Y. Taniguchi and coworkers (Taniguchi et al. 1991) used a similar approach to align the current center of the microscope, and to investigate the effect of the microscope lens aberrations on image formation.

Techniques and developments in holography have focused on resolution extension by eliminating the microscope lens aberrations as described above. Most specimens investigated have been extremely thin and hence multiple inelastic scattering was of little concern. In this paper, it is demonstrated that in addition to minimizing microscope lens aberrations, exit wave reconstruction techniques can be used to analyze the image formation for thick specimens at the level of electron-specimen interactions such as multiple scattering.

To understand the specimen exit surface wave restoration (the specimen holograph) using a through focus series, a brief review of the wave propagation theory of image formation is presented (adapted from D. van Dyck and M. Op de Beeck, 1991). The image wave $\psi(\mathbf{r},\Delta f)$, at Δf defocus, is a result of convoluting the exit wave $\phi(\mathbf{r})$ with the well-known lens aberration function $\chi(\mathbf{r},\Delta f)$. The three dimensional Fourier transform of $\psi(\mathbf{r},\Delta f)$ is given in equation (1).

$$\psi(\mathbf{k},\xi)=\int \phi(\mathbf{k})\exp(2\pi\imath\Delta f(\xi-\lambda|\mathbf{k}|^2/2))d\mathbf{k}d(\Delta f) \qquad (1)$$

where $\mathbf{k}$ and ξ are the reciprocal space coordinates related to the real space coordinates x, y and the defocus Δf. The image detected in Fourier space is simply the convolution of the image wave with itself:

$$\mathbf{I}(\mathbf{k},\xi)=\psi(-\mathbf{k},-\xi)\otimes\psi^*(\mathbf{k},\xi) \qquad (2)$$

$$= |C|^2 \delta(\mathbf{k}) + C^* \phi(\mathbf{k}) \delta(\xi - \lambda |\mathbf{k}|^2 /2) + C\phi^*(-\mathbf{k}) \delta(\xi + \lambda |\mathbf{k}|^2 /2)$$
$$+ \int_{\mathbf{k}\neq 0, \mathbf{k}-\mathbf{k}' \neq 0} \phi^*(\mathbf{k}) \phi(\mathbf{k}-\mathbf{k}') \delta\{\xi - \lambda[(\mathbf{k}-\mathbf{k}')^2 - \mathbf{k}^2]/2\} d\mathbf{k}' \qquad (3)$$

where $\mathbf{I}(\mathbf{k}, \xi)$ is the three dimensional Fourier transform of the through focus series, and δ is the Dirac delta function. The constant C defines the average illumination intensity.

The three dimensional Fourier transform of a through focus series, $\mathbf{I}(\mathbf{k}, \xi)$ contains four terms that can be understood physically. The first term in equation 3 is the unscattered wave. The second and third terms contain the exit surface wave $\phi(\mathbf{k})$ which is completely described on the Ewald sphere parabola $\delta(\xi - \lambda |\mathbf{k}|^2 /2)$. The last term in equation 3 represents the secondary interference of the scattered beam, commonly termed the non-linear imaging component, and is distributed everywhere in the three dimensional Fourier space. In this derivation, it is assumed that all the components in the exit wave propagate with the expected dependence on the wave aberration function through focus. In general this is the case for thin specimens as has been previously demonstrated (Coene et al. 1992). Since the aberrations of the contrast transfer function are implicitly incorporated in the recovery of the exit wave, the microscope lens aberrations are naturally excluded in these reconstructions.

Thick specimen imaging is dominated by multiple scattering component which are not included in the above derivation. If the multiple scattering component propagates with the same dependence on focus as the coherent component, then it would be impossible to distinguish between the two. If the multiple scattering components propagate with a different dependence on focus than expected for the coherent component, then one would expect it to

generate a separate component in the three-dimensional power spectrum, independent from the coherent (paraboloid) component. In reality, the multiple scattering component is neither completely coherent nor completely incoherent through focus. Its propagation is also not expected to be completely independent of the coherent component. Practically, however, one can separate the large portion of multiple scattering component that propagates independent from the coherent component and has a different dependence on the focus variation. Due to finite sampling and limited range of detection of the through focus series, this technique will not be able to distinguish coherent inelastic scattering (such as plasmon scattering) from coherent elastic events (such as single elastic scattering). Therefore, the chromatic aberration coming from those electrons will not be excluded from the reconstruction. Recognizing the fact that multiple scattering is difficult to model, we will take the (empirical) experimental approach to differentiate this component from the coherent component of imaging by combining the Ewald sphere approach described above with electron spectroscopic imaging.

From the experimental results in the following section we show that indeed there is a substantial component in imaging that does not propagate with the same dependence on focus variation as the coherent component. By varying the specimen thickness, we demonstrate that the proportion of the non-coherent component is directly related to the amount of inelastic scattering. Through focus series of electron energy-loss images demonstrates that this component can be directly attributed to multiple inelastic scattering.

4. Results:

Electron-specimen interactions contribute to coherent and incoherent or partially coherent imaging components. There are two sources that contribute

to the coherent imaging component: 1) single elastic scattering and 2) single inelastic plasmon loss scattering that effectively acts as a secondary beam at a shifted energy, followed by subsequent single elastic scattering. The incoherent component includes multiple elastic and inelastic scattering. Controllable electron microscope imaging parameters such as focus level, objective aperture size and energy filtering are used to investigate the different image components.

(a) The Effects of the Objective Aperture on Specimen Contrast: It is common knowledge that the use of a small objective aperture improves image contrast. This increase in contrast is generally ascribed to "classical" amplitude contrast resulting from scattering of electrons outside the aperture. Here we demonstrate that the scattering or amplitude contrast introduced by the application of a small objective aperture can be quite specimen dependent. For highly scattering thin specimens, such as poly-crystalline gold, the application of the objective aperture results in an additional (classical) amplitude contrast by contributing to an enhancement of the signal. Figure 1 shows the images of poly-crystalline gold without and with a 20 μm (~5 mrad) aperture. The image difference (pixel by pixel intensity subtraction) between the two images shows a coherent image corresponding to an enhanced signal. Furthermore, as expected this signal has a cosine dependence with respect to defocus.

This situation is quite different for thick amorphous biological specimens stained with heavy metals. The effect of inserting a small objective aperture merely results in a decrease in background intensity, and consequently an increase in the image contrast. Figure 2 shows images of a 0.3 μm thick heavy metal stained specimen without and with a 20 μm objective aperture. The difference image shows no coherent image features, demonstrating that for

typical thick biological specimens, the additional contrast gained by inserting a small objective aperture is due to a decrease in background scatter. This background signal does not depend on defocus, which is quite different from classical amplitude contrast and derives from the incoherent multiple scattering, as demonstrated in the following experiments.

(b) More Inelastic than Elastic Electrons Scatter Outside the Objective Aperture:

To investigate the contribution of scattering outside the objective aperture, electron energy loss spectra were taken for different aperture sizes using the Gatan Imaging Filter. Figure 3 shows electron energy loss spectra for a 0.7 μm thick specimen at 200 keV using 10, 30, 40 μm (~ 11, 8, 3 mrad) objective apertures and no aperture, normalized by the peak number of elastic electrons. In all cases the spectral maximum is still at the zero-loss (elastic) peak, as opposed to a 0.5 μm thick specimen at 100 keV, where the maximum is at 100 eV energy-loss (Colliex et al. 1989). From the data one can see that more inelastically scattered electrons are removed by the objective aperture than are elastic electrons. This is in agreement with the results of Reimer from angular-resolved EELS on carbon films (Reimer 1989). As a consequence, there is a relative increase in elastically scattered electrons for smaller aperture sizes, and therefore an effective increase in contrast. For these samples, differences between images taken with and without the objective aperture again show a constant background with no specimen features (Data not shown). This then implies that all the scattering outside the objective aperture is from multiple scattering, and is largely inelastic. The second maximum in the spectra is at 25 eV-loss, which corresponds to the plasmon loss scattering of carbon (epon embedding media). These electrons are single inelastically scattered, which

effectively gives rise to a secondary source at this energy-loss. Subsequent single elastic scattering events will be mutually coherent at a slight energy-loss, resulting in a focus shift compared with the elastic image due to the chromatic aberration. Because of the small energy loss due to carbon absorption (25 eV), the additional defocus induced by chromatic aberration is less than 0.1 μm, well within the accuracy of our focus level determination.

(c) Multiple Inelastic Scattering Contributes Only to Low Resolution Information:

Spectroscopic imaging was done to explicitly evaluate, for each range of energy losses, their contribution to the high resolution image. Figures 4A through 4E are energy filtered images for every 15 eV-loss intervals with 15 eV energy windows, from zero energy-loss to 75 eV-loss, using a 10 μm objective aperture. Energy loss images higher than 75 eV-loss show almost no specimen features (data not shown). High resolution substructures were only present in the elastic (zero-loss) and plasmon at (25 eV-loss) images. For all other energy ranges, the images formed contained only very low resolution information and showed no specimen substructure, resulting from the incoherent transfer of the multiple inelastic scattering component. We thus conclude from this series of experiments that multiple inelastic scattering within the objective aperture contributes largely at low resolutions ($< 0.04 nm^{-1}$).

(d) Inelastic and Elastic Mean-free-paths:

The probability of multiple scattering can be represented by the electron scattering cross section, which is inversely related to the mean-free-path. EELS experiments were performed on specimens of different thickness in order to estimate the mean-free-paths for our specimens. Figure 5 shows EELS spectra of 0.3, 0.5 and 0.7 μm thick specimens, using a 10 μm objective aperture,

normalized by the total number of electrons (A) and normalized by the zero-loss electrons (B). As expected, the fraction of zero-loss electrons decreases with increasing specimen thickness. From these spectra, for our typical specimen, the inelastic and elastic mean-free-paths were determined to be 256 nm and 468 nm respectively at 200 keV. When scaled up for 300 keV using equation 6.9 (p.189) by Reimer (Reimer 1984), the inelastic and elastic mean-free-paths are estimated to be 500 nm and 700 nm. From this, it is estimated that for specimens thicker than 0.5 μm, more than 60% of the electrons will be multiply inelastically scattered at 300 keV.

(e) Exit wavefront Reconstruction Shows A Significant Degree of Coherence for Thick Specimens:

As demonstrated above, the multiple inelastic scattering component provides little or no high resolution imaging features, and contributes only to the low resolution deterioration. Therefore, our goal is to exclude these electrons from the images by means of extracting only the coherent imaging component. The coherent component of the image can be extracted from a three-dimensional power spectrum of a through focus series, where it lies on a parabolic surface described as the Ewald sphere. By contrast, those components which are not mutually coherent through focus such as inelastic and elastic multiple scattering, do not fall on the Ewald sphere. Therefore, by analyzing the three-dimensional Fourier transforms of a through focus series, one can quantify the relative amount of coherent and incoherent scattering. In addition, reconstructions can be made by selecting only those components which fall on the Ewald sphere and back transforming.

Through focus series of 30 images (see methods) were taken for this analysis. A few representative diffractograms of the through focus series are

shown in Figure 6, demonstrating high quality of resolution transfer. Figure 7 shows a few representative cross-sections of the Ewald sphere from the exit wavefront of a 0.3 μm specimen (epon embedded, stained with uranyl acetate and lead citrate). As the schematic diagram shows, the coherent component is on the outer circle, and the noise or incoherent component is at the center. This experiment shows that for thick specimens, coherent transfer is still very significant. If we compare these results to those from thin carbon films, the prominent incoherent component at the center is markedly larger for the thick specimens as expected (data not shown). The radial average of the cross-sections through the Ewald sphere were fitted with three Gaussian functions: the coherent (parabola), incoherent and Poisson noise components. To avoid artifact at $\xi=0$, the fits were done starting at 34.3 nm^{-1}. Figure 8 shows a fitted curve of the cross-section shown in figure 7B, demonstrating the accuracy of the fit to be on average 7%. The fits were done for the entire three-dimensional power spectrum. Figure 9 shows fits for the coherent (A), incoherent (B), and Poisson noise (C) components. Figure 9A shows that the Ewald sphere is characterized by finite width with the peaks at the correct radius given the sampling through focus. The most important factor that contributes to the width of the parabola is the coarse and finite sampling of the through focus series. A much thinner parabola is observed for carbon films (data not shown). The additional width of the parabola for the thick specimen may be caused by the energy spread of the coherent component. Figure 9B shows that the central incoherent component is well fitted by a three-dimensional Gaussian. This component is clearly independent of the coherent component, and has a different dependence on focus variation. This is distinguished from the Poisson noise component which is roughly independent of focus as shown in figure 9C.

This is expected since its contribution is not related to the propagation of the exit wave.

(f) Specimen Thickness Dependence - agreement between the Exit Wavefront Reconstruction and the EELS Data:

Figure 10 shows two levels of cross-sections through the Ewald Sphere for different specimen thicknesses, demonstrating that the paraboloid (or coherent) components decrease as specimen thickness increases, and vice versa for the central (or incoherent) component. By fitting the radially averaged cross-sections for each of the thicknesses as described above, the coherent component can be quantified for each specimen thickness with an average fit error of less than 5%. Table I lists the calculated relative amount of the coherent component from the Ewald Sphere analysis and the relative amount of the elastic scattering component from the EELS analysis as a function of specimen thickness. It is clear that the relative amount of coherent scattering is proportional to the degree of elastic scattering. They are not identical because, in addition to the coherently scattered electrons, the zero loss peak in EELS contains electrons that have not been scattered at all in addition to those that have been multiply scattered. Figure 11 plots the data from Table I on a logarithmic scale as a function of thickness, showing a clear correlation between the proportion of elastic scattering with the amount of coherent imaging component. This demonstrates that through focus exit wavefront restoration can act, to some degree, as an energy filter for thick specimens by way of excluding multiple scattered inelastic electrons.

(g) Restored Exit Wave Shows Significant Phase Contrast for Thick Specimens:

Exit wavefront reconstruction was performed by complex inverse 3-dimensional Fourier Transform of those components lying on the Ewald sphere. This filter process results in real and imaginary components of the exit wave which correspond to the classical amplitude and phase components respectively. Figure 12 and 13 show the real (amplitude) and imaginary (phase) components of the restored exit wave for 0.3 and 0.7 μm thick specimens, and the corresponding diffractograms. It is clear that amplitude contrast dominates at low resolutions and phase contrast at high resolutions. Phase contrast is still dominant at high resolutions compared with amplitude contrast for a 0.7 μm thick specimen. As expected, the phase component of the 0.7 μm the specimen has less resolution extent as that of the 0.3 μm thick specimen.

For most applications such as tomography, it is impractical to restore images using 30 focus levels on a regular basis. Using the empirically derived coherent and incoherent contributions to the thick specimen image formation, these restorations can serve as standards for developing techniques to restore images using fewer focus levels (manuscript in preparation).

5. Discussion:

High resolution structures of cellular organelles and nuclear structures are most appropriately studied using transmission electron microscopy. For quantitative studies and three dimensional reconstructions, it is essential to properly relate the image intensities with the specimen mass densities. This relies on the accurate understanding of image formation mechanism of these specimens in the transmission electron microscope (TEM). The images taken in TEM are not always direct representations of the specimen mass density. In fact, there are two sources of aberrations that effect image formation: electron-specimen interactions, and microscope lens aberrations.

Because of the natural sizes of cellular organelles and nuclear structures, they are generally prepared as thick specimens for electron microscope tomography. Compared with thin specimens, thick specimens have additional multiple and inelastic scattering components that cause image aberration. Here, we have used a number of different techniques to empirically analyze the image-degrading components, and have explored the possibility of separating these components from the coherent component.

It was shown that the application of a small objective aperture increased image contrast in thick biological specimens by decreasing the amount of background noise using EELS and EELSI experiments. This noise is largely due to multiple inelastic scattering. The multiple inelastic component that scatters within the objective aperture contributes only at resolutions well below 250 angstroms. Approximate multiple inelastic and elastic scattering cross sections for our specimens at 300 keV were determined to be 500 nm and 700 nm respectively. From this, it is estimated that at 300 keV, biological specimens >0.5 μm thick will experience more than 60% inelastic scattering.

The degree of coherence for thick specimen imaging was determined using the Ewald sphere construction. Surprisingly, there is still a significant amount of coherent scattering for thick specimens. It is demonstrated that the degree of coherence decreased as specimen thickness increased; vice versa for the magnitude of the incoherent component. This result directly parallels the degree of elastic scattering as a function of specimen thickness. The incoherent component is therefore attributed to multiple inelastic scattering.

In contrast to common assumptions, the restored exit wavefront showed significant phase contrast for thick specimens. Indeed there is a large amplitude contrast component dominating at low resolutions, where the

contribution is mostly from multiple inelastic scattering. Therefore, the image intensities from amplitude contrast have a log relation to specimen mass density which is related to the inelastic scattering cross section. The high resolution phase image has a direct relation with the relative mass thickness of the specimen, and is therefore linearly related to mass density.

Since the exit wavefront restoration excludes the incoherent (multiple inelastic scattering) component, it can act in part as an energy filter for thick specimens. With the combined analysis of through focus series and energy filtering, it is demonstrated that the enhanced contrast using energy filters for thick specimen imaging is the elimination of the multiple inelastically scattered electrons. Thus the optimal restoration for thick specimens would utilize through focus series taken with only elastic electrons.

While energy filtration may be the most useful filter for thick specimen imaging, through focus restoration will always be required to eliminate microscope lens aberrations and to retrieve the phase and amplitude components of the exit wave. For electron microscope tomographic reconstructions and other analytical studies, routine restoration using 30 focus levels is impractical due to excessive beam exposure (specimen damage). Thus it is useful to develop techniques using fewer focus levels and restore images that exclude the multiple inelastic scattering component. An empirical model for multiple scattering derived from the experimental results is proposed in a separate paper (manuscript in preparation).

These results demonstrate that it is inaccurate to directly relate image intensities to specimen mass densities for thick specimen imaging, as is commonly assumed. This will have direct implications for the three-dimensional reconstructions of thick biological specimens.

Acknowledgments:

We thank M. Op de Beeck, D. Van Dyck, M. Gustafsson, W. Liu and H. de Ruijter for helpful discussions; A. Gubbens for the use of the Gatan Imaging Filter; J.Fung for providing the specimens for this study. K.F.H. is supported by the Howard Hughes Medical Institute Predoctoral Fellowship in the Biological Sciences. This work is supported by grants from the National Institutes of Health (GM 31627 for D.A.A.; GM25101 for J.W.S.) and by Howard Hughes Medical Institute. This chapter is an approved reprint of the material as it appears in K.F.Han, J.W.Sedat and D.A.Agard (1995). *J. Microscopy* **178**,107-119.

References:

Bazett-Jones, D. (1992). Electron spectroscopic imaging of chromatin and other nucleoprotein complexes. *Electron Microscopy Review,* **5**, 37-58.

Belmont, A.S., Braunfeld, M.B., Sedat, J.W. & Agard, D.A. (1989). Large-scale chromatin structural domains within mitotic and interphase chromosomes in vivo and in vitro. *Chromosoma,* **98**, 129-143.

Belmont, A.S., Sedat, J.W. & Agard, D.A. (1987). A three-dimensional approach to mitotic chromosome structure: Evidence for a complex hierarchical organization. *Journal of Cell Biology,* **105**, 77-92.

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

Chen, H., Clyborne, W., Sedat, J. & Agard, D. (1994). PRIISM: An integrated system for display and analysis of 3D microscope images. *SPIE:Biomedical Image Processing and 3-Dimensional Microscopy,* **1660**, 784-790.

Coene, W., Janssen, G., Op de Beeck, M. & Van Dyck, D. (1992). Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. *Physical Review Letters,* **69**, 3743-3746.

Colliex, C., Mory, C., Olins, A., Olins, D. & Tence, M. (1989). Energy filtered STEM imaging of thick biological sections. *Journal of Microscopy,* **153**, 1-21.

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

de Ruijter, H. (1992). Quantitative high resolution electron microscopy and holography.

de Ruijter, W.J. & Weiss, J.K. (1993). Detection limits in quantitative off-axis electron holography. *Ultramicroscopy,* **50**, 269-283.

Frank, J. & Radermacher, M. (1992). 3-Dimensional reconstruction of single particles negatively stained or in vitreous ice. *Ultramicroscopy,* **46**, 241-262.

Fung, J.C., Agard, D.A. & Sedat, J.W. (1994). Three-dimensional reconstruction of the synaptonemal complex from high-pressure frozen maize meiocytes using IVEM tomography. *Proc. 53rd Ann. Microscopy Society of America,* 14-15.

Gribelyuk, M. & Cowley, J. (1993). Determination of experimental imaging conditions for off-axis transmission electron holography. *Ultramicroscopy,* **50**, 29-40.

Gubbens, A. & Krivanek, O. (1993). Applications of a post-column imaging filter in biology and material science. *Ultramicroscopy,* **51**, 146-159.

Hohenberg, H., Mannweiler, K. & Muller, M. (1994). High pressure freezing of cell suspensions in cellulose capillary tubes. *Journal of Microscopy,* **175**, 34-43.

Joy, D., Zhang, Y., Zhang, X. & Hashimoto, T. (1993). Practical aspects of electron holography. *Ultramicroscopy,* **51**, 1-14.

Koster, A.J., Chen, H., Sedat, J.W. & Agard, D.A. (1992a). Automated microscopy for electron tomography. *Ultramicroscopy,* **46**, 207-227.

Koster, A.J. & W., d.R. (1992b). Practical autoalignment of transmission electron microscopes. *Ultramicroscopy,* **40**, 89-107.

Krivanek, O.L., Gubbens, A.J., Dellby, N. & Meyer, C.E. (1992). Design and 1st applications of a post-column imaging filter. *Microscopy Microanalysis Microstructures,* **3**, 187-199.

Langmore, J.P. & Smith, M.F. (1992). Quantitative energy-filtered electron microscopy of biological molecules in Ice. *Ultramicroscopy,* **46**, 349-373.

Lichte, H. (1986). Electron holography approaching atomic resolution. *Ultramicroscopy,* **20**, 293-304.

Lichte, H. (1993). Parameters for high-resolution electron holography. *Ultramicroscopy,* **51**, 15-20.

McCartney, M.R. & Gajdardziskajosifovska, M. (1994). Absolution measurement of normalized thickness, T/Lambda(i), from off-axis electron holography. *Ultramicroscopy,* **53**, 283-289.

Olins, A., Olins, D., Levy, H., Margle, S., Tinnel, E. & Durfee, R. (1989). Tomographic reconstruction from energy-filtered images of thick biological sections. *Journal of Microscopy,* **154**, 257-265.

Reimer, L. (1984). Transmission Electron Microscopy. Berlin, Springer-Verlag.

Reimer, L. (1989). Calculation of the angular and energy distribution of multiple scattered electrons using Fourier transforms. *Ultramicroscopy,* **31**, 169-176.

Reimer, L., Rennekamp, R., Fromm, I. & Langenfeld, M. (1991). Contrast in the electron spectroscopic imaging mode of a TEM. IV. Thick specimens imaged by the most-probable energy loss. *Journal of Microscopy,* **162**, 3-14.

Taniguchi, Y., Ikuta, T. & Shimizu, R. (1991). Assessment of image formation by three-dimensional power spectrum in transmission electron microscopy. *Journal of Electron Microscopy,* **40**, 5-10.

Turner, J.N. (1981). Introduction to Stereo Imaging. Three-dimensional Ultrastructure in Biology, Methods in Enzymology. New York, Academic Press.

Van Dyck, D. & Op de Beeck, M. (1990). New direct methods for phase and structure retrieval in HREM. *Proc. of 12th Int'l Congress for Electron Microscopy,* 26-27.

Van Dyck, D., Op de Beeck, M. & Coene, W. (1993). A new approach to object wave function reconstruction in electron microscopy. *Optik,* **93**, 103-107.

Figure and Table Captions:

Figure 1: Polycrystalline gold taken at 10 um underfocus with (A) and without (B) a 20 um (5 mrad) objective aperture. The difference image (C) shows specimen features, demonstrating an increase in signal by the application of the aperture. Scale bar: 50 nm.

Figure 2: 0.3 um thick specimen of HeLa chromatin taken at 10 um underfocus with (A) and without (B) a 20 um (5 mrad) objective aperture. The difference image (C) shows no specimen features, demonstrating a decrease in background by the application of the aperture. Scale bar: 100nm.

Figure 3: Electron energy loss spectra (EELS) for a 0.7 um specimen at 200 keV using no aperture and 40, 30, and 10 μm apertures. The spectra are normalized by the peak elastic electron counts demonstrating that more inelastic electrons scatter outside the objective aperture than do elastic electrons.

Figure 4: A through E are images taken at zero energy-loss (elastic scattering), 15eV, 30 eV, 45 eV, 60 eV and 75 eV-loss, demonstrating that only the elastic and plasmon images contain high resolution information. All other inelastic images contribute at low resolutions and are derived largely from multiple scattering. The average relative contrast is 8.73E-2, 3.95E-2, 6.00E-2, 4.41E-2, 3.92E-2 and 3.89E-2 respectively. Relative contrast is calculated as: $\{\sum | I(i,j) - mean| / mean\} / (nx*ny)\}$, summed over each pixel. Scale bar: 50 nm.

Figure 5: Electron energy-loss spectra of 0.3, 0.5 and 0.7 um thick specimens at 200 keV normalized by total electron counts (A) and normalized by peak elastic counts (B).

Figure 6: Selected diffractograms of a through focus series: 17.28 , 8.06 and 3.46 um underfocus (A, B, and C) and overfocus (D, E, and F). Resolution limit (the distance between the center to the edge of the diffractograms) : 2.68 nm^{-1}.

Figure 7: Schematic diagram and selected cross sections through the Ewald sphere of a 0.3 um thick specimen at 300 KeV, demonstrating significant coherent transfer. B: 17.3 μm^{-1}, C: 12.7 μm^{-1}, D: 6.9 μm^{-1}. Resolution limit: 2.68 nm^{-1}.

Figure 8: Fitting the coherent, incoherent and Poisson noise components for the radial average of a cross section through the Ewald sphere.

Figure 9: Resultant fits for the coherent (A), incoherent (B) and the Poisson noise (C) components for the radial averaged cross sections through the Ewald sphere. Labels in 9A, 9B, and 9C represents the coherent components of the cross sections shown in figures 7B, 7C, and 7D.

Figure 10: Two selected cross sections through the Ewald Sphere comparing specimens of 0.3 (A,E), 0.4 (B,F), 0.5 (C,G) and 0.6 um (D,H) specimen thicknesses. Resolution limit: 2.68 nm^{-1}.

Table I. For each of the thicknesses, the first column lists the specimen thickness; the second lists the percent coherent component obtained from the Ewald Sphere mapping; the third lists the percent elastically scattered electrons approximated at 300 keV; the forth lists the measured percent of elastically scattered electrons at 200 keV using the Gatan Imaging Filter.

Figure 11: Log plot of Table I, demonstrating the correlation of the degree of coherence (open circles) with the amount of elastic scattering (open squares) as a function of specimen thickness.

Figure 12: Restored amplitude (A) and phase (C) images of the exit surface wave for a 0.3 um specimen and the respective diffractograms (B & D). Scale bar is 50 nm and the resolution limit is 2.68 nm^{-1}.

Figure 13: Restored amplitude (A) and phase (C) images of the exit surface wave for a 0.7 um specimen and the respective diffractograms (B & D). Scale bar is 50 nm and the resolution limit is 2.68 nm^{-1}.

Figure 1

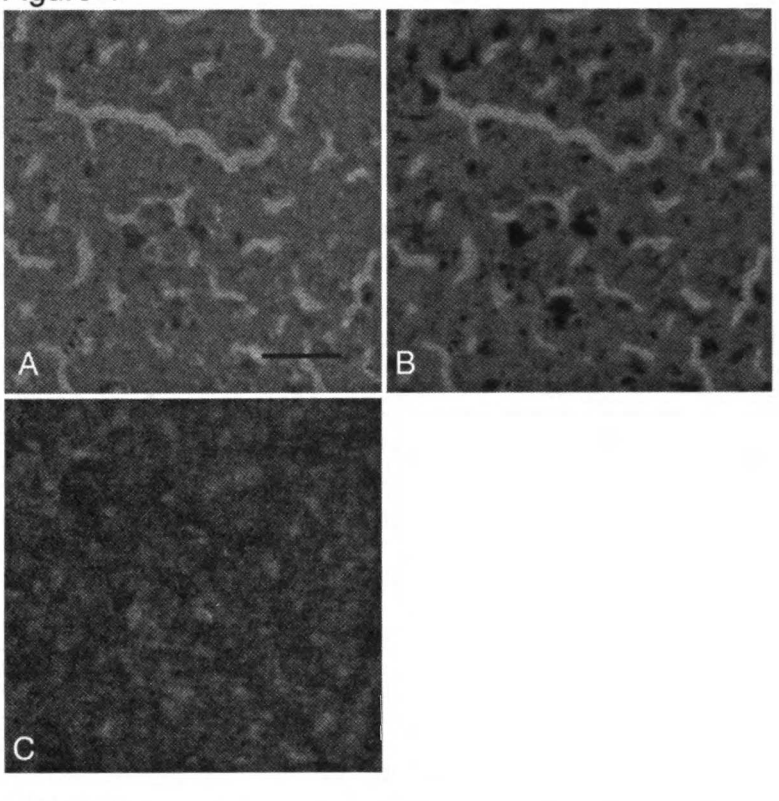


Figure 2

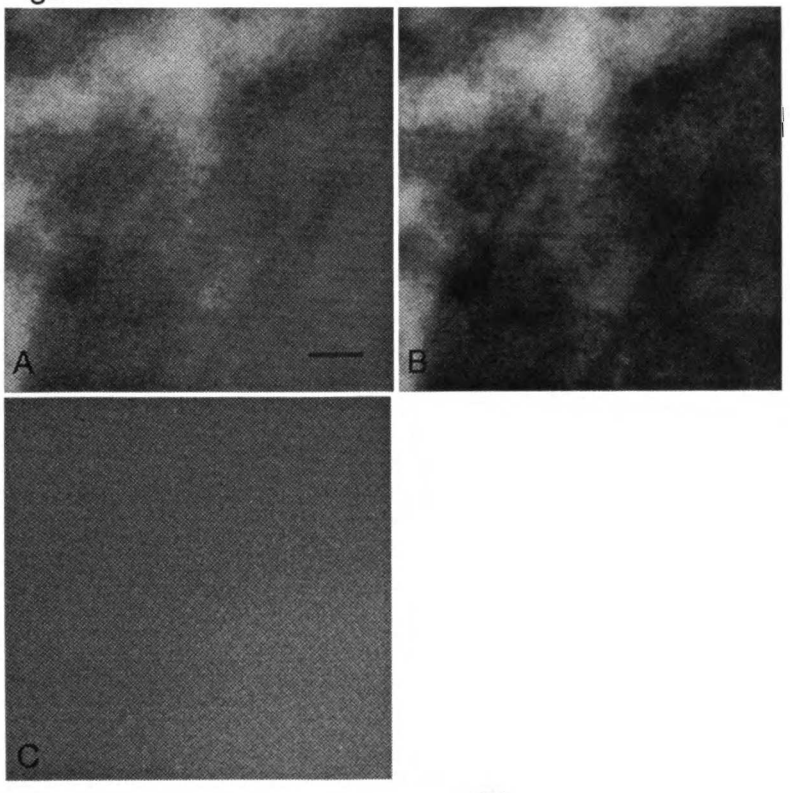


Figure 3

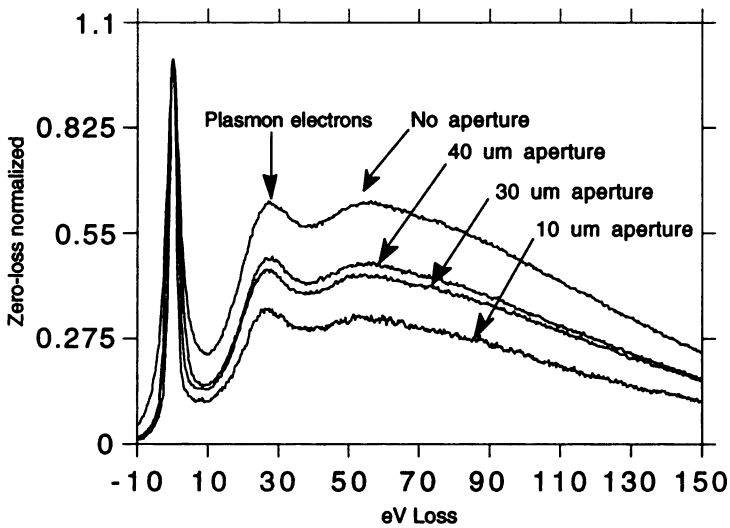

1.1
0.825
0.55
0.275
0
Zero-loss normalized
Plasmon electrons
No aperture
40 um aperture
30 um aperture
10 um aperture
-10 10 30 50 70 90 110 130 150
eV Loss

Figure 4.

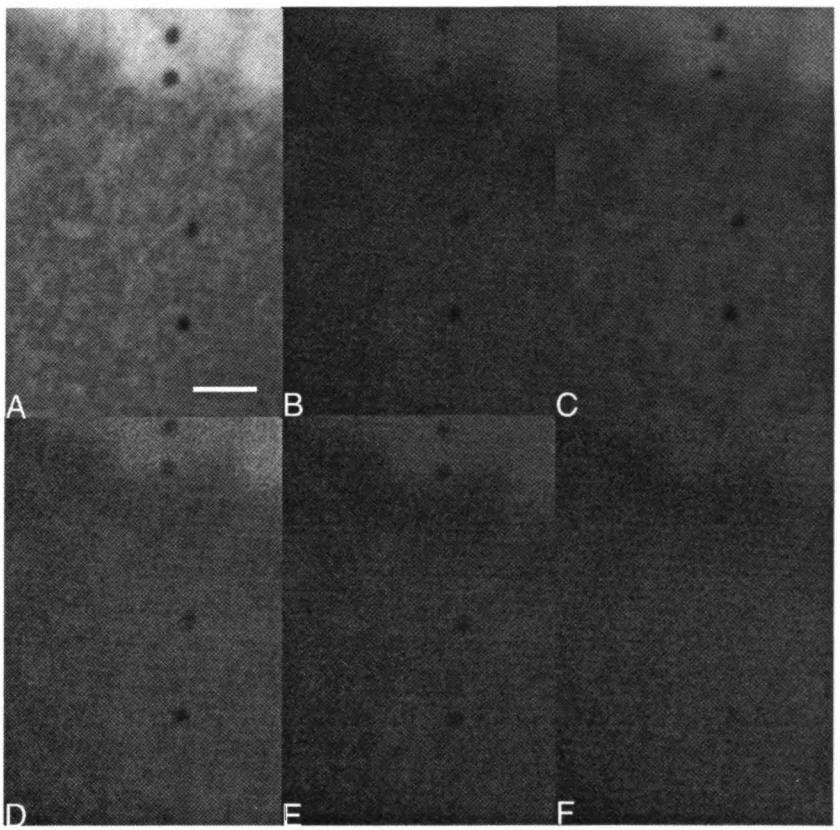


Figure 5a

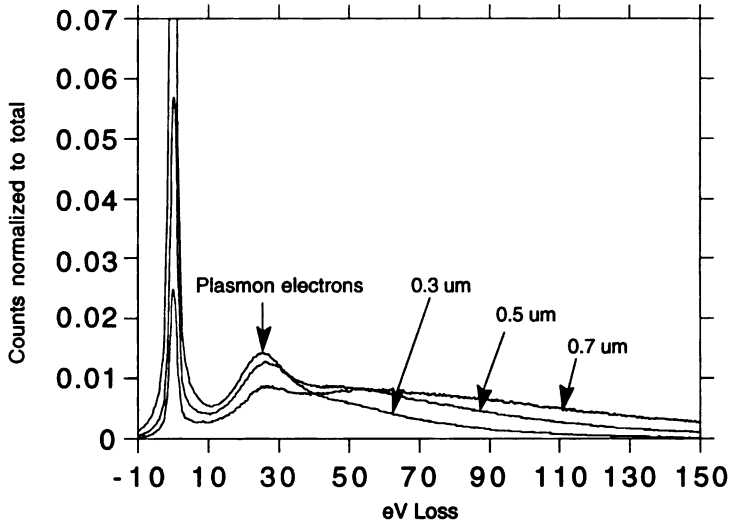


Figure 5b

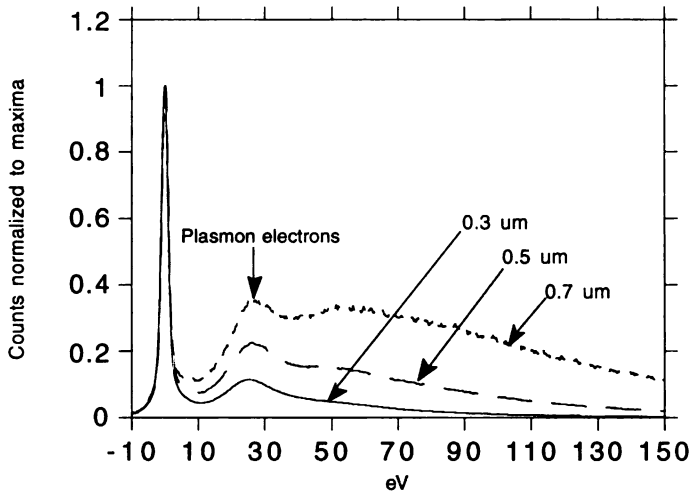


Figure 6.

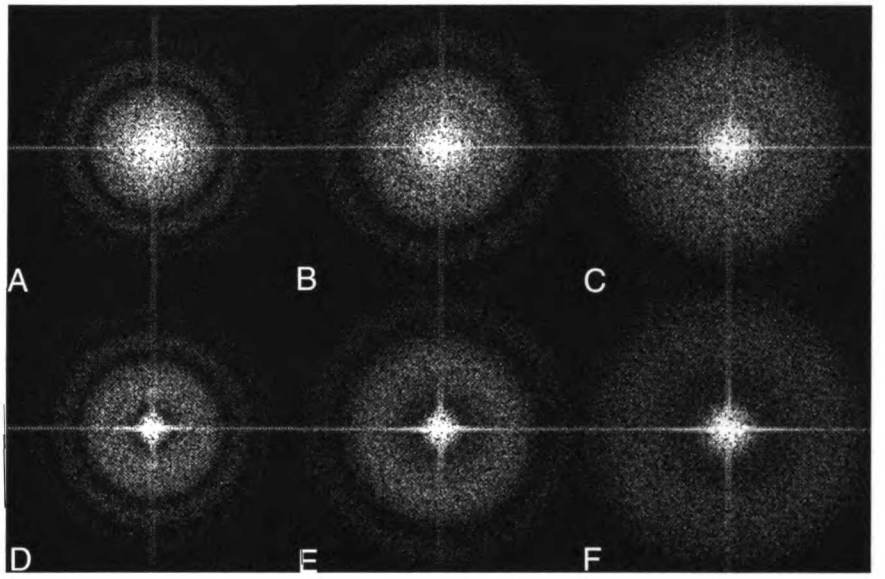


Figure 7

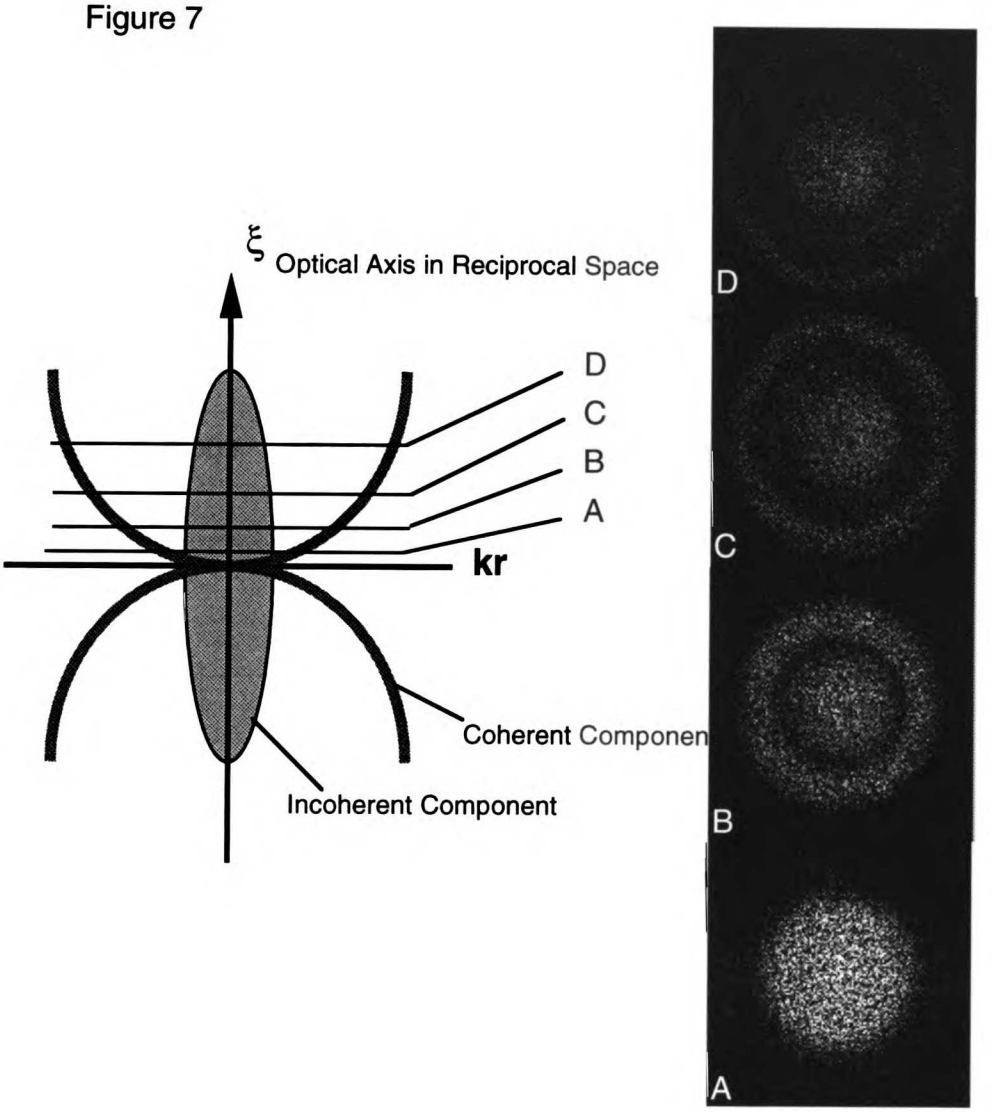

ξ
Optical Axis in Reciprocal Space
D
C
B
A
kr
Coherent Componen
Incoherent Component
D
C
B
A

Figure 8

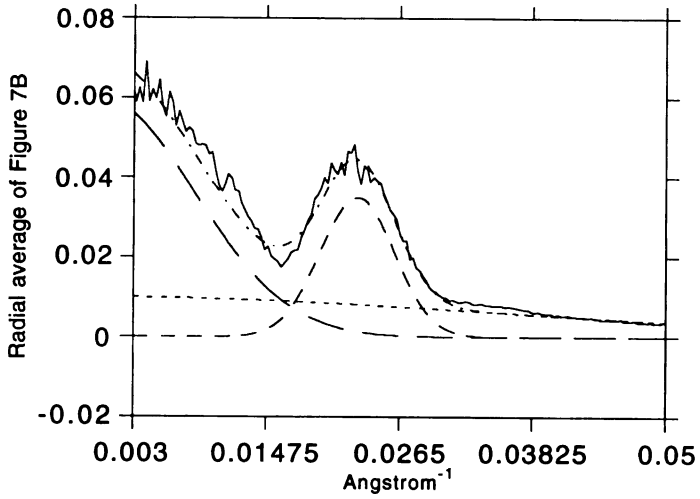


Figure 9a

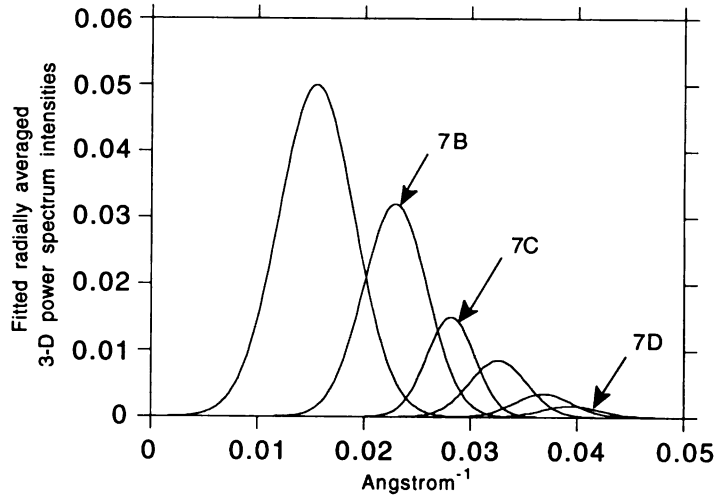


Figure 9b

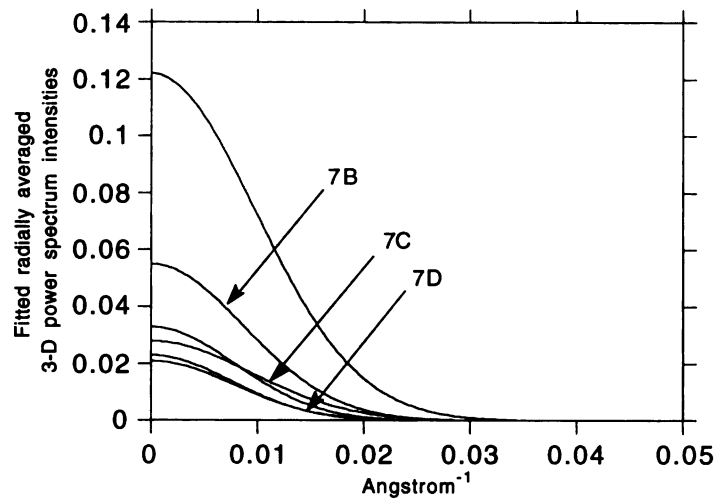


Figure 9c

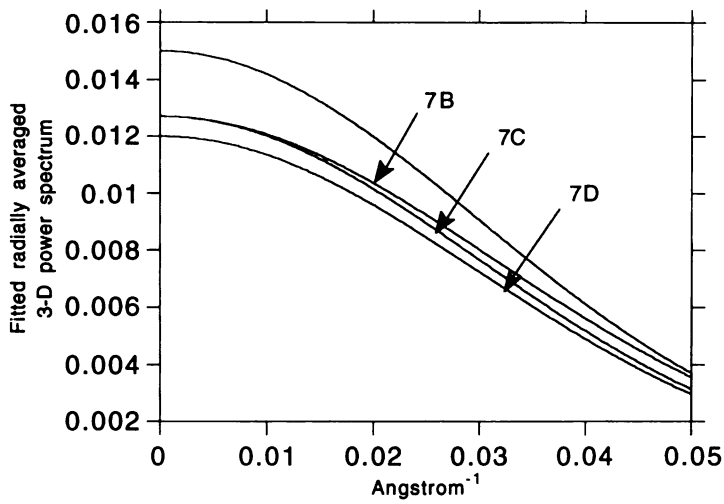
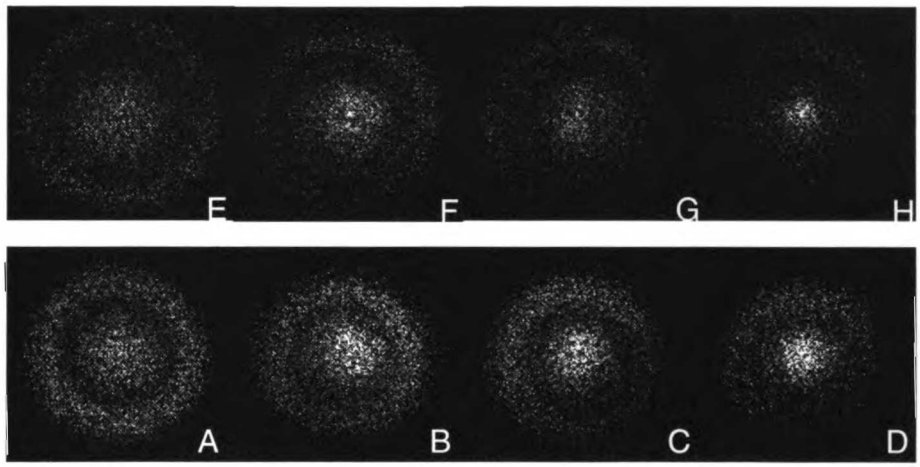


Table I

Specimen thickness (μm)	% Coherent component 300 keV	% Elastic electrons at 300 keV*	% Elastic electrons at 200 keV
0.3	22.4%	58%	39.7%
0.5	13.0%	40%	18.3%
0.6	9.8%	34%	---
0.7	7.8%	28%	8.3%

Figure 10.



0.3 um | 0.4 um | 0.5 um | 0.6 um

Figure 11

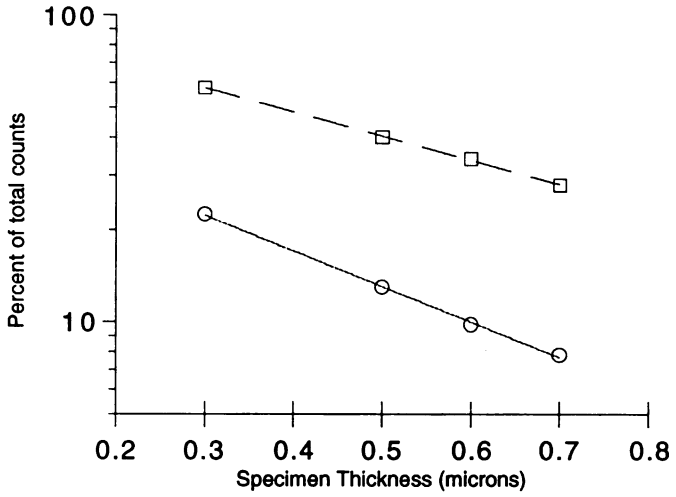


Figure 12.

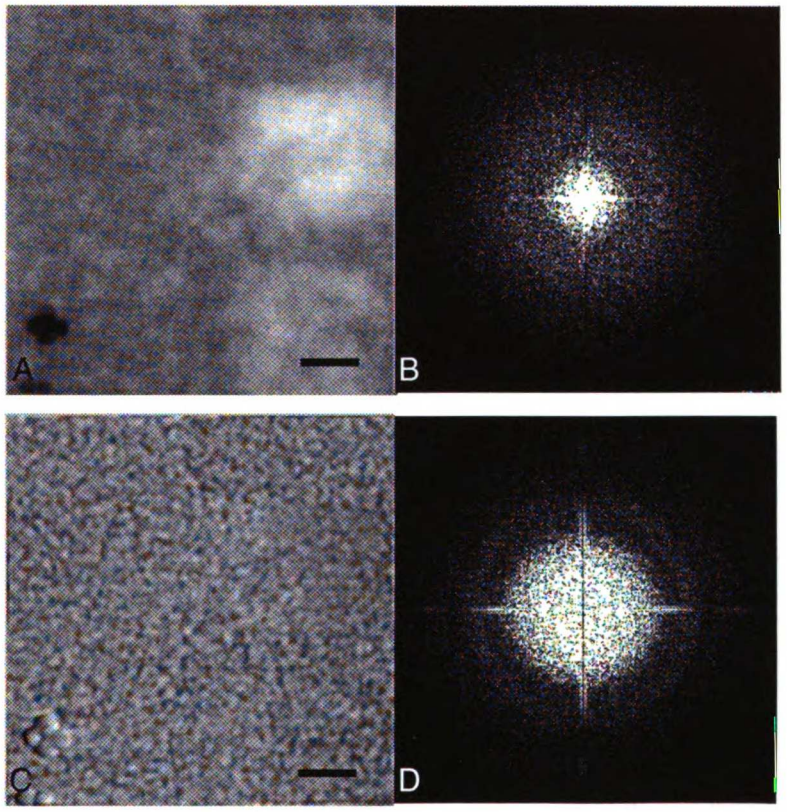
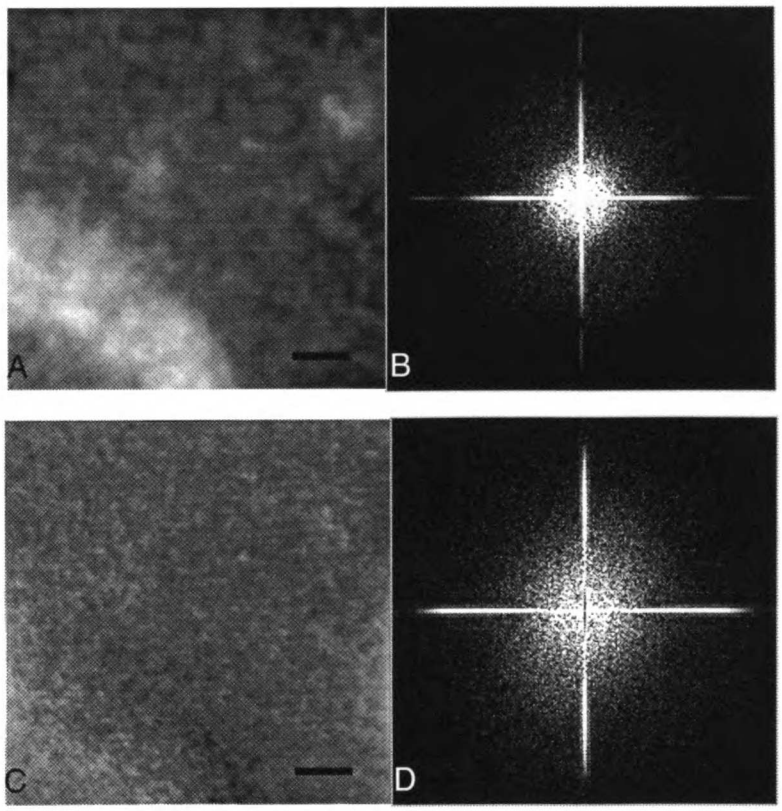


Figure 13.



Chapter 3

Optimal strategies for imaging thick biological specimens: exit wavefront reconstruction and energy-filtered imaging.

In transmission electron microscopy (TEM) of thick biological specimens, the relationship between the recorded image intensities and the projected specimen mass density is distorted by incoherent electron-specimen interactions and aberrations of the objective lens. It is highly desirable to develop a strategy for maximizing and extracting the coherent image component, thereby allowing the projected specimen mass density to be directly related to image intensities. For this purpose, we previously used exit wavefront reconstruction to understand the nature of image formation for thick biological specimens in conventional TEM. In this paper, exit wavefront reconstruction is used in combination with electron-energy filtering to quantitatively study the imaging properties of the microscope. We found that for imaging thick biological specimens (> 0.5 μm) at 200 keV, only elastically scattered electrons contribute to the coherent image component. Surprisingly little coherent transfer was seen when using energy-filtering at the most probable energy-loss (in this case at the first plasmon energy-loss peak). Furthermore, the use of zero-loss filtering in combination with exit wavefront reconstruction is considerably more effective at removing the effects of multiple (inelastic) scattering and microscope objective lens aberrations than either technique by itself. Optimization of the zero-loss signal requires operation at intermediate to high primary voltages (>200 keV). These results have important implications for the accurate recording of images of thick biological specimens as for instance in electron microscope tomography.

Introduction and background:

High resolution three-dimensional structural analysis of complex biological samples is best carried out using electron microscope tomography. This method allows the computational reconstruction of internal specimen structure in three dimensions using a large number of tilted projections (Turner 1981). Tomography has proven to be a powerful technique to study the supramolecular assemblies of cellular organelles and nuclear structures (Belmont et al. 1987; Fung et al. 1994; Horowitz et al. 1994; Ladinsky et al. 1994; Moritz et al. 1995; Olins et al. 1994; Schmekel et al. 1993). The accuracy of the reconstruction relies on the precise interpretation of the individual projected views. Due to the large depth of focus in electron microscopes, it is generally assumed that the recorded image intensities can be directly related to the projected specimen mass densities. This is only correct if the scattering were purely coherent and not affected by objective lens aberrations. Images of thin biological specimens (<0.1 μm) are generally dominated by coherent single scattering, however, significant aberrations occur for thick specimens (>0.3 μm) as a result of the large fraction of incoherent multiple elastic and inelastic scattering. In this paper, we define the coherent component as the component which follows the expected behaviour of the microscope contrast transfer function (CTF). The incoherent component not only includes the secondary interference between scattered waves (quadratic term in Eq. 1), but also the larger component contributed by multiple scattering where a systematic behaviour (or CTF) cannot be defined. For accurate quantitative analysis of specimen structure, the effects from the incoherent multiple scattering and the objective lens aberrations must be removed. As we demonstrated previously (Han et al. 1995), specimen exit surface wavefront reconstruction using a

through focus series removes the effect of the objective lens aberrations and extracts and quantifies the coherent image component. This exit wavefront reconstruction method uses a 3-D Fourier Transform of a through focus series: the coherent imaging component maps onto a paraboloid whereas the incoherent components dominate other regions, especially the center of the transform. Adopted from van Dyck et al, the relationship between the 3-D Fourier Transform and the specimen exit surface wavefront is as follows (van Dyck et al. 1993):

$$\mathbf{I}(\mathbf{k},\zeta) = |C|^2\,\delta(\mathbf{k}) + C^*\phi(\mathbf{k})\delta(\zeta - \lambda|\mathbf{k}|^2/2) + C\phi^*(-\mathbf{k})\delta(\zeta + \lambda|\mathbf{k}|^2/2)$$
$$+ \int_{\mathbf{k}\neq 0,(\mathbf{k}-\mathbf{k}')\neq 0} \phi^*(\mathbf{k})\phi(\mathbf{k}-\mathbf{k}')\delta\{\zeta - \lambda[(\mathbf{k}-\mathbf{k}')^2 - \kappa^2]/2\}d\mathbf{k}'$$

where $\mathbf{k}$ and ζ are reciprocal axes for x, y and z respectively; δ is the Dirac delta function; λ is the electron wavelength and ϕ is the specimen exit surface wavefront. By extracting the parabolic component, $\delta(\zeta \pm \lambda|\mathbf{k}|^2)$, the unaberrated exit surface wave (the coherent component) can be recovered and quantified (Figure 3a, schematic). Note that a chromatic aberration disc contribution to the additional focus spread is not included in the above equation since the through focus sampling required to resolve such a spread is impractical (Saxton 1994).

In recent years the advent of commercially available electron spectroscopic imaging (ESI) filters have made it possible to image specimens at specific energy-loss ranges and have enabled the analysis of the contribution of the various energy-loss ranges to the image formation. Currently two classes of ESI instrumentation can be distinguished: 1) 'in column' filters currently limited to operation up to 120 keV primary voltage (Zeiss 902, 912(Probst et al. 1993)) and 2) 'post-column' filters which are available for use up to 1250 keV (Gatan GIF and HV GIF (Gubbens et al. 1995; Gubbens et al. 1993; Krivanek et al.

1995; Krivanek et al. 1992)). The study described here was done with a post-column filter because, as we will show for thick biological specimens, it is optimal to image at intermediate to high accelerating voltages, and currently only the post-column filters can be used at these voltages. To do a true comparison of filtered versus unfiltered imaging and to verify that the filter does not effect the normal imaging properties of the microscope, control-experiments were performed under identical imaging conditions using only a CCD camera identical to the one used in the filter on the same TEM. We present the first comparison of this kind and the first quantitative assessment of the quality of energy filtering for thick biological specimens.

By combining the through focus exit wavefront reconstruction analysis with energy-filtering, it is possible to assess the contribution of different energy-loss ranges to the high resolution coherent and the low resolution incoherent image components. These results shown in the following section have important implications on the use of ESI for electron microscope tomography and stress the importance of combining ESI with intermediate to high acceleration voltages.

Methods:

Thick Biological Specimens: The specimens used in the following experiments are *in vitro* reconstituted centrosomes from Drosophila embryos embedded in epon and stained with uranyl-acetate and lead-citrate (Moritz et al. 1995). The microtubules are 25 nm in diameter. The specimens were cut to 0.5 and 0.7 μm thickness. The preparation of these specimens is described in detail in Moritz et. al, 1995.

Unfiltered and Energy-filtered Imaging: The unfiltered images were recorded at 200 keV with a Gatan Model 694 slow-scan CCD camera mounted

on a Philips CM200 SuperTwin TEM. The energy-filtered images were recorded at 200 keV with a Gatan Imaging Filter, Model GIF100 (Gubbens et al. 1993; Krivanek et al. 1992) mounted on the same CM200. The slow-scan CCD camera used on the GIF100 is identical to the Model 694. All images were recorded at a calibrated magnification of 40,000 times at the CCD. A 30 μm objective aperture was used for all experiments. The images were binned two times in the camera hardware resulting in an effective pixel size projected back to the specimen of 1.20 nm. The energy-window used for energy-filtered imaging was 10 eV, the energy dispersion used for recording the energy-loss spectra was 0.5 eV per CCD pixel.

Through Focus Series: Through focus series consisting of 41 images were recorded from 18.1 μm under focus to 18.1 μm over focus with a focus step size of 0.905 μm. To minimize specimen alterations and shrinkage during data collection, the specimens were stabilized by pre-irradiating with approximately 1000 e^-/nm^2 (Braunfeld et al. 1994). The individual images were aligned prior to the exit surface wavefront reconstruction using fudicial gold markers and cross-correlation. Image processing and visualization were done on a DEC VAX-9000 and a Silicon Graphics Iris workstation using Priism, the image visualization software developed in our laboratory (Chen et al. 1994).

Results:

It has been generally assumed that the mechanism of image formation for thick biological specimens is dominated by amplitude contrast. However, as we previously demonstrated (Han et al. 1995), there is still a considerable coherent image component largely due to phase contrast. Superimposed on the coherent component are the incoherent components dominated by multiple elastic and inelastic scattering. We have used exit surface wavefront

reconstructions from through focus series at different energy-loss ranges, to evaluate the coherent image component as a function of resolution for 0.5 and 0.7 μm thick biological specimens at 200 keV primary energy.

For the 0.5 μm thick specimen, coherent transfer can be observed at both the zero-loss and the plasmon energy-loss, whereas for the 0.7 μm thick specimen coherent transfer can only be observed at the zero-loss.

Figure 1 shows energy-loss spectra recorded for the 0.5 and 0.7 μm thick specimens used in this study. At 200 keV, the fraction of elastically scattered electrons is very low for both thicknesses (Table I). For the through focus series of the 0.5 μm thick specimen, diffraction rings could be observed with and without zero-loss filtering, and also weakly in the plasmon-loss series (data not shown), although they could be observed most clearly in the zero-loss filtered series. For the 0.7 μm thick specimen the diffraction rings observed in the unfiltered through-focus series were extremely weak (data not shown). By contrast, the zero-loss filtered through-focus series showed significant phase transfer (Fig. 2). Quite surprisingly, through-focus series at other energy-loss ranges all showed a complete absence of diffraction rings. Thus, at 200 keV, 0.7 μm thick biological specimens contain phase contrast information which is obscured by the large proportion of inelastically and multiply scattered electrons.

The coherent image component is significantly enhanced by zero-loss filtering for both 0.5 and 0.7 μm thick specimens.

Figures 3 and 4 show selected cross-sections of the 3-D power spectra for the unfiltered, zero-loss filtered and plasmon-loss filtered though focus series for the 0.5 and 0.7 μm specimens respectively. The paraboloid (coherent)

component is narrower and extends to higher resolution in the zero-loss filtered power spectrum compared to the unfiltered power spectrum. Virtually all of the coherent component is contributed by the zero-loss electrons as evident by the fact that the plasmon-loss and other energy-loss ranges (data not shown) contain very little coherent contribution.

It is important to realize that not all zero-loss electrons contribute to the coherent component of the images. There is still a significant amount of incoherent component in the zero-loss filtered images due to multiple elastic scattering.

For both the 0.5 and 0.7 μm thick specimens the elastically scattered electrons contribute a similar amount of coherent high resolution information.

In order to quantitatively analyze the 3-D power spectrum, cross sections of the spectra were radially averaged and then curve-fitted with three components as previously described (Han et al. 1995). The three components are the parabola (coherent), the center (partially (in)coherent), and the background (fully incoherent and random electron statistical noise) (Fig. 3 schematic and Fig. 5). Three component curve fits were obtained for all sections and all energy ranges and the results of the experiments where there was a measurable non-zero coherent component are summarized in Table I. The percentages listed in Table I represent the fraction of the total number of electrons (taken as the sum of all three components). The proportion of coherent electrons of the filtered images are comparable for both thicknesses: 13% for 0.5 μm and 9% for 0.7 μm. Although energy filtering enhances the coherent component by an overall decrease in the incoherent contribution, the change in the relative distribution of the three components are different between the two thicknesses tested. Both

the parabola and the central components vary as a function of focus (cross-section) levels, whereas the background component stays relatively constant. As with the central component, this background is also contributed by the incoherent imaging effects. In the case of the 0.5 μm thick specimen, the slight gain in the proportion of the coherent component in the zero-loss filtered image is accompanied by a larger decrease in the central component and a slight increase in the background component. Whereas for the 0.7 μm specimen, the more dramatic gain in the coherent component is accompanied by a larger drop in background and a slight gain in the central component. This may be explained, if the central and background components are due to different types of incoherent, or partially coherent scattering each dominating at different thicknesses. For both thicknesses, the plasmon-loss images show a proportionally larger background component with very little coherent component due to the overlapping contribution from the multiple inelastic scattering (the broadened second plasmon-loss in Figure 1). No coherent contribution was observed at any of the other energy ranges. It is important to note again that not all of the elastically scattered electrons contribute to the coherent component of the images.

Figure 6 shows the number of electrons contributing to the coherent component as a function of resolution. Although the zero-loss filtered images of the 0.7 μm thick specimen contain a smaller coherent component overall, the variation as a function of resolution is the same as for the 0.5 μm thick specimen. In the case of the 0.5 μm thick specimen, the coherent component is only slightly reduced in the unfiltered versus zero-loss filtered images through all resolutions. This again demonstrates that the relative enhancement in the zero-loss filtered signal is simply due to the reduction of the central and

background components. In this case, exit wavefront restoration can almost completely recover the coherent image component.

For the 0.7 μm thick specimen, the striking observation is that in the unfiltered images, the coherent component is greatly reduced beyond $(3.8nm)^{-1}$ resolution. Thus in this case, the relative signal enhancement in the filtered images is due to a combination of the reduction of the incoherent central and background components, as well as an increase in the coherent high resolution component. Here, exit wavefront reconstruction by itself can not completely recover the coherent component through the entire resolution range.

Exit wavefront restored images from a zero-loss filtered through focus series show much higher resolution. Figures 7 and 8 compare unfiltered and zero-loss filtered exit surface wavefront restored images and their respective diffractograms for the 0.5 and 0.7 μm thick specimens. As can be seen, the unfiltered restorations contain only the very low resolution components, whereas, the energy-filtered restorations show transfer up to much higher resolutions. Figure 9 plots the power spectra of scaled contrast images comparing the unfiltered and zero-loss filtered data and restorations. For each image, the contrast at each pixel was calculated

$$I'(x,y) = \frac{I(x,y) - I_{ave}}{I_{ave}}$$

and then subsequently scaled to a constant intensity range for comparison between the different images. For a 0.5 μm thick specimen, the 0.9 μm underfocused filtered data has higher contrast than the unfiltered data at the same focus level at resolutions higher than $(10\ nm)^{-1}$. With this conservative measure of image contrast, the through focus series restoration shows dramatically enhanced signal between $(14\ nm)^{-1}$ and $(4.5\ nm)^{-1}$, compared to the unrestored images. Importantly, the through focus series restoration of the

zero-loss filtered images shows a further enhancement in contrast (Fig 9A). Not surprisingly, the restoration of the unfiltered through focus series of the 0.7 μm thick specimen show little contrast enhancement compared to the data, indicative of the low coherence in the unfiltered series. While the zero-loss data does show improvement over the unfiltered data at higher frequencies, dramatic improvements are seen when energy filtering is combined with through focus restoration.

Discussion:

Electron spectroscopic imaging has been used previously to study the mechanism of image formation for thick specimens (Bazett-Jones 1992; Colliex et al. 1989; Langmore et al. 1992; Reimer et al. 1991). These studies showed that when there is little elastic scattering, such as when low accelerating voltages are used, the optimal contrast for thick specimens is obtained by imaging at the most probable energy-loss. The enhancement in resolution arises from the reduction of chromatic aberration as a result of the small energy-window. Although these images show an enhanced contrast and resolution, their image intensities do not properly relate to the projected specimen mass density, which is important for quantitative imaging such as 3-D electron tomography.

We demonstrated that for imaging thick biological specimens at 200 keV there is a significant coherent image component which can be extracted from a through focus series (Figures 7 and 8). This coherent component can be directly related to projected specimen mass density for 3-D tomographic reconstruction. Using ESI, it was shown that this coherent image component is contributed almost solely by elastically scattered electrons. The plasmon-loss electrons contribute to the coherent component only at very low resolutions.

Proper interpretation of the image intensities can therefore only be achieved using only the elastically scattered electrons. Images collected at energies other than zero-loss, for instance at the most probable energy-loss (in this case, the first plasmon), show significant loss in coherent transfer and cannot be directly interpreted as a projection of the specimen mass density. This emphasizes the importance of using higher accelerating voltages to increase the fraction of elastic scattering for even moderate resolution images of biological specimens.

All the experiments presented in this paper comparing energy-filtered and unfiltered imaging were performed with and without the post-column ESI filter attached. This allowed us to directly assess the imaging properties of the post-column filter. The filter effectively serves as an additional projector lens and indeed the experiments described here clearly demonstrated that the image formation properties are not compromised by the post-column filter. The properties of the TEM equipped with the post-column ESI filter show the same expected behavior as the TEM alone (Figure 3b). Using zero-loss filtering we were able to restore images of thick biological specimens at higher resolutions than possible without filtering.

Thus the most optimal strategy for imaging thick biological specimens, as for instance in electron microscope tomography, is combining zero-loss filtering operating at intermediate to high accelerating voltages and specimen exit surface wavefront reconstruction from a through focus series.

For example, if one is interested in substructures of a 0.5 μm thick biological specimen using tomography (effective thickness of up to 1.0 μm at 60° tilt), at resolutions lower than 5 nm, conventional intermediate voltage electron microscopy should suffice. But if higher resolution is required to

visualize fine substructure, then it is strongly recommended that zero-loss filtering be combined with through focus series restoration to obtain an accurate 3D reconstruction.

Acknowledgments:

The authors thank M. Moritz and M. Braunfeld for providing the centrosome specimens; B. Kraus for assiting in the experiments; W. Liu for helpful discussions. K.F.H. is supported by the Howard Hughes Medical Institute Predoctoral Fellowship in the Biological Sciences. This work is supported by grants from the National Institutes of Health (GM 31627 for D.A.A.; GM25101 for J.W.S.) and by the Howard Hughes Medical Institute. This chapter is an approved reprint of the material as it appears in K.F.Han, A.J.Gubbens, J.W.Sedat and D.A.Agard (1996). *J. Microscopy* (in press).

References:

Bazett-Jones, D. (1992). Electron spectroscopic imaging of chromatin and other nucleoprotein complexes. *Electron Microscopy Review,* **5**, 37-58.

Belmont, A.S., Sedat, J.W. & Agard, D.A. (1987). A three-dimensional approach to mitotic chromosome structure: Evidence for a complex hierarchical organization. *Journal of Cell Biology,* **105**, 77-92.

Chen, H., Clyborne, W., Sedat, J. & Agard, D. (1994). PRIISM: An integrated system for display and analysis of 3D microscope images. *SPIE:Biomedical Image Processing and 3-Dimensional Microscopy,* **1660**, 784-90.

Colliex, C., Mory, C., Olins, A., Olins, D. & Tence, M. (1989). Energy filtered STEM imaging of thick biological sections. *Journal of Microscopy,* **153**, 1-21.

Fung, J.C., Agard, D.A. & Sedat, J.W. (1994). Three-dimensional reconstruction of the synaptonemal complex from high-pressure frozen maize meiocytes using IVEM tomography. *Proc. 53rd Ann. Microscopy Society of America,* 14-15.

Gubbens, A., Kraus, B., Krivanek, O. & Mooney, P. (1995). An imaging filter for high voltage electron microscopy. *Ultramicroscopy,* **59**, 255-65.

Gubbens, A. & Krivanek, O. (1993). Applications of a post-column imaging filter in biology and material science. *Ultramicroscopy,* **51**, 146-59.

Han, K.F., Sedat, J.W. & Agard, D.A. (1995). Mechanism of image formation for thick biological specimens: *exit wavefront reconstruction and electron energy-loss spectroscopic imaging. J. Microscopy,* **178:2**, 107-19.

Horowitz, R.A., Agard, D.A., Sedat, J.W. & 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.

Krivanek, O., Friedman, S., Gubbens, A. & Kraus, B. (1995). An imaging filter for biological applications. *Ultramicroscopy,* **59**, 267-82.

Krivanek, O.L., Gubbens, A.J., Dellby, N. & Meyer, C.E. (1992). Design and 1st applications of a post-column imaging filter. *Microscopy Microanalysis Microstructures,* **3**, 187-99.

Ladinsky, M.S., Kremer, J.R., Furcinitti, P.S., McIntosh, J.R. & Howell, K.E. (1994). HVEM tomography of the trans-Golgi network: structural insights and identification of a lace-like vesicle coat. *J Cell Biol,* **127**, 29-38.

Langmore, J.P. & Smith, M.F. (1992). Quantitative energy-filtered electron microscopy of biological molecules in Ice. *Ultramicroscopy,* **46**, 349-73.

Moritz, M., Braunfeld, M., Fung, J., Alberts, B., Sedat, J. & Agard, D. (1995). Three-dimensional structural characterization of centrosomes from early drosophila embryos. *J. Cell Biol.,* **130**, 1149-59.

Olins, A.L., Olins, D.E., Olman, V., Levy, H.A. & Bazett-Jones, D.P. (1994). Modeling the 3-D RNA distribution in the Balbiani ring granule. *Chromosoma,* **103**, 302-10.

Probst, W., Benner, G., Bihr, J. & Weimer, E. (1993). An omega energy filtereing TEM- principles and applications. *Advanced Materials,* **5**, 297-300.

Reimer, L., Rennekamp, R., Fromm, I. & Langenfeld, M. (1991). Contrast in the electron spectroscopic imaging mode of a TEM. IV. Thick specimens imaged by the most-probable energy loss. *Journal of Microscopy,* **162**, 3-14.

Saxton, W. (1994). What is the focus variation method-is it new- is it direct. *Ultramicroscopy,* **55**, 171-81.

Schmekel, K., Skoglund, U. & Daneholt, B. (1993). The three-dimensional structure of the central region in a synaptonemal complex: a comparison between rat and two insect species, Drosophila melanogaster and Blaps cribrosa. *Chromosoma,* **102**, 682-92.

Turner, J.N. (1981). Introduction to Stereo Imaging. Three-dimensional Ultrastructure in Biology, Methods in Enzymology. New York, Academic Press.

van Dyck, D., Op de Beeck, M. & Coene, W. (1993). A new approach to object wave function reconstruction in electron microscopy. *Optik,* **93**, 103-7.

Figure and Table captions:

Figure 1: Electron energy-loss spectra for the specimens used in these experiments A: 0.5 μm, B: 0.7 μm thick specimens.

Figure 2: Selected diffractograms of zero-loss filtered through focus series of a 0.7 μm specimen. Resolution limit: $(2.4\ nm)^{-1}$. A,D: ±16.29; B,E: ±10.86; C,F: ±7.24 μm defocus.

Figure 3a: Schematic diagram of the three-dimensional power spectra of a through focus series.

Figure 3b: Selected cross-sections of three-dimensional power spectra of unfiltered and zero-loss filtered through focus series of a 0.5 μm specimen. A: $(18.1\ \mu m)^{-1}$; B: $(12.67\ \mu m)^{-1}$;C: $(5.43\ \mu m)^{-1}$;D: $(2.72\ \mu m)^{-1}$. Resolution limit: $(2.4\ nm)^{-1}$.

Figure 4: Selected cross-sections of three-dimensional power spectra of unfiltered and zero-loss filtered through focus series of a 0.7 μm specimen. A: $(12.67\ \mu m)^{-1}$; B: $(5.43\ \mu m)^{-1}$;C: $(2.72\ \mu m)^{-1}$;D: $(1.81\ \mu m)^{-1}$. Resolution limit: $(2.4\ nm)^{-1}$.

Figure 5: The radially averaged plots and curve-fits of Figure 3(D) demonstrating the quality of the fits. The coherent, incoherent and background components are as labeled. Curve fit error is 3.7%.

Figure 6: Plot of the number of parabola electrons in the 3D power spectra as a function of resolution for zero-loss and unfiltered images.

Figure 7: Exit wavefront restored filtered (A) and unfiltered (C) images and the respective diffractograms (B, D) of a 0.5 μm specimen. Scale bar is 60 nm, resolution limit is $(2.4\ nm)^{-1}$.

Figure 8: Exit wavefront restored filtered (A) and unfiltered (C) images and the respective diffractograms (B, D) of a 0.7 μm specimen. Scale bar is 60 nm, resolution limit is $(2.4\ nm)^{-1}$.

Figure 9: Plot of scaled power spectra of the contrast images (see text) comparing filtered and unfiltered data and restoration for (A)0.5 and (B)0.7 μm specimens.

Table I: Summary of the relative proportion of each of the three components as a function of thickness and energy range.

Thick-ness	Energy filter experiment	Elastic elec-trons %	Parabola (coherent) %	Central (partially (in)coherent) %	Background (incoherent & noise) %
0.5 μm	Unfiltered		10.1	33.5	56.4
	Zero-loss		13.1	23.3	63.6
		21.4			
	Plasmon		3.0	13.7	83.3
	30eV-130eV		0.0	16.2	83.8
0.7 μm	Unfiltered		2.8	10.6	86.6
	Zero-loss	8.6	9.3	19.4	71.2
	Plasmon		1.4	15.8	82.8
	30eV-130eV		0.0	24.9	75.1

Summary of the experiments (column 2) done on 0.5 and 0.7 μm specimens at 200 keV: column 3 lists the percent elastically scattered electrons; columns 4, 5 and 6 list the percent parabola (coherent), central (partially coherent and incoherent) and background (incoherent and noise) components.

Fig. 1

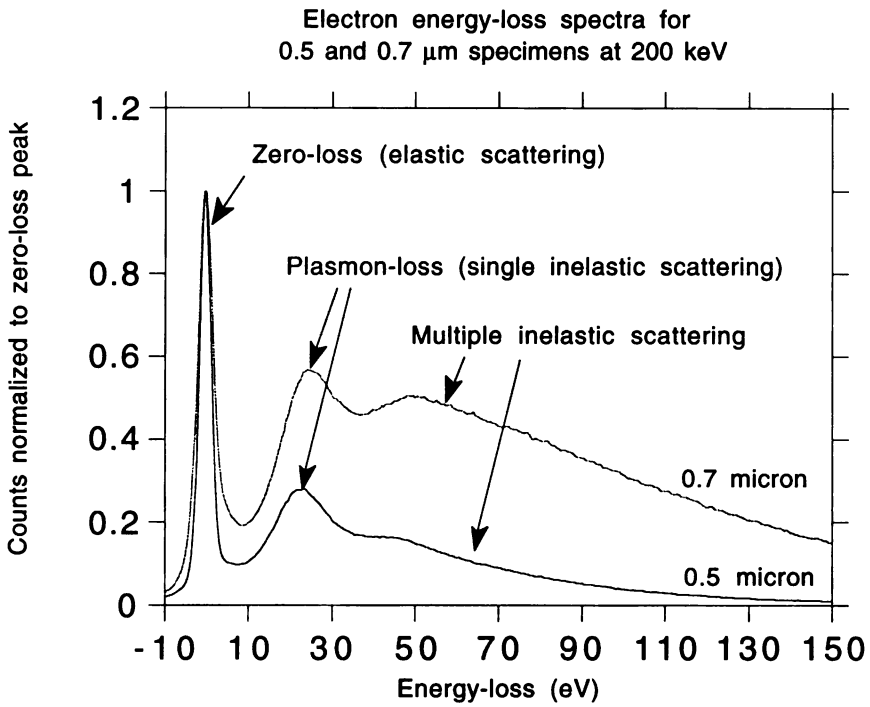
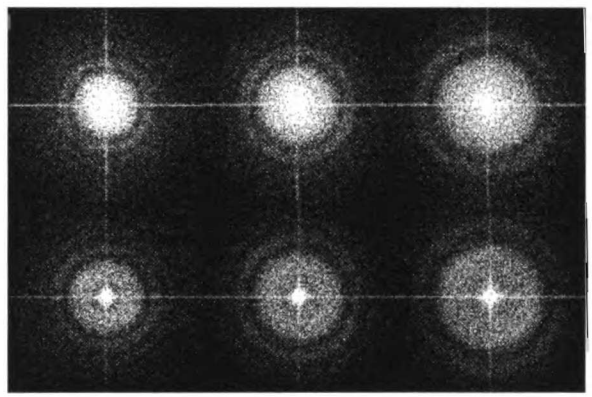


Figure 2

+/- 16.29 μm +/- 10.86 μm +/- 7.24 μm



Selected zero-loss filtered diffractograms of a 0.7 μm thick specimen

Fig. 3a, Schematic diagram of the components in the 3D power spectra of a through focus series.

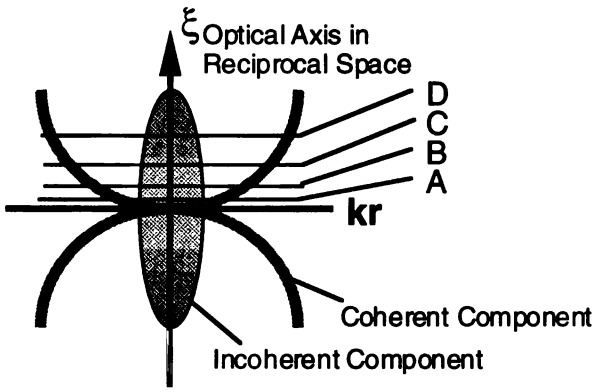
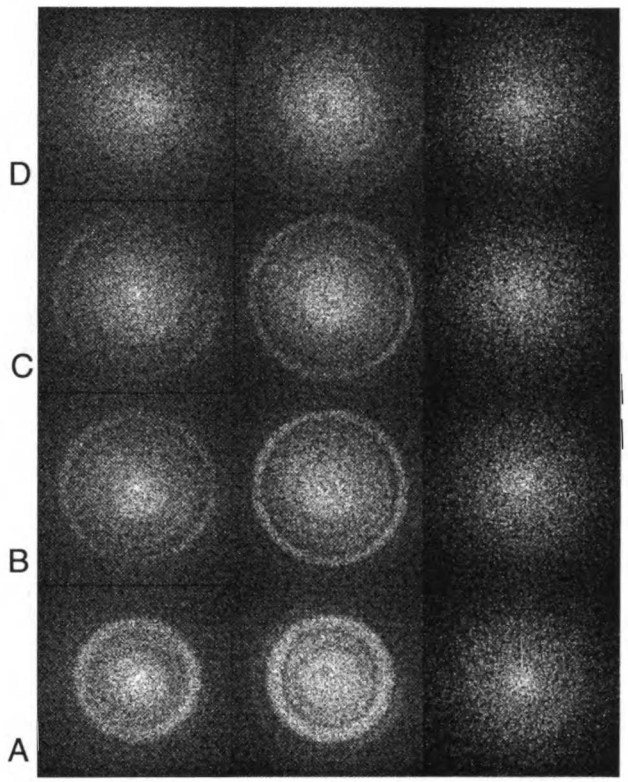


Fig. 3b

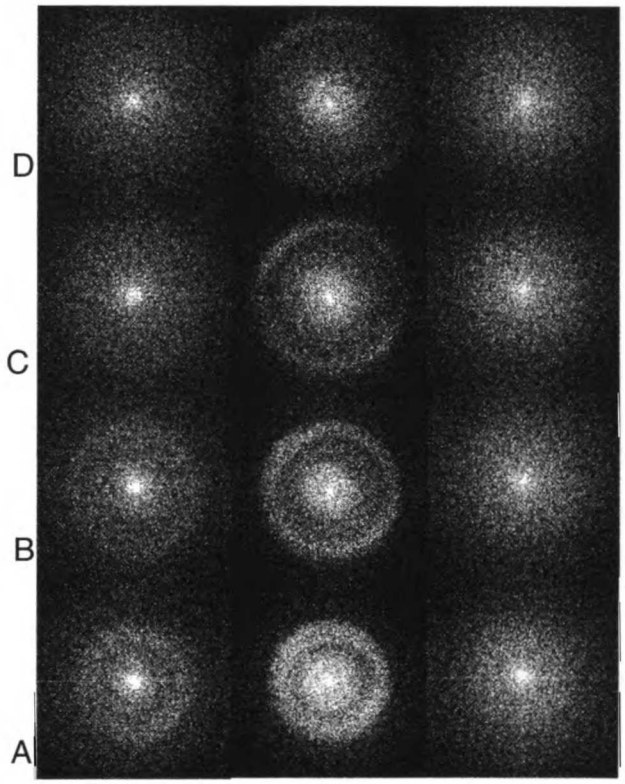
Figure 3b



Unfiltered Zero-loss filtered Plasmon-loss filtered

0.5 μm thick specimen

Figure 4



Unfiltered Zero-loss filtered Plasmon-loss filtered

0.7 μm thick specimen

Fig. 5

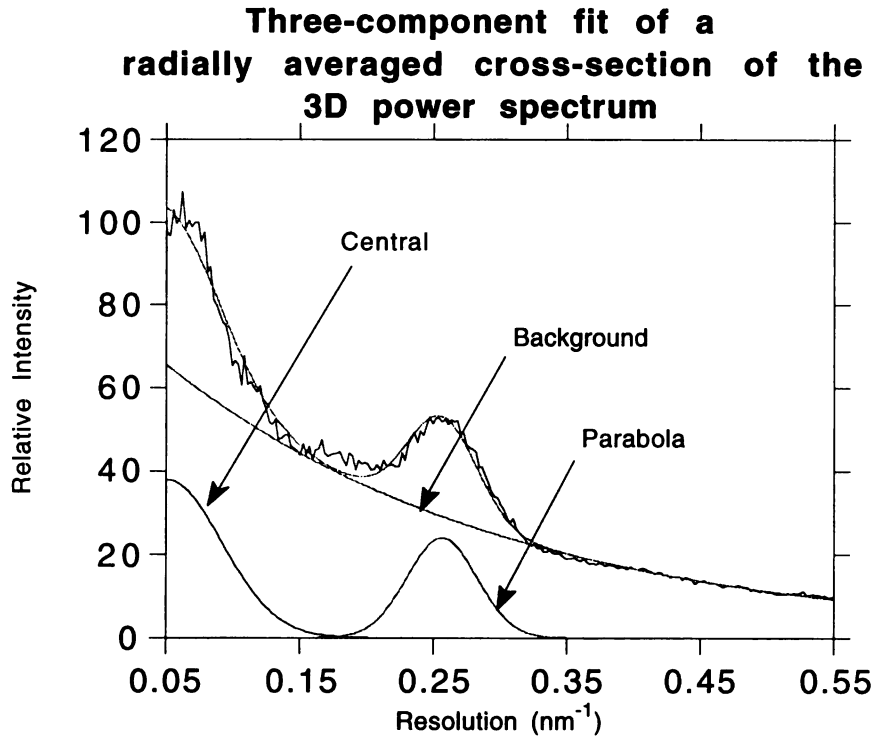


Fig. 6

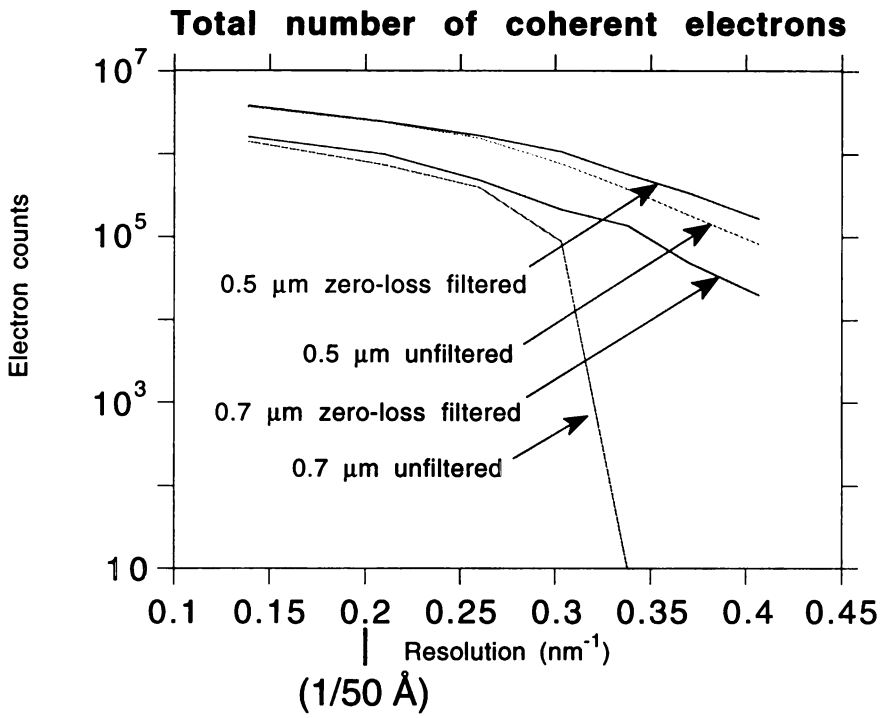


Figure 7

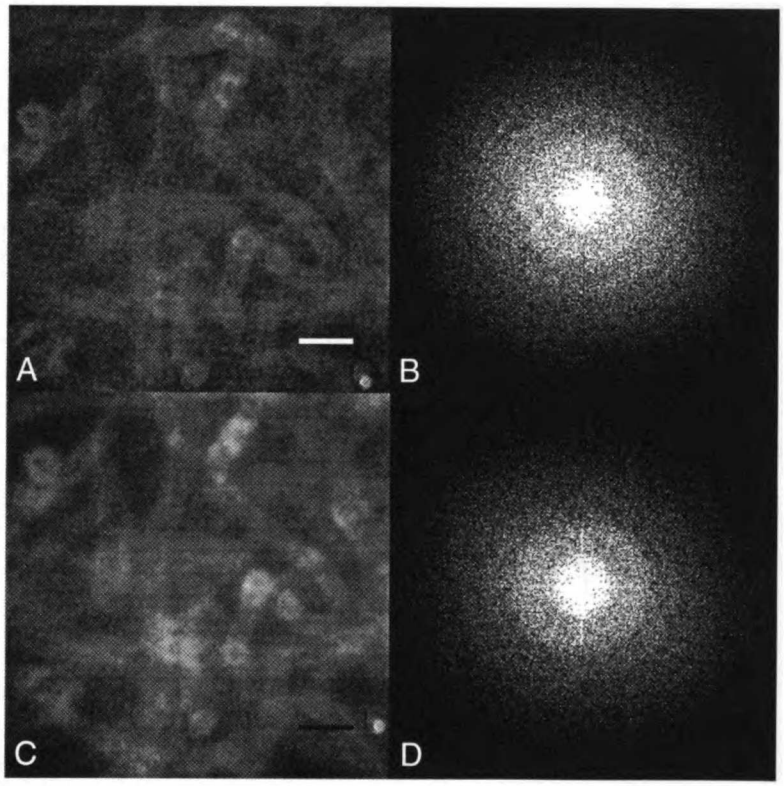


Figure 8

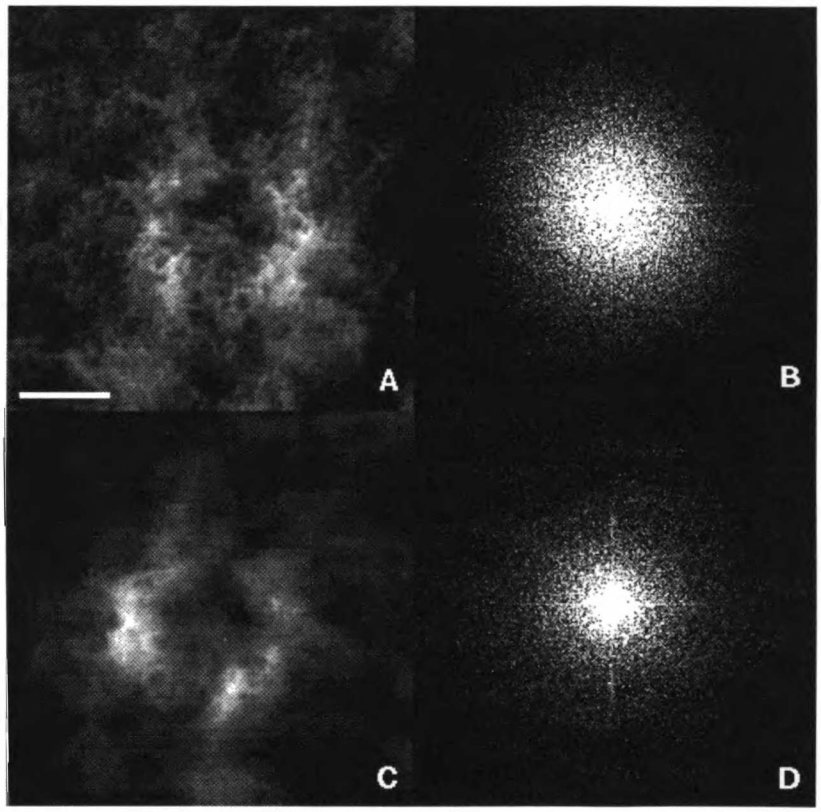


Fig. 9A

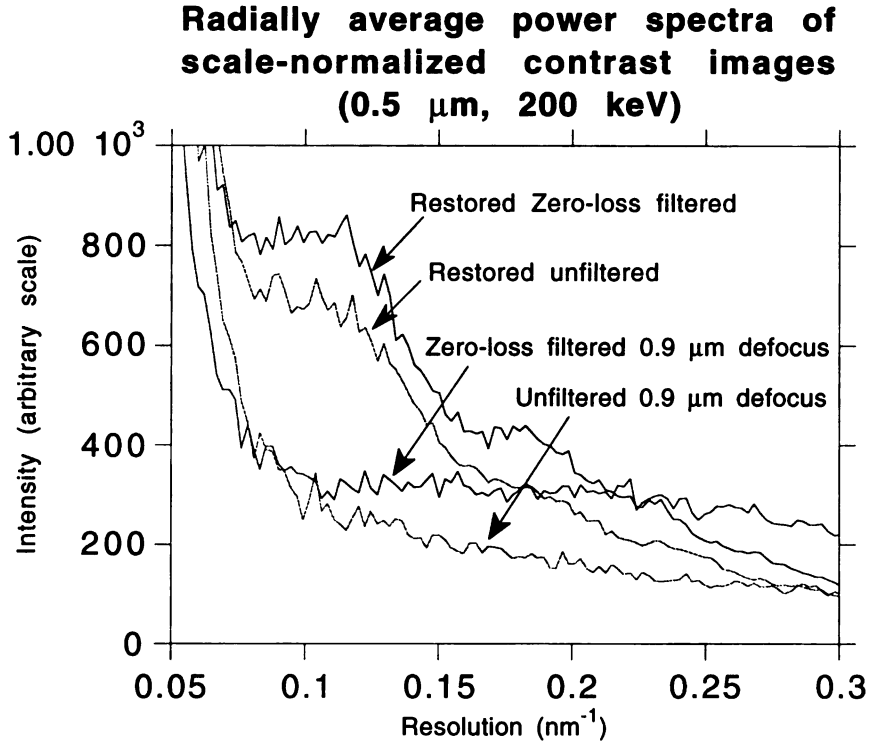
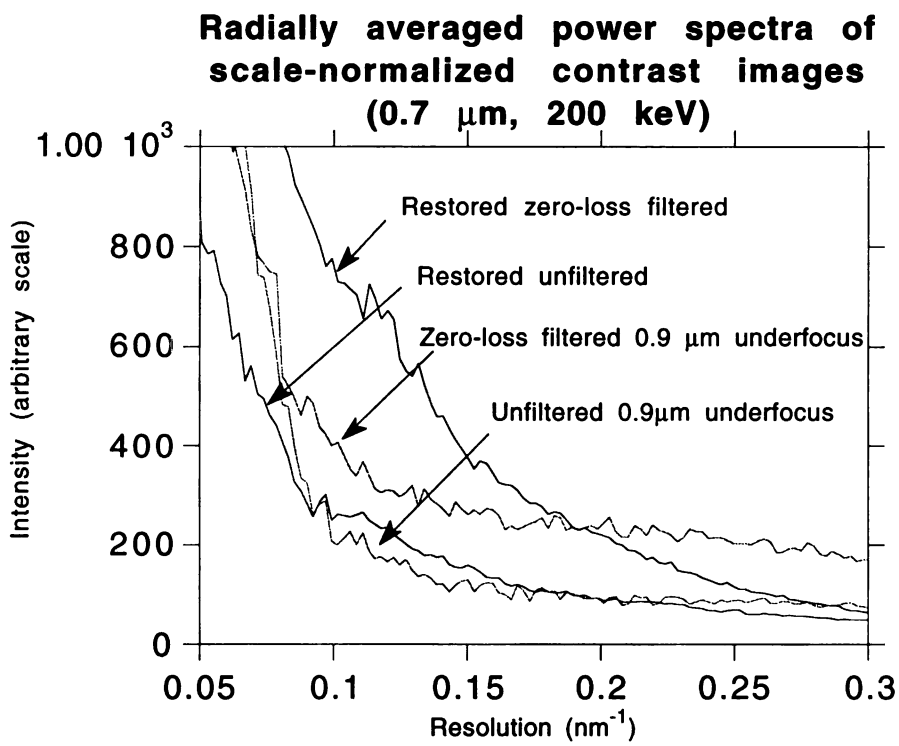

Radially average power spectra of scale-normalized contrast images (0.5 μm, 200 keV)
1.00 10^3
800
600
400
200
0
Intensity (arbitrary scale)
Restored Zero-loss filtered
Restored unfiltered
Zero-loss filtered 0.9 μm defocus
Unfiltered 0.9 μm defocus
0.05
0.1
0.15
0.2
0.25
0.3
Resolution (nm^-1)

Fig. 9B


Radially averaged power spectra of scale-normalized contrast images (0.7 μm, 200 keV)
1.00 10^3
800
600
400
200
0
Intensity (arbitrary scale)
Restored zero-loss filtered
Restored unfiltered
Zero-loss filtered 0.9 μm underfocus
Unfiltered 0.9μm underfocus
0.05
0.1
0.15
0.2
0.25
0.3
Resolution (nm^-1)

Chapter 4

Practical image restoration of thick biological specimens using multiple focus levels in transmission electron microscopy.

An accurate interpretation of image intensities in transmission electron micrographs is critical for the three-dimensional reconstruction of thick biological specimens. In addition to microscope lens aberrations, thick specimen imaging is complicated by additional distortions resulting from multiple elastic and inelastic scattering. Extensive analysis of the mechanism of image formation using electron energy-loss spectroscopy and imaging as well as exit wavefront reconstruction demonstrated that multiple scattering does not contribute to the coherent component of the exit wave (Han et al. 1996; Han et al. 1995). Although exit wavefront restored images showed enhanced contrast and resolution, that technique, which requires the collection of more than 30 images at different focus levels, is not practical for routine data collection in 3D electron tomography, where usually over 100 projection views are required for each reconstruction. Since only the coherent component exhibits the expected behavior of the transfer function, it can be restored to a good approximation using a simple restoration filter (Schiske 1968) with only 4 focus levels. We propose a new interpretation of the restored amplitude and phase components based on our previous image formation analysis, where the amplitude component is an approximation of the logarithm of the specimen mass-density, whereas the phase component is linearly related. The accuracy of limited reconstructions using 2, 4, 6, and 8 focus levels were assessed by comparing to the complete exit wave restoration. Although there was expected improvement with increased number of images used, the fractional root-mean-square

deviation between the 4-focus level and the 40-focus level restorations was only 5.1%. This compares with 25.5% deviation for the unrestored infocus image.

Introduction:

High resolution three-dimensional (3D) analyses of cellular organelles and nuclear structures is most appropriately studied using transmission electron microscopic tompgraphy of thick sections (Belmont et al. 1987; Frank et al. 1992; Fung et al. 1994; Horowitz et al. 1994; Ladinsky et al. 1994; Moritz et al. 1995). To accurately compute the three dimensional reconstructions from a set of tilted images and for quantitative analyses, it is essential to properly relate the image intensities to the projected specimen mass densities. This relies on an accurate understanding of the image formation mechanism of these specimens in the transmission electron microscope (TEM). The images taken in the TEM are not always a direct representation of the specimen mass density. There are two sources of aberration that effect image formation: electron-specimen interactions, and microscope lens aberrations. The difference between imaging of thick and thin specimens is at the level of electron-specimen interactions. For thin specimens, image formation is dominated by singly elastically scattered electrons, whereas for thick specimens, multiple elastic and multiple inelastic scattering contribute the majority of the electrons. In a detailed analysis of thick specimen image formation (Han et al. 1995), we demonstrated previously that exit wavefront reconstruction can exclude most of the multiple scattering and correct for lens aberrations. As exit wavefront reconstruction requires many (often over 30) through focus images (Coene et al. 1992; van Dyck et al. 1990), it is impractical for routine image restoration-- particularly in the application to tomographic reconstruction, where a complete tilt data set requires over 100

projection images. It is therefore important to develop a more practical restoration approach that uses fewer focus levels yet can still exclude most of the multiple scattering component and also correct for the lens contrast transfer function.

In this paper we present a quantitative analysis of the application of the Schiske formalism to the imaging of thick biological specimens. We use a complete exit wave reconstruction based on 40 images as a standard for comparison. In addition, based on our experiments on the mechanism of image formation for thick sections, we propose a new way to combine the amplitude and phase contrast components to generate the restored image. This approach better accounts for the contribution of multiple scattering at low resolution.

Materials and Methods:

Thick Biological Specimens: The specimens used in the experiments discribed here were the same as were used in the previous paper (Han et al. 1996). They are isolated centrosomes from Drosophila embryos embedded in epon and stained with uranyl-acetate and lead-citrate (Moritz et al. 1995). The diameter of the microtubules, which are 22 nm, serve as an internal standard. Specimens were cut to 0.7 μm thickness.

Microscopy and through focus series: The energy-filtered images were recorded with a Gatan Imaging Filter, Model GIF100 (Gubbens et al. 1993; Krivanek et al. 1992) mounted on a Philips CM200. All TEM images were recorded at a calibrated magnification of 40,000 times at the CCD. The images were binned twice in the camera hardware resulting in an effective pixel size projected back to the specimen of 1.20 nm. The energy-window used for energy loss-filtered imaging was 10 eV, the energy dispersion used for recording the energy-loss spectra was 0.5 eV per CCD pixel.

Through focus series consisting of 41 images were recorded from 18.1 μm under focus to 18.1 μm over focus with a focus step size of 0.905 μm. To minimize specimen alterations and shrinkage (Braunfeld et al. 1994) during data collection, the specimens were stabilized by pre-irradiating with approximately 1000 e^-/nm^2. The individual images were aligned prior to the exit surface wave front reconstruction using fiducial gold markers and cross-correlation (Koster et al. 1992). Image processing and visualization were done on a DEC VAX-9000 and a Silicon Graphics Iris workstation using Priism, the image visualization software developed in our laboratory (Chen et al. 1994).

Theory:

Many authors have presented approaches for image restoration (Coene et al. 1992; Hawkes 1980; Kirkland 1982; Saxton 1978; Scherzer 1949; Schiske 1968; Schiske 1973). Many such restorations are based on the assumption that the specimen is relatively thin and is a weak phase object. Typically, others have been interested in recovering very high resolutions ((~ .5-.1 $nm)^{-1}$ range) where the contrast transfer function (CTF) is highly oscillatory. In such cases, recovery is particularly important because contrast inversions cause the image intensities to vary dramatically, making interpretation extremely difficult. In the study of thick biological specimens, we are interested in recovering a large range of relatively low resolutions, from $(15\ nm)^{-1}$ to $(3\ nm)^{-1}$. Here it is very difficult to recover the image wave due to the large fraction of multiple inelastic scattering contributing mostly in this resolution range. Since the CTF is varying slowly at low resolutions, it is also very difficult to uniquely restore the exit wave unless large values of underfocus are used. For thick specimens, only the coherent component (single elastic scattering) exhibits the expected behavior of wave propagation through focus (Han et al.

1995). Linear exit wave restoration techniques can be used to extract the coherent component which will exhibit an enhanced contrast throughout a large resolution range. Following the restoration, real and imaginary components of the exit wave must be properly related to the specimen mass density, using the experimental results that address the nature of image formation for thick specimens.

In high resolution electron microscopy, van Dyck and coworkers have shown that the exit wavefront can be restored by extracting the coherent electrons which map on to a parabola in the three-dimensional Fourier Transform of a through focus series (>30 images)(Coene et al. 1992; van Dyck et al. 1990):

$$\hat{\mathbf{I}}(\mathbf{k},\zeta) = |C|^2\,\delta(\mathbf{k}) + C^*\psi(\mathbf{k})\delta(\zeta - \lambda|\mathbf{k}|^2/2) + C\psi^*(-\mathbf{k})\delta(\zeta + \lambda|\mathbf{k}|^2/2) + \int_{\mathbf{k}\neq 0,(\mathbf{k}-\mathbf{k}')\neq 0} \psi^*(\mathbf{k})\psi(\mathbf{k}-\mathbf{k}')\delta\{\zeta - \lambda[(\mathbf{k}-\mathbf{k}')^2 - \kappa^2]/2\}d\mathbf{k}' \tag{1}$$

where $\mathbf{k}$ and ζ are reciprocal axes for x, y and z respectively; δ is the Dirac delta function; λ is the electron wavelength and ψ is the specimen exit surface wavefront. By back transforming along the parabola, the exit wavefront can be recovered as follows:

$$\hat{\psi}_e = \exp(i\pi C_s\lambda^3\mathbf{k}^4)\frac{1}{N}\sum_{\Delta f_n=1}^{N}\hat{\mathbf{I}}(\mathbf{k},\Delta f_n)\exp(-i\pi\lambda\mathbf{k}^2\Delta f_n) \tag{2}$$

Although equation (2) is essentially the same as the simplified Schiske restoration filter (Eq. 11, see below) (Saxton 1994; Schiske 1968), the mapping of electrons in the 3D power spectra is an informative technique to select the appropriate focus levels to restore the resolution range of interest. For thick

biological specimens, it was shown that a large central (incoherent) component can be isolated from the parabolic (coherent) component. The central component is contributed largely by inelastic multiply scattered electrons (Han et al. 1996; Han et al. 1995). The derivation by van Dyck et. al. (1990) suggested that a large evenly spaced through focus series (>30 images) is most optimal for restoration, which is however impractical in routine 3D tomography. The advantage of restoration using a wide range of focus levels is, in addition to the reduction of statistical noise by $\sqrt{N}$ (where N is the total number of images used in the restoration), a better separation between the parabolic and incoherent components by reducing the width of the parabola in ζ (proportional to sinc(Z), Z is the full range of focus levels). Although, it is not required to use an evenly spaced through focus series for restoration using equation (2), it is more desirable as it facilitates the use of FFTs to speed the calculations.

A brief review of the ideal specimen exit wave recovery is presented below. The complex specimen exit surface wave function can be approximated as:

$$\tilde{\psi}_e \approx \tilde{\psi}_0 + \tilde{\psi}_{sc} \qquad (3)$$

where $\tilde{\psi}_0$ is the unscattered wave and $\tilde{\psi}_{sc}$ is the scattered wave with real and imaginary components:

$$\tilde{\psi}_{sc}(\mathbf{r}) = \tilde{\psi}_{amp}(\mathbf{r}) + i\tilde{\psi}_{phs}(\mathbf{r}) = A(\mathbf{r})\exp i\phi(\mathbf{r}) \qquad (4)$$

The unaberrated image (I_e) formed on the exit surface is:

$$I_e = |\tilde{\psi}_e|^2 = |\tilde{\psi}_0|^2 + \tilde{\psi}_0^{*}\tilde{\psi}_{sc} + \tilde{\psi}_0\tilde{\psi}_{sc}^{*} + |\tilde{\psi}_{sc}|^2 \qquad (5)$$

The wave aberration introduced by the objective lens effects the scattered wave with a known systematic dependence on the focus level, Δf, and spherical

aberration, C_s. This wave aberration function causes a well-characterized phase-shift in Fourier space:

$$\exp[i\chi(\Delta f,\mathbf{k})] = \exp[i\frac{\pi\lambda\mathbf{k}^2}{2}(Cs\frac{\lambda^2\mathbf{k}^2}{2} - \Delta f)] \tag{6}$$

The detected image wave is a result of the real-space convolution of the scattered wave with the wave aberration function (or multiplication in reciprocal-space):

$$\hat{\psi}_{sc}^{ab}(\Delta f,\mathbf{k}) = \hat{\psi}_{sc}(\mathbf{k})\exp[i\chi(\Delta f,\mathbf{k})] \tag{7}$$

where $\hat{\psi}_{sc}^{ab}(\Delta f,\mathbf{k})$ denotes the aberrated wave function in reciprocal space and $\hat{\psi}_{sc}(\mathbf{k})$ the unaberrated scattered wave. The image detected is analogous to the formation of the unaberrated image (Eq. 5) on the exit surface but with $\tilde{\psi}_{sc}^{ab}$ substituted for $\tilde{\psi}_{sc}$. Assuming $|\tilde{\psi}_0|^2$ is constant, set to unity, and the contribution by $|\tilde{\psi}_{sc}|^2$ is negligible in equation (5) the Fourier transform of the aberrated image detected at a particular defocus is given by:

$$\begin{aligned}
&\hat{I}_{\mathrm{det}ected}(\Delta f,\mathbf{k}) = \delta(\mathbf{k}) + \hat{\psi}_{sc}(\mathbf{k})\exp[i\chi(\Delta f,\mathbf{k})] + \hat{\psi}_{sc}^{*}(\mathbf{k})\exp[-i\chi(\Delta f,\mathbf{k})] \\
&= \delta(\mathbf{k}) + (\hat{\psi}_{+}(\mathbf{k}) + i\hat{\psi}_{-}(\mathbf{k}))\exp[i\chi(\Delta f,\mathbf{k})] \\
&+ (\hat{\psi}_{+}(\mathbf{k}) - i\hat{\psi}_{-}(\mathbf{k}))\exp[-i\chi(\Delta f,\mathbf{k})] \\
&= \delta(\mathbf{k}) + 2\hat{\psi}_{+}(\mathbf{k})\cos[\chi(\Delta f,\mathbf{k})] - 2\hat{\psi}_{-}(\mathbf{k})\sin[\chi(\Delta f,\mathbf{k})]
\end{aligned} \tag{8}$$

where $\delta(\kappa)$ is the Dirac delta function, and

$$\begin{aligned}
\hat{\psi}_{+}(\mathbf{k}) &= FT[\tilde{\psi}_{amp}(\mathbf{r})] = \frac{\hat{\psi}_{sc}(\mathbf{k}) + \hat{\psi}_{sc}(-\mathbf{k})}{2} \\
\hat{\psi}_{-}(\mathbf{k}) &= FT[\tilde{\psi}_{phs}(\mathbf{r})] = \frac{\hat{\psi}_{sc}(\mathbf{k}) - \hat{\psi}_{sc}(-\mathbf{k})}{2}
\end{aligned} \tag{9}$$

Thus, by solving for $\hat{\psi}_+(\mathbf{k})$ and $\hat{\psi}_-(\mathbf{k})$, the unaberrated scattered exit wave is completely recovered. From equation (8), to solve for the two unknowns, one must collect at least two images at different defocus levels to recover the exit wave. Indeed, if many images were collected at different defoci (i.e. a through focus series), this becomes an over-determined problem, and $\hat{\psi}_+(\mathbf{k})$ and $\hat{\psi}_-(\mathbf{k})$ can be solved by setting up a simple 'Ax=B' matrix problem. For example, in the case of only two focus levels:

$$\begin{bmatrix} I(\Delta f_1) \\ I(\Delta f_2) \end{bmatrix} = \begin{bmatrix} \cos\chi(\Delta f_1,\mathbf{k}) & -\sin\chi(\Delta f_1,\mathbf{k}) \\ \cos\chi(\Delta f_2,\mathbf{k}) & -\sin\chi(\Delta f_2,\mathbf{k}) \end{bmatrix} \begin{bmatrix} \hat{\psi}_+(\mathbf{k}) \\ \hat{\psi}_-(\mathbf{k}) \end{bmatrix} \tag{10}$$

The above derivation assumes that the entire scattered wave is effected by the wave aberration function in the same way. That is, the electrons contributing to the scattered wave are single elastically scattered. Unfortunately, for thick biological specimens, majority of the imaging electrons arise from multiple inelastic scattering (Han et al. 1993). The contribution of this component varies with specimen, and cannot be systematically characterized as illustrated above for the coherent component. Empirically, it was shown that the multiple elastic and inelastic scattering components contribute a central Gaussian and a large background component in the three-dimensional power spectrum of a through focus series (Han et al. 1995). This implies that a unique expression for the multiple scattering component as a function of defocus does not exist. Although we were able to quantitate the overall contribution of these components to images of thick specimens, the specific contribution for the recovery of the exit wave is very specimen-dependent, and thus cannot be easily generalized or incorporated into equation (8). Thus, the solution to equation (10) is only approximate since it does not account for all the electrons

contributing to image formation, except at high resolutions where multiple scattering does not contribute as significantly.

Using electron energy-loss spectroscopic imaging (ESI), as mentioned above, we showed that only single elastically scattered electrons contribute to the behavior of equation (6) (Han et al. 1995). By restoring this component through the resolution range of interest, the restored exit wave can then be properly related to projected specimen mass density. Schiske and other authors have optimized the statistics of the restoration (Hawkes 1980; Saxton 1978; Schiske 1968; Schiske 1973):

$$\hat{\psi}_e(\mathbf{k}) = \sum_{\Delta f_n=1}^{N} \hat{I}(\mathbf{k},\Delta f_n)\cdot r(\mathbf{k},\Delta f_n)$$

$$r(\mathbf{k},\Delta f_n) = \exp[i\chi(\Delta f_n,\mathbf{k})]\frac{\{N - \sum_{\Delta f_m=1}^{N}\exp[2i[\chi(\Delta f_m,\mathbf{k}) - \chi(\Delta f_n,\mathbf{k})]\}}{\{N^2 - |\sum_{\Delta f_m=1}^{N}\exp[2i[\chi(\Delta f_m,\mathbf{k})]]|^2\}} \qquad (11)$$

In a through focus series with equal focus level increments as it was done by van Dyck and co-workers, equation (11) reduces to equation (2). Since our goal is to restore the exit wavefront using as few as four focus levels equation (11) is the more appropriate.

Once an approximation to the exit wave has been recovered, it is necessary to properly relate the amplitude and phase components of the projected specimen mass density. Based on previous experiments aimed at understanding the mechanism of image formation for thick specimens, we have shown that the relative contribution of multiple scattering to imaging has a logarithmic relationship to specimen thickness (or mass density) (Han et al. 1996). In addition, calculating the average image intensities of the same

specimen area as a function of specimen tilt (hence thickness), showed a logarithmic relationship between intensity and thickness (data not shown). Thus, we interpret the restored amplitude component of the exit wave to arise directly from absorption by the thick specimen. As a consequence, it should have a logarithmic relationship to mass density. The phase component is interpreted to be directly related to specimen thickness and thus should follow a linear relationship. The sum of the phase component, $\phi(\mathbf{r})$, with the logarithm of the amplitude component, $A(\mathbf{r})$, is our estimate of the projected mass density distribution, $D(\mathbf{r})$, of the thick specimen:

$$D(\mathbf{r}) = \phi(\mathbf{r}) - \log[A(\mathbf{r})] \qquad (12).$$

Results:

A. Comparison of the restoration using fewer focus levels with the Ewald sphere reconstruction. Since no absolute mass standard exists for thick biological specimens, it is difficult to assess whether the restored mass-distribution is 'correct'. Although the exact dimensions and mass distribution of the microtubules are known, the images are projections of often many overlaying microtubules, making the absolute assessment of mass density distribution difficult. Nonetheless, if we assume that the exit wavefront restoration using 40 focus levels gives the closest representation of the "true" mass density distribution, we can use it as a standard for comparison with different restoration techniques. As discussed in the *Theory* section, the Schiske filter was used to restore images from a through focus series (Eq. 11)(Schiske 1968). To quantitatively compare the reconstructions while ignoring trivial differences due to different scales and background levels, the

images were scaled by matching their contrast. The contrast image I', was calculated for each image as:

$$I'(x,y) = \frac{I(x,y) - I_{ave}}{I_{ave}}$$

and subsequently scaled to a common intensity range. Fractional root-mean-square deviation ($\%D$) were used to assess the similarity between the restorations and the exit wavefront ($I_{ref}(x,y)$):

$$\%D = \frac{1}{nx \cdot ny} \sum_{nx,ny} \frac{\sqrt{(I(x,y) - I_{ref}(x,y))^2}}{I_{ref}(x,y)}$$

In comparison to the complete reconstruction using 40 images, the restorations using fewer focus levels have reasonable fractional deviations (Table I). As expected, the more focus levels included in the restoration, the better the estimate of the exit wave (see Theory, Table I, trials 1,2,5). The choice of focus levels strongly affects the restoration quality, and in the case of 4 focus level restoration, $\%D$ can range from 5% to 12% (Table I, trials 9-13). Indeed, any restoration is better than no restoration as evidenced by the high $\%D$'s of the unrestored images. As expected, the further out of focus of a single image (Table I, trials 12-15), the higher the $\%D$ due to the highly oscillatory behavior of the contrast transfer function.

B. Restoration using appropriately chosen four focus levels shows enhanced contrast. Figure 1 compares the 905nm underfocused image to the exit wavefront restored image of a 0.7 μm thick specimen (see *Methods)* using equation (2) (see *Theory*). This reconstruction is used as a standard to compare with other restoration approaches using fewer focus levels. Figure 2 plots the power spectrum of the contrast normalized images comparing the in-focus data, 8 and 4 focus-level restorations (Eq. 11) and the 40 focus-level Ewald sphere reconstruction (Eq. 2). The exit wave using fewer focus levels is

recovered by restoring the coherent component of the optimally chosen through focus series, requiring at least 2 defoci. Optimal focus levels are chosen such that the coherent component in the 3D FFT of the through focus series span the resolution range of interest shown in Figure 3. For restoration using 6 defoci, the optimal range of focus levels are chosen to evenly span the resolution range between (15 nm)$^{-1}$ to (3 nm)$^{-1}$: ± 18.10, ± 5.43, and ± 3.62 μm defoci (trial 2, Table I). The $\%D$ for this restoration is only slightly higher than that restored using 8 defoci (trial 1, Table I), with a value of 3.99%. The most optimal set of defoci for the restoration using only four focus levels are: ± 18.10 and ± 5.43 μm, with an $\%D$ value of 5.2%. Note that although it is not necessary to choose matching under- and over- foci for the restorations, doing so results in better discrimination between the amplitude and phase components (compare trials 2 and 3, 5 and 6 in Table I). Figure 4 shows the restored images using four and six focus levels with equations (11) and (12) to obtain an estimate of the projected specimen mass density. The restoration using 4 focus levels (Figure 4A) has higher contrast compared to the 905nm data (Figure 1C), although it clearly has a much higher background (noise) component compared to the restorations using 6 and 40 focus-levels (Figures 4B and 1A). In addition, the microtubule boundaries are much more clearly delineated in the resstored images (Figures 1A, 4 A and B) as compared to the unrestored (Figure 1C).

C. The effect of error in focus level on the restorations. Figure 5 plots the expected error in the contrast transfer function (CTF) for a conserved estimate of the error in focus levels. Error at each focus level is determined by:

$$error = \frac{(\frac{dCTF}{d(\Delta f)})\varepsilon_{\Delta f}}{CTF} = \frac{\pi\lambda k^2(\beta\sin\chi + \cos\chi)\varepsilon_{\Delta f}}{(\beta\cos\chi - \sin\chi) + \eta}$$

where β is the fraction of amplitude contrast (taken to be 0.15, empirically determined through curve-fits of diffractograms), $\varepsilon_{\Delta f}$ is the estimated error in

focus level determination (a conservative estimate is taken to be 10 nm), and η the estimated noise-to-signal ratio. The error in CTF is less than 2% for resolutions up to $(3\ nm)^{-1}$ except for regions where the CTF crosses zero (Fig 5, spikes). Since the restoration is an over-determined problem, at resolutions where the CTF is crossing zero for one of the focus levels, the others are well-determined. Thus, the exit wave function is solvable within 2% for the entire range of the resolutions of interest.

Discussion:

An accurate three-dimensional tomographic reconstruction relies on the assumption that the image accurately represent the projected specimen mass densities. While we have utilized a previously proposed restoration approach using only a few focus levels (Schiske 1968), it is our interpretation of the relationship between the exit wave and the specimen mass density that differs from what has been traditionally assumed. In comparison with the full exit wave reconstruction, the reduced restoration indicates that this approach recovers a good approximation to the exit wavefront. We demonstrated that the relative contrast through all resolutions is increased by removing aberrations from images of thick biological specimens. This enhanced contrast revealed substructures in these thick specimens that were otherwise not seen in the uncorrected images. From these results, we demonstrated the importance of correcting for aberrations in electron micrographs to achieve an accurate interpretation of images for 3D reconstruction. The future incorporation of this modified restoration approach for thick specimens will improve the resolution extent of the 3D tomographic reconstructions.

Acknowledgments:

The authors thank M. Braunfeld and M. Moritz for providing the centrosome specimens; A. Gubbens for assisting data collection and critical reading of this manuscript; A. Gubbens, W. Liu, M. Op de Beeck, M. Gustafsson for helpful discussions. K.F.H. is supported by the Howard Hughes Medical Institute Predoctoral Fellowship in the Biological Sciences. This work is supported by grants from the National Institutes of Health (GM 31627 for D.A.A.; GM25101 for J.W.S.) and by the Howard Hughes Medical Institute.

References:

Belmont, A.S., Sedat, J.W. & Agard, D.A. (1987). A three-dimensional approach to mitotic chromosome structure: Evidence for a complex hierarchical organization. *Journal of Cell Biology,* **105**, 77-92.

Chen, H., Clyborne, W., Sedat, J. & Agard, D. (1994). PRIISM: An integrated system for display and analysis of 3D microscope images. *SPIE:Biomedical Image Processing and 3-Dimensional Microscopy,* **1660**, 784-90.

Coene, W., Janssen, G., Op de Beeck, M. & Van Dyck, D. (1992). Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. *Physical Review Letters,* **69**, 3743-3746.

Frank, J. & Radermacher, M. (1992). 3-Dimensional reconstruction of single particles negatively stained or in vitreous ice. *Ultramicroscopy,* **46**, 241-262.

Fung, J.C., Agard, D.A. & Sedat, J.W. (1994). Three-dimensional reconstruction of the synaptonemal complex from high-pressure frozen maize meiocytes using IVEM tomography. *Proc. 53rd Ann. Microscopy Society of America,* 14-15.

Gubbens, A. & Krivanek, O. (1993). Applications of a post-column imaging filter in biology and material science. *Ultramicroscopy,* **51**, 146-59.

Han, K.F., Gubbens, A.J., Koster, A., Braunfeld, M., Sedat, J.W. & Agard, D.A. (1993). Analysis of electron-specimen interactions of thick biological specimens in transmission electron microscopy at 200 keV. *52nd Ann. Microscopy Society of America.* 204-205.

Han, K.F., Gubbens, A.J., Sedat, J.W. & Agard, D.A. (1996). Optimal strategies for imaging thick biological specimens: *exit wavefront reconstruction and energy filtering. J. Microscopy,* submitted.

Han, K.F., Sedat, J.W. & Agard, D.A. (1995). Mechanism of image formation for thick biological specimens: *exit wavefront reconstruction and electron energy-loss spectroscopic imaging. J. Microscopy,* **178:2**, 107-19.

Hawkes, P.W. (1980). Image processing based on the linear theory of image formation. Computer Processing of Eletron Microscope Images. Berlin, Pringer-Verlag.

Horowitz, R.A., Agard, D.A., Sedat, J.W. & 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.

Kirkland, E.J. (1982). Nonlinear high resolution image processing of convectional transmission electron micrographs. *Ultramicroscopy,* **9**, 45-64.

Koster, A.J., Chen, H., Sedat, J.W. & Agard, D.A. (1992). Automated microscopy for electron tomography. *Ultramicroscopy,* **46**, 207-227.

Krivanek, O.L., Gubbens, A.J., Dellby, N. & Meyer, C.E. (1992). Design and 1st applications of a post-column imaging filter. *Microscopy Microanalysis Microstructures,* **3**, 187-99.

Ladinsky, M.S., Kremer, J.R., Furcinitti, P.S., McIntosh, J.R. & Howell, K.E. (1994). HVEM tomography of the trans-Golgi network: structural insights and identification of a lace-like vesicle coat. *J Cell Biol,* **127**, 29-38.

Moritz, M., Braunfeld, M., Fung, J., Alberts, B., Sedat, J. & Agard, D. (1995). Three-dimensional structural characterization of centrosomes from early drosophila embryos. *J. Cell Biol.,* **130**, 1149-59.

Saxton, W. (1994). What is the focus variation method-is it new- is it direct. *Ultramicroscopy,* **55**, 171-81.

Saxton, W.O. (1978). Computer Techniques for Image Processing in Electron Microscopy. New York, Academic Press.

Scherzer, O. (1949). The Theoretical Resolution Limit of the Electron Microscope. *Journal of Applied Physics,* **20**, 20-26.

Schiske, P. (1968). Zur Frage der Bildrekonstruktion durch Fokusreihen. 4th European Conference on Electron Microscopy. 145.

Schiske, P. (1973). Image processing using additional statistical information about the object. Image Processing and Computer-aided Design in Electron Optics. London, Academic Press.

van Dyck, D. & Op de Beeck, M. (1990). New direct methods for phase and structure retrieval in HREM. *Proc. of 12th Int'l Congress for Electron Microscopy,* 26-27.

Figure and Table Captions:

Table I: $\%D$ difference with respect to the full exit wavefront restoration (see text) is used to compare the restorations and raw data. The first column lists the trials referred to in the text; second column, the $\%D$'s; third column, the focus levels used in the restoration.

Figure 1: Exit wavefront restoration (A) from 40 through focus images and its diffractogram (B); 905nm defocused image (C) and its diffractogram (D). Scale bar: 50nm, resolution limit is $(2.41\ nm)^{-1}$.

Figure 2: Contrast and scale normalized power spectra comparing exit wavefront restored, 8-, 4- focus-level restored and in-focus data, demonstrating enhanced contrast.

Figure 3: Plot of the coherent component for each reciprocal defocus in the 3DFFT of the through focus series used to restore Figure 1A.

Figure 4: Restorations (A, B) using 4 and 6 focus levels and their diffractograms (C, D). Scale bar: 50nm, resolution limit is $(2.41\ nm)^{-1}$.

Figure 5: Expected error of the contrast transfer function (CTF) as a function of resolution for focus levels used in the restoration described (Fig. 4) with a 10nm focus level error.

Table I

Trial	$\%D$	Defocus levels used in the restoration (μm)
1) 8 focus levels (A)	3.77%	± 18.10, ± 9.05, ± 5.43, ± 3.62
2) 6 focus levels (A)	3.99%	±18.10, ± 5.43, ± 3.62
3) 6 focus levels (B)	5.09%	18.10, 5.43, 3.62, -2.72, -4.53, -9.05
4) 6 focus levels (C)	8.05%	± 12.11, ± 7.03, ± 5.43
5) 4 focus levels (A)	5.17%	± 18.10, ± 5.43
6) 4 focus levels (B)	8.01%	18.10, 5.43, -9.05, -3.62
7) 4 focus levels (C)	9.05%	± 12.11, ± 7.03
8) 4 focus levels (D)	12.14%	± 8.57, ± 5.43
9) 2 focus levels (B)	11.22%	5.43, -4.53
10) 2 focus levels (A)	15.85%	18.10, -17.19
11) in focus data	25.47%	
12) 905 nm underfocused data	29.66%	
13) 3.63 μm underfocused data	31.44%	
14) 5.43 μm underfocused data	32.97%	
15) 18.1 μm underfocused data	36.11%	

Figure 1

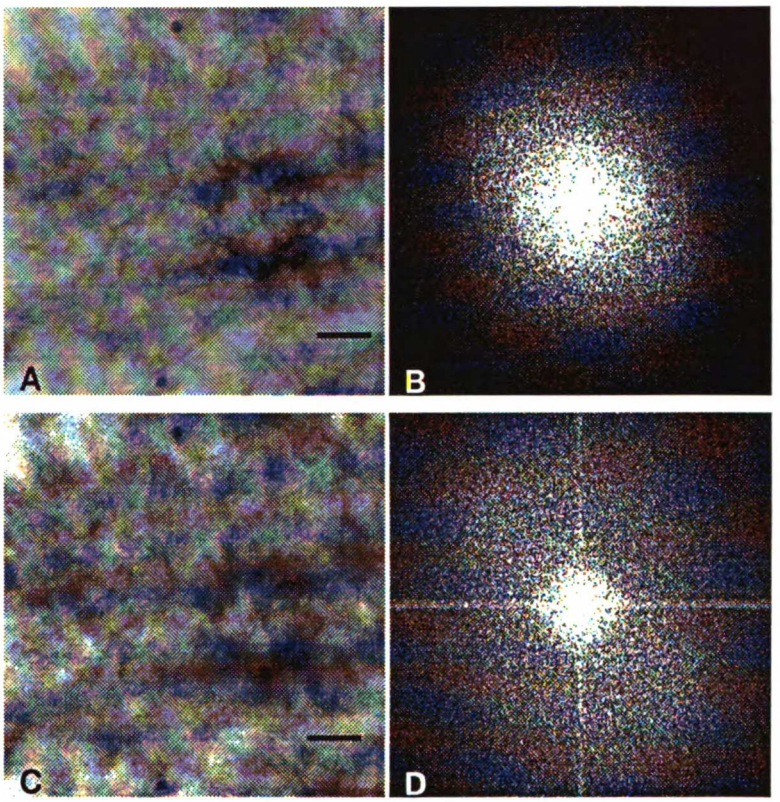


Figure 2

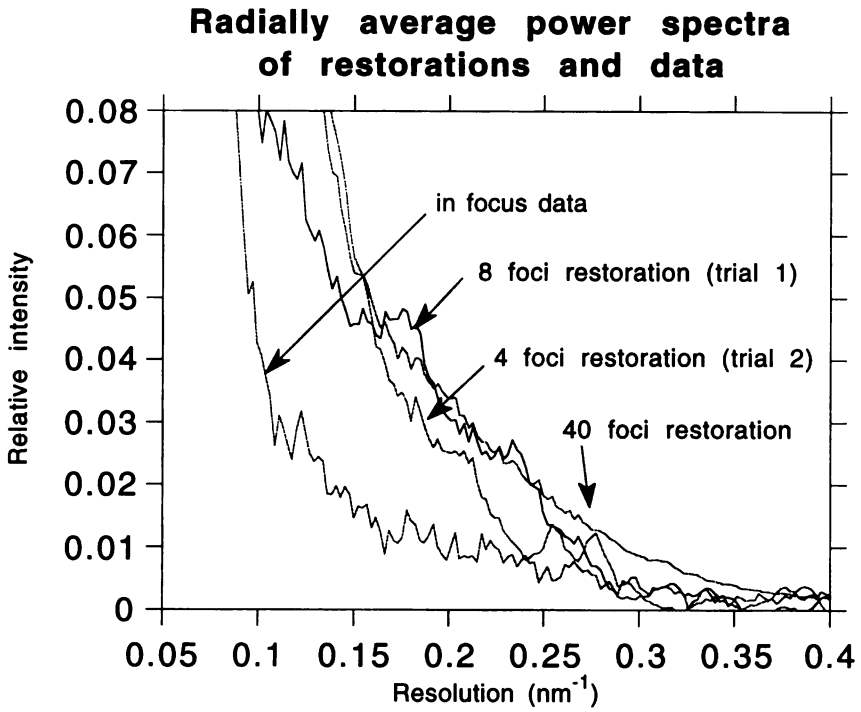


Figure 3.

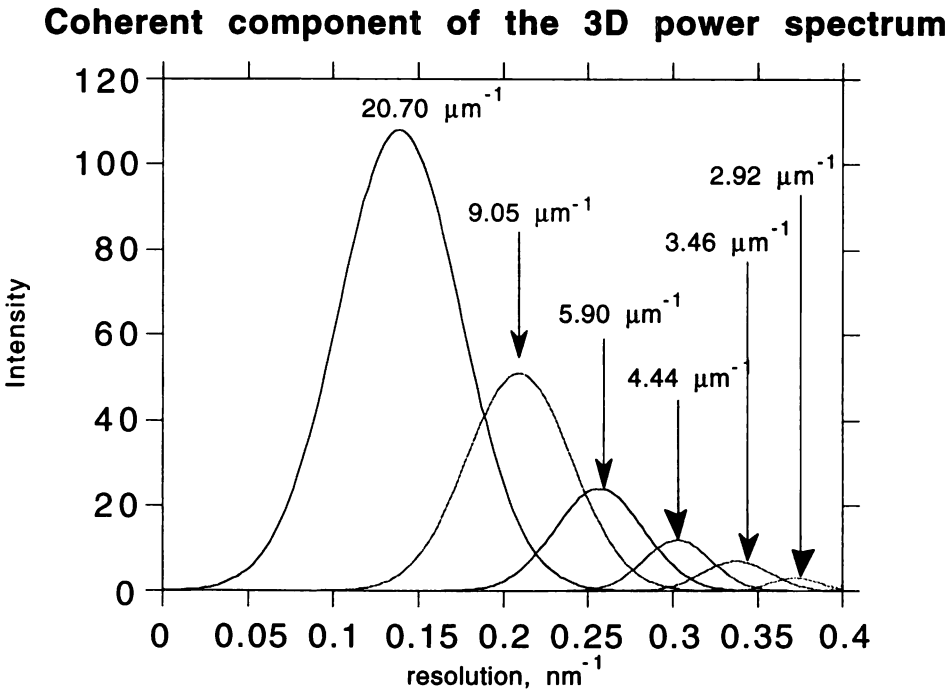


Figure 4

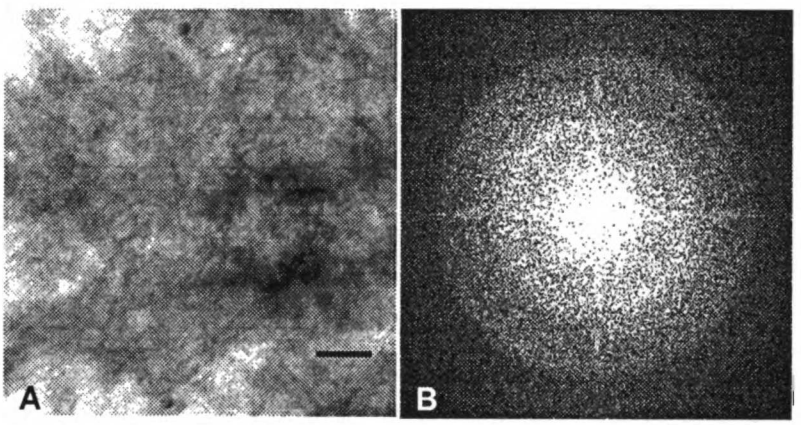
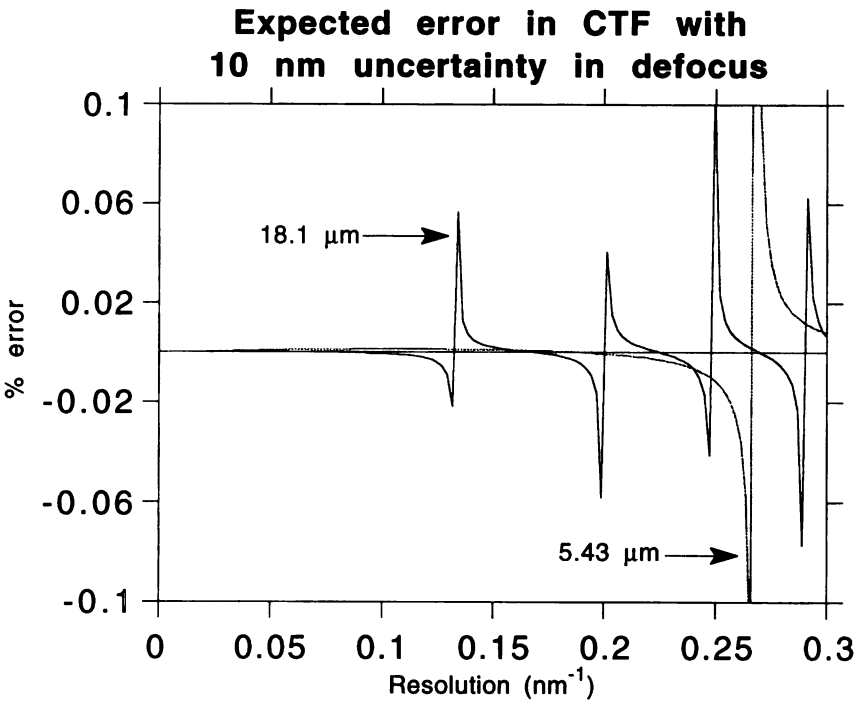


Figure 5.


Expected error in CTF with
10 nm uncertainty in defocus
0.1
0.06
0.02
-0.02
-0.06
-0.1
% error
18.1 µm
5.43 µm
0
0.05
0.1
0.15
0.2
0.25
0.3
Resolution (nm-1)

Conclusion

Three-dimensional tomographic reconstruction requires an accurate set of projected specimen mass density. In reconstructing supramolecular assemblies such as chromosomes and centrosomes, these specimens are too thick for electron microscopy, where the images are degraded due to multiple elastic and inelastic scattering. To achieve high enough resolutions (up to $(3\ nm)^{-1}$) for fiber tracing, unless these degrading components are removed, it will remain difficult to interpret the reconstructions.

Part I of this thesis described an extensive set of experiments characterizing the nature of the image degrading components, which is termed 'incoherent'. The component that is easily interpretable as projected mass density is termed 'coherent'. It was demonstrated that computationally, it is possible to separate a significant portion of the incoherent component from the coherent component.

It was shown that the optimal strategy is collect data at high primary voltages (>200 keV) and combine electron energy-loss filtering and through focus image restoration to recover the coherent component of the images in transmission electron microscopy of thick biological specimens.

After a series of experiments over sampling and using the energy filter, I have found that the intrinsic (stain) resolution for our typical specimens is just beyond $(3\ nm)^{-1}$. If other specimens are used where the suspected intrinsic resolution is higher, over sampling and a larger range of through focus series is required (Appendix A).

The quantitative analysis was done using similar specimens of different thicknesses at zero tilt. It is important to consider the focus change

perpendicular to the tilt axis at high specimen tilts for through focus series restorations. (I found that in addition to the expected focus change, there is an additional blurring that may be due to the electron optics disturbed by the tilted stage, although I haven't convinced myself that it was the case.) Thus restoration of highly tilted specimen should be verified using the procedures illustrated in Appendix A.

In future experiments, data should always be collected at high voltages with an energy filter. The focus series exit wavefront restoration should be done prior to the three-dimensional tomographic reconstruction. In theory, the corrected images already represent the projected specimen mass densities. Therefore, scaling image intensities between specimen tilts should be reduced only to correct for fluctuations in the beam current during data collection.

The newly interpreted images should result in a more accurate reconstruction, and hopefully will aid in the modeling of such complex supramolecular assemblies such as chromosomes.

Part II: Inverse sequence structure mapping reveals novel structure motifs in proteins

Summary:

Structure analysis provides the ultimate understanding of the mechanisms by which proteins function. New proteins are rapidly determined and their biochemical roles in cell biology elucidated. But the exact mechanisms of function await the difficult experimental determination of these structures by either X-ray crystallography or multidimensional NMR. These high resolution structures have in many cases aided in our understanding of the results of mutational and biochemical analyses. In addition, the precise mechanistic determinations have pharmaceutical applications such as designing specific agonists or antagonists to these enzymes or signal transducers. It is thus of interest to understand how proteins fold to predict the 3D structure of proteins based on their primary amino acid sequences.

It is generally assumed that the native fold of a given protein sequence represents the most stable state with the lowest free-energy. Thus, a given sequence would have essentially a single unique fold, but which can share a similar architecture of another unrelated sequence. Due to the limited structural database, the traditional approach to studying the relationship between sequence and structure is by first identifying the structural environments (such as helix, sheet, or turn), and then tabulating the propensities of each of the amino acid types that were found in these environments (Chou and Fasmann).

An alternative approach to study the relationship between sequence and structure is to utilize a vast database of multiple sequence alignments where allowable substitutions in a given structure are explored. Chapter 5 presents the statistical method of classifying recurrent multiple sequence patterns (profiles) in proteins. Extensive analysis was done to show that indeed there are patterns that are common to many different protein families which are

functionally unrelated. Chapter 6 shows that many of these patterns are well correlated with structural properties. These recurring profiles (or allowable substitution patterns) have similar structural constraints in their local environments. In addition, using the inverse mapping approach described in Chapter 5, the ambiguity in mapping can be explicitly addressed. These analysis explains why there has been an upper-limit to secondary structure prediction of 70% or less. Indeed, only 44% of all positions in the sequence database maps to a single structure type. Chapter 7 discusses the three-dimensional context in which these motifs occur, focusing on the newly identifies structural motifs such as buried helices, N-terminal capping of sheets, C-terminal capping of sheets and helices, and helix-turn-sheet transition motifs. The global context including solvent accessibility and packing properties explains the preferential allowable substitutions.

Some new motifs identified are putative folding initiation sites for protein folding.

Chapter 5

Recurring local sequence motifs in proteins

We describe a completely automated approach to identifying local sequence motifs which transcend protein family boundaries. Cluster analysis is used to identify recurring patterns of variation at single positions and in short segments of contiguous positions in multiple sequence alignments for a non-redundant set of protein families. Parallel experiments on simulated data sets constructed with the overall residue frequencies of proteins but not the inter-residue correlations show that naturally occurring protein sequences are significantly more clustered than the corresponding random sequences for window lengths ranging from one to thirteen contiguous positions. The patterns of variation at single positions are not in general surprising: chemically similar amino acids tend to be grouped together. More interesting patterns emerge as the window length increases. The patterns of variation for longer window lengths are in part recognizable patterns of hydrophobic and hydrophilic residues, and in part less obvious combinations. A particularly interesting class of patterns features highly conserved glycine residues. The patterns provide a means to abstract the information contained in multiple sequence alignments and may be useful for comparison of distantly related sequences or sequence families and for protein structure prediction.

Introduction

Are there recurring local patterns in the amino acid sequences which encode proteins? Global similarity is often used to classify sequences into families; are there local patterns which transcend family boundaries?

Given that all viable protein sequences must be such that the proteins they encode can fold and have at least marginal stability, it is reasonable to expect that not all 20^N amino acid sequences of length N are equally probable. There are far too few distinct protein families to tabulate meaningful statistics on the frequencies of occurrence of the different peptides of length N for N greater than two (Gonnet et al. 1994). An alternative approach is to use cluster analysis to identify recurring sequence patterns. This requires a suitable measure of similarity between two sequences.

Global sequence comparisons almost always rely on amino acid substitution matrices compiled by averaging over large sets of related sequences. The disadvantages of using a single substitution matrix have been pointed out on numerous occasions (Johnson et al. 1993; Risler et al. 1988). The major problem is that at different positions in protein structures, different sets of amino acid sequences are likely to substitute for one another. In other words, there is no single and universally applicable set of distances (or similarities) between the 20 amino acids. Rather, similarity can be quite context-dependent.

A more natural measure, which does not require the assumption of a single substitution matrix, is available for comparison of protein families if there are a number of sequences in each family. For each position in a set of multiply aligned sequences, one can calculate the frequency of occurrence of each of the amino acids. The resulting sequence of frequency distributions is often called a profile (Gribskov et al. 1990). To evaluate the distance between two aligned profile segments, one can compare the frequency distributions at corresponding positions.

Here we use such a distance measure in conjunction with cluster analysis to identify patterns which occur frequently in multiple sequence alignments for proteins of known structure. Because only one multiple sequence alignment is included for each family, the patterns are necessarily common to many different protein families and are distinct from the family specific patterns compiled in the Prosite database (Bairoch et al. 1994). Because the patterns are universal but still fairly detailed, they present a possible route to overcoming some of the limitations of the global amino acid substitution matrices used in sequence comparisons and the individual residue secondary structure and solvent accessibility propensities used in local protein structure prediction. The work described in this paper is a first step towards correlating local sequence patterns with local structural motifs.

Results

If there are a finite number of distinct chemical environments in proteins, there should be a finite number of patterns of variation in sets of multiply aligned sequences. Here we use cluster analysis to identify recurring patterns of variation at single positions and in short segments of contiguous positions in multiple sequence alignments. A non-redundant set of global multiple sequence alignments for proteins of known structure was extracted from the HSSP database (Sander and Schneider, 1991) as described in the Methods section. After excluding positions in which fewer than 20 sequences contributed to the alignment, the data set contained approximately 20,000 individual columns from 154 protein families.

A. Patterns at single positions

The frequencies of occurrence of the 20 amino acids at each position were calculated, and the K-means algorithm was used to group similar frequency distributions using the simple "city block" metric (*d1*, see Methods).

The amino acid groupings obtained (Table I) are consistent with expectation. The mean of the frequency distributions belonging to a given cluster provides a convenient summary statistic. To save space, the mean values of each of the 20 amino acids in each cluster are not shown, instead only the amino acids whose mean frequency of occurrence in a cluster is greater than 0.1 (upper case) or between 0.07 and 0.1 (lower case) are listed (Table I, column 3). The degree of conservation of these primary components is reflected in the variability index (column 4), which gives the number of amino acid components whose mean frequency of occurrence is greater than .05.

The patterns generally fall into either hydrophobic (clusters 1, 2 and 3) or polar (clusters 4 through 8) classes (Table I, column 6). However, the different clusters contain different combinations of hydrophobic and hydrophilic groups. For example, cluster 1 contains primarily V, I, and L while cluster 2 contains primarily I, L and M. Cluster 3 contains only aromatic residues while cluster 6 contains only negatively charged residues. Amino acid residues with special structural properties are prominent in clusters 9 (P) and 10 (G). Although the RMS deviation of points within a cluster is not dramatically less than that of points in the entire dataset (see Methods), the products of the variances are considerably lower in the former than in the latter (Table I, column 6). As outlined in the Methods, the patterns were independent of the choice of starting cluster centers implicit in the K-means algorithm. Patterns similar to those in

Table I were obtained in a Dirichlet mixture decomposition of multiple sequence alignments (Brown et al. 1993).

The first ten patterns in Table I are the result of a low resolution subdivision of sequence space (ten classes were allowed). More subtle patterns are revealed when the number of classes is increased (see Methods). For example, in cluster 11, primarily L, R and K, the common feature is the long aliphatic side-chain common to the three residues. Pattern 13 is dominated by the beta branched residues V, I and T. A cluster with conserved cysteine residues also emerges when more classes are allowed. Thus, although hydrophobicity appears to be the major feature distinguishing the largest clusters, other chemical properties are often important in the smaller clusters.

How clustered are the frequency distributions in sequence space? The K-means algorithm can always subdivide a set of points into convex subsets and does not depend on the "clumpiness" of the data. To investigate this question, random data sets were generated using the individual residue frequency distributions of the HSSP database but lacking the inter-residue correlations (see Methods). The HSSP data set and a simulated data set were subjected to the same clustering procedure and the results are compared in Figure 1.

As described in Methods, no single statistic adequately captures the spread of points within a cluster embedded in a high dimensional space. With two statistics one can do much better. We have used V, the within cluster variance per dimension, and D, the dimension of the smallest subspace that contains the cluster. Each cluster is represented as a point in Figure 1.

The most striking aspect that distinguishes the results of application of the K-means algorithm to the real (Figure 1A, open triangles) and simulated

(Figure 1A, closed triangles) data sets is the smaller number of dimensions in the former. There is also significantly greater variation in the number of dimensions per cluster in the real data set. The clusters in the random set appear to have roughly similar shapes and volumes, as expected in a relatively uniform distribution. In contrast, the size and shapes of the clusters obtained for the real data set vary considerably, presumably because different sequence patterns in protein families are constrained to different extents.

B. Comparison of weighting schemes and distance measures

Frequency distributions from multiple sequence alignments can be taken as estimators of the "true" probability distributions for substitution of the 20 amino acids at a given position in a protein, but there are two important caveats. First, there are a limited number of sequences in each family, so that observed frequencies may be inaccurate estimates because of small sample size effects. We have dealt with this problem by excluding poorly represented families and positions from the analysis. Second, and perhaps more serious, the different sequences in a family are not independent observations. Rather, they are highly correlated. Frequency distributions derived from sets of evolutionarily related sequences may be heavily biased. A particular amino acid may be highly represented in a particular position simply because it was present in a common ancestor, and not because of any underlying structural constraint.

A number of different weighting schemes have been proposed for compensation of the heavily biased sampling in evolutionarily related sequence sets (Vingron et al. 1993). We experimented with 1) a weighting scheme similar to that described in (Altschul et al. 1989) and (van Ooyen et al. 1990) in which weights are derived from a tree constructed from pairwise distances between the aligned sequences, 2) the self-consistent weighting scheme of Sander and

Schneider (Sander et al. 1991), and 3) the Monte Carlo approach to estimating Voronoi volumes described by Sibbald and Argos (Sibbald et al. 1990). Frequency distributions were recalculated for each of the weighting schemes and subjected along with corresponding simulated data sets to the K-means clustering procedure.

Space limitations prohibit the display of scatter plots for each of the weighting schemes. However, the essence of these plots can be roughly captured by the mean and variance of D, the cluster dimensionality (Figure 1). The results obtained with frequency distributions weighted using scheme (1) were very similar to those obtained with the unweighted distributions (Figure 1, compare circles to triangles).

The average cluster dimensionality was very similar for all the weighted data sets (Fig. 1C, column 3), indicating that the interrelationships among the frequency distributions are not substantially changed by the different weighting schemes. Furthermore, the resulting sequence patterns were not greatly altered by any of the weighting schemes (Fig. 1C, column 2). Since both the relative weight on a particular sequence and the probability of misalignment increase with sequence divergence, attempts at correcting the biased sampling through unequal sequence weighting may increase noise from misalignment errors. Because of the lack of dependence of the results on the weighting scheme, unit weights were used for simplicity in the experiments described in the following sections.

A similar approach was used to evaluate alternative distance measures. The Euclidean distance metric gave results very similar to that of the city block metric *d1* (data not shown). Because differences between amino acid frequencies of 0.8 and 0.6 are likely to be less significant than differences

between frequencies of 0.2 and 0.0, we experimented with the somewhat *ad hoc* distance measure *d2* which effectively down-weights differences of the former type. Again, the clusters obtained with distance measures *d2* had similar overall properties to those obtained with *d1* (Figure 1B). We also experimented with a PAM(250) matrix based distance measure and with the use of the overall covariance matrix as well as individual cluster covariance matrices to adjust for the different frequencies of the different amino acids and to relax the assumption of spherical clusters implicit in the K-means algorithm (see Methods for details).

As summarized in Figure 1C, the different distance measures gave qualitatively similar results, with the real data set consistently more clustered than the random data set (Fig. 1C, columns 2 and 4). The simplicity of the city block metric and the Euclidean metric makes them preferable over the other distance measures. Because of complications associated with the use of the Euclidean metric for clustering frequency distributions (see Methods), the city block metric was chosen for the studies described in Tables I and III. The lack of sensitivity to the details of the weighting scheme and distance measure argue that the groupings shown in Table I are inherent in the data and not simply imposed by the clustering algorithm, a conclusion supported by the degree to which the patterns agree with intuition.

C. Results of contiguous position classification

The clustering procedure can be readily generalized to treat segments of contiguous positions as described in the Methods. To investigate the types of patterns occurring on different length scales, the clustering procedure was repeated for segment lengths ranging from 3 to 15 residues using a fixed number (200) of clusters. Table II lists the average cluster dimensionality per

position for both the real and simulated data sets. As the window length increased, the variation in the average number of dimensions increased (Table II, column 4). In contrast, the variation for the simulated data set was relatively constant (Table II, column 6). Thus, the clusters adopt a wider range of shapes at larger window lengths.

Space limitations preclude the description of the patterns for each segment length. Instead, the following analysis is focused on the results for segment length nine. A detailed description of all patterns for window lengths two to fifteen can be obtained from the authors.

D. Patterns for nine consecutive positions

The distribution of clusters obtained for segment length nine is shown in Figure 2 for both the real (open triangles) and simulated (closed triangles) data sets. As in Figure 1, the clusters in the real data set are consistently lower in dimensionality than those in the random data set. The former also have a greater variety of dimensions and shapes.

Several of the patterns for window length nine are described in detail in Table IIIa along with relevant statistics. Space limitations preclude the description of even a modest number of clusters in this detail, instead we have adopted a more compact representation (Table IIIb) to show a number of common patterns found in three separate classifications using different random starting cluster means. Because the distance calculation assumes a one to one correspondence between the positions of the segments being compared, frame shifted patterns are frequently observed in which for example positions 1 to (N-1) of pattern 1 are very similar to positions 2 to N of pattern 2. To save space such frame shifted patterns are shown only once. Clusters containing less than 25 members are omitted.

As expected, many substitution patterns at individual positions are similar to those observed in the single position clustering (compare the single position substitution patterns in brackets of Table IIIb to Table I). However, because the averaging is also constrained by neighboring sequence patterns, there appear to be more subtle patterns in the contiguous sequence clusters (e.g. compare positions 1 and 3 in cluster 40).

The patterns fall roughly into three broad categories which are illustrated by the examples in Table IIIa. The first and largest category consists of patterns with pronounced amphipathicity. The first cluster in Table IIIa belongs to this category; a number of additional patterns are shown in more condensed form in Table IIIb (section G). In some positions--those in which the average hydrophobicity is either very high or very low, but the variability index is high--a simple H/P reduced code is clearly sufficient. For example, most positions in cluster 3 in Table IIIa are strongly hydrophobic but eight amino acid residues occur more than average. In contrast, the relative hydrophobicity index in some positions is at one extreme or the other, but only particular residues are allowed. For example, position 4 in cluster 44 tolerates only aromatic residues, while position 1 in the same cluster prefers V and I. In some cases, side chain size appears to be important, perhaps because of packing effects. Patterns 19, 22, and 23 contain positions in which small (A), medium (L) and large (F) hydrophobic side chains are conserved. The hydrophobicity patterns neighboring these conserved positions are in many cases quite distinctive.

The second category consists of patterns with highly conserved residues (Table IIIb, sections A-E). Interestingly, only a subset of the amino acids are absolutely conserved in any of the patterns. Clusters with conserved glycine residues are particularly common (20% of all clusters). Because of the

conformational flexibility of glycine residues, these patterns may be advantageous in local structures with unusual backbone torsion angles. Several clusters have more than one conserved glycine. For example, pattern 2 (Table IIIa) contains two consecutive conserved glycines, and pattern 1 has a GXXG motif. In pattern 6, there is a proline four positions prior to a conserved glycine, with preferences for hydrophobic residues in the two positions following the proline. In pattern 3 the aromatic residues Y and F are favored five residues prior to a conserved glycine. Other clusters containing conserved glycines and highly constrained neighboring substitution or hydrophobicity patterns are listed in Table IIIb section A.

Prolines also have unique structural characteristics. Again, there are a number of patterns with conserved prolines (Table IIIb section B) and these have additional positions with distinctive substitution and hydrophobicity profiles.

Conserved charged amino acids may be involved in metal chelation, salt bridges, or catalysis. Interestingly, patterns with conserved charged residues often have strong preferences at additional positions. For example, patterns 14 and 18 contain conserved aspartic acids with strongly hydrophobic substitution patterns at different relative positions. In pattern 13, a position rich in V, I, and L occurs three residues prior to a conserved asparagine.

The third category consists of patterns which have similar substitution patterns at all positions. For example, in Table IIIa pattern 3 has preference for I, L, and F, pattern 4 is glycine rich and pattern 5 is dominated by T and S. Fairly strong structural constraints such as the requirement for flexibility may give rise to these repetitive patterns.

It is instructive to compare the patterns described in Table III to the patterns in the Prosite database. There are a number of key differences. First, patterns listed in Table III are common to multiple protein families: the proteins around which the different multiple sequence alignments in the starting dataset are based have less than 25% sequence identity. Families with particularly divergent sequences are represented several times (there are four globin chains and three immunoglobulin chains in the set), but since most of the clusters have on the order of fifty members, a particular pattern would have to occur in more than ten different places within a single protein for a single family to contain the majority of instances of the pattern. In contrast, Prosite patterns most often characterize single protein families. Second, the patterns in Table III contain no gaps (perhaps the major shortcoming of the current approach), while Prosite patterns can extend for substantially longer stretches. Third, the patterns are obtained in quite a different way. The patterns in Table III are generated completely automatically without any information other than the amino acid sequence, while the patterns of Prosite depend on the prior classification of sequences into functional or structural groups.

Primarily because of the first factor, there is not a large overlap between the patterns contained in the two sets. This reflects a more fundamental difference: the conserved patterns in Prosite reflect either functional constraints or quite specific structural constraints, while the patterns of Table III probably arise from more general structural constraints or properties common to many different classes of proteins.

E. Variation patterns and substitution matrices

It is interesting to compare the association of amino acids in clusters with conventional substitution matrices which estimate the cost of substituting one

residue type for another. One of the most powerful current substitution matrices is the BLOSUM62 matrix which was generated from the Blocks database of multiple sequence alignments (Henikoff et al. 1992). The relationship between the BLOSUM substitution matrix and the clusters of Table I is simple: the value of a particular cross term in the substitution matrix is a function of the (weighted) average probability that two residues will be in the same cluster. It should be pointed out that our analysis relies on the alignments contained within the HSSP database, which were generated using a conventional substitution matrix (McLachlan 1971).

There are instructive differences in the performance of the PAM(250) (Gonnet et al. 1992) based distance measure *d3* (see Methods) for single positions versus strings of contiguous positions. As shown in Figure 1C, use of the PAM matrix in clustering of single positions gives results quite similar to those of the simple distance measure (1). However, for segment length nine, many of the patterns which contain highly conserved residues were not present when the PAM matrix was used and there were many fewer patterns overall (data not shown). The averaging involved in the use of a substitution matrix, although not detrimental for the patterns in Table I, which in any event are averages over large numbers of different local contexts, results in considerable loss of sensitivity for comparisons between segments of contiguous positions.

It is clear from Table III that substitution patterns are position-dependent. There have been numerous proposals for grouping the 20 amino acids into smaller numbers of sets in order to make the analysis of sequence to structure mapping more manageable. One approach groups amino acids according to their similarity based on standard substitution matrices. For example, the sub-groupings 1) I,L,M,V; 2) F,Y; 3) H,K,R; 4) A,P,S,T; and 5) D,E,N,Q were derived

from analysis of the Dayhoff substitution matrix (Risler et al. 1988). Mixed codes based on chemical properties of the amino acids have also been proposed (French et al. 1983), the suggested groupings were 1) L,M,I,V,F; 2) R,K,E,D; 3) Q,N,T,S. The wide variety of groupings shown in Table III suggests that any reduced code will have limited generality.

Discussion

We have described a completely automated approach to identifying recurring sequence motifs in protein families. The patterns identified here (see Tables I and III) probably include most of the local motifs which transcend protein family boundaries for proteins of known structure. Because of the numerous factors which enter into the determination of protein structures, the data set is probably somewhat biased and there may well be additional patterns in the large number of protein families for which structures are not available.

The clustering procedure used here, although simple, appears to be quite adequate for modeling the data-- the local covariances of residue occurrences found in multiple sequence alignments. First, the independence of the results from the choice of starting cluster centers required for the K-means algorithm attests to the numerical stability of the procedure. Second, the results are surprisingly robust to changes in the distance metric and sequence weighting schemes (Table I and Figure 2). Third, most of the patterns obtained for individual positions (Table I) and many of the patterns obtained for segments of contiguous positions (Table III) are consistent with expectation (the division between hydrophobic and polar patterns in Table I is perhaps the simplest example).

Our results permit limited but significant generalizations about the distribution of protein amino acid sequences in sequence space. The

robustness of the results suggests that the majority of the patterns are reasonably well separated from one another. Furthermore, the distribution of sequences in protein families appears to be considerably more "clumpy" than random distributions. The clusters obtained for the real protein sequence data are consistently lower in dimensionality than those identified in applications of the same clustering procedure to random datasets (Figures 1,2, Tables I-III).

The classification of positions into different clusters provides a simple yet potentially powerful means to abstract the information contained in multiple sequence alignments into a higher level representation. A multiple sequence alignment can be replaced by a sequence of cluster numbers with relatively little loss of information. The resulting higher level sequences can be subjected to much the same types of analysis as normal amino acid sequences in efforts to correlate sequence with structure(Rost et al. 1993).

Our results may have useful applications for sequence comparisons, in particular for the identification of distant homologues for newly determined sequences. It is well established that searches with profiles constructed from sets of aligned sequences are considerably more sensitive to distant homologues than searches with single sequences. The reason for this is simple: a sequence profile contains at each position family specific information about the likelihood of different amino acids to substitute at that position, while a search with a single sequence typically uses the same global substitution matrix at each position. As mentioned in the introduction, the use of a single substitution matrix may average out weak but important similarities, whereas our clusters are in fact strings of distinct substitution matrices. One can imagine using the clusters as "generalization rules" whereby the substitution matrices

generated from the closest cluster or clusters to each segment of a query sequence are used for scoring sequence alignments.

A similar strategy may facilitate extrapolating from a small number of aligned sequences. The idea is that given a small sample of the variation possible at a given position, the closest clusters can be identified to predict the variation likely to be observed in new members of the same family. Generalization in this fashion may permit the power of profile-based searching to be employed with only a few examples from the sequence family (or perhaps from only one example).

One way to implement the strategy described in the previous paragraph would be to use the variation patterns of Table III to generate a rough profile or sets of profiles for new sequences which have no close relatives: for each segment of nine residues in the sequence, select the closest pattern (or a weighted average of nearby patterns) and build a profile by splicing together the variation patterns for the different segments. Next, search the database with this inferred profile. This procedure potentially circumvents the limitations associated with using the same substitution matrix at each position of a sequence. The method may also be useful for generalizing from a small number of aligned sequences, but once there are more than 5-10, the substitution patterns are probably better inferred directly from the aligned sequence set.

There are also potential applications to protein structure prediction. There is a significant correlation between the local structures adopted by members of a given cluster, although the extent of correlation varies from cluster to cluster. For example, more than 80% of the occurrences of the first two patterns in Table IIIa in known protein structures are in alpha helices.

Intriguingly, the conserved charged residues in patterns 13, 15 and 16 in Table IIIb are buried in more than 70% of the occurrences of the patterns. Pattern 7 in Table IIIa is very similar to the Schellman helix C-terminal capping motif (Aurora et al. 1994) and as expected occurs frequently in helix caps. A more extensive analysis of the structural correlates of the sequence patterns will be presented elsewhere. The tracing of the structural correlates of sequence patterns is essentially the complement of the more standard (and very powerful) procedure of tabulating the frequencies of occurrence of the 20 amino acids in different structural environments (Bowie et al. 1991; Chou et al. 1978).

Finally, we should note that the results described here are highly dependent on the quality of the starting multiple sequence alignments. As the amount of sequence data increases and multiple sequence alignment algorithms are improved, approaches similar to the one described here should become increasingly powerful.

Methods

A. The Data: Multiple sequence alignments for proteins of known structure were taken from a non-redundant subset (PDB select 25; (Rost et al. 1993) of the HSSP database (Sander et al. 1991). No two multiple sequence alignments in this subset have parent sequences with greater than 25% identity. Because of the wide degree of sequence variation in families such as the globins and the immunoglobulins, the PDB select 25 list does include more than one chain per family in several cases (there are four globin chains and three immunoglobulin chains, for example). To reduce the problems associated with small sample size, families with fewer than 20 members were excluded from the analysis. Insertions common to less than 20 members of larger families were also excluded (the HSSP database consists of global sequence

alignments). The final data set included 154 protein families with an average of 98 sequences per family.

B. Distance Measure: Cluster analysis requires a metric on the space to be clustered. An advantage of using multiple sequence alignments is that there is a natural choice of metrics: the difference in the frequency distributions. A particularly simple choice is the "city block" metric:

$$d1(i,j) = \sum_{k=1}^{20} |F(i,k) - F(j,k)| \qquad (1)$$

where *d1(i,j)* is the distance between frequency distributions *i* and *j* and *F(i,k)* is the frequency of occurrence of the kth amino acid at position *i*, $\sum_{k=1}^{20} F(i,k) = 1$.

A distance measure for comparing single positions can be readily generalized to treat strings of contiguous positions. The distance between one segment of a multiple sequence alignment and a second segment of the same length is conveniently defined to be the sum of the distances between each of the corresponding positions:

$$d_N(i,j) = \sum_{n=0}^{N-1} d(i+n, j+n) \qquad (2)$$

where N is the length of the window, *i* and *j* are the starting positions of the first and second segments, and *d(i+n,j+n)* is for example distance measure *d1* above.

C. Cluster Analysis. The data set consists of roughly 20,000 frequency distributions. Most clustering algorithms become extremely time consuming with data sets of over 1000 members. The K-means algorithm is one of the few that can be used with extremely large data sets. In brief, a set of K initial cluster centers are chosen at random and each data point is assigned to the closest

center. New cluster centers are then determined by taking the mean of all of the data points in each cluster, and each data point is reassigned to the closest center in another pass through the data set (Everitt 1993). This simple iterative scheme of recalculating cluster means and reassigning data points to clusters is repeated until no data points are moved from cluster to cluster.

For technical reasons, the city block metric is somewhat preferable to the Euclidean metric for clustering frequency distributions using the K-means algorithm. Viewed as vectors in a 20N dimensional space, the frequency distributions vary widely in absolute magnitude (for window length one, a position in which only one amino acid occurs is represented by a vector of length one, while a position in which all twenty amino acids occur with equal probabilities is represented by a vector with length $[20 \times (1/20)^2]^{1/2} = 0.22$). The Euclidean distance between a position in which ten of the amino acids occur with equal frequencies and a position in which the other ten amino acids occur with equal frequencies is .45, while the distance between two positions in which different residues are absolutely conserved is 1.4. The city block distance between the two sets of positions is the same (1.0) in both cases, a more satisfactory results since no residues are in common in either pair. To avoid the problems associated with the use of the Euclidean metric with variable magnitude frequency vectors, the frequency vectors can be normalized to unit magnitude. However, the updating procedure basic to the K-means algorithm also changes the absolute magnitude of the cluster centers. The latter can be kept fixed, but this requires a somewhat awkward renormalization step after each reassignment of vectors to clusters in the K-means procedure.

D. Error measures. How is the extent of clustering best evaluated? An explicit example illustrates the difficulties with evaluating different clustering

strategies in high dimensional spaces, and in particular with data of the type involved here. Consider a position which can tolerate either of two amino acids, for example valine and isoleucine. With a small and possibly biased sample, the frequency of occurrences of the two residues may range from 0.0 to 1.0; the constraint being that the variation is contained within a two dimensional subspace of the entire 20 dimensional space (only valine and isoleucine are allowed). The maximum distance between two points in this subspace is the same as the maximum distance between two points in the entire 20 dimensional space (2 in both cases). The mean distance of the members of a cluster from the cluster mean is clearly a poor measure of the dimensionality of the cluster.

Two statistics which have proved useful for capturing the distribution of points within a given cluster are D, the number of dimensions for which the cluster mean exceeds 0.02 (chosen empirically), and V, the average variance in these dimensions. D clearly indicates the dimensionality of the subspace in which the cluster lies, and V, the average spread within this subspace.

To assess the extent of clustering of the sequence data, parallel experiments were carried out on simulated data sets. To construct these sets, the frequency distributions for each of the 20 amino acids were evaluated and then used to generate randomized versions of the HSSP database. The statistics of the simulated data sets are essentially those of the HSSP database with all covariances between substitutions at particular positions or between nearby positions set to zero. For each weighting scheme, a separate simulated dataset was generated based on the amino acid frequency distributions of the corresponding weighted dataset. We note that the more standard procedure of

randomization by shuffling does not apply here since we are not seeking family specific patterns.

A single composite statistic--the product of the variances of the individual residue frequencies--is also given in several of the tables to facilitate comparison between different positions within the same cluster. This crude volume measure is normalized by division by the corresponding quantity for the whole data set:

$$Volume(l) = \frac{\prod_{k=1}^{20} \frac{1}{M_l} \sum_{j=1}^{M_l} |F_l(j,k) - \langle F_l(j,k) \rangle|}{\prod_{k=1}^{20} \frac{1}{S} \sum_{j=1}^{S} |F(j,k) - \langle F(j,k) \rangle|}$$

where $F_l(j,k)$ is the frequency of the *kth* amino acid in the *jth* distribution in cluster l, $\langle F_l(j,k) \rangle$ is the center of the *lth* cluster, and $\langle F(j,k) \rangle$ is the center of the entire data set. M_l is the number of vectors (or distributions) in cluster l, and S, the number in the whole dataset. To reduce the effects of small sample size artifacts, .001 is added to the terms in the product in the numerator (again, the value of .001 was determined empirically).

E. Numerical stability, alternative distance measures and the K-means algorithm. A disadvantage of the K-means algorithm is that both the number of clusters and the starting cluster centers must be specified in advance. In practice, use of more than the natural number of groupings results in the subdivision of several of the larger clusters. This is easily recognized, and each pattern is shown only once in Tables I-IV. The numerical stability of the algorithm and the dependence of the results on the starting cluster centers were assessed by carrying out multiple independent calculations using different sets of starting centers. Only the recurrent clusters are reported in the tables.

A potential disadvantage of distance measure *d1* (eq.1) is that a difference in frequency of .1 is treated similarly regardless of whether the difference is between 0.7 and 0.6 or between 0.1 and 0.0. Because of lineage effects, the former is likely to be less informative than the latter. A simple exponential scaling was used to emphasize differences of the latter type:

$$d2(i,j) = \sum_{k=1}^{20} |\exp[-F(i,k)] - \exp[-F(j,k)]| \qquad (3)$$

The K-means algorithm implicitly assumes the clusters to be spherical. If several variables are highly correlated or have significantly different variances, clusters may resemble prolate ellipsoids more closely than spheres. Non-spherical clusters can be accommodated by calculating the within-cluster covariance matrix and using the generalized Mahalonobus distance given by equation (4) when assigning data points to clusters (Everitt 1993):

$$d3(i,j) = [|\mathbf{F}_i - \mathbf{F}_j|]\mathbf{M}[|\mathbf{F}_i - \mathbf{F}_j|] \qquad (4)$$

where $\mathbf{F}_i = F(i,k)$ and **M** is the inverse of a covariance matrix.

If the number of dimensions is of the same order as the number of data points in individual clusters, the matrix inversion required is not possible. In this case the inverse of the covariance matrix can be approximated by inverting the diagonal elements (the variances) and setting off-diagonal elements to zero. The modified K-means method in this case leads to minimization of the effective volume of the clusters rather than the average within-cluster distances.

Distance measure *d3*, with M equal to an amino acid substitution matrix such as a PAM matrix weights differences according to the likelihood of substitution of one residue type for another (Dayhoff et al. 1972). This is a simple generalization of the similarity measure used in comparing single

sequences that are distantly related. This measure, essentially a return to the single substitution matrix approach mentioned in the introduction, is clearly only useful in the limit of small numbers of sequences per family.

Acknowledgments:

We thank H. Schneider for the HSSP database; D. A. Agard for encouragement and computational resources; N. Hunt for compiling the PDB/ϕ-ψ dataset; S. Henikoff, J. Henikoff, S. Pietrokovski, D. Yee, K. Zhang, N. Hunt, T. Defay, M. Robinson, D. Teller, S. Karlin, L. Brocchieri, and members of the Agard lab for critical reading of the manuscript. K. F. Han is supported by the Howard Hughes Medical Institute Predoctoral Fellowship. This work was partially supported by the National Science Foundation, Science and Technology Center Cooperative Agreement BIR-9214821 and young investigator awards to D.B. from the NSF and the Packard Foundation. This chapter is an approved reprint as it appears in K.F.Han and D.Baker (1995), *J. Mol. Biol.*, **251**,176-87.

References

Altschul, S.F., Carroll, R.J. & Lipman, D.J. (1989). Weights for Data Related by a Tree. *J. Mol. Biol.,* **20**, 647-53.

Aurora, R., Srinivasan, R. & Rose, G.D. (1994). Rules for alpha-helix termination by glycine. *Science,* **264**, 1126-30.

Bairoch, A. & Bucher, P. (1994). PROSITE: recent developments. *Nucl. Acids. Res.,* **22**, 3583-89.

Brown, M., Hughey, R., Krogh, A., Mian, I., Sjolander, K. & Haussler, D. (1993). Using Direchlet Mixture Priors to Derive Hidden Markov Models for Protein Families. First International Conference on Intelligent Systems for Molecular Biology.

Dayhoff, M.O., Eck, R.V. & Park, C.M. (1972). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. Washington, D.C., National Biomedical Research Foundation.

Everitt, B. (1993). Cluster analysis. New York, Halsted Press.

French, S. & Robson, B. (1983). What is a conservative substitution? *J. Mol. Evol.,* **19**, 171-75.

Gonnet, G., Cohen, M. & Benner, S. (1992). Exhaustive matching of the entire protein sequence database. *Science,* **256**, 1443-45.

Gonnet, G., Cohen, M. & Benner, S. (1994). Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix. *Biochem. Biophys. Res. Commun.,* **199**, 498-96.

Gribskov, M., Luthy, R. & Eisenberg, D. (1990). Profile analysis. *Methods in Enzymology,* **183**, 146-59.

Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U S A,* **89**, 10915-19.

Johnson, M., Overington, J. & Blundell, T. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.,* **231**, 735-52.

McLachlan, A.D. (1971). Identification of common molecular subsequences. *J. Mol. Biol.,* **61**, 409-24.

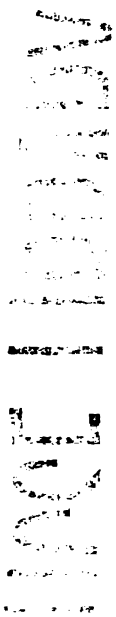
Risler, J.L., Delorme, H.D. & Henaut, A. (1988). Amino Acid Substitutions in Structurally Related Proteins, A Pattern Recognition Approach. *J. Mol. Biol.,* **204**, 1019-29.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.,* **232**, 584-99.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins,* **9**, 56-68.

Sibbald, P. & Argos, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.,* **216**, 813-18.

van Ooyen, A. & Hogeweg, P. (1990). Iterative character weighting based on mutation frequency: a new method for constructing phyletic trees. *J. Mol. Evol.,* **31**, 330-42.



Vingron, M. & Sibbald, P.R. (1993). Weighting in Sequence Space: A comparison of methods in terms of generalized sequences. *Proc. Natl. Acad. Sci. USA,* **90**, 8777-81.

Cluster#	# of members	Dominant Substitutions	Variability Index	Hydrophobicity	Relative Cluster Volume
1	2449	V,I,l	3	.832	2.3e-4
2	1971	L,i,m	5	.853	5.4e-4
3	1521	Y,F,w	4	.818	1.6e-3
4	1166	N,H,d	4	.151	7.4e-4
5	2263	R,K,q	4	.163	5.8e-3
6	2396	D,E	4	.148	2.5e-3
7	1401	T,s	3	.237	2.4e-4
8	1412	S,a,t	3	.199	3.3e-4
9	2214	P,A	3	.538	1.2e-3
10	1349	G	2	.166	4.3e-4
11	84	L,R,K	4	.450	2.8e-3
12	150	G,N,k	4	.101	2.4e-6
13	114	V,I,T	4	.687	1.1e-5

Table I: Recurrent patterns at individual positions. The amino acids which occur with frequencies greater than 0.1 are shown in upper case, those which occur at frequencies between 0.07 and 0.1, in lower case (column 3). The number of amino acids which occur at frequencies greater than 0.05 is given in column 4. The average summed frequency of occurrence of the amino acids A, V, I, L, M, F, W, C is listed in column 5.

Table II: Results for clustering segments of contiguous positions into 200 classes. Note that as the window length increases, the total number of frequency vectors decreases slightly as N-1 positions are lost from each end of each sequence, where N is the segment length.

Window Length	Number of Vectors	Real Data Cluster Dimensionality		Simulated Data Cluster Dimensionality	
		Mean	Variance	Mean	Variance
3	21146	12.07	.57	14.79	.16
5	20812	13.51	.66	15.76	.15
7	20483	14.22	.85	16.36	.12
9	20157	14.50	1.42	16.48	.11
11	19833	13.75	2.35	16.72	.18
13	19517	14.79	2.61	16.80	.15
15	19204	14.69	3.39	16.95	.17

Table IIIa

Cluster Number	Size	V	D	Prominent AA	Variability Index	Hydrophobicity	Relative Cluster Volume
1	134	.94	15.3				
				VIL	5	.73	5.3E-5
				TSd	6	.19	2.1E-5
				paRkDE	8	.26	4.4E-2
				AskDE	6	.20	2.4E-3
				QDE	3	.15	8.4E-2
				AVIL	6	.72	6.2E-2
				AKdE	8	.30	1.4E-2
				AqrkE	8	.30	3.4E-1
				aViL	6	.68	4.9E-2
2	210	.95	14.4				
				aKDE	6	.21	2.0E-2
				AkDE	8	.28	5.2E-2
				aVIL	5	.77	5.1E-2
				ArKE	8	.32	1.2E-2
				AKDE	9	.22	3.1E-2
				AL	9	.55	8.0E-2
				VILf	6	.83	5.4E-2
				aqRKE	7	.24	6.3E-2
				ArKdE	9	.25	3.2E-2
3	148	.39	11.7				
				avILFt	8	.70	3.8E-12
				AILFw	8	.74	4.1E-10
				GAiLFts	8	.64	2.7E-13
				GAvILF	9	.73	1.1E-10
				GAvILFts	8	.66	2.4E-10
				GAvILFts	8	.70	2.9E-13
				gAvILF	8	.76	2.0E-10
				avILF	6	.79	1.1E-10
				gAiLF	9	.68	2.2E-11

4	133	1.02	15.6				
				GAvts	7	.41	1.7E-3
				GAs	9	.40	4.2E-4
				GAvs	7	.44	2.1E-1
				GAvs	7	.41	1.4E-1
				GAvs	7	.41	5.3E-1
				GAls	7	.41	4.9E-1
				GAls	7	.41	2.6E-1
				GpAs	8	.39	3.1E-1
				GAvls	8	.40	4.2E-1
5	107	.67	16.1				
				TS	4	.25	3.5E-7
				avTS	8	.41	2.1E-8
				gaTSD	7	.25	3.2E-6
				gpatSnD	8	.23	1.9E-8
				gtSd	10	.33	5.1E-5
				paTSnq	8	.30	8.6E-5
				aTSn	7	.30	5.5E-5
				PaTSn	7	.25	1.3E-4
				gptSnd	7	.24	5.3E-4
6	74	.67	16.1				
				Gpavlk	8	.39	8.4E-5
				Vlf	9	.58	1.6E-3
				ilTr	9	.45	1.1E-2
				P	2	.20	2.0E-4
				VILe	6	.66	2.4E-3
				ailyFe	8	.67	3.0E-3
				aVltSn	7	.45	6.1E-3
				G	1	.06	9.4E-6
				GalfSk	7	.39	9.2E-4
7	67	.88	15.3				
				AkDE	8	.30	1.1E-4
				AILfk	7	.55	2.2E-2
				AL	5	.59	2.1E-2
				ARKe	7	.27	7.1E-3
				gaKE	8	.29	5.1E-3
				aLyf	8	.56	2.7E-2
				G	1	.04	7.7E-4
				aVIL	6	.62	1.0E-2
				tskDe	7	.28	1.6E-2

Table IIIb

	Size	V	D
A. Conserved glycine			
1. [GAv] [G] . . [G] . . . [gaV]	146	0.87	15.2
2. . [G] [G] . . . π . π	100	0.86	14.8
3. [Yf] . π . . [G] . [LsD] .	69	0.72	13.8
4. π . . π [G] . [sDe] [IYF] π	37	0.91	15.4
5. . . . [VIL] . . [G] [gAS] .	228	0.85	15.7
6. . . . [P] [VIL] ϕ . [G] .	74	0.85	14.3
7. . [AvL] [ArK] π [ALy] [G] . π [VIL]	161	0.89	14.6
8. [ARK] . [AVI] [AL] π π . [G] [AVI]	167	0.95	15.0
B. Conserved proline			
9. [VIL] [P] . . . π π . .	147	0.76	15.2
10. π ϕ . [VIL] [P]	101	0.72	14.3
11. . ϕ . . [Ats] [P] . . [aVI]	152	0.81	16.1
12. [PLF] . [GAV] ϕ [P] π . . .	54	0.75	12.1
C. Conserved polar residues			
13. [VIL] . . [N] . . π . .	61	0.81	14.0
14. . . [D] . . . [iln] [TSh] π	54	0.76	14.4
15. . . [VIL] [D] . . π π π	138	0.91	15.7
16. . [AVt] . . [D] . . . π	149	0.84	15.4
17. . [T] π . . [AvT] . . .	90	0.76	14.9
18. P P [qDe] [LYF] [vLF] X X [D] X	76	0.81	13.8
D. Conserved nonpolar residues			
19. [iLF] . π [A] ϕ π π . π	136	0.89	14.6
20. . . π . [iLm] . . [A] [ViL]	126	0.73	14.3
21. . [AtS] [A] ϕ . [ALY] [AVF] [LmQ] π	69	0.76	14.6
22. [AvL] π π [L] . π . [vIL] .	228	0.92	15.9
23. π π [F] π π . π π .	93	0.84	14.4
E. Conserved arginine/lysine			
24. . . . [Rk] [LFw] . . . [IF]			
25. [ASD] . [AVT] π . [RK] . [vIL] .	74	0.81	12.7
26. [ThR] . [RK] [LFK] [VIL] [VIL] [AvI] [AY] .	63	0.81	12.2
27. . π [RK] [gPA] π [Hde] [AVI] . [AvI]	35	0.79	13.1
F. Threonine and serine			

28. . [aiT] [PTS] π π . π . [iLm]	59	0.70	14.1
29. [TSd] π π [QDE] ϕ . π [AVL] .	145	0.91	14.6
30. π [TSQ] . [aTS] . [GLN] . ϕ [Vlw]	26	0.62	13.0
31. . [iTS] [TS] π π π [aTS] [ATS] π	65	0.76	14.0
32. . [gAS] π [ATS] ϕ [iL] [gAT] . .	95	0.76	15.2
33. [vwT] . [tSq] π . . [vIL] [Re] π	54	0.93	13.8
G. Alternating hydrophobic-polar			
34. [VIL] [TSd] π π [QDE] [AVIL] π π ϕ	134	0.94	15.3
35. [aKD] π [aVI] π π . [VIL] π π	210	0.95	14.4
36. π ϕ . [vIL] [aVT] [VIL] [PAS] . [VIL]	69	0.65	12.2
37. . [ViL] π [viL] . π . [VIL] .	63	0.78	15.0
38. [GND] . π [VIL] . [VIL] . . π	99	0.87	15.2
39. . π π π π [VIL] . [VIL] .	122	0.84	15.0
40. [VIT] π [VIL] [Vil] . π π π π	111	0.86	14.4
41. . . [VIt] . [aVI] [VIT] . [Pa] π	70	0.81	14.3
H. Miscellaneous			
42. ϕ [GAs] . ϕ . . . . [gAS]	29	0.76	14.2
43. [PVL] [VIL] [ViL] [gAl] [AVY] . [PNH] . .	58	0.70	12.2
44. [VI] π . [YFW] . . [WTR] π .	43	0.77	14.5
45. π ϕ . [GYs] [NHR] [PYF] [iLF] . [gAR]	33	0.89	11.2
46. ϕ ϕ . ϕ ϕ ϕ ϕ [avl] ϕ	148	0.39	11.7

Table III. Recurring patterns for nine consecutive positions. A. Detailed statistics for several selected clusters. B. Condensed representation of selected patterns. Positions with variability indices less than six are described by amino acids in brackets (upper and lower cases are as in Table I). The remaining positions are represented by 'ϕ' (average hydrophobicity greater than 0.65), 'π' (average hydrophobicity less than 0.35) or '.' (average hydrophobicity between 0.35 and 0.65).

Figure Legends

Figure 1: Comparison of different weighting schemes and distance measures for both the real and random data sets. Each symbol represents a single cluster; the x axis is the number of non-zero dimensions, the y axis, the average variance. A. Comparison between unweighted (triangles) and weighted (scheme 2, circles) data sets. B. Comparison between the city block metric (triangles) and distance measure *d2* (circles). Open symbols, clusters generated from the real data set; closed symbols, clusters generated from a simulated data set. The clusters for the real set are numbered as in Table I. Note that the weighting scheme changes the residue frequency distributions such that the within-cluster variance is higher for both real and simulated data sets. C. Summary of statistics for the different weighting schemes and distance measures. Column 1 describes the trial, column 2 lists the fraction of patterns that were found in the unweighted data set clustered using the city block metric (trial 1), columns 3 and 4 list the mean and variance of the cluster dimensionality for the real data set; columns 5 and 6, the same quantities for the simulated data set. The city block metric was used for the comparison of weighting schemes (trials 1-4), and the unweighted data set was used for the comparison of distance measures (trials 5-8). The weighting scheme trials are: 1. no weights, 2. tree-based weights, 3. self consistent weights, 4. Voronoi weights (see text for more description). The distance measure trial 5 utilized *d2* and trials 6-8 utilized *d3* with the matrix M the PAM(250) substitution matrix, the overall covariance matrix, and within cluster covariance matrices respectively (see Methods). For trial 8, covariance matrices were calculated for each of the clusters generated using the standard procedure (trial 1) and used for a second round of clustering as described in the Methods.

Figure 2: Scatter plots of the number of non-zero dimensions (D) and average variance (V) for each cluster obtained in the nine-consecutive position classification. A. Real data set, B. Simulated data set.

Figure 1a

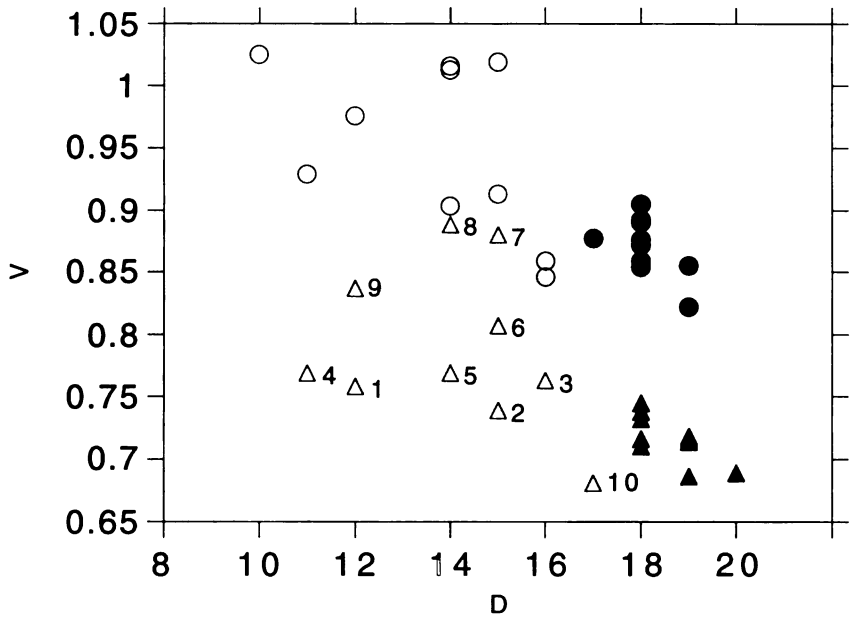


Figure 1b

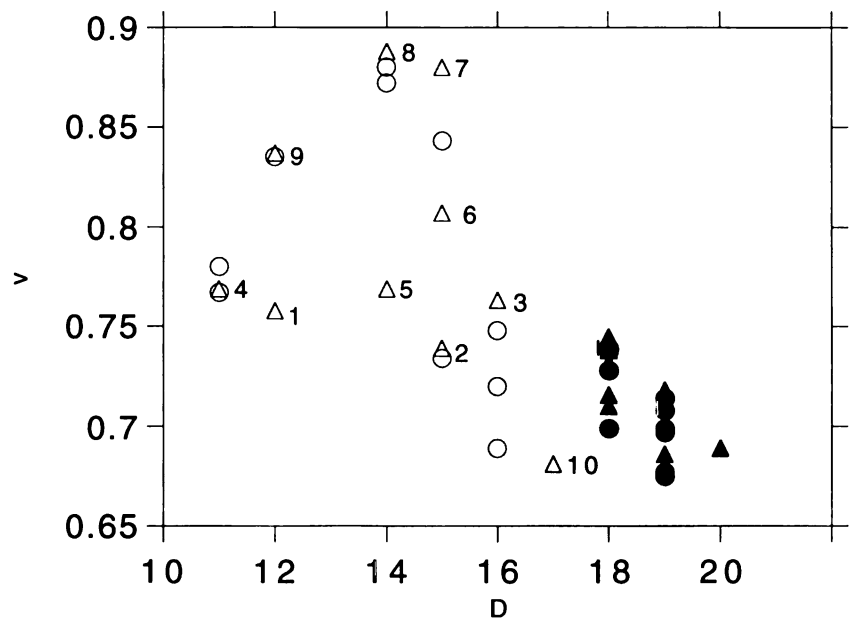


Figure 1c. Summary statistics

	Pattern Conservation	Real Data Cluster Dimensionality		Simulated Data Cluster Dimensionality	
		mean	variance	mean	variance
Weighting Scheme					
1.	1.0	11.8	.54	16.3	.17
2.	.9	11.8	.54	16.5	.15
3.	.9	12.1	.47	16.5	.18
4.	.7	12.6	.46	16.5	.20
Distance Measure					
5.	.8	11.8	.45	17.5	.06
6.	.9	11.8	.51	16.7	.08
7.	.7	11.1	.84	14.9	.25
8.	1.0	11.3	.39	13.5	.02

Figure 2a

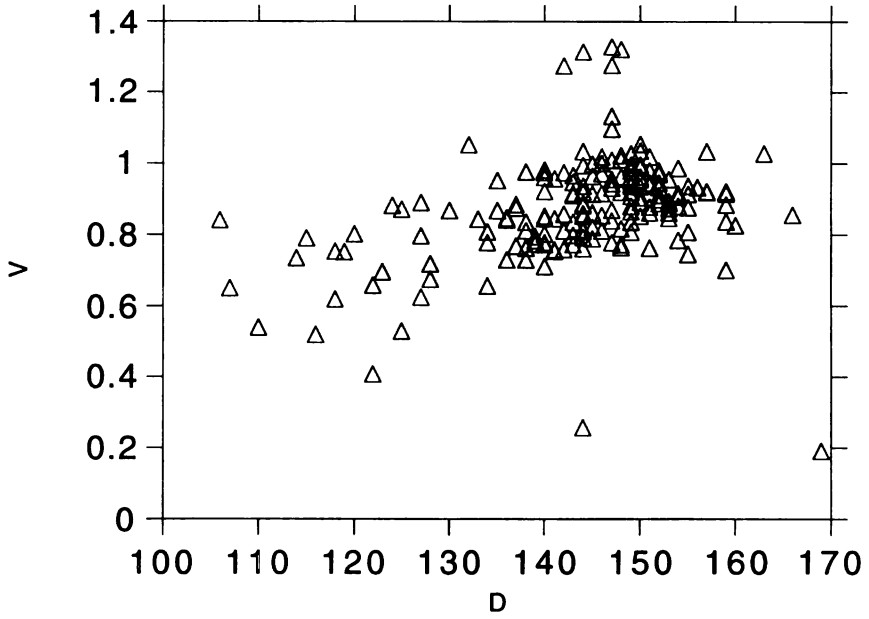
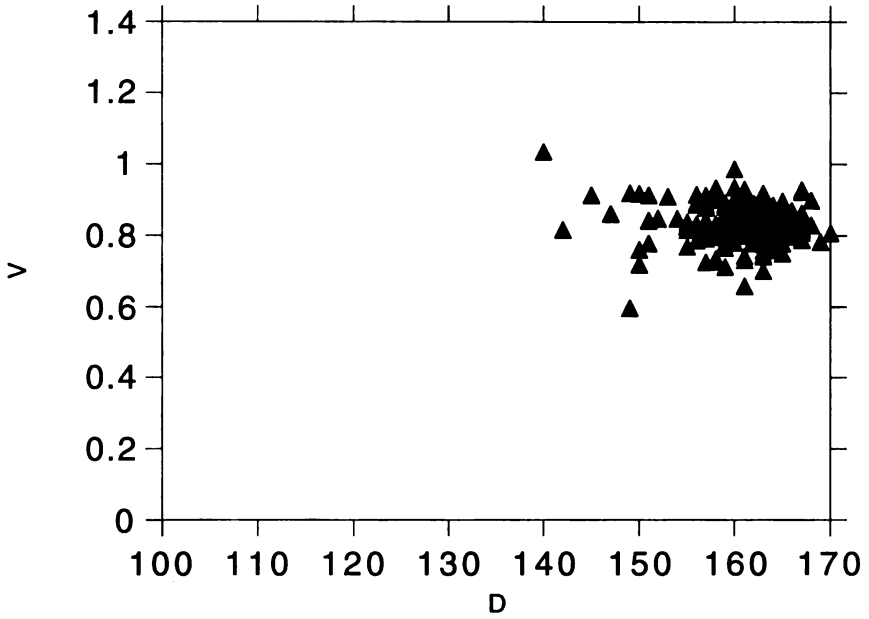


Figure 2b



Chapter 6

Global properties of the mapping between local amino acid sequence and local structure in proteins.

Local protein structure prediction efforts have consistently failed to exceed ~70% accuracy. We characterize the degeneracy of the mapping from local sequence to local structure responsible for this failure by investigating the extent to which similar sequence segments found in different proteins adopt similar three dimensional structures. Sequence segments 3-15 residues in length from 154 different protein families are partitioned into neighborhoods containing segments with similar sequences using cluster analysis in conjunction with a measure of sequence similarity. The consistency of the sequence-to-structure mapping is assessed by comparing the local structures adopted by sequence segments in the same neighborhood in proteins of known structure. 45% and 28% of the positions in our protein database occur in neighborhoods in which one and two local structures predominate, respectively. The sequence patterns which characterize the neighborhoods in the first class probably include virtually all of the short sequence motifs in proteins that consistently occur in a particular local structure. These patterns, many of which occur in transitions between secondary structural elements, are an interesting combination of previously studied and novel motifs. The identification of sequence patterns which consistently occur in one or a small number of local structures in proteins should contribute to the prediction of protein structure from sequence.

Introduction

Most studies of local sequence-structure relationships have involved the tabulation of statistics on sequences which occur in structural motifs of interest ((1, 2, 3, 4), Figure 1A). Our approach (Figure 1B) is essentially the inverse. Instead of investigating the sequence patterns found in predefined local structural environments, we first identify recurring sequence patterns and then investigate their structural correlates.

It is well established that the local sequence-to-structure mapping is not one to one over all of sequence space: identical pentapeptide sequences exist in completely different tertiary structures in proteins (5). Furthermore, local structure prediction efforts consistently fail to exceed ~70% accuracy (6), suggesting that the mapping from local sequence to structure is likely to be degenerate for a significant fraction of sequence space. In this paper we characterize the degeneracy of the mapping by determining the number of sequence segments in neighborhoods (regions of sequence space) in which the sequence-to-structure mapping is one to one, one to two, and one to three (Figure 1B).

The definition of local sequence neighborhoods requires a measure of distance between short sequence segments. Most sequence comparison methods rely on a single global substitution matrix compiled by averaging over all positions in a large set of aligned protein sequences (7). However, at different positions in proteins, different amino acid residues are likely to substitute for each other, and thus the use of a global substitution matrix is potentially problematic. These problems can be circumvented if the two segments being compared are both derived from protein families with multiple members: sequence profiles (8) constructed from sets of aligned sequences contain position-specific information on amino acid substitution patterns.

In previous work (9), we utilized a measure of the distance between sequence profiles generated from multiple sequence alignments to identify sequence patterns that transcend protein family boundaries. A similar distance measure is used in this paper, and the term "segment" below refers to a segment of a profile generated from a multiple sequence alignment. The earlier work focused on the identification and characterization of recurring sequence patterns; the focus of the current paper is on the structural correlates of these patterns.

Methods

The clustering procedures have been described in detail (9). In brief, 29921 segments of profiles derived from a non-redundant set (PDB-select 25 (10)) of the HSSP database (11) of multiple sequence alignments were subdivided into 1200 neighborhoods containing sets of related segments using the K means algorithm (12) and the city block metric

$$d(i,j) = \sum_{n=1}^{N} \sum_{k=1}^{20} |F_i(k,n) - F_j(k,n)|$$

where $F_i(k,n)$ (a profile segment) is the frequency of the kth amino acid in the nth position of segment i and N is the segment length. Because the PDB-select 25 subset contains very few pairs of alignments from even distantly related families, segments in a given neighborhood are necessarily derived from quite different protein families. To capture patterns of different lengths, the procedure was repeated for segment lengths ranging from three to fifteen residues. Frequently, segments of length nine to fifteen which belonged to neighborhoods with strong sequence to structure correlations contained shorter segments which also belonged to such neighborhoods. To avoid over counting, the statistics in the tables for a given segment length exclude positions already included in the statistics for a longer segment length.

Secondary structure and solvent accessibility data for each of the segments in each of the neighborhoods were extracted from the HSSP database using previously described simplifications (6). The average consistency of secondary structure within a neighborhood was evaluated using the simple formula:

$$\frac{\sum_{i=1}^{N} \max(p_{i,helix}, p_{i,strand}, p_{i,turn})}{N}$$

where the p_i are the frequencies of occurrence of the indicated secondary structure among the segments in the neighborhood at position i and N is the segment length. For N greater than seven, the lowest scoring position was excluded from the average to allow for ambiguities in secondary structure assignments in transition elements.

To test the statistical significance of the results with the HSSP data set, simulated sequences were generated with the average occurrences and variances of each of the amino acid residues in the HSSP data set, but not the inter residue correlation's (9). To preserve the non-random sequential correlation's in secondary structure elements, the secondary structure assignments were not shuffled in the simulated set.

The secondary structure consistency within the neighborhoods generated from the HSSP data set frequently exceeded 80%, but almost never reached 80% in the simulated data set. A consistency threshold of 80% is used throughout the paper: for example, the sequence-to-structure mapping was considered to be one to one if the agreement in secondary structure among the segments within a neighborhood averaged 80% or greater over the length of the segments.

Tables III and IV include only a subset of the patterns; unabridged versions are available from the authors by electronic mail.

Results

Sequence segments ranging from three to fifteen residues in length from a non-redundant subset of the HSSP database of multiple sequence alignments were partitioned into neighborhoods using the K-means algorithm. Since the HSSP database includes at least one sequence of known three dimensional structure per multiple sequence alignment, the structure adopted by each of the segments in each neighborhood is known with reasonable certainty (12).

Approximately 44% of the positions in the input set of multiple sequence alignments fell into a neighborhood in which a single local structure predominated (Table I). For segment lengths thirteen and fifteen, these predominant local structures are primarily helix caps; for segment lengths seven to eleven, helices; and for segment lengths three and five, turns and loops. Although considerably less frequent than the patterns found in helices and turns, a number of patterns were found in turn to sheet transitions for segment lengths seven and nine, and in beta strands for segment lengths three and five.

To determine the number of distinct structural elements in the neighborhoods in which the sequence-to-structure mapping was not one to one, the K-means algorithm was used to subculture the segments in each neighborhood into different structural classes (Table II legend). A substantial fraction of the neighborhoods contained two different types of local structures (Table II, row 2). To assess the statistical significance of the results, parallel experiments were carried out on a simulated data set in which the sequence-structure relationships of the individual segments were randomized.

Importantly, sequence segments in the same neighborhood are restricted to one, two, or three local structures far more often in the HSSP database than in the simulated database (Table II). Thus, the sequence-structure relationships we observe are distinctly non-random.

The sequence patterns strongly associated with particular local structures are an interesting combination of previously studied and new motifs (Table III). Familiar motifs include amphipathic patterns with hydrophobic residues separated by two or three positions almost exclusively found in α helices (Table III, patterns 1 and 2), or with hydrophobic residues separated by one position very frequently occurring in surface β strands (Table III, patterns 3 and 4). A less strongly amphipathic pattern (pattern 5) was found in somewhat buried helices. A number of short patterns with conserved glycine and proline residues occur predominantly in turns as expected (Table III, patterns 6-9; (13)). Pattern 10 is a serine-rich turn. Pattern 11 is similar to a classic N-terminal helix cap motif (3) and indeed is found predominantly in helix N caps. Pattern 12 is close to the Schellman helix C-cap (2, 14) and is found predominantly at the C termini of α helices.

Several patterns extend and/or refine previously characterized motifs. Pattern 13 is an extension of the Schellman motif; following the characteristic helix--turn transition is a hydrophobic stretch that is almost always part of a β sheet. Pattern 14 is very similar to a previously described motif (the α-L motif, (2)), but surprisingly it appears primarily in strand C-caps rather than in the helix C-caps where it was originally described.

A number of the patterns which correlate very strongly with local structure have not been explicitly singled out in the literature. A strongly hydrophobic stretch in pattern 15 is almost always found in buried β strands (note low

average solvent accessibility in column 'SA'). Patterns 16 and 17 are found in transitions from amphipathic helices through an exposed loop to a buried β strand. Pattern 18 is a helix C-cap with a conserved glycine, but otherwise different than the Schellman motif. Pattern 19 is a helix C-cap with turn-favoring residues (S,N,K) instead of a conserved glycine. Patterns 20 and 21 are found in transitions from turns to strands, and 22, in transitions from strands to turns. The two latter classes of patterns link well-studied short reverse turns with specific types of β strands. Analysis of the three dimensional contexts in which these patterns occur is currently under way and should yield insights into the specific interactions responsible for the prevalence of particular local structures.

Because non-local interactions play an important role in protein three dimensional structures, local sequence--structure relationships are not absolute. It should be noted that with the 80% consistency threshold used here, up to 20% of the sequence segments in the neighborhoods described in Tables I and II may adopt local structures different from that of the majority of sequence segments in the neighborhood. Furthermore, the ~20% of positions in the HSSP data set not accounted for in Table II belong to neighborhoods in which the consistency of the local sequence-to-structure mapping is not significantly greater than that observed in the simulated data set.

Discussion

Our approach uses the vast amount of available sequence data as a guide to identify natural structural groupings which otherwise may be hidden by the complexity of protein three dimensional structures. The two major results are the description of the overall features of the local sequence-to-structure mapping (Tables I and II) and the identification of most of the sequence patterns

in proteins that consistently occur in a particular type of local structure (Table III). The identification of sequence patterns which correlate strongly with structure has proceeded in a rather piecemeal fashion in the past (most studies have focused on a particular type of local structure and sought to determine whether the sequences found in the structural element in proteins have any distinguishing features); our automated approach has in one pass probably identified virtually all of such patterns.

An important issue for approaches to protein structure prediction is the extent to which a local "stereochemical code" operates between sequence and structure(2). Our results have both positive and negative implications for the success of such a code. First, we do find a number of patterns which correlate with local structure, and have not been heretofore described (the α-turn-β motif in Table III, for example). Since the patterns were generated using unsupervised learning methods, they are probably not optimal for the classification problem (12), but refinement of neighborhood boundaries using structural information could yield some improvement in local structure prediction. However, Table III shows that currently well-studied motifs dominate the set of patterns that correlate strongly with structure, suggesting that recent success with helix capping motifs (2, 3) may not generalize to a large fraction of other local structure elements. Secondary structure prediction efforts have traditionally had more difficulty with β strands, presumably because of their greater dependence on non-local interactions and indeed, β strands are conspicuously underrepresented in the set of patterns which correlate strongly with structure (Table I). With regard to the question of the contribution of hydrophobicity patterns alone to sequence--structure relationships, we found that considerable resolution was lost, particularly in the case of structural

transitions, when sequences were represented using a two letter hydrophobic-polar code (data not shown).

The explicit treatment of the ambiguity of the local sequence--structure mapping could have useful application to the prediction of tertiary structure from primary sequence. Examples of sequence patterns which are found in two distinct local structures are shown in Table IV. Most work on local structure prediction has sought to specify uniquely the local structure of a protein segment given the sequence. The demonstration that sequence patterns may correlate with two specific local structures out of a larger set of possible structures has immediate relevance to the global protein structure prediction problem since it suggests a means to greatly reduce the size of conformational space. Such a reduction in the size of the space could readily be incorporated into a search procedure in which only a limited number of local conformations are allowed as a global energy function involving non-local interactions is minimized.

The use of sequence patterns to identify structural motifs opens a new paradigm for studies of protein structure. The amount of available sequence data is vast and growing rapidly, and one dimensional sequences are much more amenable to pattern recognition approaches than are three dimensional protein structures. The striking correlation we observe between a number of sequence patterns and local protein structure is probably only the first indication of the power of such "inverse" approaches.

Acknowledgments

We thank H. Schneider and C. Sander for the HSSP database, D. A. Agard for encouragement and computational resources, S. Henikoff, J. Henikoff, S. Pietrokovski, K. Zhang, D. Gerloff, N. Hunt, T. Defay, R. Klevit, and members

of the Baker laboratory for critical reading of the manuscript. KFH is supported by a Howard Hughes Medical Institute Predoctoral Fellowship. This work was partially supported by the National Science Foundation, Science and Technology Center Cooperative Agreement BIR-9214821, the Merck Research Laboratories, and young investigator awards to DB from the NSF and the Packard Foundation. This chapter is an approved reprint as it appears in K.F.Han and D.Baker (1996),*Proc. Nat'l Academy of Sciences USA*, in press.

References:

1. E. T. Harper, G. D. Rose, (1993) *Biochemistry* **32**, 7605-9.

2. R. Aurora, R. Srinivasan, G. D. Rose, (1994) *Science* **264**, 1126-30.

3. S. Presnell, B. Cohen, C. FE, (1992) *Biochemistry* **31**, 983-93.

4. H. X. Zhou, P. Lyu, D. E. Wemmer, N. R. Kallenbach, (1994) *Proteins* **18**, 1-7.

5. W. Kabsch, C. Sander, (1984) *Proc. Nat'l Acad. Sci, USA* **81**, 1075-78.

6. B. Rost, C. Sander, (1993) *J. Mol. Biol.* **232**, 584-99.

7. M. O. Dayhoff, R. V. Eck, C. M. Park, in *Atlas of Protein Sequence and Structure*

M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, D.C.,

1972), pp. 89-99.

8. S. Henikoff, J. G. Henikoff, (1992) *Proc. Natl. Acad. Sci. U S A* **89**, 10915-19.

9. K. F. Han, D. Baker, (1995) *J. Mol. Biol.* **251**, 176-87.

10. U. Hobohm, C. Sander, (1994) *Protein Sci.* **3**, 522-24.

11. C. Sander, R. Schneider, (1991) *Proteins* **9**, 56-68.

12. R. O. Duda, P. E. Hart, *Pattern classification and scene analysis* (Calif. Artificial Intelligence Group, Stanford Research Institute, Menlo Park, 1970).

13. F. Cohen, R. Abarbanel, I. Kuntz, R. Fletterick, (1986) *Biochemistry* **25**, 266-75.

14. C. Schellman, in *Protein folding : proceedings of the 28th Conference of the*

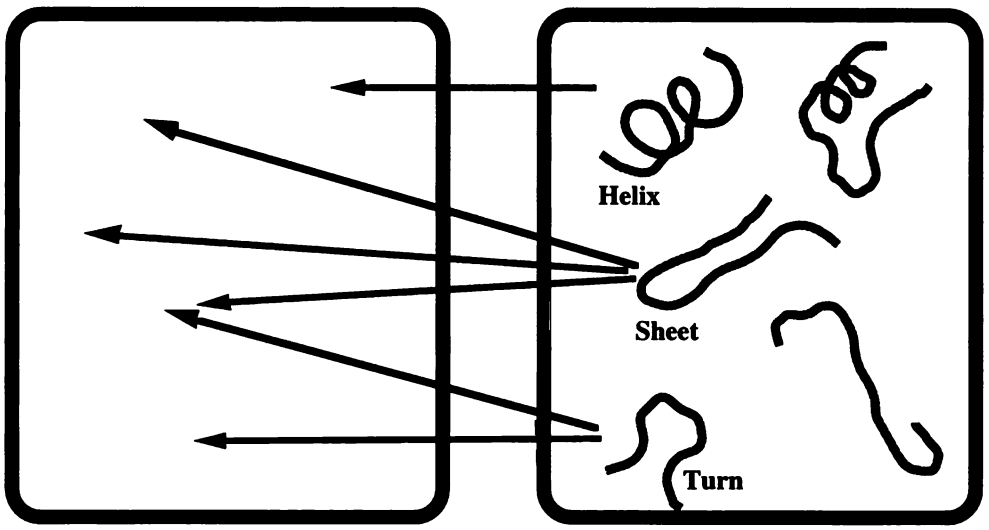
German Biochemical Society R. Jaenicke, Ed. (Elsevier/North-Holland, New York,

1980), pp. 53-61.

15. R. Unger, J. L. Sussman, (1993) *J Comput Aided Mol Des* **7**, 457-72.

Figure 1.

A: Determination of amino acid propensities for predefined local structures



B: Determination of structural correlates of sequence patterns

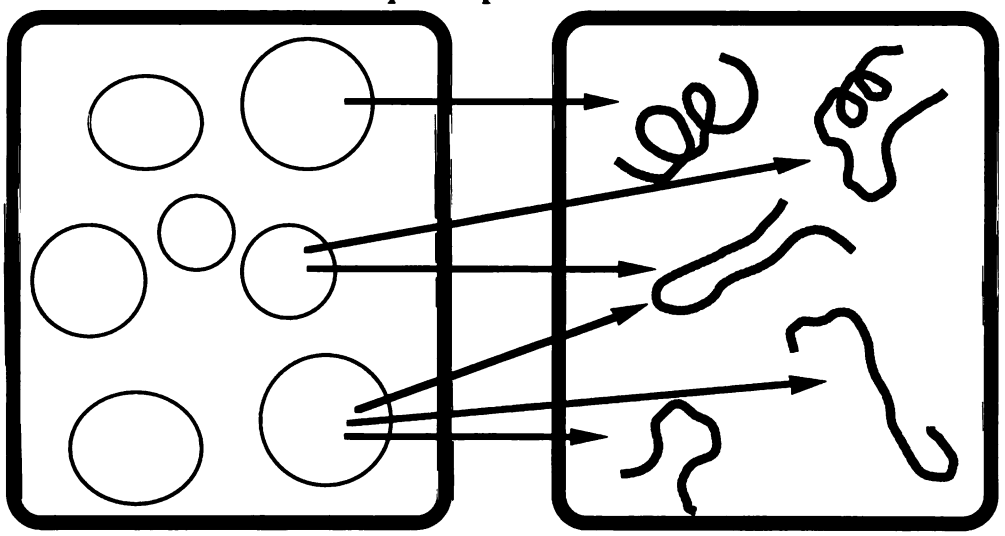


Table legends

Table I. Overall distribution of sequence patterns for which a single local structure predominates. The total number of positions in neighborhoods in which the consistency of the sequence-to-structure mapping was greater than 80% (column 2) and their distribution among different local structures (H, S, T: helix, sheet or turn throughout the segment; HT, TH, TS: helix-turn, turn-helix, and sheet-turn transitions) is given for different segment lengths (column 1). The choice of local structure groupings is primarily for convenience of presentation; other choices would include the 3-D building blocks of Sussman and coworkers (15). No ST dominated neighborhoods were found.

Table II. Distribution of sequence segments among neighborhoods in which the sequence structure mapping is one-to-one, one-to-two and one-to-three. To identify neighborhoods which contained two or three different local structures, the segments within a neighborhood were subdivided into two or three groups using the K-means algorithm (12) and the distance measure

$$d_s(i,j) = \sum_{n=1,N} \sum_{k=helix,strand,turn} |S(n,i,k) - S(n,j,k)|$$

where S(n,i,k) is the frequency of occurrence of secondary structure type k at position n in segment i, and N is the window length. Column 2 lists the percentage of positions in neighborhoods in which the overall secondary structure consistency within 1 (row 1), 2 (row 2), or 3 (row 3) subgroups was greater than 80% . Comparison with the simulated data set showed that for segment lengths of less than nine, the one to three mapping had little statistical significance, and thus only positions in segments of at least nine residues are included in the 1-3 mapping statistics. Statistics on positions in neighborhoods with one to two and one to three sequence to structure mappings exclude positions falling into neighborhoods with one to one and one to two mappings,

respectively. Rows 5-8 give the results of applying the same procedures to a simulated data set.

Table III. Selected sequence patterns which occur predominantly in a single type of local structure. For each neighborhood, the first row gives the identifier and the number of segments in the neighborhood; the subsequent rows contain summary statistics on each position. Letters within brackets indicate the prominent amino acids at the corresponding position in the neighborhood: capitals indicate frequencies greater than 0.1, lower case letters, frequencies between 0.07 and 0.1. For example, the third position in the nine residue pattern characterizing neighborhood 1 is rich in alanine, arginine, and lysine. Positions at which more than 7 different amino acids occurred with frequencies greater than 0.05 are represented by 'π', 'ϕ', and '.' for average hydrophobicities of less than 0.35, greater than 0.65 and between 0.35 and 0.65, respectively. 'Hϕ' is the sum of the frequencies of occurrence of A, V, I, L, M, P, F and W. Solvent accessible surface areas ('SA') were taken directly from the HSSP files, and then normalized by the exposed area of amino acids in A-X-A tripeptides. Residues with less than 16% of their surface exposed were considered buried. Columns 'H', 'S' and 'T' are the number of segments in the neighborhood that are in helix, strand or turn/loop configurations. Patterns 13, 14, 16, and 22 have consistency scores slightly below the 80% threshold.

Table IV. Selected sequence patterns with two prominent local structures. Abbreviations are as in Tables I and III.

Table I.

Length	# positions	H	S	T	HT	TH	TS
15	300	0	0	0	0	300	0
13	2847	691	0	0	1393	763	0
11	3399	1973	0	0	840	586	0
9	2609	1376	0	0	433	411	359
7	1711	819	0	559	0	71	262
5	1327	208	231	888	0	0	0
3	958	0	103	855	0	0	0
Total	13151	5067	334	2302	2666	2131	621

Table II.

	positions (% total)	H	S	T	HT	TH	ST	TS	HTS	TST
HSSP 1-1	43.9	17.0	1.1	7.7	8.9	7.1	0.0	2.1	0.0	0.0
HSSP 1-2	27.7	11.0	1.1	10.0	1.9	1.3	0.7	0.9	0.8	0.0
HSSP 1-3	8.3	5.0	0.0	2.0	0.2	0.7	0.0	0.0	0.0	0.4
Total	79.9	33.0	2.2	9.7	11.0	9.1	0.7	3.0	0.8	0.4
SIM 1-1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SIM 1-2	2.0	0.4	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0
SIM 1-3	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	2.5	0.9	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0

Table III.

Patterns	Hϕ	SA	H	S	T
Amphipathic helices					
#1 25					
[GPa]	.41	.24	23	0	2
[Alr]	.38	.76	24	0	1
[Ark]	.38	.72	23	0	2
[aVI]	.83	.12	24	0	1
[Av]	.81	.28	25	0	0
π	.19	.84	25	0	0
.	.35	.56	24	0	1
[vLY]	.88	.24	25	0	0
.	.44	.72	21	0	4
#2 23					
.	.39	.65	19	0	4
[Anq]	.19	.96	22	0	1
π	.26	.70	22	0	1
[gAv]	.75	.09	21	0	2
[RKd]	.24	.74	22	0	1
π	.24	.91	22	0	1
[ViL]	.86	.26	22	0	1
[Ay]	.75	.13	21	0	2
[gAq]	.19	.91	20	0	3
Amphipathic strands					
#3 58					
π	.22	.82	6	9	43
[G]	.05	.75	4	5	49
π	.12	.79	3	10	45
π	.16	.70	4	33	21
[Vil]	.66	.15	4	46	8
π	.35	.50	3	50	5
[VIL]	.74	.10	5	50	3
#4 39					
π	.19	.64	0	28	11
[VI]	.65	.15	0	34	5
[TsKE]	.11	.66	0	38	1
[Vi]	.71	.10	1	36	2
π	.24	.69	2	23	14
Less amphipathic helix					
#5 28					
.	.52	.50	23	1	4
[gAs]	.74	.18	23	1	4
[vLF]	.89	.18	24	1	3
.	.54	.46	24	1	3
[AvlT]	.59	.29	25	1	2
[AS]	.76	.21	24	1	3
[aVIL]	.70	.36	25	1	2
[Alsk]	.53	.57	22	0	6
[gAvL]	.61	.43	16	0	12

Patterns	Hϕ	SA	H	S	T
Turns/coils with conserved glycines and prolines					
#6 24					
[P]	.07	.61	4	6	21
[AVTR]	.33	.48	4	4	23
π	.17	.77	4	3	24
π	.24	.80	2	2	27
[G]	.02	.64	2	0	29
#7 33					
[G]	.04	.48	2	4	27
.	.38	.73	2	3	28
π	.22	.58	1	2	30
.	.55	.42	3	2	28
[P]	.06	.49	7	2	24
#8 27					
[P]	.05	.70	0	2	25
.	.42	.81	0	0	27
π	.35	.67	0	2	25
[P]	.10	.78	3	2	22
π	.31	.63	6	3	18
#9 24					
.	.37	.54	0	2	22
[Pf]	.23	.50	0	2	22
[GAVf]	.46	.63	0	1	23
[P]	.10	.58	2	2	20
π	.32	.92	3	2	19
.	.45	.75	4	2	18
[PaLT]	.28	.83	2	1	21
[VTS]	.48	.63	2	1	21
[PLSD]	.29	.75	2	1	21
π	.17	.67	2	1	21
[GpvI]	.47	.67	3	3	18
[PYFE]	.58	.54	3	4	17
Other turn/coil					
#10 32					
[PAtS]	.14	.75	7	1	24
[GASd]	.11	.78	6	0	26
[AS]	.17	.84	4	0	28
[S]	.08	.81	2	0	30
[tSnq]	.11	.71	2	3	27

Patterns	Hϕ	SA	H	S	T
Helix N-cap					
#11 61					
[GkdE]	.16	.75	9	7	45
π	.16	.80	6	8	47
.	.41	.41	4	9	48
[TSND]	.06	.83	4	7	50
π	.27	.77	44	1	16
[AkDE]	.16	.95	51	1	9
[qDE]	.08	.78	54	3	4
[aVIL]	.64	.13	56	2	3
π	.28	.63	53	2	6
π	.19	.91	51	2	8
[AVLM]	.52	.26	50	1	10
[ILmf]	.49	.42	47	0	14
π	.16	.88	43	0	18
Schellman Helix C-cap					
#12 68					
π	.25	.70	53	2	13
[aILf]	.59	.16	58	1	9
[AVL]	.47	.35	58	1	9
[rKDE]	.12	.79	60	1	7
π	.28	.58	59	1	8
[L]	.69	.16	57	0	11
[ARKE]	.22	.82	52	1	15
π	.16	.89	48	2	18
[ALRk]	.32	.69	27	2	39
[G]	.06	.70	11	3	54
[aviL]	.49	.44	13	8	47
Schellman Helix-Turn-Sheet					
#13 67					
[ArKd]	.33	.75	53	7	7
.	.52	.40	56	7	4
[AL]	.82	.16	58	4	5
[AiLn]	.41	.75	56	4	7
[AqKd]	.27	.88	49	3	15
.	.47	.72	17	2	48
[G]	.07	.79	1	1	65
[AVI]	.68	.33	2	4	61
[arKD]	.17	.74	4	24	39
[VIL]	.62	.37	7	43	17
[VIL]	.72	.31	7	48	12
[GVIL]	.60	.36	6	49	12
ϕ	.70	.30	6	45	16

Patterns	Hϕ	SA	H	S	T
αL,Strand C-cap					
#14 48					
π	.25	.58	2	21	25
.	.40	.54	1	33	14
[VIl]	.88	.15	1	41	6
[pAVi]	.63	.25	1	40	7
[VIl]	.78	.19	2	41	5
[VIF]	.66	.29	2	34	12
[GA]	.30	.19	3	23	22
.	.43	.38	3	15	30
[G]	.05	.52	2	9	37
π	.33	.62	6	6	36
[GPVn]	.34	.58	6	7	35
Buried strand					
#15 37					
[AVIl]	.82	.16	5	29	3
[ViLm]	.75	.13	4	32	1
[IL]	.88	.08	4	32	1
[VILf]	.81	.16	3	30	4
[G]	.06	.21	2	23	12
Other Helix-Turn-Sheet					
#16 34					
.	.54	.44	26	2	6
[aVIL]	.65	.32	25	4	5
.	.40	.73	24	4	6
[ALsK]	.44	.74	25	4	5
[aViL]	.69	.47	22	3	9
.	.42	.76	21	1	12
π	.34	.85	16	1	17
π	.26	.94	12	1	21
π	.30	.73	4	2	28
[PaVi]	.50	.41	5	3	26
[aDE]	.14	.77	5	6	23
[aVIL]	.72	.18	5	22	7
[VIL]	.85	.15	5	25	4
[AVIL]	.71	.17	5	27	2
.	.60	.26	5	26	3

Patterns	Hϕ	SA	H	S	T
#17 41					
[ASRK]	.22	.87	34	1	6
π	.27	.90	30	0	11
[AiLy]	.61	.46	17	0	24
[G]	.03	.85	3	0	38
[gAV]	.72	.29	3	5	33
π	.20	.75	4	20	17
[VIL]	.77	.36	4	29	8
Other helix C-cap					
#18 56					
[VILf]	.64	.07	49	3	4
π	.25	.44	51	2	3
[AKDE]	.11	.87	52	2	2
.	.43	.23	53	2	1
[ILM]	.59	.10	49	2	5
[atRK]	.13	.78	46	2	8
π	.15	.87	36	1	19
[aLf]	.44	.62	20	1	35
[G]	.04	.83	5	0	51
.	.37	.58	6	2	48
[sNKD]	.09	.76	9	7	40
#19 41					
.	.41	.56	34	0	7
π	.24	.68	35	1	5
[aLd]	.72	.22	38	1	2
[VILy]	.82	.17	37	1	3
[AqrK]	.28	.81	38	2	1
[lqRk]	.26	.68	36	2	3
[IL]	.81	.22	34	1	6
[LRK]	.40	.66	30	1	10
π	.19	.85	24	1	16
[SNKD]	.22	.76	15	0	26
π	.24	.71	9	2	30
.	.37	.81	10	4	27

Patterns	Hϕ	SA	H	S	T
Sheet N-cap					
#20 44					
π	.17	.86	3	6	35
[G]	.09	.84	1	3	40
π	.16	.72	0	6	38
[VTrK]	.19	.81	2	30	12
[VI]	.71	.20	2	35	7
[AViT]	.37	.47	2	37	5
[VILw]	.74	.11	3	39	2
[ViFT]	.47	.31	3	36	5
[GAiS]	.33	.43	3	28	13
#21 41					
[Gk]	.06	.78	5	3	33
[As]	.72	.29	5	5	31
[avkD]	.25	.68	5	17	19
[aVir]	.57	.29	4	32	5
[VI]	.84	.26	4	34	3
[VILs]	.74	.29	4	35	2
[VIL]	.81	.17	4	33	4
Sheet C-cap					
#22 38					
[ViLy]	.74	.15	3	30	5
[AVi]	.73	.18	3	31	4
[VI]	.88	.15	3	31	4
[VILF]	.80	.13	4	27	7
[G]	.05	.15	3	11	24
[Gasd]	.23	.26	3	6	29
[Gvs]	.30	.50	7	7	24

Table IV.

	Pattern	H	S	T	HT	TH	ST	TS
1	[GaqE] ϕ [ILT] [VILF] [AV] [AVI] [aLmt] [AV] π	19	18					
2	. π . [GD] [AII] [ALmE] [AVLE] [iL] [YhRKE]	15		10				
3	[LsR] π [iL] [ATs] π . ϕ π π	15		11				
4	. [N] π . [GVsR] ϕ . π π	11		17				
5	[As] [qhRKE] . [pVIL] [aviLF] π . [pLqe] [aRK]	18			14			
6	π π [LF] . π [AsKDE] [vSnkDe] π π π	11			10			
7	. [pAV] [Lf] [ILf] [GAS] [GAs] [VLh] [G] [G]	10					14	
8	[aVi] . [VIL] . . [VIL] π π [G]	7					10	
9	[gAT] π [GIkdE] π π [IL] [VILk] [aRkE] [aVIk]	16						8
10	[gAs] [GSDE] [IWRk] [VILe] [VIL] [GAs] ϕ [nD] .	14						19
11	π [yF] [TND] [PA] π [AVsR] . [PAVi] π			11		11		
12	π . [AVL] . [lyFw] [NrD] P π [ALSQ]			19			13	
13	π . [ytnDe] π [PsnD] [G] . [VIY] π			20				25
14	[ASK] [G] [SNK] [yFT] π π [AVL] [vIL] [ViL]					11		11
15	. π [G] [nD] [aLQ] . [GAS] [ALs] [aLmF]					10		13

Chapter 7

Inverse squence mapping approach reveals novel structure motifs in proteins

We have used unsupervised learning methods to identify sequence patterns which transcend protein family boundaries. A subset of these patterns occur predominantly in a single type of local structure. Here we characterize the three dimensional structures and contexts in which selected patterns in the latter class occur, with particular attention to the interactions responsible for their striking structural selectivity. The results form the basis for a set of rules linking specific sequences with specific local structures.

Introduction

The traditional approach to characterizing the mapping between amino acid sequence and local structural properties is to decide first on the important structural properties and then investigate their associated amino acid probability distributions. The classic example of this approach is the prediction of protein secondary structure. A striking feature evident after determination of the first few protein structures was the prevalence of simple secondary structure elements: helices, sheets, and turns. Workers such as Chou and Fasman (Chou et al. 1978) tabulated the probabilities of occurrence of each of the amino acids in each of the secondary structure types, and used these probabilities to try to predict the secondary structure adopted by new sequences. This basic procedure has been refined with the application of neural networks and other more sophisticated methods (Presnell et al. 1992; Rost et al. 1993; Sasagawa et al. 1993).

A second example of this approach began from the realization that hydrophobicity patterns can be important in specifying structure. Mutational analysis has shown in many cases that the hydrophobicity of a residue is the best indicator of whether it would be tolerated in a particular position. Both theoretical and experimental approaches have explored the degree to which the pattern of hydrophobic residues specifies particular folds (Dill 1990; Kamtekar et al. 1993). The most simple way to reduce a three dimensional structure to a less complex representation in this regard is to consider each position to be either buried or exposed, and then to score hydrophobicity patterns (Bowie et al. 1990). A more refined approach is to define more specific environments based on secondary structure, local polarity and solvent accessibility. The probabilities of occurrence of different amino acids in each of the environments can then be determined from the database of known structures as in the case of secondary structure prediction (Bowie et al. 1991).

The underlying approach in the above examples is to learn the rules connecting sequence with predefined structural properties using the database of sequences whose structures are known, and then to use the rules to predict the structural characteristics of new sequences. This is supervised learning where correlations between two variables are sought from a large set of examples. An alternative approach is unsupervised learning where patterns are sought in a data set without reference to correlations with other variables. Such an approach is less useful for prediction since groupings are not chosen to optimize the prediction of the second variable from the first. However, unsupervised learning has the advantage that the important properties need not be specified in advance and thus new patterns and groupings can be more readily identified.

We have used unsupervised learning methods to identify recurring amino acid sequence patterns. By examining the secondary structures adopted in different instances of the same sequence pattern, we characterized the degeneracy of the local sequence to secondary structure mapping responsible for the limited success of protein secondary structure prediction. In the course of this study, sequence motifs which occur predominantly in a single type of secondary structural element were identified. However, this connection of sequence motifs with secondary structure patterns does not fully capitalize on the power of our unsupervised learning approach noted in the previous paragraph: the potential to identify new structural properties and groupings.

Towards this end, in this paper we investigate the three dimensional structures adopted by a particularly interesting subset of the sequence motifs. We find that many of the motifs not only occur in well defined three dimensional structures, but also in well defined protein contexts. Interactions between conserved residues that likely give rise to the pronounced structural selectivity of the patterns are identified. The results form the basis for a set of rules linking specific sequences with specific local structures.

Results

Eleven of the motifs identified in our previous study were selected for more detailed analysis. The sequence patterns and their associated secondary structure propensities are listed in Table I (extracted from Table III of reference 3).

Why would a particular set of sequences all adopt the same structure in a protein? The possibilities would include 1) distinctive amphipathicity patterns matching a particular type of secondary structure context (i.e. buried helix), 2)

particular conserved atomic interactions and 3) conformational constraints consistent with a particular structure (i.e. glycine, proline).

To illustrate the approach, we begin with motif I, the well studied N terminal helix cap. Table I lists the locations 67 occurrences of this motif in the sequences of proteins in the pdb-select 25 dataset. To examine the 3D structures adopted by these segments, 30 randomly chosen segments were superimposed (Figure 1a). The density of protein atoms surrounding the segments is depicted in Figure 1b. Figure 1c shows a specific example of a typical segment in this sequence class, high-lighting the reciprocal backbone-backbone interactions between capping threonine and the glutamate in the first turn of the helix. Because this motif has been well characterized by others, these results are not particularly novel; rather, they serve to illustrate that our purely sequence based approach identifies previously characterized structural motifs.

The 29 instances of Motif II (Table II) occur predominantly in buried helices. Figure 1a shows an overlay of the Cα backbones of these segments and their global context down the axis of the buried helices. There is an intriguing pattern of small residues on one side of the helix and large residues on the other. Indeed, side chain size rather than amphipathicity appears to be the distinguishing feature in this motif. The identification of this motif is interesting in the light of the results of Benner et. al. (7) who showed that buried helices are particularly difficult to distinguish from surface ones.

The 41 instances of motif III occur predominantly in helix C-terminal caps. The motif is distinct from previously described capping motifs (8,9). This motif is an amphipathic helix terminated by a strongly polar segment. Positions 4 and 7 have conserved non-polar sidechains. Instead of a non-polar residue at

position 10 or 11, continuing the amphipathic pattern, positions 9 through 11 are strongly polar, or have non-polar sidechains which are out-of-register with the preceding turn of the helix, or contain a proline. In each case, formation of an additional turn of helix would not be favorable. However, there are no conserved interactions within the polar segment 9 - 15, which takes a wide variety of forms. Figure 3a shows a superposition of the Cα atoms of twenty of the instances. There is considerably more variation in the turn than in the helix. The polar residues in this motif are often involved in salt bridges, although the positions of interaction are not conserved (data not shown). The motif has a pronounced tendency to occur at protein surfaces (Figure 3b).

Most strands in proteins are buried forming a part of a sheet making surface strands very special (7). Table IVa shows the consensus sequence for surface sheets and lists the segments in this sequence class (Fig. 4a). As evident from the consensus sequence pattern, the amphipathicity pattern is typical of sheets, but the specific preference for only valine and isoleucine distinguishes this motif from buried ones. These strands are straighter and less twisted than usual. The backbone ϕ and ψ angles of the β-branched hydrophobic residues are restricted by steric interactions between backbone and gamma-carbons, when the sidechain is in its predominant rotamer conformation (χ=-60 degrees). Given this limitation, a narrow hydrophobic contact between alternate isoleucine or valine γ-carbons occurs only when their χ1 bonds are nearly parallel. A twist in the chain at the intervening position would break this contact, which is conserved in 80% of the cases. Exceptions occur when proline intervenes or when α-helix periodicity is present (i.e. a hydrophobic at position 5).

Motif IVb is a short segment of amphipathic strand which is similar to structural motif IVa, except that these are found to be buried (Fig. 4d). The β-branched non-polar sidechains at positions 3 and 5 restrict the backbone dihedral angles and provide one conserved non-polar contact. The degree to which the great majority (90%) of the segments superimpose is remarkable (RMSD is 0.9 Å for all backbone atoms, Fig. 4c). Spatial neighbors of the superimposed segments superimpose well enough to easily see the position of the pairing β-strands. To fall into this sequence cluster, the segment must not have fallen into any highly predictive clusters of greater length. This may mean that it lacks strong structural determinants on either side. But when alternating [VI] and polar residues occur in longer motifs (V, VI, VIII and IX discussed below) they also take on a β-strand conformation, providing more evidence for the surprising self-sufficiency of the short motif.

Strand capping motifs are relatively poorly defined. Some classes of strands contain specific turn motifs that have specific interactions which distinguishes them from other strand endings. Table V lists the consensus sequence for a Glycine-conserved N-terminal capping of strands. As evident in backbone overlay (Figure 5a), register of the turn to strand motif is not very precise, although the overall shape of a 'cane' is clear. Ninety-six percent of the segments are in β-strands at positions 6 - 11; 72% are N-terminated at position 5, a conserved glycine. Of these, all but one case (94%) are accounted for by one or both of the structural consequences of having a glycine at position 5: the absence of a sidechain breaks the chain of non-polar contacts that characterize beta strands (54%), and promotes tight- turn or L-shaped bend formation (69%). Glycine may appear in any of the four positions in a beta-turn, but in this motif it is most often found in the fourth position The conserved glycine leaves a hole

which is often filled by sidechains of the preceding N-terminal polar segment. The N-terminal segment is always found on the surface (Figure 5b), but beyond that lacks conserved structural features.

Motif VI is a β-strand C-terminated at or near a glycine, followed by a polar segment (Fig. 6a). The N-terminal seven residue segment is β-strand (90%) except when glycine intervenes, and is usually buried (85%). As in the previous motif, the absence of a non-polar sidechain at the glycine position removes a possible favorable contact with the conserved non-polar two positions before it, and gives the chain the flexibility to assume a variety of forms. The C-terminal segment generally does not fold back on the strand, again probably due to the lack of non-polar residues to make favorable contacts. Curiously, this motif (Table VI) resembles the sequence motif for the previously described alternative Glycine-conserved C-capping of helices, the αL motif (Aurora et al. 1994). See discussion for further comments.

Motif VII is a β-strand (80%) C-capped by a four-residue segment containing a conserved glycine, which unlike the previous motif, usually forms a beta-turn or similar structures (66%) (Table VII, Fig. 7a). Other strand C-caps include a conserved proline (75%) motif (Table VIII, Fig. 8a). The strong periodicity of hydrophobics in the first six positions favors b-strand (90%). A conserved polar at position 7 breaks the pattern. A tight turn often forms (60%) with the conserved proline in the second position. L-shaped turns also form at that position. Specific conserved interactions (75%) are between the polar residues two position C-terminal to the proline with the polar positions two or one position N-terminal (Fig. 8c). Solvent accessibility show that the strands in this motif tended to be on the surfaces of proteins. Often, the turn diverges rather than forming a beta hairpin. This is due to an interaction between the

polar sidechain one position after the glycine and the residue two positions before it (Fig. 8c).

A number of examples of helix--strand transitions are shown to be specific (Table IX, Fig. 9a). Over 80% of this structural motif were found where the a helices are at the surface of the protein, and the strands following the turn motif form part of a buried b sheet (Fig. 9b). In ~17% of the cases, coils where found in place of strands and are relatively buried. This sequence motif strongly resembles the known Schellman a helix C-terminal capping motif. There are at least five specific conserved contacts surrounding the conserved glycine at position 7. Backbone nitrogens at positions 7 and 8 make hydrogen-bonds with the backbone oxygen of position 3. The conserved non-polar sidechain at position 8 interacts with the conserved non-polar sidechain at position 3 and sometimes with the sidechain at position 6. The conserved non-polar sidechain at position 10 interacts with the non-polar sidechains of positions 3 and 4, creating a small hydrophobic cluster around position 3. Glycine at position 7 allows the backbone to adopt a left-handed curve at the end of the helix ($\phi > 0$)(84%). This happens a significant percentage of the time even when position 7 is not a glycine (38%). This is perhaps the tightest possible turn which places a non-polar sidechain of a β-strand (position 10) between two non-polar sidechains of a preceding α-helix (positions 3 and 4). Only five residues intervene, one of which prefers a glycine with a positive ϕ angle. In fact, when the backbone ϕ angle at position 7 is negative, the non-polar trio (sidechains 3,4,10) is not found. Figure 9c shows examples of conserved specific interactions between the hydrophobic residues in the sheet, and those in the helix.

A smaller, but significant number of examples (34 patterns from 23 protein families) fall into a different sequence class containing a similar structural motif of turns connecting a helices and β strands. Table X lists the proteins and position where this consensus sequence pattern were found in the multiple sequence alignments. Instead of a conserved glycine, this motif has a preference for proline-containing turns. Structurally, the turns linking the surface helices (positions 1 - 6) and buried strands (positions 11 - 15) are more open comparing to the previous, showing a slight cavity (Figure 10a). Sixty-five percent of the cases the strand folds back to interact with α–helix. Position 2, 5, 12 and 14 prefer non-polar sidechains, while positions in the turn region (7 - 9) prefer polar groups. The turn is longer than that of motif IX, with about 6 residues between the last non-polar sidechain of the α-helix (position 5) and the first non-polar sidechain of the β-strand (position 12). Figures 10a and 10b show structural overlay and the global context in which this sequence pattern was found. The helix is terminated by no specific interaction but by the absence of a non-polar group at position 9 to pair with position 5 and/or by the presence of a proline at position 10 (Figure 10c). The proline-containing turns constrains the trajectory of the β strand more stringently than the previous class, resulting in a fewer number of examples where coil were found in place of the strand.

These α-turn-β motifs are found largely in the α/β protein class. The topology is generally right-handed, as expected from the turn motif. Specific hydrophobic packing between the residues on the helix and those in the strand are often found, particularly in the second sequence class (Fig. 10c). A fewer number of examples show specific salt-bridge or hydrogen-bonding interactions between polar residues in the helix and the strand. These interactions are considered long-range in that it involves interactions between secondary

structure elements that are spaced more than 6 residues apart. This is a first description of an intermediate-range structural motif that is highly predictive (>80%).

Methods

The clustering procedures have been described in detail(Han et al. 1995a; Han et al. 1995b). In brief, 29921 segments of profiles derived from a non-redundant set (PDB-select 25 (Hobohm et al. 1994)) of the HSSP database (Sander et al. 1991) of multiple sequence alignments were subdivided into 1200 neighborhoods containing sets of related segments using the K means algorithm (Duda et al. 1970) and the city block metric

$$d(i,j) = \sum_{n=1}^{N} \sum_{k=1}^{20} |F_i(k,n) - F_j(k,n)|$$

where $F_i(k,n)$ (a profile segment) is the frequency of the kth amino acid in the nth position of segment i and N is the segment length. Because the PDB-select 25 subset contains very few pairs of alignments from even distantly related families, segments in a given neighborhood are necessarily derived from quite different protein families.

The Cα overlay were calculated using the standard RMSD, with a threshold of 1.8 Å. The global contexts are displayed by overlaying the Cβ's of the rest of the proteins not including those within the segments selected. Thus, the dense areas are those in which the segments are buried. The structures were visualized using Insight II ®.

Discussion

We have previously identified recurring sequence patterns in multiple sequence alignments by dividing sequence space into mutually exclusive "neighborhoods" (1). Because there is at least one high-resolution structure for

each family of alignment, the structural correlates in each sequence "neighborhood" can be readily investigated. As previously reported, the general features of sequence to structure mapping include one to one, one to two and one to three categories (2). In this paper, we investigate 11 motifs in detail and their rules linking sequence to structure are identified.

Transitions such as turn elements have shown to be particularly difficult in secondary structure prediction. Turns are important in defining the boundary and the resulting trajectory of a secondary structure element, which subsequently effect the prediction of the tertiary fold. One of the sources of failure in secondary structure prediction is the ability to determine where the transition elements occur (10). Helix capping motifs are examples of transition elements that help determine the termination of helices. Our technique was able to extract sequence rules of well-known motifs (helix N-caps), and identified new ones in which the structures were almost always found in transition at either N- or C- termini of helices or strands or specific turns linking helices and strands.

To stress the advantage and importance of the use of multiple sequence alignments, we showed that 'rules' derived from single sequences are subject to many exceptions. For example, Motif VI contains a set of sequence segments that resembles the rules for the αL helix C-cap motif described by Aurora et al (Aurora et al. 1994). But, there were far more examples of these sequences that adopt C-terminal capping of strands rather than helices. Indeed, the protein segments that fell into this sequence class is not the same set as those that were listed in Aurora et al. In fact, most segments listed in Aurora et al were found in different sequences classes other than an those listed in Table VI. For example, helix 280-296 in *3tln* fell into a sequence class which was only 62%

predictable for helix C-cap, and its consensus profile sequence only loosely resemble the αL motif which does not contain a conserved glycine. In fact, the strongest motif for that segment is the relatively conserved leucine at position 295. Similarly, helix 12-29 in *3rnt*, and helix 234-245 in *3tln* fell into the same sequence class, with the strongest motif-determining factor being also the conserved leucine in positions 26 and 243 respectively. In these cases, the consensus sequence in this low predictable class (63%) resembled more closely to the αL criteria (again without conservation for glycine), and is different from the motif shown in Table VI. Interestingly, helix 251-264 in *1gd1* fell into the sequence class that was a very good predictor of helix C-caps (79%), but with a much different profile of consensus sequence compared to the αL motif. In fact, the parent sequence does not at all resemble the consensus profile:

π . [VIL] . π . [VLYF] $\pi\,\pi$ [aL] ϕ π π π π

parent sequence:

K AAA E<u>G</u>E

where the K260 and E264 of the parent sequence are aligned with the hydrophobic positions 7 and 11 of the consensus motif, and G265 falls in a position of variable polar, not conserved glycine as proposed by Aurora et al. These results reflects the advantage of using profiles from multiple sequence alignments rather than single sequences for classification to obtain rules for structural motifs.

Although good predictors, there are exceptions to the structures which fall into the sequence classes illustrated (i.e. the rules are not 100% predictable). The nature of the degeneracy in sequence--structure mapping for these motifs are shown in Table XI. Alternative folds that exist for these motifs almost always are due to the more favorable stabilization from long-range

interactions. In these cases, the only method to distinguish which conformation is more stabilizing clearly depends on the tertiary fold of the protein.

The novel motifs described here were a natural result of our inverse sequence--structure mapping approach. Other recurring patterns in the database from different protein families have less obvious structural correlates (such as secondary structure elements). In those cases, specific tertiary structure environment are likely to have been the constraint for the particular sequence motif. These newly defined environments can be readily used in protein structure prediction algorithms either as a score for whether the model structure was a likely candidate for the sequence, or, as a part of an energy function which is minimized in the three-dimensional folding procedures.

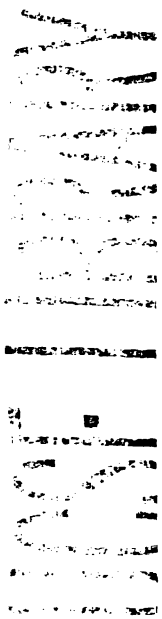
Acknowledgements:

We thank H. Schneider and C. Sander for the HSSP database, D. A. Agard for encouragement and computational resources. KFH is supported by a Howard Hughes Medical Institute Predoctoral Fellowship. This work was partially supported by the National Science Foundation, Science and Technology Center Cooperative Agreement BIR-9214821, the Merck Research Laboratories, and young investigator awards to DB from the NSF and the Packard Foundation.

Reference:

Aurora, R., Srinivasan, R. & Rose, G.D. (1994). Rules for alpha-helix termination by glycine. *Science,* **264**, 1126-30.

Bowie, J.U., Clarke, N.D., Pabo, C.O. & Sauer, R.T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins,* **7**, 257-64.



Bowie, J.U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164-70.

Chou, P. & Fasman, G. (1978). Empirical predictions of protein conformation. *Ann. Rev. Biochem.,* **47**, 251-76.

Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry,* **29**, 7133-55.

Duda, R.O. & Hart, P.E. (1970). Pattern classification and scene analysis. Menlo Park, Calif. Artificial Intelligence Group, Stanford Research Institute.

Han, K. & Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *PNAS,* (in press).

Han, K.F. & Baker, D. (1995b). Recurring local sequence motifs in proteins. *J. Mol. Biol.,* **251**, 176-87.

Hobohm, U. & Sander, C. (1994). *Protein Sci.,* **3**, 522-24.

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. & Hecht, M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science,* **262**, 1680-5.

Presnell, S., Cohen, B. & FE, C. (1992). A segment-based approach to protein secondary structure prediction. *Biochemistry,* **31**, 983-93.

Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A,* **90**, 7558-62.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins,* **9**, 56-68.

Sasagawa, F. & Tajima, K. (1993). Prediction of protein secondary structures by a neural network. *Comput Appl Biosci,* **9**, 147-52.

Table and Figure Legends

Table I: Helix N-cap (SXXE) motif (67 segments in 40 protein families)

Table II: Buried helices (575) (29 segments from 25 protein families)

Table III: Helix C-terminal capping favoring polar residues S,N,K,D. (41 segments from 25 protein families)

Table IVa: Surface sheets (22 segments in 19 protein families)

Table IVb: Buried sheet cluster (37 segments from 27 protein families)

Table V: Gly conserved N-capping of sheets (698) (31 segments in 23 protein families)

Table VI: αL sequence motif are found more frequently in C-capping of sheets (48 segments from 39 protein families).

Table VII: Glycine-conserved sheet C-cap (33 segments from 27 protein families)

Table VIII: Proline rich C-terminal capping of sheets (23 segments in 21 protein families)

Table IX: Helix to strand with a glycine-conserved turn (HTS-Gly) (67 segments in 39 protein families).

Table X: Helix to strand with a proline-conserved turn (HTS-Pro) (34 segments in 23 protein families).

Table XI: Alternative structure subclasses for HTS-gly, HTS-pro

Figures 1a through 10a are Cα overlay with RMSD of 1.8 Å threshold. Figures 1b through 10b are Cβ overlay ('+') of all the positions different from those positions in the motif. This is to explore the global packing density and solvent accessibility properties of these motifs. Buried areas are where Cβ is more dense, and sparse regions represents solvent accessible positions. The solid filled model is a representative member in the cluster. Residues colored in red

are those that are conserved. Figure 1c through 10c are selected examples of specific conserved interactions in the motif.

Figure 1a,b&c: Helix N-cap of proteins listed in Table I.

Figure 2a&b: Buried helix of Table II.

Figure 3a&b: Helix C-cap of Table III.

Figure 4a&b: Surface strand of Table IVa.

Figure 4c&d: Buried strand of Table IVb.

Figure 5a&b: Glycine conserved strand N-cap of Table V.

Figure 6a&b: Glycine conserved (αL) strand C-cap of Table VI.

Figure 7a&b: Strand C-Cap of Table VII.

Figure 8a,b&c: Proline-conserved strand C-cap of Table VIII.

Figure 9a,b,&c: Helix to strand with a glycine conserved turn of Table IX.

Figure 10a,b&c: Helix to strand with a proline conserved turn of Table X.

Table I

π . [pVIL] [TSD] . [ADE] [nQDE] [AVIL] . [Aqkd] [AVIL] . π . [AviL]

Protein	Chaintype	Pos
1avh	A	115.
1avh	A	165.
1baa	_	4.
1bmd	A	312.
1cpc	B	17.
1crl	_	351.
1ctd	A	19.
1glc	G	193.
1gpl	A	48.
1hdd	C	6.
1hdd	C	24.
1hdx	A	348.
1hle	A	69.
1hle	A	265.
1hle	A	319.
1ipd	_	51.
1ipd	_	155.
1mct	A	161.
1ndc	_	45.
1nip	A	257.
1pfk	A	194.
1pfk	A	289.
1pii	_	93.
1pii	_	139.
1pii	_	191.
1pii	_	216.
1ppn	_	46.
1scm	B	33.
1scm	B	49.
1scm	B	82.
1scm	B	102.
1scm	B	118.
1scm	C	26.
1scm	C	42.
1scm	C	79.
1scm	C	99.
1scm	C	115.
1spa	_	136.
1spa	_	198.
1tbp	A	216.
1tnd	A	55.
1tnd	A	235.
1tpl	A	190.
1tre	A	178.
1trk	A	83.
1wsy	B	325.
2acq	_	65.
2acq	_	262.
2ctc	_	212.
2hmz	A	37.
2ihl	_	1.
2ihl	_	21.
2mge	_	79.
2tpr	A	173.
2tpr	A	247.
2tpr	A	374.
2tpr	A	460.
3chy	_	109.
4gpb	_	357.
4gpb	_	511.
4gpb	_	711.
4ts1	A	271.
5p21	_	123.
5p21	_	145.
8atc	A	13.
8cat	A	242.
8tln	E	296.

Notations for the concensus sequence in Tables I - XI, where positions with specific substitutions are shown within brackets: lower-case letter represent amino acid occurence frequencies between .07 and 0.1; upper-case letters are those over 0.1. Positions which are underlined is the constraining motif or predominant interactions, '.' is variable, 'π' is variable polar, and 'ϕ' is variable hydrophobic.

Table II

[gAS] . [AVIL] . [gAS] [AVIL] .

Protein	Chaintype	Pos
1bmd	A	248.
1cpc	B	128.
1eaa	_	473.
1eco	_	71.
1glc	G	379.
1hdx	A	274.
1ipd	_	289.
1len	A	113.
1lga	A	135.
1onc	_	42.
1pfk	A	18.
1pii	_	323.
1ppn	_	29.
1prc	M	211.
1prc	M	219.
1prc	M	267.
1rnd	_	52.
1s01	_	175.
1tnd	A	79.
1tpl	A	363.
1trk	A	510.
1trk	A	638.
1wsy	B	120.
2ccy	A	34.
2ccy	A	44.
2ctc	_	292.
4gpb	_	219.
8tln	E	287.
9rnt	_	18.

Table III

. π [aLd] [VILy] [AqrK] [lqRk] [IL] [LRK] π [SNKD] π . . π .

Protein	Chaintype	Pos
1aak	_	10.
1aak	_	127.
1ak3	A	70.
1apm	E	126.
1apm	E	266.
1apm	E	270.
1apm	E	285.
1cau	B	403.
1cau	B	407.
1crl	_	272.
1crl	_	410.
1hdd	C	14.
1hle	A	35.
1hle	A	97.
1hle	A	131.
1hle	A	269.
1hmy	_	312.
1mct	A	100.
1mct	A	103.
1pii	_	13.
1ppn	_	69.
1s01	_	5.
1s01	_	228.
1s01	_	231.
1scm	C	69.
1spa	_	54.
1spa	_	313.
1trk	A	42.
1tys	_	5.
1tys	_	57.
1tys	_	113.
1wht	A	200.
1wsy	B	29.
2bop	A	339.
2plv	3	97.
2tpr	A	342.
3chy	_	66.
3cla	_	122.
4gpb	_	584.
5p21	_	17.
5p21	_	94.

Table IVa

π [VI] π [VI] . . π

Protein	Chaintype	Pos
1atr	_	102.
1bmd	A	3.
1cau	A	163.
1cgm	E	93.
1cgt	_	547.
1cgt	_	561.
1gd1	O	54.
1gd1	O	70.
1hdx	A	25.
1hpl	A	419.
1mct	A	82.
1onc	_	85.
1pkp	_	71.
1rnd	_	105.
1spa	_	320.
1tre	A	60.
1vil	_	28.
2aza	A	19.
2mip	A	63.
2tpr	A	133.
2tpr	A	387.
4gcr	_	34.

Table IVb

. π [VI] [aTS] [V]

Protein	Chaintype	Pos
1ahc	_	221.
1alk	A	95.
1atr	_	101.
1bbt	2	105.
1bbt	2	156.
1bbt	2	179.
1bet	_	34.
1bmd	A	4.
1bmd	A	121.
1bmd	A	123.
1cgt	_	560.
1cgt	_	637.

Protein	Chaintype	Pos
1cob	A	3.
1fkb	_	20.
1gd1	O	23.
1gd1	O	235.
1hdx	A	288.
1hle	A	214.
1ipd	_	207.
1len	A	114.
1ms2	A	16.
1ms2	A	60.
1pdg	A	68.
1pdg	A	85.

Protein	Chaintype	Pos
1pfk	A	242.
1r1a	2	106.
1s01	_	26.
1slt	A	85.
1trk	A	611.
2bpa	1	408.
2bpa	2	138.
2cas	_	80.
2dnj	A	87.
2hmz	A	48.
2tpr	A	211.
2tpr	A	386.
4gpb	_	563.

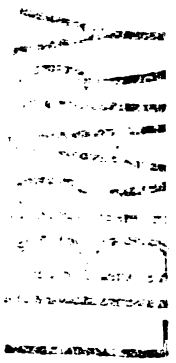
Table V

π [LCTn] [gaTE] π [G] [TsK] [gVIY] . [VILF] [TSd] .

Protein	Chaintype	Pos
1aaj	_	82.
1ab2	_	27.
1ab2	_	39.
1ab2	_	62.
1alk	A	312.
1atr	_	199.
1cgm	E	131.
1cgt	_	522.
1cob	A	12.
1cob	A	87.
1ctm	_	184.

Protein	Chaintype	Pos
1eaa	_	549.
1fkb	_	29.
1fxi	A	8.
1hdx	A	122.
1hmy	_	94.
1php	_	230.
1pmy	_	68.
1ppl	_	20.
1ppl	_	239.
1tnr	R	32.

Protein	Chaintype	Pos
1tnr	R	54.
1tnr	R	119.
1tys	_	19.
2cas	_	53.
2cas	_	260.
2cas	_	527.
2cpl	_	105.
2ctc	_	163.
3aah	A	514.
5nn9	_	231.



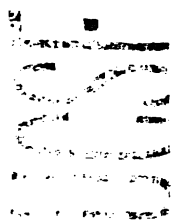


Table VI

π . [VII] [pAVi] [VII] [VIF] [GA] . [G] π [GPVn]

Protein	Chaintype	Pos
1ak3	A	7.
1alk	A	415.
1atr	_	193.
1bmd	A	2.
1cau	B	343.
1cgt	_	512.
1cgt	_	550.
1crl	_	116.
1eaa	_	586.
1fkb	_	2.
1gpr	_	140.
1hdx	A	71.
1hdx	A	193.
1hpl	A	244.
1hsb	A	10.
1mct	A	134.
1mct	A	204.
1mct	A	210.
1nip	A	31.
1nip	A	145.
1ofv	_	81.
1pfk	A	215.
1php	_	53.
1pmy	_	29.
1ppn	_	159.
1rla	2	28.
1rnd	_	104.
1s01	_	194.
1tpl	A	279.
1trk	A	555.
1wht	B	388.
1wsy	B	105.
1wsy	B	226.
1wsy	B	251.
1wsy	B	370.
2acq	_	10.
2dnj	A	64.
2ech	_	17.
2snv	_	225.
2tpr	A	4.
2tpr	A	132.
2tpr	A	188.
4fxn	_	80.
4gpb	_	196.
5nn9	_	112.
5p21	_	4.
6fab	L	93.
8atc	A	122.

Table VII

. [GK] [VItK] . [VIL] ϕ . [G] . [GaiK] [acsk]

Protein	Chaintype	Pos
1alk	A	43.
1bbt	2	28.
1caj	_	144.
1cau	B	304.
1crl	_	329.
1fas	_	31.
1hdx	A	286.
1hpl	A	243.
1ipd	_	1.
1nbt	A	34.
1pfk	A	216.
1pkp	_	33.
1pkp	_	42.
1pmy	_	3.
1ppl	_	161.
1slt	A	28.
1smr	A	161.
1tbp	A	118.
1tbp	A	155.
1tbp	A	209.
1tgx	A	30.
1tre	A	123.
1trk	A	149.
1wht	A	170.
1wsy	B	128.
1wsy	B	371.
2btf	A	39.
2btf	A	41.
3aah	A	159.
5nn9	_	422.
5p21	_	3.
5p21	_	108.
8tln	E	116.

Table VIII

[aILF] . [VIL] π [VIL] . π [Pn] π [GPAv] [GtSk]

Prot.	Chain	Pos.	Prot.	Chain	Pos.	Prot.	Chain	Pos.
1aak	_	36.	1cob	A	93.	1slt	A	94.
1ahc	_	211.	1crl	_	98.	1spa	_	126.
1avh	A	80.	1crl	_	360.	2acq	_	204.
1baa	_	157.	1fxi	A	3.	2ctc	_	47.
1caj	_	45.	1hle	A	282.	3aah	A	139.
1cdh	_	114.	1mct	A	158.	4blm	A	245.
1cdh	_	126.	1ppl	_	153.	8atc	A	100.
1cgt	_	548.	1prc	M	89.			

Table IX

[ArKd] . [AL] [AiLn] [AqKd] . [G] [AVI] [arKD] [VIL] [VIL] [GVIL]

Protein	Chaintype	Pos	Protein	Chaintype	Pos	Protein	Chaintype	Pos
1aaj	_	11.	1ofv	_	107.	1wsy	B	243.
1alk	A	134.	1pfk	A	25.	1xya	A	123.
1alk	A	227.	1pfk	A	87.	2acq	_	146.
1bmd	A	165.	1pfk	A	179.	2acq	_	168.
1cgt	_	123.	1php	_	45.	2bop	A	366.
1cgt	_	271.	1php	_	251.	2mip	A	21.
1cob	A	85.	1php	_	355.	2tpr	A	20.
1ctm	_	183.	1pii	_	74.	2tpr	A	21.
1eaa	_	455.	1pii	_	124.	2tpr	A	239.
1fdd	_	22.	1pii	_	151.	2tpr	A	379.
1fxi	A	26.	1pii	_	171.	2tpr	A	441.
1gd1	O	82.	1pii	_	414.	3chy	_	96.
1gd1	O	106.	1pkp	_	113.	3pgm	_	99.
1hle	A	40.	1ppn	_	73.	3pgm	_	165.
1hle	A	301.	1prc	M	133.	4fxn	_	21.
1hmy	_	26.	1s01	_	112.	4fxn	_	101.
1ipd	_	119.	1slt	A	47.	4gpb	_	698.
1ipd	_	171.	1tpl	A	202.	5nn9	_	291.
1ndc	_	30.	1trk	A	205.	5p21	_	132.
1ndc	_	85.	1trk	A	454.	8atc	A	61.
1nip	A	24.	1wsy	A	116.	8atc	A	200.
1nip	A	59.	1wsy	B	147.	8cat	A	234.
1nip	A	198.						

Table X

. [aVIL] . [ALsK] [aVIL] . π π π [PaVi] [aDE] [aVIL] [VIL] [AVIL] .

Protein	Chaintype	Pos
1ak3	A	97.
1ak3	A	190.
1atr	_	322.
1cpc	B	55.
1cpc	B	157.
1crl	_	373.
1gd1	O	16.
1gd1	O	80.
1gd1	O	133.
1gd1	O	296.
1gpr	_	116.
1hmy	_	63.
1hmy	_	104.
1ipd	_	198.
1nip	A	108.
1ofv	_	38.
1pfk	A	271.
1pii	_	39.
1pii	_	72.
1pii	_	364.
1ppl	_	245.
1ppn	_	120.
1s01	_	110.
1s01	_	138.
1spa	_	56.
1spa	_	239.
1tre	A	29.
1trk	A	546.
2acq	_	19.
2acq	_	235.
2ctc	_	21.
2tpr	A	268.
3chy	_	42.
8atc	A	90.

Table XI

Pattern	HTS	HT
Gly-HTS	55	12
Pro-HTS	28	6

'HTS' represents helix - turn - strand structure class, and 'HT' represents helix to turn structure class.

Figures 1a &1b

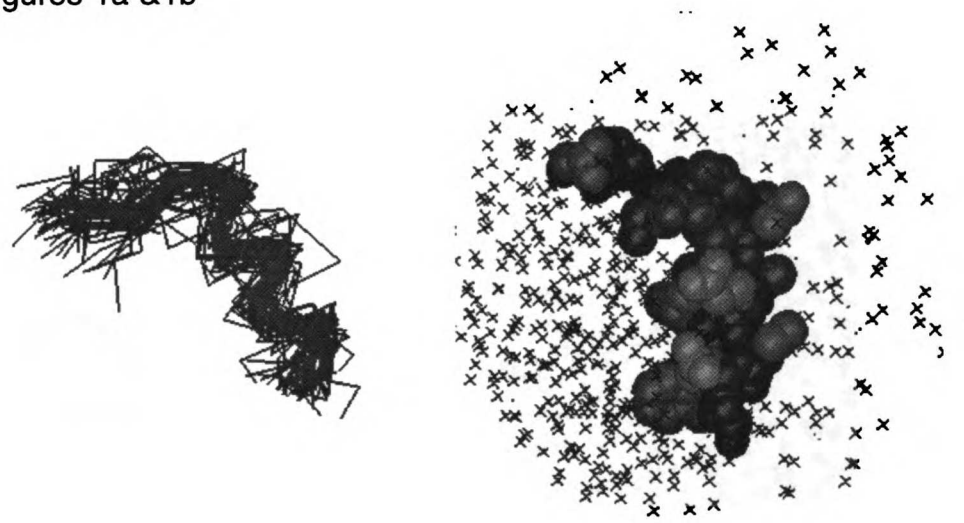


Figure 1c

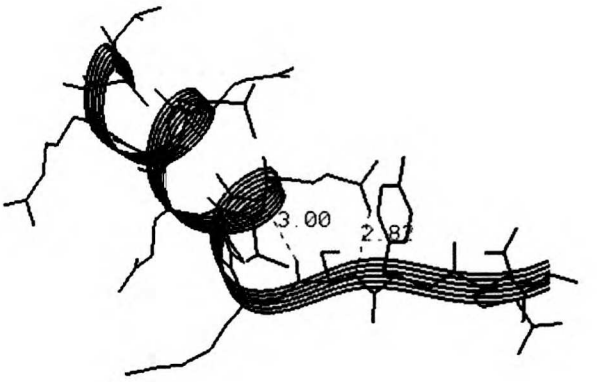
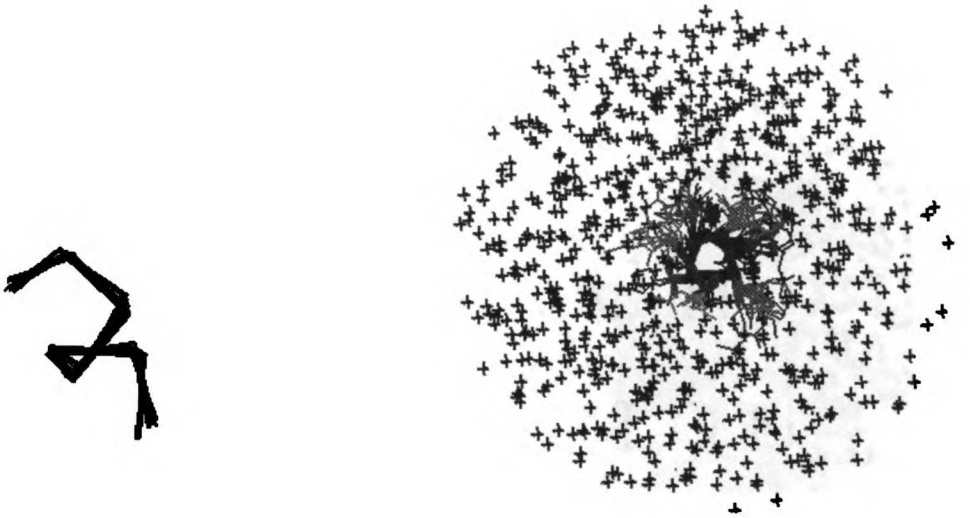


Figure 2a & 2b



Figures 3a & 3b

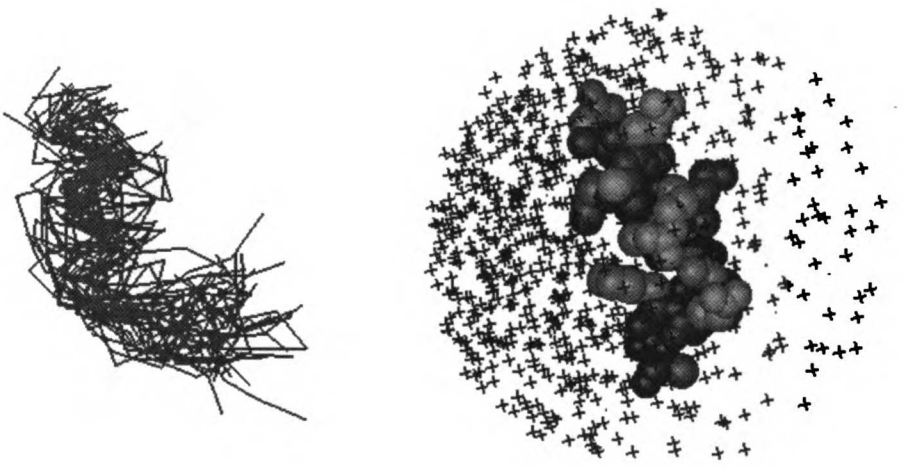


Figure 4a & 4b

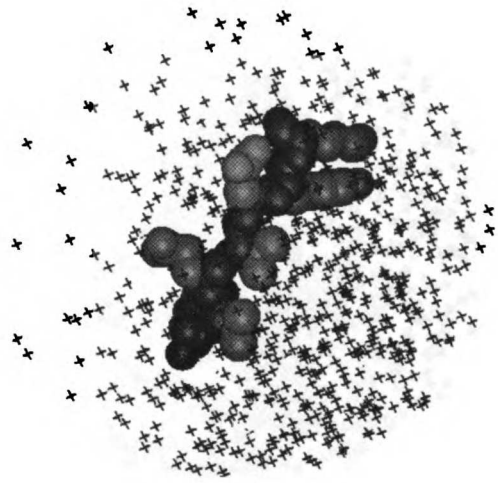


Figure 4c & 4d

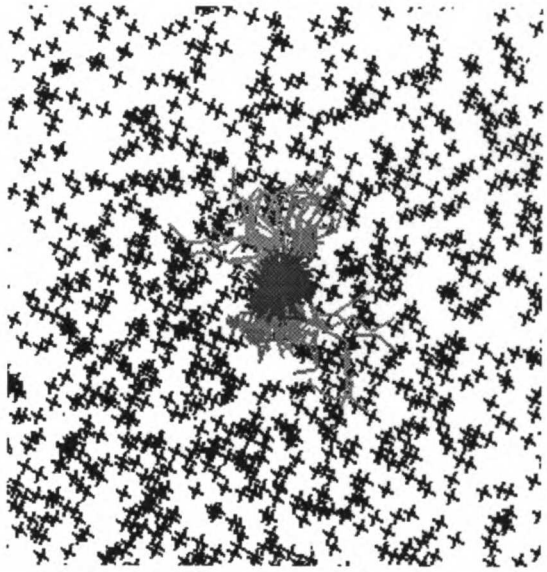


Figure 5a & 5b

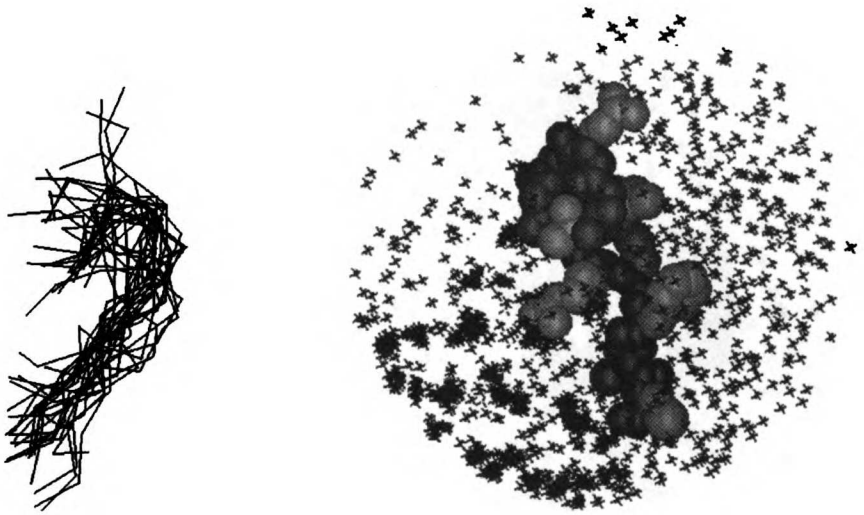


Figure 6a & 6b

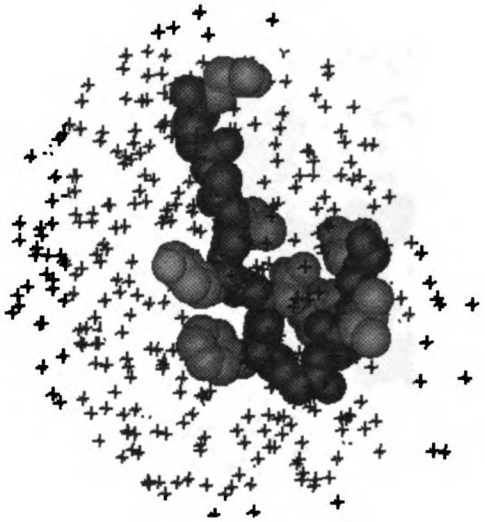


Figure 7a & 7b

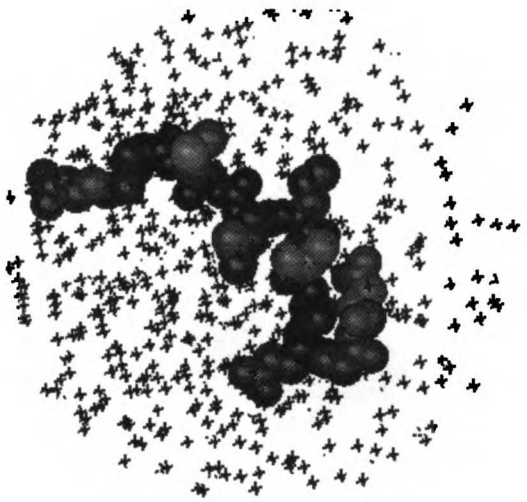


Figure 8a & 8b

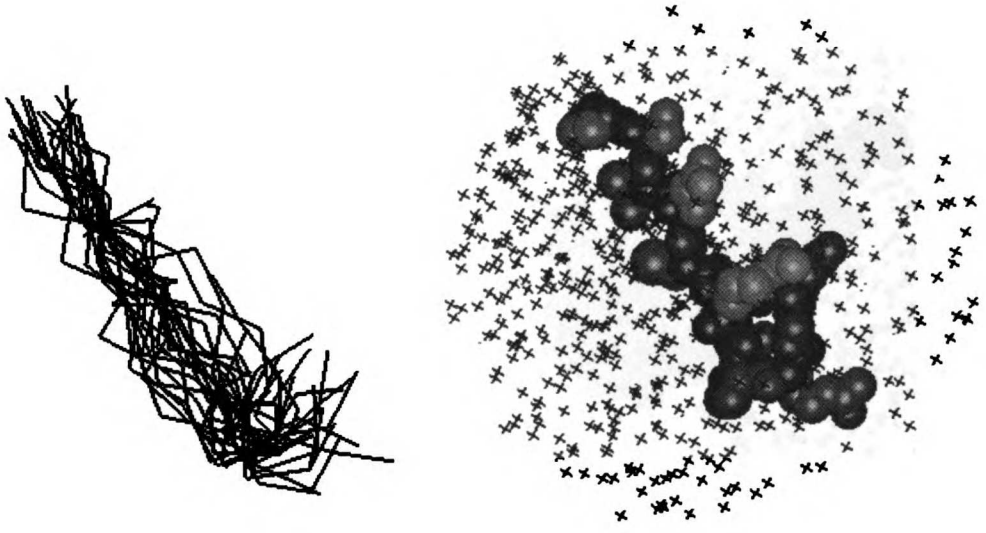


Figure 8c

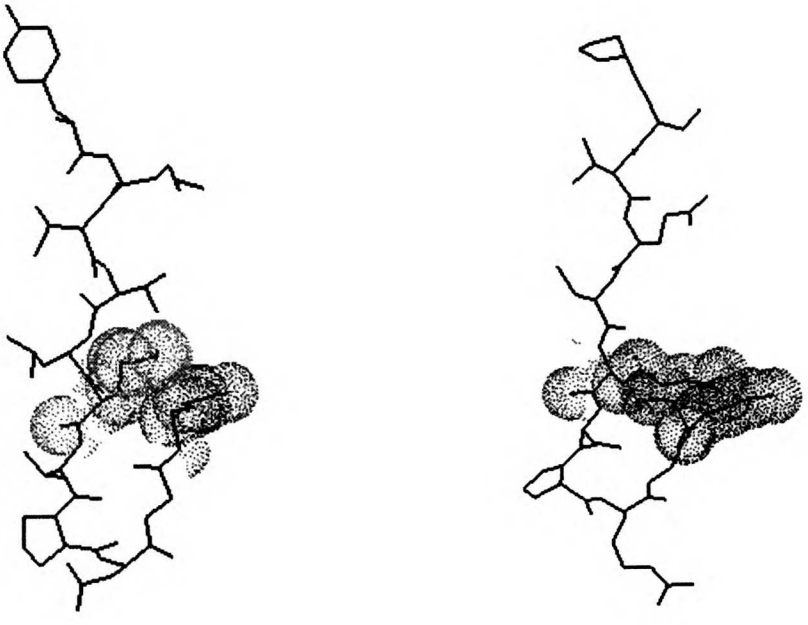


Figure 9a & 9b

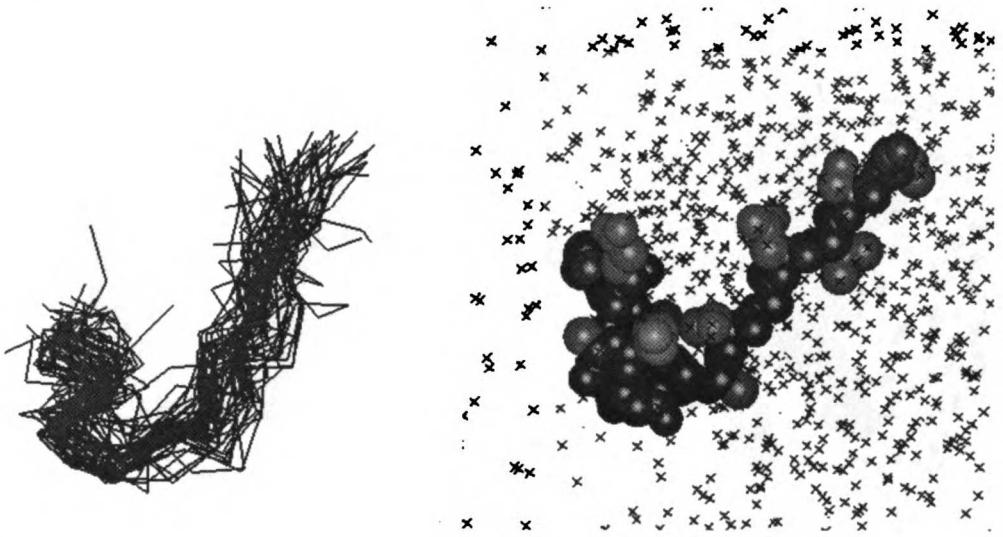


Figure 9c

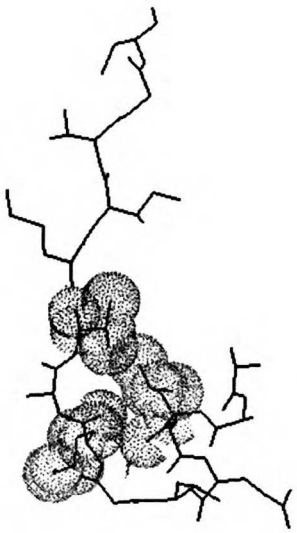


Figure 10a & 10b

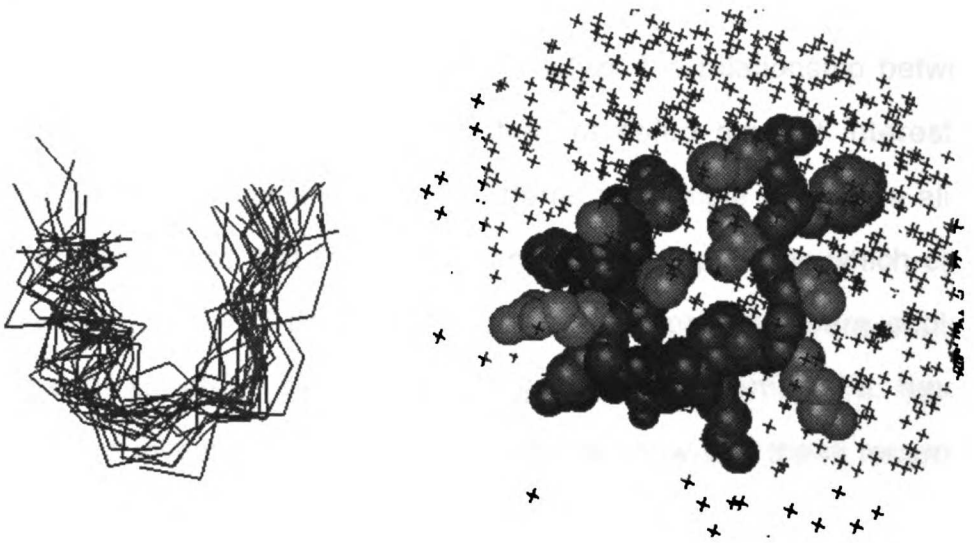
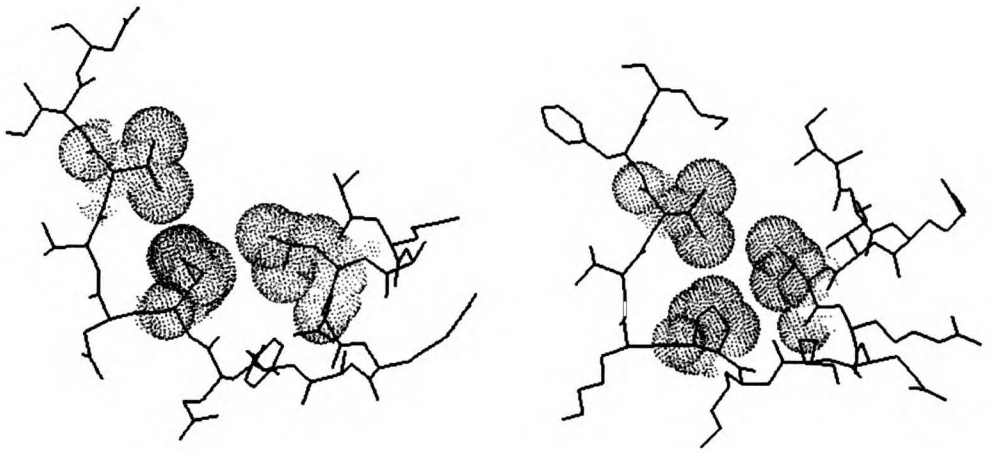


Figure 10c



Conclusion

Using the unique approach to analyze the relationship between protein primary sequence to tertiary structure revealed several interesting points. Firstly, based on classification of patterns in multiple sequence alignments, it was demonstrated that there are many recurring patterns which cross protein family boundaries. The significance of these patterns were explored in the context of structural similarity, although other properties (i.e. functional) still await to be investigated. Preliminary results show that these recurring patterns performed better than other substitution matrices (i.e. PAM, BLOSUM, etc...) used to classify newly discovered sequences into known protein families (Baker and Robinson). This is not surprising since it is shown that different substitution patterns are prominent in different context as demonstrated in chapter 5.

On a global investigation, the problem of an 'upper-limit' secondary structure prediction was addressed (chapter 6). It was shown that there were only a selected number of patterns that correlate well with a single structural state. Whereas, the majority of the sequence patterns map ambiguously to two or three structural classes. Including this concept into tertiary folding algorithms, preliminary studies were able to fold all helical proteins up to the size of myoglobin, de novo, with an RMSD of less than 4 Å (Baker and Simon).

The caveat remains with the prediction of sheet proteins. It is clear from the sequence analysis that there were not many patterns which contain good correlation with sheets. In fact, from the global tabulation of structures, it is not very well represented at all. The reason is that sheet segments often contain sequence patterns stronger for other secondary structure types, and end up as the ill-predicted portion of the neighborhood. For example, a cluster may be

predominantly in helical structures that is 85% predictable, but the other 15% may easily have been in strands, and that 15% is then lost in the noise.

Most strand patterns are non-specific. Indeed, the strongest strand patterns have been associated with a specific turn or bend. In those cases, it seems that the specific turn (usually tight turns) favor the strand (extended) conformation rather than helical conformation. Another explanation is that strands tend be stabilized in the core of the protein structure (thus buried majority of the time) via long-range interactions. Therefore, the classification of local sequence patterns will not be able to capture the long-range motifs. All the possible insertions and deletions of surface loops make the register of strand stabilizing long-range interactions in the core difficult to characterize based on sequence alone. There are only few examples where more local constraints result in a strand conformation. A demonstration of that are the two helix to strand transition motifs (Chapter 7), showing specific stabilizing long-range conserved interactions between the strand and the helix N-terminal to it.

Incorporating the 'hydrophobic zipper' concept of Dill, Fiebig and Chan, the folding initiation motifs can serve as anchors for such an approach. The tertiary protein folding using either a genetic or Monte Carlo algorithm, will be first to identify the anchoring sights, followed by the segments in which one-to-one mapping is prominent. And for all other segments, the structures will be changed based on the degree of mapping ambiguity. For each cycle, energy is minimized.

The difficulty is once again the long range potential with which energy is minimized. The same type of analysis that's been done here can also be applied to refining an empirical long range potential. One can use contact maps to determine favored interactions, much the same way it has been done

traditionally, but with the added substitution profile of the multiple sequence alignment. It is expected that there will be discrete ambiguity from conformational constraints, analogous to what's been found at the sequence level.

Future directions thus include the conclusion of applying the sequence motifs for pattern searches as well as for tertiary de novo protein structure prediction. Before pursuing the tertiary folding project, a better refined empirical long-range interaction potential using the multiple sequence alignments should be explored. In conjunction with the motifs found using the sequence classification approach, the conformational search space for protein folding will have been greatly reduced to a more solvable problem.

Appendix A:

Summary of the approach to chapters 3 and 4

Thick specimen imaging is necessary to investigate the intact objects with relatively large (>0.5 μm) natural dimensions such as cellular organelles. For amorphous specimens, tomographic techniques are required to achieve moderate resolutions ((3-5 nm)$^{-1}$). In any quantitative analysis such as tomography, it is necessary to correctly relate the detected image intensities to the projected specimen mass density. As a result of the objective lens aberrations and the electron specimen interactions, this relation is distorted. Correcting for this requires an accurate understanding of the process of image formation. For thick specimens, the distortion is largely due to (multiple) inelastic scattering and can be very specimen-dependent. Although the general properties of image formation in thick amorphous specimens have been thoroughly investigated (Han et al. 1995; Han et al. 1996), it is recommended that a reduced but similar analysis be done on your typical specimen to properly interpret your images.

Equipment

Transmission electron microscope (TEM) operating at intermediate to high primary voltages (>200 keV). Preferably, use a TEM equipped with a slow-scan CCD camera to ensure optimum image quality and the convenience of direct digital data which facilitates the image manipulation.

Instrumentation for electron spectroscopic imaging (ESI), either in column or post-column. If available, operating at the selected primary voltage on the TEM. It is strongly recommended that energy filtering is used, but it is not essential for specimens less than 0.5 μm thick.

Specimen

If available, add fudicial markers such as gold beads to your specimen.

Methods

The goal is to evaluate the fraction (if any) of coherently scattered electrons for your specimen . Steps 6 and 7 below are not necessary if your specimen is a typical biological specimen embedded in epon and stained with lead and unranyl. The extensive analysis for these types of specimen have already been done and could be referred to in (Han et al. 1995).

1. Standard alignment of the TEM as recommended by the manufacturer. If ESI capability is available, be sure that voltage center is aligned and do the following procedures using zero-loss filtering. Note that for thicker specimens a correct alignment on the objective lens' voltage center is more important than on the current center. Misalignment on the voltage center will cause streaking in the image proportional to the range of energy-losses allowed to contribute to the image. If on-line diffractogram calculation is available, check the stigmation by making sure that the diffractogram rings of the collected high underfocused (~10 μm) image is symmetrical.

2. Choose an optimal condenser (intensity) setting such that the illumination angle does not degrade your resolution. This typically means to image at a high defocus conditions while not compromising signal to noise ratio. Again, if on-line diffractogram calculation is available, then adjust the condenser setting such that the diffractogram rings of the underfocused image covering the resolution range of interest are most visible.

3. Select and align the smallest objective aperture that 1) is not limiting your resolution of interest, and 2) that does not drift during the collection of the through focus series.

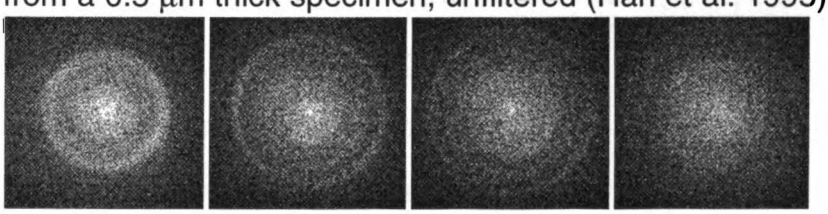
4. At the appropriate magnification, collect a large through-focus series above and below Gaussian focus at equal intervals. Depending on the resolution of interest , the range and step size could vary (van Dyck et al. 1987):

$$\varepsilon \leq \frac{k}{\pi g^2}$$

where ε is the defocus interval, k is the wave-vector and g is the resolution. For example, at 300 keV, k= 1/.0197Å = 50.8Å^{-1} and resolutions up to 30 Å^{-1}, g= 1/30Å = 0.033Å^{-1}, the defocus interval has to be at least 1.454 μm. Practically, we over-sample at 1.0 μm interval from 20 μm under- to over- focus.

5. Align the through focus series using cross-correlation or fudicial marker alignment schemes (Koster et al. 1993). Be sure to correct for the change in magnification with defocus (Typke 1992).

6. Take the three-dimensional power spectrum (amplitudes from 3DFFT) of the aligned through focus series. In each section, you will see a ring of intensity representing the coherent (interpretable) imaging component, and a sizable central (incoherent or partially coherent) component. For example, the following Figure 1 shows selected cross-sections of the 3D power spectrum from a 0.5 μm thick specimen, unfiltered (Han et al. 1995):

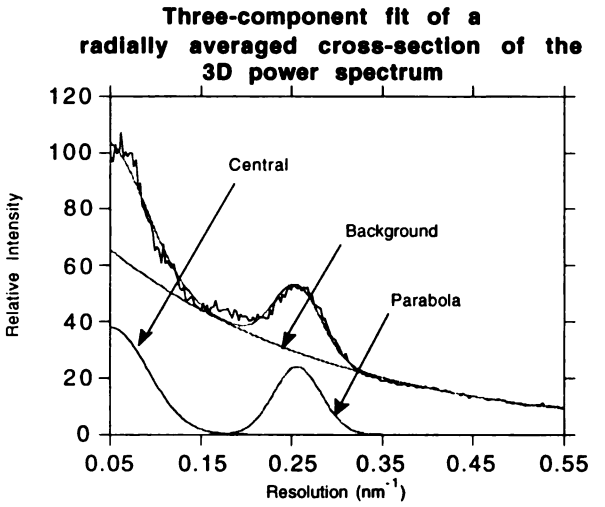


From left to right, they correspond to 18.1, 12.67 , 5.43 and 2.72 μm focus levels (resolution limit: 24 Å^{-1}).

7. Calculate the proportion of the coherent component by applying curve fit to the radially averaged image of each section in the 3D power spectrum. As shown in Figure 2, the three components to fit are a Gaussian at the center (incoherent component), an offset Gaussian (coherent component) and a

background component (exponential). The resulting proportion of the coherent component gives the amount of electrons that can be recovered using restoring filters derived from linear imaging theory. This is the limit to which one would be able to recover computationally.

Fig. 2:



8. Use the following formula to restore the amplitude and phase components of the exit wavefront from this large through focus series (van Dyck et al. 1990):

$$\hat{\psi}_e = \exp(i\pi C_s \lambda^3 \kappa^4) \frac{1}{N} \sum_{\Delta f_n=1}^{N} \hat{\mathbf{I}}(\kappa, \Delta f_n) \exp(-i\pi\lambda\kappa^2 \Delta f_n)$$

where $\hat{\psi}_e$ and $\hat{\mathbf{I}}$ are the exit wavefront and detected images in reciprocal space respectively, λ is wavelength, Δf is the defocus level, and N is the total number of images in the through focus series.

9. To relate the image intensity to specimen mass density, measure the relative drop in intensity as a function of specimen thickness. This can be achieved by cutting the specimen into the appropriate thicknesses, or preferably, by tilting the specimen stage to vary the effective thicknesses. If the fitted curve best matches a logarithmic function, then an absorption model should be applied when interpreting the amplitude image (i.e. the logarithm of the amplitude component is added to the phase component). If the fitted curve depends

linearly on thickness, then the amplitude image is directly related to specimen mass density.

10. If ESI capability is available, repeat the same analysis (steps 4-8) at the most probable energy-loss (the maximum of the energy-loss spectrum), with the widest energy-selecting slit where the chromatic aberration is not limiting the desired resolution. Determine the contribution of the most probable loss electrons to the coherent imaging component by using the exit wavefront reconstruction. If the contribution is only at low resolutions (i.e. below 200 $Å^{-1}$), then restoration of the unfiltered through focus series will contain only the elastically scattered component which is readily interpretable. If the most probable loss electrons contribute significantly to the higher resolutions, then this contribution cannot be eliminated through linear restoration. In this case, the restored exit wave of the unfiltered through focus series must be more carefully examined using simulations of electron-specimen interactions.

11. Once it is determined that linear restoration is valid for your specimen, to routinely reconstruct using fewer focus levels, it is recommended to use a more commonly used restoring filter derived by Schiske (Schiske 1973):

$$\hat{\psi}_e(\kappa) = \sum_{\Delta f_n=1}^{N} \hat{I}(\kappa, \Delta f_n) \cdot r(\kappa, \Delta f_n)$$

$$r(\kappa, \Delta f_n) = \exp[i\chi(\Delta f_n, \kappa)] \frac{\{N - \sum_{\Delta f_m=1}^{N} \exp[2i[\chi(\Delta f_m, \kappa) - \chi(\Delta f_n, \kappa)]\}}{\{N^2 - |\sum_{\Delta f_m=1}^{N} \exp[2i[\chi(\Delta f_m, \kappa)]]|^2\}}$$

where $\exp[i\chi(\Delta f, \kappa)] = \exp[i\frac{\pi\lambda\kappa^2}{2}(Cs\frac{\lambda^2\kappa^2}{2} - \Delta f)]$

The focus levels can be chosen by the curve from step 7, selecting the focus levels which the coherent component will cover the appropriate range of

resolution. The interpretation of the restored amplitude and phase images is a result of the analysis done in step 9.

This appendix is an approved reprint of the material as it appears in K.F.Han, Procedures in Electron Microscopy, Chapter 17 Module 5.

References

Han, K.F., Sedat, J.W. & Agard, D.A. (1995). Mechanism of image formation for thick biological specimens: *exit wavefront reconstruction and electron energy-loss spectroscopic imaging. J. Microscopy,* **178:2**, 107-19.

Han, K.F., Gubbens, A.J., Sedat, J.W. & Agard, D.A. (1996). Optimal strategies for imaging thick biological specimens: *exit wavefront reconstruction and energy filtering. J. Microscopy,* submitted.

Koster, A., Braunfeld, M., Fung, J., Abbey, C., Han, K., Liu, W., Chen, H., Sedat, J. & Agard, D. (1993). Towards Automated Three-Dimensional Imaging of Large Biological Structures Using Intermediate Voltage Electron Microscopy. *MSA Bulletin,* **23**, 176-88.

Schiske, P. (1973). Image processing using additional statistical information about the object. Image Processing and Computer-aided Design in Electron Optics. London, Academic Press.

Typke, D.H., R; Kleinz, J (1992). Image restoration for biological objects using external TEM control and electronic image recording. *Ultramicroscopy,* **46**, 157-173.

van Dyck, D. & Coene, W. (1987). A new procedure for wave function restoration in high resolution electron microscopy. *Optik,* **77**, 125-28.

van Dyck, D. & Op de Beeck, M. (1990). New direct methods for phase and structure retrieval in HREM. *Proc. of 12th Int'l Congress for Electron Microscopy,* 26-27.

Figure 3: General schematic approach to quantitate the coherent component in thick specimen imaging