COMPUTATIONAL CHALLENGES IN HIGH-RESOLUTION CRYO-ELECTRON MICROSCOPY

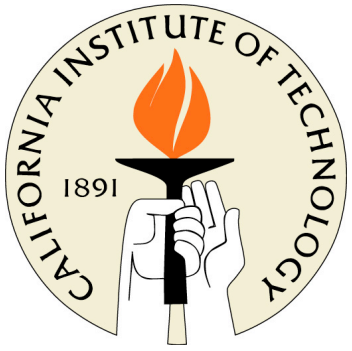
Thesis by

Peter Anthony Leong

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2009

(Defended Aug 04, 2008)

To God, who made all this possible

Acknowledgements

I would like to thank Prof. Grant Jensen for being an excellent advisor and role model for me during my time at Caltech. He has always shown tremendous understanding, kindness, and encouragement throughout, and supported me to the fullest extent in my research, career, and personal development — making this a wonderful life experience.

I am also very indebted to Drs. Bernard Heymann and Andrew Rawlinson. Bernard, who was a mentor to me when I first joined the lab, taught me much about computer hardware and software. Andy, who worked closely with me during the middle of my thesis work, helped me greatly in our discussions about the mathematics and physics related to our research work. Their mentoring has been very significant in my development as a scientist. In addition, I would also like to thank all my other lab mates, both past and present, who have been wonderful colleagues and friends. This thesis work could not have been completed without their help.

I would also like to thank my thesis and candidacy committee members Profs. Scott Fraser, Douglas Rees, Brent Fultz, Robert Phillips, and Z. Hong Zhou. Their advice and feedback about my research projects and about academics in general have been extremely helpful.

I would also like to thank my collaborators Profs. Hong Zhou, Wen Jiang, and Nikolaus Grigorieff, and their respective lab members, especially Drs. Xuekui Yu and Weimin Wu, for supporting me and helping in the completion of my thesis work.

Lastly and most importantly, I would like to thank my family, especially my parents, for all the love, support, and encouragement they have always shown me.

Abstract

To avoid the challenges of crystallization and the size limitations of NMR, it has long been hoped that single-particle cryo-electron microscopy (cryo-EM) would eventually yield atomically interpretable reconstructions. For the most favorable class of specimens (large icosahedral viruses), two of the key obstacles are the large computational requirements of high-resolution reconstructions and the curvature of the Ewald sphere, which leads to a breakdown of the projection theorem used by conventional 3D reconstruction programs. Here, two solutions to these obstacles are presented.

First, a simple distributed processing system named Peach was developed to meet the rising computational demands of modern structural biology (and other) laboratories without additional expense by using existing hardware resources more efficiently. A central server distributes jobs to idle workstations in such a way that each computer is used maximally, but without disturbing intermittent interactive users. As compared to other distributed systems, Peach is simple, easy to install, easy to administer, easy to use, scalable, and robust. While it was designed to queue and distribute large numbers of small tasks to participating computers, it can also be used to send single jobs automatically to the fastest currently available computer and/or survey the activity of an entire laboratory's computers. Tests of robustness and scalability are reported, as are three specific cryo-EM applications where Peach enabled projects that would not otherwise have been feasible without an expensive, dedicated cluster.

Second, an iterative refinement reconstruction algorithm, *Prec*, is described that overcomes the curvature of the Ewald sphere resolution limitation by averaging information from images recorded from different points of view, as are present in typical micrographs. *Prec* was implemented in the popular software packages IMIRS, EMAN, and Bsoft. In preliminary tests with both simple and multi-slice simulated images, *Prec* overcame the curvature problem even in the presence of noise. *Prec* was then used to refine the three recently published, ~ 4 Å resolution, icosahedral virus reconstructions from experimental cryo-EM images, but unfortunately no significant improvements in resolution were realized. Further simulations showed that limitations other than the Ewald sphere curvature problem must still be dominant in these experimental studies.

Table of Contents

List of Figures and Tables

Chapter 1

Introduction

1.1 Structural Biology

Structural biology is the approach to understanding cell biology through determining the structures of objects found in the cell. These objects range from proteins and molecular machines to organelles. To accommodate the difference in scales of these objects, which span from nanometers to microns, a variety of complementary imaging techniques are used. The imaging techniques, together, determine the structures of molecular machines and cellular structures and provide information about their quantity, distribution, and location. Also, real-time information about processes within cells, sometimes in their native states, can be extracted.

1.2 Structure Determination Techniques

The main techniques used in structural biology are X-ray crystallography (XRC), nuclear magnetic resonance spectroscopy (NMR), light microscopy (LM), computational biology and cryo-electron microscopy (Cryo-EM). These methods work together in a complementary way to reveal information about a variety of structures in different physical conditions.

As of June 2008, XRC has produced by far the largest number of atomic models of proteins as compared to NMR and Cryo-EM according to the Protein Data Bank. XRC works well for proteins that can be crystallized and the structures often reach atomic

resolution. The difficulty with this technique is that the crystallization process requires trying numerous conditions of temperature, pH, and buffer concentrations to produce a crystal that diffracts to sufficiently high resolution. These conditions result in structures of the proteins in non-native states. Once such crystals can be grown, X-ray diffraction patterns are then recorded, giving the Fourier amplitudes of the crystal. Next, the phases need to be determined ("phase problem") before the structures can be obtained.

NMR also produces atomic resolution structures but is limited to molecular masses of less than 50 kDa, which includes only the smaller proteins. On rare occasions, larger protein structures may be determined, for example, an 82-kDa enzyme in 2005 (Tugarinov, Choy et al. 2005).

LM allows for real-time imaging of live cells. Traditionally, this technique was limited in resolution by the wavelength of light and thus could not reveal the workings of the cell to higher resolutions. Recently, "super-resolution" techniques have been developed to surpass the diffraction limit as described in a recent review (Hell 2007) and have reached sub 100-nm resolutions (Juette, Gould et al. 2008; Schmidt, Wurm et al. 2008).

Computational biology techniques include comparative structure prediction, where protein structures are predicted using known structures as a reference, and *de novo* predictions in which no assumptions are made about the structures.

1.3 Cryo-Electron Microscopy

Cryo-EM delivers structures that span the resolution and size range between the atomic models provided by XRC or NMR, and the imaging of entire cells by LM. Its advantages are that samples are easily obtained, and when used in conjunction with plunge freezing (Dubochet and Mcdowall 1981) using a Vitrobot (Iancu, Tivol et al. 2006), the proteins or cells can be studied in their near-native state. This is achieved by first having the sample in a buffer which is spread onto a carbon film. The film is then plunged into liquid ethane, which cools the sample quickly enough so that the water in the sample is frozen in vitreous form (Angell 2004). This prevents the crystallization of water, which would damage the sample. The sample is then inserted into the microscope and imaged with electrons, which are scattered and then focused by electron lenses to form an image that is recorded on film or on a digital camera such as a charged-coupled device (CCD) or CMOS detector. An advantage of cryo-EM over XRC is the recording of images instead of just amplitudes. However, cryo-EM samples are limited to a thickness of ~ ½ micron (Lucic, Forster et al. 2005) to prevent multiple scattering of electrons within the cell. Also, the electron beam causes significant damage to the sample and thus the electron dose has to be kept low in order to reduce damage. This low dose results in images with low signal-to-noise ratios (SNRs).

There are several cryo-EM techniques available. Electron crystallography (EC) is used when 2D crystals of proteins, which are one unit cell thick, can be formed. In such situations, near-atomic resolution has been achieved (Henderson, Baldwin et al. 1990).

Similarly, the imaging of helical or tubular crystals also allows for atomic structures to be determined (Unwin 2005).

Electron cryo-tomography (ECT) is a technique which allows for the study of large structures and even entire small cells (Henderson and Jensen 2006). ECT can image the sample to high resolution in its native state, which is not possible with XRC, NMR, and LM. ECT complements LM because cells can be first observed *in vivo* with LM and then plunge-frozen to be imaged by ECT (Briegel, Ding et al. 2008). The ECT technique images cells from various tilt angles along one or more tilt axes. In theory, this technique would allow for a full reconstruction of a cell if the tilt angles ranging from -90° to +90° could be used. In practice, a maximum tilt of about ±65° is used, resulting in an artifact known as the "missing wedge or pyramid" (Iancu, Wright et al. 2005) in reconstructions of the cell. This artifact arises due to a wedge or pyramid of missing information in Fourier space. Another limitation of this technique is that the maximum dose to which the sample can be exposed has to be shared by all images of the tilt series in order to prevent information loss due to structural damage by the beam.

Lastly, Single particle analysis (SPA) is a technique in which many identical copies of a specimen are imaged. The particles in solution are applied to a grid and plunge-frozen. These grids are imaged resulting ideally in random views of these particles from all angles, although certain types of particles have preferred orientations. The images obtained from electron microscopes are noisy due to the low electron dose that can be tolerated by the sample. Fortunately, the information from these views can be averaged

to improve the SNR and produce high-resolution reconstructions of particles through Fourier reconstruction techniques (Crowther, Amos et al. 1970).

1.4 Reconstruction Theory

The reconstruction process can be simplified into three main stages (Figure 1-1). First, information about the object to be reconstructed is obtained in the form of raw projection images in various orientations, which are described by Euler angles and determined by the common-line method (Fuller, Butcher et al. 1996) for particles of high symmetry, or by 3D projection matching (Penczek, Grassucci et al. 1994). Secondly, corrected images are produced by the correction of raw images, which removes artifacts that were introduced during the imaging process due to the point spread function (PSF). This process is called contrast transfer function (CTF) correction and is performed by taking the 2D Fourier transform (FT) of a raw image and dividing it by the CTF, which is the FT of the PSF, before taking the inverse FT to get a corrected image. Thirdly, a 3D real-space reconstruction of the object is determined by a reconstruction algorithm.

To a good approximation, corrected images are projections of the object, which are equivalent to the inverse FT of central slices in the 3D FT of the object being reconstructed (Bragg 1929):

$$\begin{aligned} p(x,y) &= \int \rho(x,y,z)dz \\ &= \int \iiint F(X,Y,Z)e^{i2\pi(xX+yY+zZ)}dXdYdZdz \\ &= \iiint F(X,Y,Z)e^{i2\pi(xX+yY)}\delta(Z)dXdYdZ \\ &= \iint F(X,Y,0)e^{i2\pi(xX+yY)}dXdY \end{aligned} \tag{1}$$

where $p(x,y)$ is a projection of the object along the z-axis, $\rho(x,y,z)$ is the density of the object and $F(X,Y,Z)$ is the 3D FT of the object. This derivation can be generalized for projections in all possible directions and is called the projection theorem.

Using the property above, the 3D FT of the object can be determined by adding many central slices with different orientations using Whittaker-Shannon interpolation (Whittaker 1915; Shannon 1949) or by Fourier-Bessel synthesis (Klug, Crick et al. 1958). Once the 3D FT has been sufficiently sampled, the inverse FT can be calculated to give the reconstruction of the object.

1.5 Resolution Measures

When discussing resolution, a high resolution (or spatial frequency) corresponds to the resolvability of features separated by small distances, while a low resolution (or spatial frequency) corresponds to the resolvability of features separated by large distances; Atomic resolution refers to the resolvability of the distances between atoms while near-atomic resolution, which is slightly lower, implies that atomic models can be fit with the help of additional information such as the protein sequence.

In SPA, the quality of a reconstruction is measured in terms of the resolution achieved, which can be measured numerically or visually. Both these methods are subjective and can be manipulated to provide better or worse results by adjusting certain parameters.

The most commonly used numerical resolution measure is the Fourier shell coefficient (FSC) (Harauz and Van Heel 1986). In order to calculate the FSC, a data set consisting of a large number of images is split randomly into two halves. Independent reconstructions of each half of the data set are generated. The two reconstructions are then compared by calculating the value of the FSC at each spatial frequency

$$FSC(|s|) = \frac{\sum_i |F_1^i||F_2^i| Cos(\phi_1^i - \phi_2^i)}{\sqrt{\sum_i |F_1^i|^2 \sum_i |F_2^i|^2}} \qquad (2)$$

where i enumerates the set of points found at spatial frequency s in the 3D FTs of the two reconstructions, F_1^i and F_2^i represent the values of the Fourier coefficients for each half of the data set and ϕ_1^i and ϕ_2^i represent their phases.

A variety of factors (van Heel and Schatz 2005) can affect the value of the FSC resolution, such as the number of additional voxels in the reconstruction which are in excess to the object being reconstructed. Changing the size of the volume containing the reconstruction adjusts the amount of additional voxels. Other factors that affect the measured resolution include the types of masks and how sharp these masks are, and most importantly the FSC threshold value, which indicates the maximum resolution of the reconstruction.

The resolution of a reconstruction can be determined visually if the resolution is sufficiently high. This has recently been possible with high-resolution reconstructions of icosahedral virus particles at ~ 4 Å resolution (Jiang, Baker et al. 2008; Yu, Jin et al. 2008; Zhang, Settembre et al. 2008). Table 1-1 gives a list of biological structural features that can be observed at various resolutions. High-resolution details can be enhanced to a certain extent by applying an "inverse" B-factor to the reconstructions, which adjusts the weighting of higher-resolution information by multiplication with the following factor:

$$e^{Bs^2} \tag{3}$$

where B is the B-factor and s is the spatial frequency.

However, it is important to note the FSC is not affected by the B-factor:

$$\begin{aligned} FSC_B(|s|) &= \frac{\sum_i |F_1^i e^{Bs^2}||F_2^i e^{Bs^2}|Cos(\phi_1^i - \phi_2^i)}{\sqrt{\sum_i |F_1^i e^{Bs^2}|^2 \sum_i |F_2^i e^{Bs^2}|^2}} \\ &= \frac{\sum_i |F_1^i||F_2^i|Cos(\phi_1^i - \phi_2^i)}{\sqrt{\sum_i |F_1^i|^2 \sum_i |F_2^i|^2}} \left(\frac{e^{Bs^2}}{e^{Bs^2}}\right)^2 \\ &= FSC(|s|) \end{aligned} \tag{4}$$

In addition, since the FSC calculation uses half datasets while the visually determined resolution uses the entire dataset, the latter gives a higher measure of resolution.

1.6 Resolution Limitations

There are two sets of resolution limitations involved in the SPA process. The first set consists of instrumentation limitations. These include incoherent beam sources, specimen preservation during the imaging process, and specimen charging by the electron beam, among others. The second set of resolution limitations consist of processing limitations, which include orientation, origin, and defocus determination and lack of computational power. There also exists the depth of field or equivalently the Ewald sphere curvature problem, which can be solved both computationally and instrumentally. Further discussion of these resolution limitations can be found in cryo-EM reviews (Baker, Olson et al. 1999; van Heel, Gowen et al. 2000).

1.7 Instrumentation Progress

Better electron sources and energy filters, more stable cooling stages, and larger, more sensitive CCD cameras have allowed structure determination by cryo-EM to approach near-atomic resolution by improving the recording of higher-resolution information with fewer artifacts and increasing data throughput.

In modern electron microscopes, the electron beam source is a highly coherent field emission electron gun (FEG). The FEG consists of a pointed field emission tip placed near a positive electrode. This causes a strong electric field to form which allows electrons to overcome the work function of the filament (usually tungsten) and be emitted. FEGs are better than previous electron sources, such as the thermionic W or LaB_6 and Schottky ZrO/W guns. They are spatially and temporally more coherent

because they produce better point electron sources and are colder, which reduces the thermal energy spread, leading to more monochromatic beams, respectively. The electron beams are focused with improved electron lenses that have lower spherical aberrations than previously. Samples are cooled by liquid nitrogen in ECT (Iancu, Wright et al. 2006) and by liquid helium (Fujiyoshi, Mizusaki et al. 1991) in SPA (van Heel, Gowen et al. 2000) and EC (Hite, Raunser et al. 2007) to reduce beam damage. In addition, energy filters are used to ensure that only elastically scattered electrons are recorded on the CCD. Furthermore, the entire data collection process can be automated (Potter, Chu et al. 1999).

1.8 Progress in Processing Techniques

Although the fundamentals of the reconstruction process are still the same, there now exist several popular software packages that are used in the reconstruction of virus particles by single particle analysis. For example, IMIRS (Liang, Ke et al. 2002) utilizes the Fourier-Bessel synthesis method and was written for Microsoft Windows XP, while EMAN (Ludtke, Baldwin et al. 1999), FREALIGN (Grigorieff 2007) and Bsoft (Heymann 2001) are Cartesian-coordinate, UNIX-based packages which use a variety of interpolations which are approximations of a full 3D Fourier interpolation (Whittaker 1915; Shannon 1949).

Fundamental improvements to the reconstruction process include CTF correction of images and more sophisticated orientation determination algorithms, among others.

Improvements in computer hardware have also allowed for larger reconstructions to be computed because of 64-bit memory addressing and faster CPU speeds.

1.9 Approaching Atomic Resolution by Cryo-EM

With these advances, near-atomic resolution of biological structures was first achieved using EC (Henderson, Baldwin et al. 1990) and then by helical or tubular reconstructions (Unwin 2005). Thus the next technique by Cryo-EM that will approach these high resolutions is SPA. The alignment and orientation determination process, which is not required for EC and helical reconstructions, is non trivial, but using particles with large masses lessens this obstacle. In addition, high physical symmetry allows for fewer particles to be used in the reconstruction process. Thus large icosahedral virus particles are the best candidates for SPA to achieve atomic models.

1.10 Icosahedral Virus Structures

Virus capsids are composed of many identical copies of one or a few different capsid proteins, and as a result, the genetic material of the virus can be smaller and the production of a complete virus capsid quicker (Crick and Watson 1956; Caspar and Klug 1962). This use of identical proteins usually results in capsids of helical symmetry, the best known example being the tobacco mosaic virus (Bloomer, Champness et al. 1978), or icosahedral symmetry, for example, the herpes simplex virus (Zhou, Dougherty et al. 2000). Icosahedral symmetry is the naturally preferred structure for containing the virus genome because it provides the largest volume using the fewest capsid units possible. Each of the 20 triangular faces of the icosahedral structure consists of three asymmetric

units. Furthermore, each of these asymmetric units can be composed of a number of either identical or different subunits. The triangulation (T) number (Caspar and Klug 1962) specifies the number of subunits in each asymmetric unit.

Any image of an icosahedral virus particle can be used 60 times in the reconstruction process because icosahedral virus particles possess 60-fold symmetry. Alternatively, only $1/60^{th}$ of the total information is required to reconstruct a virus particle. The latter approach is more difficult to achieve in reconstruction algorithms but some progress has been made towards it with the Fourier-Bessel reconstruction algorithm (Crowther, Amos et al. 1970) which uses $1/10^{th}$ of the information by aligning the 5-fold axis along the z-axis and utilizing 2-fold symmetry which results in information being required only between the azimuthal angles of 0° and 36° in a cylindrical coordinate system. Likewise, orientation determination of icosahedral particles is also easier due to the symmetry which allows for the use of the common-line method (Fuller, Butcher et al. 1996), which compares intersections of the 60 central slices from each image to derive the correct orientation.

1.11 Viruses

Virus structures are being intensively researched, as shown by a recent PubMed search for “virus structure”, which yielded over 37,000 hits. An old review of solved icosahedral virus structures listed over 175 reconstructions (Baker, Olson et al. 1999), further underlining the effort being invested.

Viruses consist of genetic material enclosed in capsids, with or without envelopes. A classification scheme was proposed (Baltimore 1971) which separated viruses into classes depending on the type of genetic material contained within the capsids. Viruses infect host cells either by being transported through the cellular membranes, or by injecting their genetic material, in the form of DNA or RNA, into the cell. If viral DNA is introduced into the cell, it is transcribed to produce RNA. The viral RNA is subsequently translated into proteins that form the virus capsid. Despite detailed understanding, there is still much to learn and exploit, for example, targeted viruses can be used to cause cancer cells to kill themselves (Ito, Aoki et al. 2006).

Viruses cause a wide range of diseases, such as AIDS (human immunodeficiency virus), cold sores (herpes virus) and even cancer (papilloma virus) (zur Hausen 2002). Greater understanding of viruses aids us in our attempts to cure or prevent certain diseases, which in turn would allow us to improve or save the lives of millions of people. While reconstructions that achieve a resolution of ~ 3.5 Å allow atomic models to be fit within the density, higher resolutions of ~ 2 Å allow predictions of the behavior and location of the interaction surfaces of virus capsids, which in turn guide drug design in producing drugs that target these surfaces by disrupting the original interaction surface properties, thereby disrupting assembly of capsids.

In addition, the study of viruses as simplified cellular machines continues to improve our understanding of evolution, for example, by understanding that viruses may be agents in

horizontal gene transfer. These studies have also improved our knowledge of cell biology.

1.12 Approach Atomic Resolution by Single Particle Analysis

3D reconstructions of virus particles from electron micrographs by Fourier synthesis were first accomplished in 1970 (Crowther, Amos et al. 1970). Since then, reconstruction algorithms have improved and matured, resulting in sub-nanometer resolution in 1997 (Bottcher, Wynne et al. 1997; Conway, Cheng et al. 1997; Trus, Roden et al. 1997).

According to Glaeser (Glaeser 1999), achieving atomic resolution, which requires the determination of orientations from 10^6 images, would require an estimated 10^{23} floating point operations, which would take the world's fastest super computer with a maximum processing power of 1.375 PFlops (June 2008, www.top500.org) over two years to complete.. Fortunately, the 60-fold symmetry of icosahedral viruses reduces that number by nearly two orders of magnitude.

When I first began my thesis work, several factors that limited the resolution of SPA reconstruction had not been addressed. I attempted to address two of these challenges, namely the lack of computing power in reconstruction algorithms and the depth of field or equivalently, the Ewald sphere curvature problem (DeRosier 2000).

The resolutions of SPA reconstructions have improved significantly in the last few years and towards the end of my thesis work in 2008, three structures reached near-atomic resolution (Jiang, Baker et al. 2008; Yu, Jin et al. 2008; Zhang, Settembre et al. 2008).

1.13 Computational Complexity of 3D Reconstruction Algorithm

3D reconstructions are highly computationally and memory intensive. Despite the increasing amounts of memory available, increasing speeds of processors, and the increase in number of cores and processors per computer, the computation requirements are still very high when trying to perform reconstructions of very large viruses to high resolutions.

The basic reconstruction algorithm requires that the 3D FT be held in memory as samples are applied to it, which results in a $O(n^3)$ memory requirement where n is the length of one side of the transform. Due to the large memory requirements, it is necessary that the computer performing the reconstruction possess enough RAM to meet this requirement. Computers lacking the necessary RAM will require swapping of memory, a process that utilizes the hard disk as additional memory. As hard disk access is several orders of magnitude slower than RAM access, the resulting computation would not be completed in a reasonable amount of time. The number and size of images being used in reconstructions are very large when high-resolution reconstructions are required, due to the smaller pixel sizes and the higher sampling of images. In practice, for a reconstruction of a virus particle using 1k x 1k images, the memory requirements would

be approximately 16, 20, and 30 GB for EMAN (Ludtke, Baldwin et al. 1999), Bsoft (Heymann 2001), and FREALIGN (Grigorieff 2007), respectively. IMIRS (Liang, Ke et al. 2002), which is highly optimized, would require less than 2GB. Currently, 64-bit systems allow for access of sufficient memory for even the largest of virus particles. Thus, memory requirements are a cost issue, which can be overcome with purchasing of sufficient RAM.

The computation of the basic reconstruction algorithm consists of applying the value of each pixel of the 2D FT of the images to the 3D FT making this a $O(m\,n^2)$ computation problem where m is the number of images and n is the length of one side of the 2D FT of an image. While the problem is tractable, it does take a significant amount of time for high-resolution structures of large virus particles, once again, due to the larger images used in the reconstruction process. While it may seem that purchasing faster computers can likewise solve the computation problem, it is not a good solution because CPU speeds have already started to plateau. Fortunately, the computation problem is trivially parallelizable for the most part and thus parallel and distributed computation are possible solutions to solve the problem efficiently.

1.14 Parallel Computation

One approach is the parallelization of the reconstruction process, which allows for the utilization of multiple cores or processors on a single computer or supercomputer that has shared memory and fast access to this memory. Parallelization takes advantage of the recent trend by CPU chip manufacturers to increase the number of cores per CPU instead

of increasing the speed of the processors. A program that is multi-threaded will be able to process multiple calculations simultaneously and would take advantage of these additional resources. This multi-threaded approach which utilizes shared memory would require only one copy of the 3D FT to be stored in memory while allowing for the computation time to be reduced due to the increased number of threads performing calculations on the various processors or cores without any significant additional memory requirements.

The most significant drawback to this approach is that when too many threads are utilized, a bottleneck of the process occurs in the write access to the large memory holding the 3D FT of the object. The number of memory accesses, due to sample values being applied to the 3D FT, would be $O(m n^2)$ where m is the number of images used in the reconstruction and n is the length of one side of the image. These memory accesses are essentially random in their access pattern of the memory and thus require that the shared memory be locked before changes are made to it to prevent race conditions where changes are inadvertently lost when multiple threads access the same memory location at the same time. This bottleneck is encountered when the rate of samples calculated, which scale linearly with the number of cores, exceeds the rate at which samples are applied to the 3D FT, which is limited to the RAM access rate that is a constant. At this point, additional cores cannot accelerate the reconstruction process any further because the additional threads would spend increasing amounts of time waiting for access to the shared memory.

Implementations of parallel optimizations to the reconstruction algorithms using multi-threading libraries, such as the pthreads library, are described in Chapter 3.

1.15 Distributed Computation

The second available approach to reducing computation time is by distributed computation. This means that individual processes, which are executed on multiple computers with the necessary memory requirements, can take a subset of the data and perform independent reconstructions that are later combined to produce the full reconstruction. It is also possible to execute multiple processes on a single machine with the requisite number of processors and the required multiples of RAM.

Many structural biology laboratories possess mixtures of heterogeneous workstations purchased individually or in small sets for laboratory personnel, which constitutes a wealth of underutilized computation capacity. This is an ideal situation for this using the distributed approach to solve the computational problem.

This untapped resource was previously unworkable because of the effort required of researchers to log in to multiple computers and manually distribute jobs across computers with different operating systems. In addition, custom scripts were needed to submit jobs one after another through the night or weekend and watch for their completion. Lastly, computer usage had to be coordinated with laboratory colleagues so as not to impede their own computation efforts. Despite such efforts on the parts of some researchers,

most workstations were still only used to a small fraction of their capacity due to the difficulty of manually managing multiple tasks on multiple workstations.

In 2003, only a few distributed systems were available, including Open PBS from Veridian Systems, Condor (Tannenbaum and Litzkow 1995), and BOINC, the Berkeley Open Infrastructure for Network Computing, which mediates the SETI@home project (Anderson, Cobb et al. 2002). These systems did not meet all our requirements for processing jobs that had extensive read, write, and memory requirements, were computationally intensive; had little or no fault tolerance, needed no changes to source code, and enabled desktop harvesting.

Peach, a distributed computation system, which is described in detail in Chapter 2, was developed in order to meet those requirements and also be simple to use and administer, scalable, secure, robust, and as compatible as possible with the existing hardware and software in structural biology. Essentially, Peach allows for multiple jobs to be submitted to a heterogeneous cluster of computers and utilizes clock cycles of idle computers. This distributed approach requires many powerful computers with sufficient RAM when used in the reconstruction of large virus particles. Furthermore, distributed computation is also applicable to a wide range of tasks in image processing.

The combination of the information, after the independent reconstructions are completed, requires $O(\log(c)n^3)$ steps, where c is the number of separate computers used in the reconstruction and n is the length of one side of the reconstruction, which is independent

of the total number of images. This combination by binary merging is significantly quicker than in the parallel approach because results have already been accumulated by the individual reconstructions before being combined and can be combined in parallel, i.e., n reconstructions can be merged by $\frac{1}{2}n$ individual processes repeatedly until the final reconstruction is left and thus require only $\log(c)$ stages of combinations. If there is availability of computers with sufficient RAM, then distributed computation is a better solution than the parallel approach because it does not encounter the memory access bottleneck.

1.16 Hybrid Approach

A hybrid approach, using both parallel and distributed approaches together, would be the best solution in the reconstruction of large viruses as it utilizes computing resources maximally by using all available cores on all available computers. This approach is feasible with new implementations (Chapter 3) of the reconstruction algorithms in Bsoft (Heymann 2001) and EMAN (Ludtke, Baldwin et al. 1999) that possess capabilities for both parallel and distributed computation, and which may be used in conjunction with a suitable distributed computation system such as Peach (Chapter 2) or by processing on several multi-core nodes of a supercomputer.

1.17 Depth of Field and Ewald Sphere Curvature

As mentioned above, one of the resolution limitations of SPA of large virus particles is the depth of field problem, or equivalently, the Ewald sphere curvature. The depth of field, which is the distance over which the sample is in focus, is sometimes mistakenly

called the depth of focus, which corresponds to the distance over which the recorded image is in focus (Fultz and Howe 2002). The depth of field can be geometrically calculated according to the following formula:

$$D = \frac{d}{\alpha} \tag{5}$$

where D is the depth of field, d is the resolution, and α is the aperture angle of the lens. For a typical transmission electron microscope, the aperture angle α is ~ 10^{-3} rad and the resolution d ~ 5 Å giving a depth of field D of ~ 5000 Å or a ½ micron.

The geometric estimate, however, cannot be applied to high-resolution phase contrast information because small defocus changes Δd, on the order of 10^2 Å, affect the image intensity distribution (Reimer 1997). This effect is due to the wave aberration

$$\chi = \frac{\pi}{2} C_s \lambda^3 s^4 - \pi \Delta f \lambda s^2 \tag{6}$$

where C_s is the spherical aberration, λ is the electron wavelength, Δf is the defocus value, and s is the spatial frequency. We find that $\Delta d \leq \frac{1}{\lambda s^2}$ when setting the change in the wave aberration to be less than π. For a resolution of 3.8 Å at 300 kV, where λ ~ 0.02 Å and s ~ 0.263 Å^{-1}, Δd is ~ 720 Å which is approximately the diameters of the CPV, ε15, and DLP capsids.

The defocus gradient and Ewald sphere curvature problems were shown to be equivalent first in 1978 (Amelinckx, Gevers et al. 1978), then in 2000 (DeRosier 2000) and again in 2004 (Wan, Chiu et al. 2004). Further elaboration about their equivalence qualitatively and quantitatively is provided below.

Firstly to understand the situation qualitatively, consider the Ewald sphere in XRC. Reciprocal lattice points have dimensions that are inversely proportional to the size of the crystal. If the crystal thickness in one direction is large, then the dimension of the reciprocal lattice point in that direction becomes small. Likewise, if we have a thin crystal, then the dimension of the reciprocal lattice point in that direction becomes very long and is known as a reciprocal rod or "rel-rod". The intersections of the Ewald sphere and reciprocal lattice points are where scattering occurs. Take the situation where reciprocal lattice points lie along the XY plane. If the incoming beam is along the Z-axis, then at high resolutions along the plane, there will be reciprocal lattice points which do not intersect the Ewald sphere. If the crystal is thin, then the rel-rods stretch and intersect the Ewald sphere. Alternatively, if instead a higher voltage is used, the Ewald sphere flattens or has a larger radius. In this situation, the reciprocal lattice points also intersect the Ewald sphere without needing to be rel-rods. Thus, in this situation having a crystal thin enough will render the Ewald sphere curvature negligible. Conversely, if a crystal is thick enough, then the Ewald sphere curvature cannot be neglected at high resolution.

The XRC example can be viewed as a simple case of what occurs in electron microscopy (EM). The difference is that in EM, the sample is not crystalline and thus, the Fourier

amplitudes vary continuously in all directions, as opposed to discrete reciprocal lattice points in crystallography, and scattering occurs at all points on the Ewald sphere. Analogously, the variations in the FT are quicker with thicker EM samples and slower for thinner EM samples in the direction of the thickness. Thus, when larger virus particles are imaged, the effect of the Ewald sphere is significant and should not be ignored. Alternatively, if a small virus particle is imaged, the variations of the FT are slower, so the Ewald sphere curvature is less significant.

This can also be explained quantitatively. First let us take a point $(X_0,Y_0,Z(X_0,Y_0))$ from the 3D FT of an object of radius R, where $Z(X,Y) \approx \frac{1}{2}\lambda(X^2+Y^2)$. The object can be broken down in real space as a set of thin slabs at different defocus values. During the recording of the image, all the slabs contribute to $(X_0,Y_0,Z(X_0,Y_0))$ but with different defocus values or, equivalently, different phase delays due to the wave aberration χ (equation 6). The difference in the defocus values, Δd, from the center defocus, result in phase delays of $-\pi\Delta d\lambda s^2$ with respect to the defocus value at the center of the object. Thus contributions from the top slab of the object will have an additional phase delay of $\pi R\lambda s^2$, as compared to the slab at the center.

Taking the alternative view, we assume a single defocus for the entire object. Then, the slabs have contributed to $(X_0,Y_0,Z(X_0,Y_0))$ without additional phase delays as all slabs are of the same defocus. However, for each slab to have no additional phase delays, the slabs would have to be located at the center of the object. Since the slabs are physically located away from the center, each slab has a phase shift due to its location, and thus a

phase delay according to the Fourier shift theorem, which states that $F(X,Y,Z)$ becomes $F(X,Y,Z)e^{i2\pi zZ}$ when an object is shifted in position by a value z. The separate slabs, with physical shifts z, thus have phase delays of $2\pi zZ \approx 2\pi z(\frac{1}{2}\lambda s^2) = \pi z\lambda s^2$ where Z is due to the curvature of the Ewald sphere. Once again, a phase delay of $\pi R\lambda s^2$ occurs between contributions at the top slab of the object and the center. This phase delay would not have existed if the Ewald sphere curvature were negligible or equivalently when Z is set to 0.

The phase delays are identical in both cases. This indicates that considering the defocus gradient over an object is equivalent to taking into account the Ewald sphere curvature while assuming a single defocus value. Alternatively, if an object possesses an insignificant defocus gradient, then curvature of the Ewald sphere can be ignored.

1.18 Virus Structures Limited by Ewald Sphere Curvature

Various studies have shown that the Ewald sphere curvature is significant for particles ~ 700 Å or greater in diameter, at near-atomic resolution. In 2008, three virus structures of this diameter were reconstructed to near-atomic resolution of ~ 4 Å (Jiang, Baker et al. 2008; Yu, Jin et al. 2008; Zhang, Settembre et al. 2008). According to Jensen and Kornberg's envelope function (Jensen and Kornberg 2000), half of the signal in a conventional reconstruction of such a large virus at 300 kV would be lost due to curvature of the Ewald sphere by ~ 3.5 Å resolution. Likewise, DeRosier's formula (DeRosier 2000) predicts that the curvature problem in this same situation would become significantly limiting by ~ 3.3 Å resolution.

Thus, the Ewald sphere curvature will be most significant for three families of large icosahedral viruses, namely, the adenoviridae, herpesviridae and reoviridae, as their diameters are large enough that the curvature of the Ewald sphere will become significant at near-atomic resolution (Table 1-2). These families are medically important as they are responsible for a large range of diseases; for example, respiratory tract infections, conjunctivitis, hemorrhagic cystitis, and gastroenteritis (Adenoviridae), oral and genital herpes, chickenpox, and shingles (Herpesviridae), and human infantile gastroenteritis (Reoviridae). For instance, the 1250 Å diameter herpes simplex virus (HSV) (Zhou, Dougherty et al. 2000), which is currently present in over 60% of the US population, is responsible for herpes, cowpox, cancer, and many other dangerous diseases.

To overcome the Ewald sphere curvature resolution limit, the paraboloid reconstruction (Prec) algorithm for Cryo-EM, was developed to correct for the effects of the Ewald sphere curvature in the context of 3D reconstructions. Details of the algorithm are discussed in Chapter 3.

1.19 References

Amelinckx, S., R. Gevers, et al. (1978). Diffraction and imaging techniques in material science. Amsterdam ; New York, Elsevier North-Holland.

Anderson, D. P., J. Cobb, et al. (2002). "SETI@home - An experiment in public-resource computing." Communications of the Acm **45**(11): 56−61.

Angell, C. A. (2004). "Amorphous water." Annual Review of Physical Chemistry **55**: 559−583.

Baker, T. S., N. H. Olson, et al. (1999). "Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs." Microbiology and Molecular Biology Reviews **63**(4): 862−922.

Baltimore, D. (1971). "Expression of Animal Virus Genomes." Bacteriological Reviews **35**(3): 235−241.

Bloomer, A. C., J. N. Champness, et al. (1978). "Protein Disk of Tobacco Mosaic-Virus at 2.8-a Resolution Showing Interactions within and between Subunits." Nature **276**(5686): 362−368.

Bottcher, B., S. A. Wynne, et al. (1997). "Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy." Nature **386**(6620): 88−91.

Bragg, W. L. (1929). "The determination of parameters in crystal structures by means of fourier series." Proceedings of the Royal Society of London Series A-Containing Papers of a Mathematical and Physical Character **123**(792): 537−559.

Briegel, A., H. J. Ding, et al. (2008). "Location and architecture of the Caulobacter crescentus chemoreceptor array." Molecular Microbiology **69**(1): 30−41.

Caspar, D. L. D. and A. Klug (1962). "Physical Principles in Construction of Regular Viruses." Cold Spring Harbor Symposia on Quantitative Biology **27**: 1−24.

Conway, J. F., N. Cheng, et al. (1997). "Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy." Nature **386**(6620): 91−94.

Crick, F. H. C. and J. D. Watson (1956). "Structure of Small Viruses." Nature **177**(4506): 473−475.

Crowther, R. A., L. A. Amos, et al. (1970). "3 Dimensional Reconstructions of Spherical Viruses by Fourier Synthesis from Electron Micrographs." Nature **226**(5244): 421−425.

DeRosier, D. J. (2000). "Correction of high-resolution data for curvature of the Ewald sphere." Ultramicroscopy **81**(2): 83−98.

Dubochet, J. and A. W. Mcdowall (1981). "Vitrification of Pure Water for Electron-Microscopy." Journal of Microscopy-Oxford **124**(DEC): RP3−RP4.

Fujiyoshi, Y., T. Mizusaki, et al. (1991). "Development of a Superfluid-Helium Stage for High-Resolution Electron-Microscopy." Ultramicroscopy **38**(3−4): 241−251.

Fuller, S. D., S. J. Butcher, et al. (1996). "Three-dimensional reconstruction of icosahedral particles - The uncommon line." Journal of Structural Biology **116**(1): 48−55.

Fultz, B. and J. M. Howe (2002). Transmission electron microscopy and diffractometry of materials. Berlin ; New York, Springer.

Glaeser, R. M. (1999). "Review: Electron crystallography: Present excitement, a nod to the past, anticipating the future." Journal of Structural Biology **128**(1): 3-14.

Grigorieff, N. (2007). "FREALIGN: High-resolution refinement of single particle structures." Journal of Structural Biology **157**(1): 117−125.

Harauz, G. and M. Van Heel (1986). "Exact Filters for General Geometry 3-Dimensional Reconstruction." Optik **73**(4): 146−156.

Hell, S. W. (2007). "Far-field optical nanoscopy." Science **316**(5828): 1153−1158.

Henderson, G. P. and G. J. Jensen (2006). "Three-dimensional structure of Mycoplasma pneumoniae's attachment organelle and a model for its role in gliding motility." Molecular Microbiology **60**(2): 376−385.

Henderson, R., J. M. Baldwin, et al. (1990). "Model for the Structure of Bacteriorhodopsin Based on High-Resolution Electron Cryomicroscopy." Journal of Molecular Biology **213**(4): 899-929.

Henderson, R., J. M. Baldwin, et al. (1990). "Model for the Structure of Bacteriorhodopsin Based on High-Resolution Electron Cryomicroscopy." Journal of Molecular Biology **213**(4): 899−929.

Heymann, J. B. (2001). "Bsoft: Image and molecular processing in electron microscopy." Journal of Structural Biology **133**(2−3): 156−169.

Hite, R. K., S. Raunser, et al. (2007). "Revival of electron crystallography." Current Opinion in Structural Biology **17**(4): 389−395.

Iancu, C. V., W. F. Tivol, et al. (2006). "Electron cryotomography sample preparation using the Vitrobot." Nature Protocols **1**(6): 2813−2819.

Iancu, C. V., E. R. Wright, et al. (2005). "A "flip-flop" rotation stage for routine dual-axis electron cryotomography." Journal of Structural Biology **151**(3): 288−297.

Iancu, C. V., E. R. Wright, et al. (2006). "A comparison of liquid nitrogen and liquid helium as cryogens for electron cryotomography." Journal of Structural Biology **153**(3): 231−240.

Ito, H., H. Aoki, et al. (2006). "Autophagic cell death of malignant glioma cells induced by a conditionally replicating adenovirus." Journal of the National Cancer Institute **98**(9): 625−636.

Jensen, G. J. and R. D. Kornberg (2000). "Defocus-gradient corrected back-projection." Ultramicroscopy **84**(1−2): 57−64.

Jiang, W., M. L. Baker, et al. (2008). "Backbone structure of the infectious epsilon 15 virus capsid revealed by electron cryomicroscopy." Nature **451**(7182): 1130−1134.

Juette, M. F., T. J. Gould, et al. (2008). "Three-dimensional sub-100 nm resolution fluorescence microscopy of thick samples." Nature Methods **5**(6): 527−529.

Klug, A., F. H. C. Crick, et al. (1958). "Diffraction by Helical Structures." Acta Crystallographica **11**(3): 199−213.

Liang, Y. Y., E. Y. Ke, et al. (2002). "IMIRS: a high-resolution 3D reconstruction package integrated with a relational image database." Journal of Structural Biology **137**(3): 292−304.

Lucic, V., F. Forster, et al. (2005). "Structural studies by electron tomography: From cells to molecules." Annual Review of Biochemistry **74**: 833−865.

Ludtke, S. J., P. R. Baldwin, et al. (1999). "EMAN: Semiautomated software for high-resolution single-particle reconstructions." Journal of Structural Biology **128**(1): 82−97.

Murray, P. R., K. S. Rosenthal, et al. (2005). Medical microbiology. Philadelphia, Elsevier Mosby.

Penczek, P. A., R. A. Grassucci, et al. (1994). "The Ribosome at Improved Resolution - New Techniques for Merging and Orientation Refinement in 3d Cryoelectron Microscopy of Biological Particles." Ultramicroscopy **53**(3): 251−270.

Potter, C. S., H. Chu, et al. (1999). "Leginon: a system for fully automated acquisition of 1000 electron micrographs a day." Ultramicroscopy **77**(3−4): 153−161.

Reimer, L. (1997). Transmission electron microscopy : physics of image formation and microanalysis. Berlin ; New York, Springer.

Schmidt, R., C. A. Wurm, et al. (2008). "Spherical nanosized focal spot unravels the interior of cells." Nature Methods **5**(6): 539−544.

Shannon, C. E. (1949). "Communication in the Presence of Noise." Proceedings of the Institute of Radio Engineers **37**(1): 10−21.

Tannenbaum, T. and M. Litzkow (1995). "The Condor Distributed-Processing System." Dr Dobbs Journal **20**(2): 40−48.

Trus, B. L., R. B. S. Roden, et al. (1997). "Novel structural features of bovine papillomavirus capsid revealed by a three-dimensional reconstruction to 9 angstrom resolution." Nature Structural Biology **4**(5): 413−420.

Tugarinov, V., W. Y. Choy, et al. (2005). "Solution NMR-derived global fold of a monomeric 82-kDa enzyme." Proceedings of the National Academy of Sciences of the United States of America **102**(3): 622−627.

Unwin, N. (2005). "Refined structure of the nicotinic acetylcholine receptor at 4 angstrom resolution." Journal of Molecular Biology **346**(4): 967−989.

Unwin, N. (2005). "Refined structure of the nicotinic acetylcholine receptor at 4 angstrom resolution." Journal of Molecular Biology **346**(4): 967-989.

van Heel, M., B. Gowen, et al. (2000). "Single-particle electron cryo-microscopy: towards atomic resolution." Quarterly Reviews of Biophysics **33**(4): 307−369.

van Heel, M. and M. Schatz (2005). "Fourier shell correlation threshold criteria." Journal of Structural Biology **151**(3): 250−262.

Wan, Y., W. Chiu, et al. (2004). "Full contrast transfer function correction in 3D cryo-EM reconstruction". IEEE Proceedings of ICCCAS 2004 Chengdu, Sichuan, China.

Whittaker, E. T. (1915). "On the Functions which are Represented by the Expansion of Interpolation Theory." Proceedings of the Royal Society of Edinburgh **35**: 181−194.

Yu, X. K., L. Jin, et al. (2008). "3.88 angstrom structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy." Nature **453**(7193): 415−419.

Zhang, X., E. Settembre, et al. (2008). "Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction." Proceedings of the National Academy of Sciences of the United States of America **105**(6): 1867−1872.

Zhou, Z. H., M. Dougherty, et al. (2000). "Seeing the herpesvirus capsid at 8.5 angstrom." Science **288**(5467): 877−880.

zur Hausen, H. (2002). "Papillomaviruses and cancer: From basic studies to clinical application." Nature Reviews Cancer **2**(5): 342−350.

1.20 Figures and tables

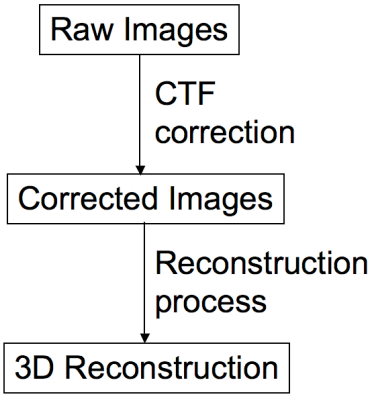


Figure 1-1. **Flow chart of simplified reconstruction process.** The reconstruction process consist of three stages: (1) Raw images from electron micrographs, (2) Corrected images produced by CTF correction of Raw images, (3) 3D real-space reconstruction generated by reconstruction algorithm using corrected images

Biological Structural Features	**Approximate Resolution**
α-helices	~ 7Å
Main chain	~ 4Å
Side chains	~ 3Å
Atomic details	~1–2Å

Table 1-1. **Table of biological structural features observable at different resolutions.** Visual resolution of a reconstruction can be determined by the observation of various structures common to biological samples

Viruses Shown to Infect Humans	**Size (Å)**
Adenoviridae	
Human Adenovirus Serotypes 1–47	700–900
Herpesviridae	
Herpes Simplex Virus Type 1 (HSV-1)	~ 1,500
Herpes Simplex Virus Type 2 (HSV-2)	
Varicella-Zostrer Virus	
Epstein-Barr Virus	
Cytomegalovirus (CMV)	
Human Herpesvirus 6 (Roseola Infantum)	
Human Herpesvirus 7	
Reoviridae	
Reovirus 1, 2, 3	600–800
Colorado Tick Fever Virus	
Rotavirus Groups A, B, C	

Table 1-2. **Table of viruses known to infect humans.** Viruses known to infect humans (Murray, Rosenthal et al. 2005) for which the correction of the curvature of the Ewald sphere will be required to derive atomic models by cryo-EM

Chapter 2

Peach: A simple Perl-based system for distributed computation and its application to cryo-EM data processing

Peter A. Leong[1#], J. Bernard Heymann[2#], and Grant J. Jensen[3*]

[1]Department of Applied Physics, California Institute of Technology, 1200 E. California Blvd., Pasadena, California 91125

[2]Laboratory of Structural Biology Research, National Institute of Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland 20892

[3]Division of Biology, California Institute of Technology, 1200 E. California Blvd., Pasadena, California 91125

[#] These authors contributed equally

* Corresponding author, email address **jensen@caltech.edu**, 626-395-8827

2.1 Summary

A simple distributed processing system named "Peach" was developed to meet the rising computational demands of modern structural biology (and other) laboratories without additional expense by using existing hardware resources more efficiently. A central server distributes jobs to idle workstations in such a way that each computer is used maximally, but without disturbing intermittent interactive users. As compared to other distributed systems, Peach is simple, easy to install, easy to administer, easy to use, scalable, and robust. While it was designed to queue and distribute large numbers of small tasks to participating computers, it can also be used to send single jobs automatically to the fastest currently available computer and/or survey the activity of an entire laboratory's computers. Tests of robustness and scalability are reported, as are three specific electron cryomicroscopy applications where Peach enabled projects that would not otherwise have been feasible without an expensive, dedicated cluster.

2.2 Introduction

The availability of ever-faster computers continues to open new possibilities throughout science and in structural biology in particular. This leads us to plan increasingly demanding projects and gather the computational resources needed. In many structural biology laboratories, the mixtures of heterogeneous workstations purchased individually or in small sets for laboratory personnel in recent years constitute a wealth of underutilized capacity. Here we report the development of a Perl-based package called "Peach" that efficiently distributes computational tasks across such workstations without disturbing interactive users.

The motivation for this work arose out of our own structural biological studies in electron cryomicroscopy (cryo-EM). Modern cryo-EM has three distinct modalities: (1) "two-dimensional crystallography", in which many crystals of a specimen only a single unit cell thick are imaged at various tilt angles with respect to the beam; (2) "single particle analysis," in which thousands of fields of randomly oriented particles are imaged in projection; and (3) "tomography," in which a single, unique object is imaged iteratively while being incrementally tilted about some axis. In each case, the resulting images are merged to produce a three-dimensional reconstruction of the specimen, and the process involves a large number of small, easily separable, independent calculations (for recent reviews and some descriptions of the computational challenges in this field, see ((Walz and Grigorieff 1998; van Heel, Gowen et al. 2000; Fernandez, Lawrence et al. 2002; Frank 2002; Sali, Glaeser et al. 2003; Frangakis and Forster 2004; Orlova and Saibil 2004; Subramaniam and Milne 2004)). Glaeser has presented a "straw man argument"

stating that solving the structure of a large protein complex by single particle analysis to near-atomic resolution with current algorithms would take even a state-of-the-art teraflop computer something like a year (Glaeser 1999). Even though we are still far away from this resolution goal for various reasons, in a typical cryo-EM laboratory today, computer power is already at a premium, and represents a real limitation. Researchers lose time logging in to multiple computers, manually distributing jobs across computers with different operating systems, generating custom scripts to submit jobs one after another through the night or weekend, watching for their completion, and coordinating computer usage with laboratory colleagues. Despite such efforts, most workstations are still only used to a small fraction of their capacity due to the difficulty of manually managing multiple tasks on multiple workstations.

To improve this situation, we searched for an inexpensive system to distribute jobs efficiently, easily, and securely across our set of workstations. Only a few options were available, including Open PBS from Veridian Systems and Condor (Tannenbaum and Litzkow 1995). While we have a running version of Open PBS on our Linux cluster, it has no "desktop harvesting", or in other words, it was not designed to take advantage of unused time on interactive workstations, as we desired. We downloaded and installed Condor, but disliked its complexity, as it required installation of separate executables for each platform, a large number of different types of daemons, over twenty-five different programs, and special submission description files. Further, source code was not available and the documentation warned of known security issues. Another well-known package is BOINC, the Berkeley Open Infrastructure for Network Computing, which

mediates the SETI@home project (Anderson, Cobb et al. 2002). BOINC was designed for "public-resource" (as opposed to in-house, or "grid") computing, in which participants are random individuals who donate time on their personal computers, connected to the internet via telephone or cable modems or DSL. While wonderful for certain applications, BOINC would not be attractive for structural biological applications because of the large amounts of data needing to be transferred, the need for accuracy (which in public-resource systems is achieved by redundant computing or some kind of post-verification), and the large amounts of memory often required.

Not finding a suitable alternative, we developed Peach, a simple Perl-based distributed computation system. The small number of scripts that constitute the system are easy to install and require no compilation. Peach is easy to use, easy to administer, free, robust, scalable, secure, and immediately compatible with almost any Unix operating system and non-interactive executable. We have installed it in two laboratories, where it has accelerated routine work and brought several structural biology projects to success that would not otherwise have been feasible without the purchase of an expensive, dedicated computer cluster.

2.3 Design

2.3.1 Design Philosophy — From the user's point of view, the goal was to develop a system that would accept anywhere from one to thousands of jobs and automatically process them as fast as possible using the existing workstations in the laboratory, but without disturbing interactive users. Two scenarios were envisioned: (1) when one or

more workstations were idle, in which case a new job would be sent immediately to the fastest one suitable, and (2) when all workstations were busy, in which case submitted jobs would be queued and distributed later. Further, the system needed to be simple to use and administer, scalable, secure, robust, and as compatible with the existing hardware and software in structural biology as possible.

2.3.2 Implementation — Peach was implemented following a client-server model, in which a single job "server" daemon runs on one workstation and maintains a queue of jobs to be done, while job "client" daemons run on all the other workstations, periodically reporting their state and running the jobs assigned to them. Three simple "access" clients constitute the complete user interface: (1) *psubmit*, which when given any executable file with flags and options as arguments, submits that job to the system; (2) *pview*, which generates reports on the status of the participating computers and submitted jobs; and (3) *pkill*, which terminates and/or removes jobs. Only clients initiate communication, so new clients can join and others terminate without disrupting the server.

2.3.3 Information Flow — Work begins when a user submits a job with the *psubmit* access client. *psubmit* writes an "execution script" on a shared disk which contains paths to an appropriate executable for each participating operating system. Next the *psubmit* client sends a message to the job server with the name of the execution script and the identity of the user, plus optional information about preferred processors and email addresses for reporting. The job server stores this information in memory and on disk. Meanwhile the job clients on all the participating workstations periodically report their

status to the job server. When the job server has a job in the queue and a suitable processor is reporting that it is idle, the server responds to the corresponding client with the name and path of the execution script. The job client, which is owned by root, forks a child process whose ownership is changed to the submitting user. Then the child process runs the execution script with "niced" priority (i.e., a low priority to allow other, interactive users better access) and writes the standard output and error files. After the job terminates, the job server sends an e-mail message to the user if requested. If during execution an interactive user begins to use the console, the job client immediately suspends active Peach jobs for a configurable time period. If that period will exceed the time the job had already been processing, the job client "releases" the job back to the job server for reassignment elsewhere. If a process fails (returns a non-zero value to the operating system), it is reassigned, but if it fails again, it is removed from the queue and the owner is notified.

2.3.4 The Job Server — The job server acts as a job broker, storing information about all the submitted jobs and processors participating in the system and matching them efficiently. The server also writes state and log files about transactions, job completions, and job failures. In the event the server crashes, it can be easily restarted (or even automatically restarted if desired) on any workstation, where it will read the state files and proceed without affecting current or waiting jobs. Clients automatically find the new IP address and port number of the server in the configuration file on the shared disk. The server has several built-in mechanisms to handle unexpected states appropriately.

2.3.5 The Job Clients — Each of the participating computers (regardless of the number of processors on the computer) has exactly one job client running at all times, which cycles automatically every few seconds to (1) monitor processor usage, (2) gather information about the status of current jobs, (3) suspend active jobs if a user begins interactive use at the console, (4) make decisions about whether to "release" suspended jobs back to the server, (5) report to the job server, and (6) launch newly assigned jobs from the server.

2.3.6 Use of Existing Capabilities — The system was written in Perl, which is installed by default on almost all Unix-variants. It makes use of only standard components available in recent distributions (Perl 5.8, March 2000, or later). This ensures cross-platform compatibility and ease of installation, since no compilation is required. For simplicity, data exchange across platforms is managed through existing TCP/IP and NFS services by mounting on all participating computers at least one shared disk where the Peach scripts, some configuration/state files, executables for each platform, and data are located. Additional shared data disks can be mounted on some or all of the participating computers (Figure 2-1). In this way large data files are not copied, even temporarily, to local disks, but rather are read from and written to a shared, central disk system. All messages are passed in standard XML format to increase compatibility with other software and in anticipation of future developments.

2.3.7 Security — Administrators and users are registered with password-protected accounts internal to Peach, allowing users access to all the participating computers without the requirement of accounts for all the users on all the computers. Each client

(such as *psubmit* or *pview*) requires a valid username and password. Messages carry unique signatures formed through a digest (a transformation of the text such that it cannot be decoded and read) of the username, the password, the message itself, and a unique authorization string provided by the job server. The recipient verifies that signature using it's knowledge of the unique string and its own database of registered users and passwords, preventing unauthorized messages. Finally, while Peach job clients are owned by root, the ownership of their child processes which actually launch all the jobs are changed to the submitter, and no jobs are allowed to run as root. Thus damage from poorly designed or malicious jobs is limited to the submitting account.

2.3.8 Peach Administration — A primary design goal of Peach was that it be simple, both for users and for the administrator. To install Peach, a simple script copies all the required files (seven executable Perl scripts and seven supporting files) to a program directory on a shared disk that must be available to all participating computers. No programs have to be recompiled: any existing Unix command, script, or program can be immediately submitted as a job. The only requirement is that these programs are available, either as common utilities on all computers, or more typically, as executables on a shared disk. During setup, the job server and job client daemons must be launched and user accounts with names and passwords must be established. New shared disks and workstations can be added easily. The appropriate configuration files are generated during installation, but various parameters can be specially configured if desired. Typical Unix conventions for file locations and configuration have been followed as far as possible to facilitate administration.

Peach was designed to have robust, independent modules. Thus if the job server dies, it can be restarted on any other machine without disruption to current or waiting jobs. If a job client dies (for instance if a workstation is rebooted), there is no impact on any other client, and a new job client can be re-started and join the system at any time. If network delays slow communication, or a job client stops reporting for any reason, Peach self-recovers as soon as conditions improve. Peach does depend on a commonly shared disk for access to programs and data, however, which is a limitation we accepted to keep the system simple and avoid the complexities of copying large amounts of data across a network.

2.4 Tests and Results

2.4.1 Installation and Test Environments — Peach has been installed and tested now in two separate laboratories at the California Institute of Technology (Caltech) and the National Institutes of Health (NIH). At Caltech it was developed on an existing heterogeneous set of 17 computers including four Macintosh dual-G5s (2.0 GHz, 2.5 GB RAM), 12 PCs running Linux (2.2–2.4 GHz, 1.0–4.0 GB RAM, 1–4 processors each) and 1 SGI Fuel (0.6 GHz, 2.0 GB RAM). At the NIH, Peach distributes jobs to 11 heterogeneous computers, including 6 Macintosh dual-G5s (2.0 GHz, 5 GB RAM), 2 dual-MIPSpro SGI Octanes (0.25 GHz, 1 GB RAM) and 3 HP Alphas (0.7 GHz, 1.5–5 GB RAM, 1–4 processors). In both laboratories, a central shared disk was available to all the participating workstations, but various additional shared "data" disks came on- and off-line during the testing period. The local networks supported 1 Gbit/s Ethernet for

communication and typically performed at 5–10% of nominal capacity. The Bsoft package (Heymann 2001) used for image processing was installed on the central shared disk with compiled versions for each operating system located in different directories. Peach was developed in the short span of a few months at Caltech within a network and computer setup configured for its use. The installation at the NIH, however, represented a useful test of how readily usable Peach would be by other groups whose hardware was not set up specifically for it. The main hurdles at the NIH were to arrange for a central disk that all the computers could access and to make all the users' individual and group identification numbers consistent across the set of participating computers. Further configuration entailed compiling all the required executables for image processing for the different platforms and installing those on the shared disk. After that, Peach was installed and configured in less than an hour.

2.4.2 Cryo-EM Applications — Peach has now been used for several of our electron cryomicroscopy projects. Three examples will be described. We have recently explored the potential benefit of cooling frozen-hydrated samples with liquid helium instead of liquid nitrogen in the context of electron cryotomography. In one test we recorded full tilt series of fields of a purified protein complex, the molluscan hemocyanin from Megathura crenulata, with total doses ranging from 10 to 300 electrons/Å^2, at each of the two temperatures. From each tilt series a three-dimensional reconstruction ("tomogram") of the field of particles was calculated, and individual hemocyanin molecules were manually identified. To measure the overall quality of the tomograms at each dose and temperature, approximately 100 hemocyanin molecules were aligned to the known 12 Å

structure (Mouche, Zhu et al. 2003) using the program bfind (Bsoft) (Figure 2-2). Thus a three-dimensional translation and orientation search was performed for ~ 1,400 cube-shaped volumes of 64^3 voxels each.

In a second, related example, we recorded multiple, iterative images of fields of frozen hemocyanin particles using 10 electrons/$Å^2$ for each image. As more and more images were recorded, the structure of the particles degraded due to radiation damage. We measured the rate of degradation by picking 100 hemocyanin particles out of each image in the series and using them to calculate a three-dimensional reconstruction, which was compared to the known higher-resolution structure. By recording such "dose series" of many fields cooled by either liquid nitrogen or liquid helium and plotting the resolution of the resulting reconstructions as a function of dose, we were able to test whether deeper cooling with liquid helium delayed radiation damage as hoped (data not shown). We used Peach throughout this project to manage the literally hundreds of "single-particle" reconstructions involved. During a 23-day period a total of 2,146 jobs related to these hemocyanin projects were run on Peach, using 322 days of CPU time. This accounts for approximately 80% of the capacity of our 17 workstations during those days, all obtained without disturbance to the intermittent interactive users.

As a third example, we have simulated images of protein complexes embedded in vitreous ice under different imaging conditions using the so-called "multi-slice" algorithm (Cowley and Moodie 1957). Three-dimensional reconstructions were calculated with various alignment errors to explore their effect on resolution. The most

computationally intensive part of this work is the atom-by-atom calculation of the atomic potential of each simulated cube of water and protein. In one recent batch of simulations, we used Peach to manage the calculation of 1947 images over a period of 3.6 days, logging 96 days of actual CPU time (Figure 2-3).

2.4.3 Robustness — The most common computer failures in our experience are stalled computers, disk problems, and network delays. Peach was designed to be as tolerant of these disruptions as practically possible, and several robustness tests were performed. In the first test, the job server daemon was terminated while managing a long queue of active and waiting jobs, as would happen, for instance, if the workstation hosting the job server hung or had to be rebooted. Active jobs continued without disturbance and began completing successfully. After two hours, the job server was restarted on its original host computer, and all the job clients re-initiated communications and began reporting and/or receiving new jobs as normal. In the second test, the job server daemon was again terminated while managing a long queue, but this time it was restarted on a different host computer. Again, no delays or complications were experienced, as the existing job clients and future access clients all found the new IP address and port number of the server from the configuration file and proceeded as normal. For the third test, a job client daemon was terminated. As expected, it was first listed by *pview* as missing, and then after one hour it was removed from the list of job clients and the jobs that had been assigned to it were re-queued and later distributed to other machines.

Without specific tests, we have observed the behavior of Peach under other challenging conditions. During periods of network delays, job clients were unable to report punctually to the server. This had little consequence, however, since active jobs continued running and only the brief breaks between jobs were extended. Of course data transfer to and from the shared disk was also delayed by network slowdowns, so network reliability and speed are areas for improvement. In one instance the central shared disk was inaccessible for several minutes, but normal communication and file transfer resumed once the disk became accessible again.

2.4.4 Scalability — It is important that distributed computation systems such as Peach maintain efficiency if more processors are added. Because Peach only distributes completely independent jobs, rather than interdependent parts of single jobs, the main bottleneck that arises when more processors are introduced is the response of the job server to each job client's report. Bottlenecks can also arise in accessing shared disks, but there is no explicit limit to the number of shared data disks that can be added to the system. While only one job client was intended to ever be running on any given computer, in order to explore Peach's scalability with our present hardware, we ran tests in which progressively larger numbers of job clients were added to the system by simply launching additional job clients on one of six chosen workstations at the rate of one additional client per minute. The corresponding server response times are plotted in Figure 2-4 for various settings of the job client reporting interval (the configurable time between when a job client receives a response from the job server and when that job client initiates its next report). For each reporting interval, the plot shows three distinct

regions. Initially, the job server is unsaturated and responds immediately to all job clients. As the number of reporting clients increases, eventually the socket queues begin to fill, and the response time increases linearly. Because the server can no longer respond to the job clients' reports as fast as they come, one might expect the socket queues and therefore the response time to lengthen steadily, even in between the additions of new clients. What was observed, however, was that a new, stable response time was reached after each additional client entered the system. This happened because job clients do not initiate a new report until *after* they receive a response. Thus new reports replace old ones on the socket queue only as fast as the old ones are served with a response. This equilibrium becomes impossible, however, in the third region, after so many job clients are added that the number of reports waiting in the queue exceeds the number of connections available (a parameter set in the operating system kernel), and reports start to be refused. Thus with our current configuration of hardware and the default five-second job client reporting interval, Peach's job server can manage up to approximately 200 participating computers reliably. Arbitrarily larger clusters can be serviced simply by increasing the reporting interval appropriately in the main Peach configuration file.

2.5 Discussion

Among large computational tasks in structural biology (as well as all science), some are not easily separated into small, independent tasks. Instead, these require intensive communication between processes and rely on large, homogeneous clusters (so-called "supercomputers") that optimize inter-node communication speeds. There are also, however, a vast number of tasks which are trivially parallelizable. This is especially true

in our field of electron cryomicroscopy, where the large number of individual images in almost every project leads naturally to easy separation. Here we have described and demonstrated a simple Perl-based distributed computation system called Peach, designed to distribute large numbers of independent jobs across the kinds of heterogeneous computer clusters commonly found in structural biology laboratories.

Here at Caltech, we have at present roughly twenty workstations scattered throughout the laboratory for interactive use. When fully loaded with jobs during normal weekdays, we found that Peach was able to use on average 69% of the capacity of these personal workstations, without disturbing interactive users. Whenever someone began using the console, even for undemanding applications such as word processing, Peach immediately suspended its jobs until the computer was once again idle. If a Peach application consumes all a client's memory, or worse causes major swapping, an annoying delay could be experienced as it is moved to the background. While we have not yet encountered this problem, we expect it would be similar to the delay caused by a complex, memory-intensive screen saver. The fact that Peach still took advantage of over two-thirds of the workstations' total potential is easily rationalized by recognizing that a regular "full-time" job accounts for only about one-fifth of the hours of a year, and further considering that the average researcher spends a great deal of time away from his/her desk even during workdays. In addition to the personal workstations, we also have some processors assembled as "compute clusters" with no monitor. Peach used these simultaneously with the personal workstations to 99% efficiency, demonstrating its ability to pool the power of such dedicated machines with the others in the laboratory.

By facilitating the use of all the available computer power, Peach has allowed us to finish projects that would otherwise have required expensive new hardware. In addition, Peach has accelerated our routine work and distributed resources more equitably by running each job on the fastest available processor, regardless of whose desk it is sitting on.

Peach is distinguished from other distributed systems by its simplicity and ease of use. There are only three user commands: one to submit jobs, one to monitor the status of jobs and processors, and one to kill jobs. A job is submitted simply by listing it, along with necessary flags and options, as arguments to *psubmit*. Peach is immediately compatible with any non-interactive command-line executable including scripts and, notably, commands in all the commonly used cryo-EM image processing packages. Installation is accomplished by running a single script which copies Peach onto a shared disk, launching the server and client daemons, and registering the users. No compilations or special libraries are required, and Peach will run on any Unix machine with a recent (less than five years old) version of Perl. Because Peach uses the modular client-server approach, it is robust to most common computer failures including loss of any of the processes, loss of any of the workstations, and delays in network communications. It remains sensitive, however, to failures of the shared disks, so choosing a reliable disk server is important.

Access to the Peach system is controlled by registration and passwords. To avoid interception, passwords are never sent in a clear text form. User registration also allows Peach to run jobs on computers without the need for user accounts on those machines, as long as the shared disk is mounted and the user has permission to read and write to the

shared disk. We have not discovered any security loopholes thus far, and believe that the code's shortness and simplicity reduce vulnerability as compared to other existing packages.

We anticipate that Peach will be used on large clusters of computers. To assess its ability to serve such large clusters, we ran simulations where hundreds of job clients were launched. These demonstrated that up to a thousand computers can be handled well by a single job server through the adjustment of one parameter, the job client reporting interval. Thus Peach can handle even the largest modern clusters. Faster computers in the future will increase the capacity of the server, and configurations with multiple servers could be used to further extend the scale, if ever necessary.

One of the design goals was that Peach be immediately compatible with the hardware and software resources of typical structural biology laboratories. While Peach does work with any command-line Unix executable, the GUIs and command-line interpreters present in many packages would have to be adapted to take advantage of Peach's distributing potential. Among the most common packages used for cryo-EM-based single particle analysis are, for example, Spider, Imagic, and EMAN (Frank, Radermacher et al. 1996; van Heel, Harauz et al. 1996; Ludtke, Baldwin et al. 1999). Spider batch jobs, which are launched from the command line, could be distributed as a single job by Peach, which would help in a situation where multiple batch jobs were being submitted simultaneously within a laboratory. In particular, Peach would make it easy to send jobs away from the computer being used to submit the job, preventing slow-

downs. Similarly, Imagic's batch accumulation mode assembles a c-shell script which could be distributed by Peach, as could EMAN script files or individual EMAN programs. The command-line interpreters and GUIs asociated with these packages, however, would have to be modified before the jobs they launched could be managed by Peach. Spider, Imagic, and EMAN already provide powerful built-in capacities to exploit homogeneous clusters. Peach's ability to distribute jobs across heterogeneous clusters should be viewed as complementary. The ideal system would efficiently access all resources (homogeneous and heterogeneous clusters) through all interfaces (command-line executables, command-line interpreters, and GUIs). As long as computational tasks did not require inter-process communication, but instead could be broken down into a large number of separate small processes, the principles we used to develop Peach could be used to achieve this. Command-line interpreters and GUIs would have to be modified to submit jobs to a Peach-like system, and Peach would have to be modified to parse large scripts defining entire image processing pipelines and launch jobs sequentially or in parallel, as appropriate. The Peach package is freely available at http://www.jensenlab.caltech.edu, or upon request to the authors.

2.6 Acknowledgements

We thank C. Iancu for her willingness to test and use Peach during development stages; P. Ober for the early development of ideas for distributed processing; and W. Tivol, S. Tivol, and D. Morris for reviewing the manuscript. This work was supported in part by NIH Grant PO1 GM66521 to GJJ, DOE grant DE-FG02-04ER63785 to GJJ, the

Beckman Institute at Caltech, and gifts from the Ralph M. Parsons Foundation, the Agouron Institute, and the Gordon and Betty Moore Foundation to Caltech.

2.7 References

Anderson, D. P., J. Cobb, et al. (2002). "SETI@home - An experiment in public-resource computing." Communications of the Acm **45**(11): 56−61.

Cowley, J. M. and A. F. Moodie (1957). "The Scattering of Electrons by Atoms and Crystals .1. a New Theoretical Approach." Acta Crystallographica **10**(10): 609−619.

Cowley, J. M. and A. F. Moodie (1957). "The Scattering of Electrons by Atoms and Crystals. I. A New Theoretical Approach." Acta Cryst. **10**: 609-619.

Fernandez, J. J., A. F. Lawrence, et al. (2002). "High-performance electron tomography of complex biological specimens." Journal of Structural Biology **138**(1−2): 6−20.

Frangakis, A. S. and F. Forster (2004). "Computational exploration of structural information from cryo-electron tomograms." Current Opinion in Structural Biology **14**(3): 325−331.

Frank, J. (2002). "Single-particle imaging of macromolecules by cryo-electron microscopy." Annual Review of Biophysics and Biomolecular Structure **31**: 303−319.

Frank, J., M. Radermacher, et al. (1996). "SPIDER and WEB: Processing and visualization of images in 3D electron microscopy and related fields." Journal of Structural Biology **116**(1): 190−199.

Glaeser, R. M. (1999). "Review: Electron crystallography: Present excitement, a nod to the past, anticipating the future." Journal of Structural Biology **128**(1): 3−14.

Heymann, J. B. (2001). "Bsoft: image and molecular processing in electron microscopy." J Struct Biol **133**(2-3): 156-69.

Lowe, J., D. Stock, et al. (1995). "Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution." Science **268**(5210): 533-9.

Ludtke, S. J., P. R. Baldwin, et al. (1999). "EMAN: Semiautomated software for high-resolution single-particle reconstructions." Journal of Structural Biology **128**(1): 82−97.

Mouche, F., Y. Zhu, et al. (2003). "Automated three-dimensional reconstruction of keyhole limpet hemocyanin type 1." J Struct Biol **144**(3): 301-12.

Orlova, E. V. and H. R. Saibil (2004). "Structure determination of macromolecular assemblies by single-particle analysis of cryo-electron micrographs." Current Opinion in Structural Biology **14**(5): 584−590.

Sali, A., R. Glaeser, et al. (2003). "From words to literature in structural proteomics." Nature **422**(6928): 216−225.

Subramaniam, S. and J. L. S. Milne (2004). "Three-dimensional electron microscopy at molecular resolution." Annual Review of Biophysics and Biomolecular Structure **33**: 141−155.

Tannenbaum, T. and M. Litzkow (1995). "The Condor Distributed-Processing System." Dr Dobbs Journal **20**(2): 40−48.

van Heel, M., B. Gowen, et al. (2000). "Single-particle electron cryo-microscopy: towards atomic resolution." Quarterly Reviews of Biophysics **33**(4): 307−369.

van Heel, M., G. Harauz, et al. (1996). "A new generation of the IMAGIC image processing system." Journal of Structural Biology **116**(1): 17−24.

Walz, T. and N. Grigorieff (1998). "Electron crystallography of two-dimensional crystals of membrane proteins." Journal of Structural Biology **121**(2): 142−161.

2.8 Figures

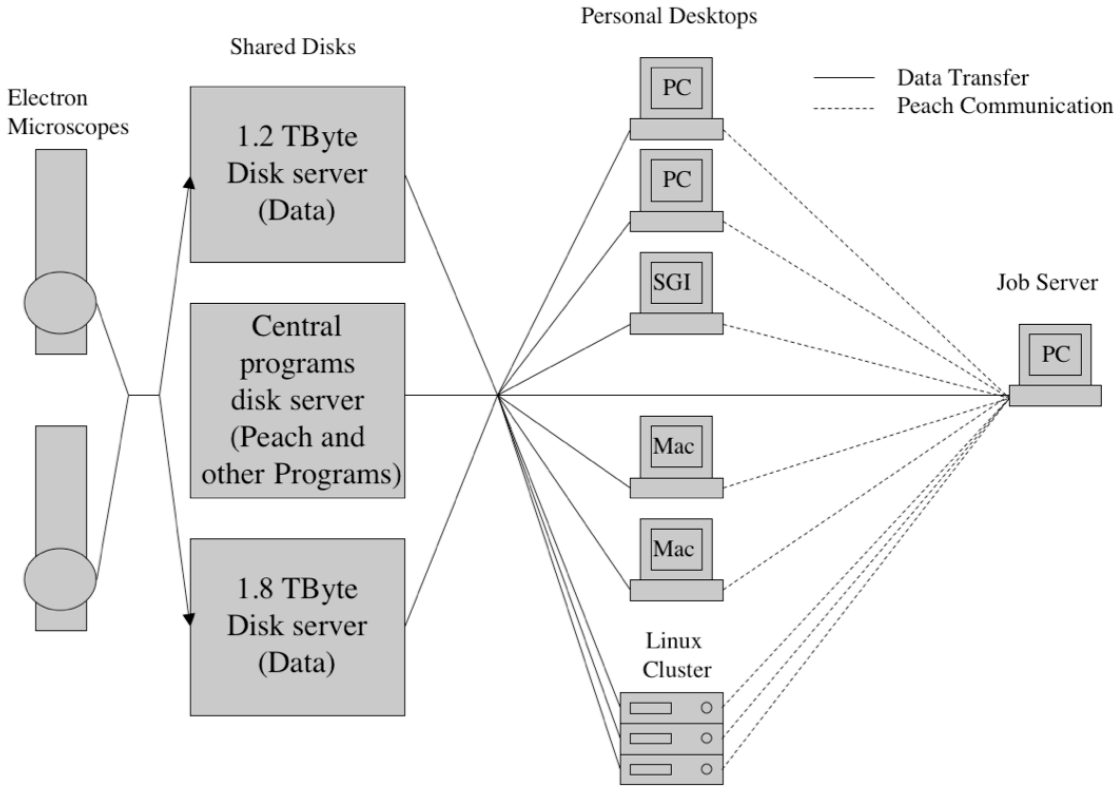


Figure 2-1. Schematic drawing of the setup and information flow in the testing of Peach. Image data was collected on two electron microscopes and transferred to two shared data disks. All the personal workstations located on desks throughout the laboratory and the several processors of a monitor-less compute cluster were configured to mount a central shared programs disk and the two data disks. Any particular workstation could host the job server. Users submitted jobs to Peach from their personal workstations. Information about each job was passed to the job server, which distributed jobs to idle workstations. Workstations retrieved job data from and wrote results to the shared disks. Solid lines represent job data transfer and dotted lines represent Peach network messages

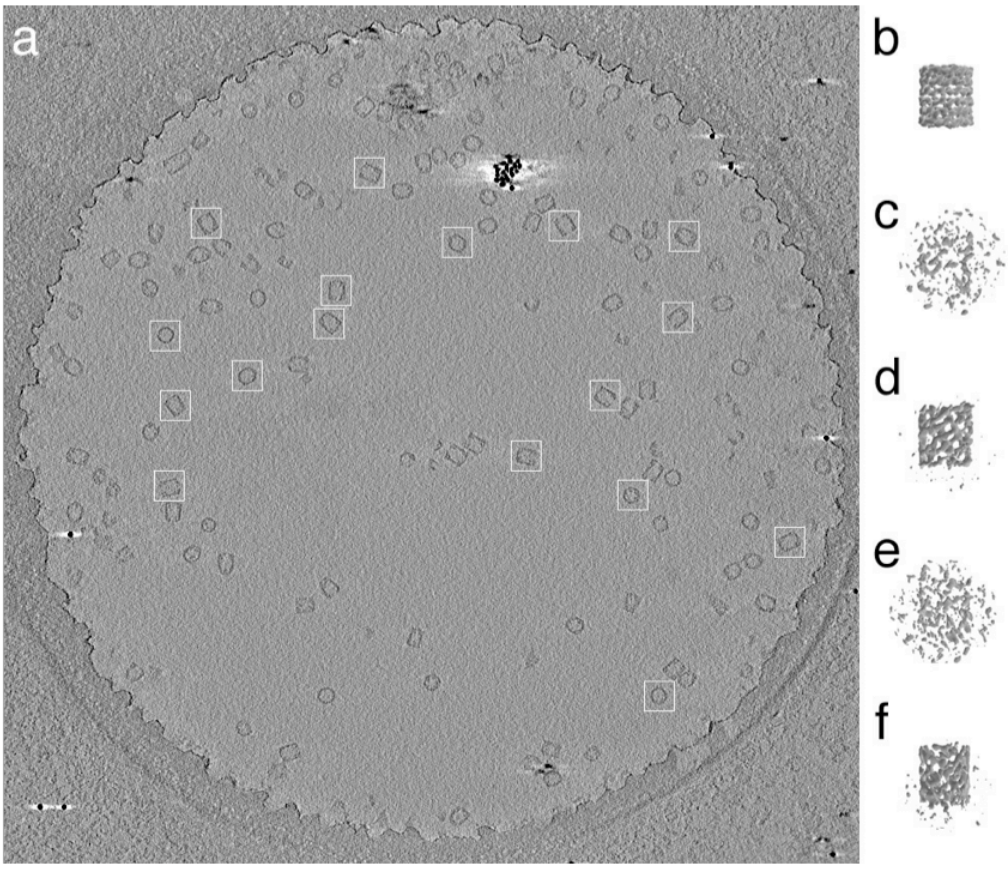


Figure 2-2. An example cryoEM image processing project made feasible by Peach. Peach managed extensive calculations comparing electron tomograms recorded with different electron doses and different sample temperatures. The sample was the 35 nm long, barrel-shaped protein complex hemocyanin, purified and suspended within a thin film of vitreous ice across circular holes in a supporting carbon film. (a) A single section through a tomogram, where several individual hemocyanin molecules are marked with square boxes. The small black dots are colloidal gold fiducial markers. (b) 12 Å structure of hemocyanin (Mouche, Zhu et al. 2003) used as template. (c–f) Representative three-dimensional reconstructions of individual hemocyanin molecules, extracted from tomograms recorded at liquid nitrogen (c,d) or helium (e,f) temperature, with doses of 10 (c,e) or 120 (d,f) electrons/$Å^2$, oriented using the template in (b)

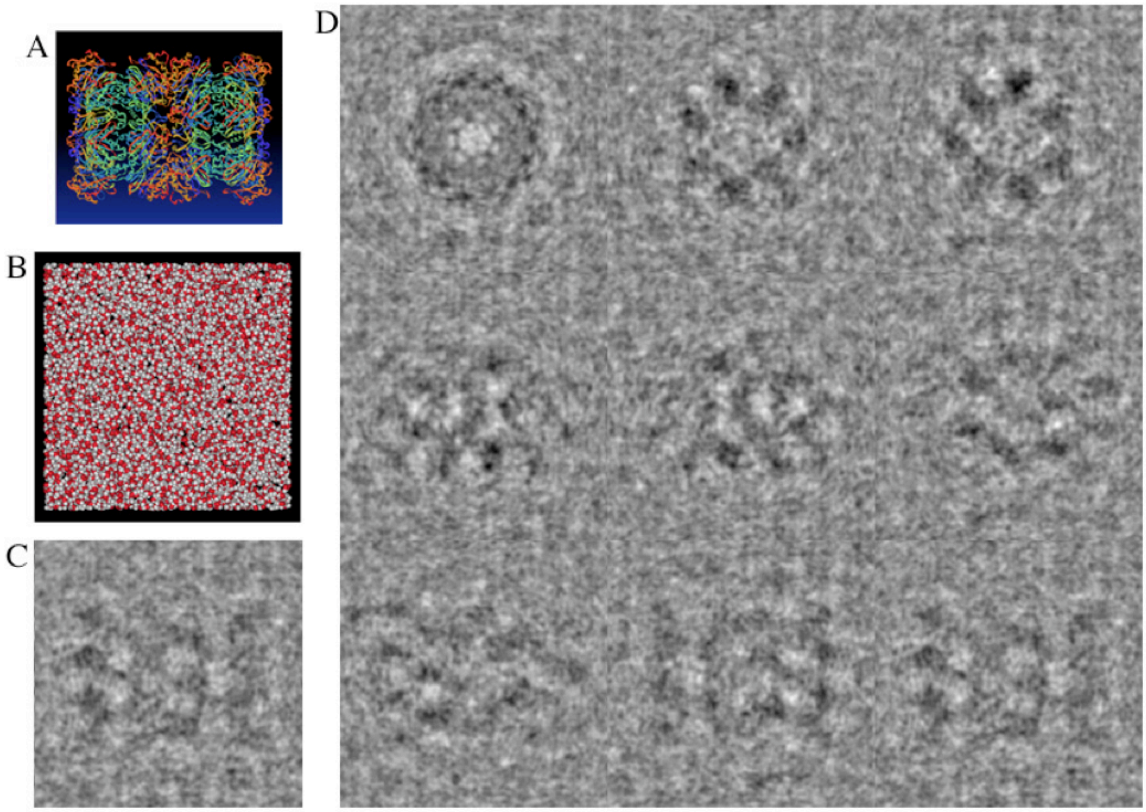


Figure 2-3. An example image simulation project managed by Peach. Peach was used to simulate thousands of cryo-EM images of a water-embedded protein from different points of view and under different imaging conditions using a multi-slice algorithm (Cowley and Moodie 1957). (a) A ribbon diagram of the test protein, the 20S proteasome (Lowe, Stock et al. 1995). (b) Block of water used to embed the test protein. (c) Simulated cryo-EM image of the 20S protein embedded in water from the same point of view as in (a). (d) Montage of nine other simulated images, showing the 20S protein from various points of view

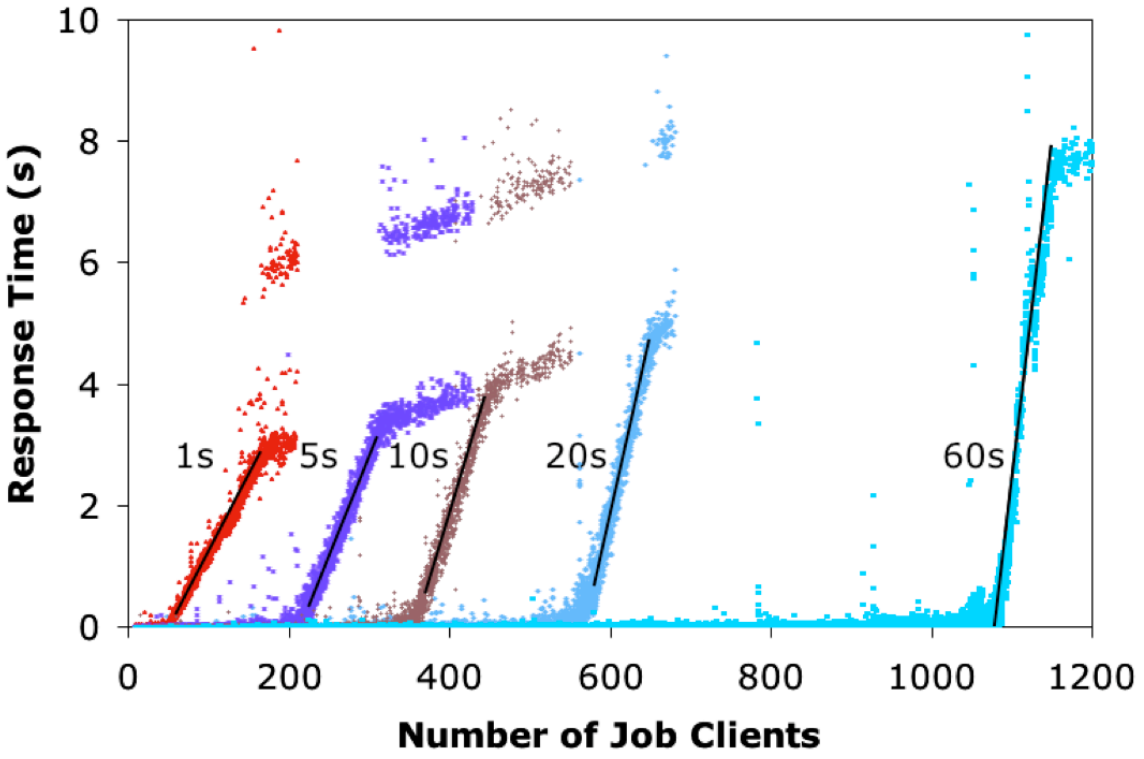


Figure 2-4. Scalability. The ability of Peach to manage large numbers of computers was tested by adding job clients to the system incrementally while measuring the delay between job client reports and the job server's response. The results from five separate tests are shown, in which the job client reporting interval (the time each job client waited before sending its next report) was set to 1, 5, 10, 20, and 60 s. Each graph shows three distinct regions. In the first region, the job server is unsaturated and responds to job clients immediately. As additional job clients are added, the server eventually becomes saturated, socket queues begin to fill, and the response time increases linearly. Finally, socket queues also become saturated and some connections are refused, generating erratic response times. For these tests, the job server was a 2.4 GHz IBM PC with 1.5 gigabytes of memory running Redhat Linux

Chapter 3

Prec: an iterative reconstruction method for correction of the Ewald Sphere

Peter A. Leong[a], Xuekui Yu[b], Z. Hong Zhou[b], Grant J. Jensen[c*]

[a]Department of Applied Physics, California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125, USA

[b]Department of Microbiology, Immunology & Molecular Genetics and The California NanoSystems Institute, 615 Charles E. Young Dr. S, BSRB 237; University of California Los Angeles, Los Angeles, CA 90095-7364, USA

[c]Division of Biology, California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125, USA

*To whom correspondence should be addressed: **jensen@caltech.edu**, 626-395-8827 (phone) 626-395-5730 (fax)

3.1 Abstract

To avoid the challenges of crystallization and the size limitations of NMR, it has long been hoped that single-particle cryo-electron microscopy (cryo-EM) would eventually yield atomically interpretable reconstructions. For the most favorable class of specimens (large icosahedral viruses), one of the key obstacles is curvature of the Ewald sphere, which leads to a breakdown of the projection theorem used by conventional 3D reconstruction programs. Here an iterative refinement reconstruction algorithm, *Prec*, is described that overcomes this limitation by averaging information from images recorded from different points of view, as are present in typical micrographs. *Prec* was implemented in the popular software packages IMIRS, EMAN, and Bsoft. In preliminary tests with both simple and multi-slice simulated images, *Prec* overcame the curvature problem even in the presence of noise. *Prec* was then used to refine the three recently published, ~ 4 Å resolution, icosahedral virus reconstructions from experimental cryo-EM images, but unfortunately no significant improvements in resolution were realized. Further simulations showed that limitations other than the Ewald sphere curvature problem must still be dominant in these experimental studies.

3.2 Introduction

X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) were the first techniques to reveal the atomic structures of biological macromolecules. Electron crystallography then followed, first on "two-dimensional" crystals (crystals one unit cell thick) (Henderson, Baldwin et al. 1990; Kuhlbrandt, Wang et al. 1994) and then on helical (tubular) crystals (Unwin 2005). To avoid the challenges of crystallization and the size limitations of NMR, it has long been hoped that single-particle cryo-electron microscopy (cryo-EM) would eventually also produce atomically interpretable maps. Steady progress towards this goal has been made (Frank 2002), led by reconstructions of large icosahedral viruses, whose 60-fold symmetry, size, and rigid architecture all facilitate precise image alignment. In just the past few months the structures of three such viruses — cytoplasmic polyhedrosis virus (CPV) (Yu, Jin et al. 2008), epsilon15 virus (ε15) (Jiang, Baker et al. 2008), and rotavirus (DLP) (Zhang, Settembre et al. 2008) — have been reconstructed to ~4 Å.

Previous analyses (DeRosier 2000; Jensen and Kornberg 2000; Zhang, Settembre et al. 2008) suggest that curvature of the Ewald sphere (or equivalently, the depth of field (Zhou and Chiu 1993)) may have been one of the principal resolution limitations in these recent studies. Conventional methods assume that EM images are true projections, but in fact they are not: the information delivered by the microscope is actually a mixture of information belonging to a curved surface within the three-dimensional (3D) Fourier transform of the specimen called the Ewald sphere. The mixing occurs when the complex electron wave functions are measured by the CCD or film to produce real

images. The severity of the problem increases with specimen thickness, resolution, and electron wavelength.

A method for recovering the full, complex electron wavefunction from focal series was proposed by Schiske in 1968 (Schiske 1968). Further discussion then followed through 1990, when the method was re-proposed using a different, more intuitive approach (Van Dyck and Op de Beeck 1990). Saxton, who referred to this class of approaches as the paraboloid method (PM), later showed it to be equivalent to the original (Saxton 1994). More recently, the problem was discussed in the context of 3D reconstruction by DeRosier, who outlined four basic strategies to recover all the unique Fourier coefficients by merging focal pairs, images at different tilt angles, or images of ordered (crystalline or helical) objects in reciprocal space (DeRosier 2000). A different idea for addressing the problem in real space was proposed by Jensen and Kornberg (Jensen and Kornberg 2000), followed by additional analyses and suggestions by Wan et al. (Wan, Chiu et al. 2004).

Unfortunately, none of these efforts produced an efficient, practical program ready for use within the software packages available for the calculation of high-resolution reconstructions of viruses from experimental images. Here we describe a version of the PM we call *Prec* (for paraboloid reconstruction), which iteratively retrieves the information lost due to curvature of the Ewald sphere, and its implementation into three modern software packages. First, two Cartesian-coordinate-based versions of *Prec* were implemented in Bsoft (Heymann 2001) and EMAN (Ludtke, Baldwin et al. 1999) to

facilitate development and testing. Next a cylindrical-coordinate-based version was implemented in IMIRS (Liang, Ke et al. 2002), a commonly used software package for high-resolution icosahedral reconstructions which exploits the advantages of cylindrical coordinates and Fourier-Bessel transforms (Klug, Crick et al. 1958). Using simulated images, we show that all three implementations relieve the resolution limitations of the Ewald sphere, but surprisingly do not substantially improve the resolution of the three recent near-atomic-resolution reconstructions from experimental cryo-EM images. We conclude that other factors (besides the curvature problem) are still principally limiting. During the course of this effort, Wolf et al. (Wolf, DeRosier et al. 2006) implemented a comparable version of the PM in the also popular, Cartesian-coordinate-based FREALIGN package (Grigorieff 2007) and tested its efficacy on simulated images. Differences in the algorithms and performance of the *Prec* and FREALIGN implementations are discussed.

3.3 Results

3.3.1 The Ewald Curvature Problem and Symbols Used — To introduce needed symbols, we will follow DeRosier's derivation of the effects of the Ewald sphere curvature closely (DeRosier 2000), except that here all Fourier coefficients F are complex and amplitude contrast is included explicitly. Beginning first with the effect of a sample on an incident electron wave and its weak-phase approximation,

$$\frac{A_t(x)}{A_0} = e^{-(\alpha+i\beta)\rho(x)} \approx 1-(\alpha+i\beta)\rho(x) \qquad (1)$$

where $A_t(x)$ is the transmitted wave, A_0 is the incoming wave, α is the amplitude contrast value, $\beta = \sqrt{1-\alpha^2}$ is the phase contrast value (Erickson and Klug 1971), $\rho(x)$ is the density of the sample, and i is an imaginary number with magnitude 1; the diffracted wave $F(X)$ takes the form

$$F(X) = FT[1-(\alpha+i\beta)\rho(x)] = \delta(X) - (\alpha+i\beta)F_\rho(X) \tag{2}$$

where $F_\rho(X)$ is the Fourier transform (FT) of our sample density.

Considering the sum of a single, symmetric pair of diffracted beams represented by Fourier coefficients F_L and F_R on an Ewald sphere (Figure 3-1), whose additional path length through the lens with respect to the unscattered beam adds an additional phase shift of $e^{i\chi}$, we have:

$$F(X) = \delta(X) - (\alpha+i\beta)F_L e^{i\chi}\delta(X+X_a) - (\alpha+i\beta)F_R e^{i\chi}\delta(X-X_a) \tag{3}$$

where χ is the wave aberration function at X_a and is defined as

$$\chi(s) = \tfrac{\pi}{2}C_s\lambda^3 s^4 - \pi\Delta f\,\lambda s^2 \tag{4}$$

in which λ is the electron wavelength, s is the spatial frequency, C_s is the spherical aberration coefficient, and Δf is the defocus.

The interference of these beams will produce a single complex fringe with a periodicity of $\frac{1}{X_a}$ whose amplitude, $\sigma(x)$, will be

$$\sigma(x) = FT^{-1}[F(X)] = 1 - (\alpha + i\beta)F_L e^{i\chi} e^{-2\pi i x X_a} - (\alpha + i\beta)F_R e^{i\chi} e^{2\pi i x X_a} \tag{5}$$

The intensity of the wave is recorded as our image

$$\begin{aligned} \left|\sigma(x)\right|^2 \approx 1 &- [(\alpha + i\beta)F_L e^{i\chi} + (\alpha - i\beta)F_R^* e^{-i\chi}]e^{-2\pi i x X_a} \\ &- [(\alpha + i\beta)F_R e^{i\chi} + (\alpha - i\beta)F_L^* e^{-i\chi}]e^{2\pi i x X_a} \end{aligned} \tag{6}$$

where the F^2 terms can be ignored due to the weak phase approximation.

The FT of our image $F_{obs}(X)$ is then

$$\begin{aligned} F_{obs}(X) = \delta(X) &- [(\alpha + i\beta)F_L e^{i\chi} + (\alpha - i\beta)F_R^* e^{-i\chi}]\delta(X + X_a) \\ &- [(\alpha - i\beta)F_L^* e^{-i\chi} + (\alpha + i\beta)F_R e^{i\chi}]\delta(X - X_a) \end{aligned} \tag{7}$$

We see that $F_{R_{obs}}$, the observed Fourier value on the right side at $X = X_a$, is

$$F_{R_{obs}} = -F_L^*(\alpha - i\beta)e^{-i\chi} - F_R(\alpha + i\beta)e^{i\chi} \tag{8}$$

Because of the curvature of the Ewald sphere, F_L and F_R are not a Friedel pair (i.e., not complex conjugates), but rather independent Fourier coefficients, mixed by the process of

image formation. Thus conventional methods, which treat $F_{R_{obs}}$ as if it were the sum of a Friedel pair F_L and F_R, do progressively worse as F_L and F_R diverge at higher resolutions.

3.3.2 The Paraboloid Method in the Context of 3D Reconstruction — The original Fourier coefficients can be recovered by averaging information from multiple images, which each contain different combinations of the unique coefficients. First, images are corrected for the contrast transfer function (CTF). This is performed by multiplying each term F_{obs} by $-(\alpha - i\beta)e^{-i\chi}$. Unlike conventional CTF corrections, where values around CTF zeros are discarded, here there is no such requirement, since this "complex" CTF-correction (cCTF) is a multiplication by a factor of magnitude 1 rather than a division by a number potentially close to zero. Thus $F_{R_{corr}}$, the cCTF-corrected coefficient on the right side, is

$$F_{R_{corr}} = -F_{R_{obs}}(\alpha - i\beta)e^{-i\chi} = F_R + F_L^*(\alpha - i\beta)^2 e^{-i2\chi} \tag{9}$$

Because each $F_{R_{corr}}$ is the sum of the correct F_R and a phase-shifted, complex-conjugated F_L, at this point it becomes clear how by averaging $F_{R_{corr}}$ from a number of different images, each measuring the same F_R but different F_Ls, the F_Rs will add coherently but the sum of F_Ls will diminish in comparison. At low resolution, however, where $F_L^* \approx F_R$,

$$F_{R_{obs}} \approx -F_R(\alpha - i\beta)e^{-i\chi} - F_R(\alpha + i\beta)e^{i\chi} = -2F_R(\alpha\cos\chi - \beta\sin\chi) \tag{10}$$

The cCTF correction then leads to wrong values

$$F_{R_{corr}} = F_R + F_R(\alpha - i\beta)^2 e^{-i2\chi} \tag{11}$$

since χ does not vary quickly, causing the second terms to also add coherently and introduce a significant error. Thus at low resolution, it is better to use the simpler, real CTF correction (rCTF), where F_{obs} is divided by the factor $-2(\alpha\cos\chi - \beta\sin\chi)$. A practical transition point can be found as the spatial frequency at which the cCTF-corrected and the rCTF-corrected reconstructions match best (as demonstrated in the CPV reconstruction below).

After CTF-correcting the raw images, the often described paraboloid method (PM) places the F_{corr} values in their correct position in Fourier space on the Ewald sphere:

$$F_{R_{PM}} = \tfrac{1}{N}\sum_k^N F_{R^k_{corr}} = \tfrac{1}{N}\sum_k^N F_{R_k} + \tfrac{1}{N}\sum_k^N F^*_{L_k}(\alpha - i\beta)^2 e^{-i2\chi_k} \tag{12}$$

$$F_{L_{PM}} = \tfrac{1}{N}\sum_k^N F_{L^k_{corr}} = \tfrac{1}{N}\sum_k^N F_{L_k} + \tfrac{1}{N}\sum_k^N F^*_{R_k}(\alpha - i\beta)^2 e^{-i2\chi_k} \tag{13}$$

where N is the total number of images (indexed by k) which contribute to each point.

3.3.3 The Prec *Algorithm* — In essence, the PM therefore "splits" the observed values F_{obs} into estimates of F_R and F_L by averaging information from a set of images. Once initial estimates are obtained, they can be refined through iteration, since knowledge of any particular coefficient will affect how all the sums it is involved in should be split. In *Prec*'s iterative refinement loop, each F_{obs} of each image is compared to the expected ("calculated") value $F_{R_{calc}}$ that is obtained by combining Ewald sphere-related Fourier coefficients from a previous reconstruction:

$$F_{R_{calc}} = F_{R_j} + F^*_{L_j}(\alpha - i\beta)^2 e^{-i2\chi} \tag{14}$$

where the index j represents the j^{th} iteration of the reconstruction. The difference between the CTF-corrected observed value for image k and this calculated value is stored as the "error" $2F_{\Delta_k}$:

$$F_{R^k_{corr}} - F_{R^k_{calc}} = 2F_{\Delta_k} \tag{15}$$

Half of these errors are then added as a refinement to the Fourier component on the right:

$$F_{R_{j+1}} = F_{R_j} + \tfrac{1}{N}\sum_k^N F_{\Delta_k} \tag{16}$$

The correction can also be immediately added to the left side:

$$F^*_{L_{j+1}}(\alpha - i\beta)^2 e^{-i2\chi} = F^*_{L_j}(\alpha - i\beta)^2 e^{-i2\chi} + F_\Delta \tag{17}$$

which, after rotation, complex conjugation, and summation of corrections, simplifies to:

$$F_{L_{j+1}} = F_{L_j} + \tfrac{1}{N}\sum_k^N F^*_{\Delta_k}(\alpha - i\beta)^2 e^{-i2\chi_k} \tag{18}$$

In the special (initial) case where the reconstruction to be refined consists completely of a set of zeroes, the calculated value, $F_{R_{calc}}$, is also zero and thus the correction applied to the left and right Fourier components (F_{R_0} and F_{L_0}) can be shown to be equivalent to the PM, scaled by a simple factor of $\frac{1}{2}$:

$$F_{R^k_{corr}} = 2F_{\Delta_k} \tag{19}$$

$$F_{R_0} = \tfrac{1}{N}\sum_k^N F_{\Delta_k} = \tfrac{1}{N}\sum_k^N \tfrac{1}{2} F_{R^k_{corr}} = \tfrac{1}{2} F_{R_{PM}} \tag{20}$$

$$\begin{aligned} F_{L_0} &= \tfrac{1}{N}\sum_k^N F^*_{\Delta_k}(\alpha - i\beta)^2 e^{-i2\chi_k} = \tfrac{1}{N}\sum_k^N \tfrac{1}{2} F^*_{R^k_{corr}}(\alpha - i\beta)^2 e^{-i2\chi_k} \\ &= \tfrac{1}{N}\sum_k^N \tfrac{1}{2} F_{L^k_{corr}} = \tfrac{1}{2} F_{L_{PM}} \end{aligned} \tag{21}$$

The effect of iterating turns out to be small, however. Take for example any Fourier coefficient F_{R_0} and the contributions to it:

$$F_{R_0} = \frac{1}{N}\sum_{k}^{N} F_{R_k} + \frac{1}{N}\sum_{k}^{N} F_{L_k}^{*}(\alpha - i\beta)^2 e^{-i2\chi_k} \tag{22}$$

where N is the number of images that measured F_R.

This can be recast as

$$F_{R_0} \approx \overline{F}_R + \varepsilon \tag{23}$$

where $\overline{F}_R$ is the average F_{R_k} and ε is the residual error which consists of the average of the $F_{L_k}(\alpha - i\beta)^2 e^{-i2\chi_k}$ terms, which is a random walk with step size of approximately $\overline{|F_{L_k}|}$. As such, after the first refinement cycle the residual error falls off as $\sim\frac{1}{\sqrt{N}}$, so that for large numbers of images, only small improvements can be expected from iteration.

3.3.4 Implementation of the Prec *Algorithm* — Three versions of *Prec* were implemented, one each in the software packages Bsoft, IMIRS, and EMAN, which each have all the functionality required to produce high-resolution reconstructions from raw cryo-EM images. While the mathematical theory is as described above, key differences exist in how the interpolations are handled in the different coordinate systems. Bsoft and EMAN use a Cartesian coordinate system. Starting with raw cryo-EM images, the Bsoft and EMAN implementations of *Prec* begin by calculating the images' 2D FTs, multiplying

them by the cCTF, and then calculating the "z-" coordinate (height up the Ewald sphere) for each Fourier coefficient. Taking into account the projection direction, the coefficients from the image are then added to the nearest corresponding lattice points of the "reconstruction" 3D FT with appropriate phase factors. In the Bsoft version, the standard interpolation procedure with weight $w = 1 - d$ (where d is distance in pixels from the measurement to the 3D lattice point) is used. In the EMAN version, any of its various built-in interpolation procedures can be used. After all the data are added to the "reconstruction" 3D FT, each amplitude is divided by the total weight of all the measurements that contributed, and a density map is produced through an inverse 3D FT. Refinement cycles, implemented in Bsoft, loop through each coefficient of each corrected image transform. The expected value is calculated by summing the coefficients at the nearest corresponding lattice points of the 3D FT of the current reconstruction with appropriate phase factors and complex conjugation. Half the difference between this expected value and the (CTF-corrected) observed value is added to each contributing coefficient.

A different version of *Prec* was implemented within IMIRS. IMIRS uses a cylindrical coordinate system for the reconstruction process where the 3D reconstruction and its FT are expressed as expansions of cylinder functions, as proposed by Klug et al. (Klug, Crick et al. 1958). We follow the notation used by Crowther et al. (Crowther, Derosier et al. 1970). The 2D FTs of the raw images are calculated and multiplied by the cCTF as before. The 3D FT of the object is represented in cylindrical coordinates, Z, R, and Φ. The Ewald sphere of measurements recorded in each image will in general intersect each

ring of coordinates in two places. For each intersection of an image Ewald sphere and a ring of the 3D FT, a Fourier coefficient for that location is estimated from the pixels of the FT of the image through bilinear interpolation. Once all the estimates on a particular ring have been calculated, all of them are used to determine the cylindrical expansion terms, $G_n(R,Z)$ through a least-squares fit which differs from the conventional IMIRS reconstruction in that the magnitude of the cCTF term is 1 and therefore is not a factor in the weighting of terms. A Fourier-Bessel transform is used next to obtain the $g_n(r,Z)$ terms, which are then used to generate the density map.

Because in this case the F_L that pairs with each F_R of a randomly spaced intersection of an image Ewald sphere and a Fourier ring does not generally fall upon any ring, a 3D nearest neighbor interpolation was required to estimate its value. Our tests (see below) suggested that the losses due to this less-accurate nearest-neighbor interpolation outweighed the gains obtained by iteration, so that iteration of the cylindrical-coordinate-based version of *Prec* is not recommended. In addition, astigmatism correction capabilities were added to both the conventional and *Prec* IMIRS reconstruction programs to accommodate the DLP dataset (see below).

3.3.5 Tests on Simulated Images — In order to explore the problems caused by Ewald sphere curvature and verify *Prec's* ability to solve them, a large number of images of the moderate-sized (~ 300 Å diameter) foot-and-mouth disease virus (FMDV) (Fry, Acharya et al. 1993) were simulated with different methodologies, voltages, and signal-to-noise ratios. A complete pdb was generated using the VIPERdb (Shepherd, Borelli et al. 2006)

and then its density was sampled to produce a reference volume using a modified version of *bgex* of the Bsoft package. Two types of simulated images were then calculated. The first, "Ewald projection" method produced images by simply summing Fourier coefficients on Ewald spheres using equation 8 and a complete 1D Whittaker-Shannon interpolation (Whittaker 1915; Shannon 1949) in the Z direction, followed by an inverse 2D FT. In order to produce a second, methodologically independent and more accurate set of simulated images, we used the multi-slice algorithm (Cowley and Moodie 1957). This well-established method tracks the dynamic scattering events that are increasingly important for thick samples, and was implemented in Bsoft by Heymann and Jensen with the assumption that scattering is completely elastic (*manuscript in preparation*). The sample is considered as a stack of equally thick slices. The effect of each slice on an incident plane wave is tracked by multiplying the slice's projected density (treated as a phase grating) with the wave function. The propagation of the wave between slices is calculated by convolution with a "propagator" function, so that the effects of Ewald sphere curvature arise naturally as the incident wave passes through the slices. After interaction with the final slice, the multi-slice image is generated by convolving the exit wave function with a complex contrast transfer function representing the lens.

As a first test, the simpler Ewald projections with varying acceleration voltages were used to study the effect of the electron wavelength on the maximum achievable resolution. Six data sets of 5000 Ewald projections each, with acceleration voltages of 15, 25, 50, 100, 200, and 300 kV, respectively, were calculated. FMDV reconstructions were then calculated from each data set using the conventional reconstruction programs

in Bsoft, IMIRS, and EMAN, which do not correct for curvature of the Ewald sphere. The resolution of each reconstruction was measured by its correlation with the original reference density map in Fourier shells (FSC) and confirmed visually (Figure 3-2, Bsoft results only). The large number of images ensured that Fourier space was well sampled. The expected increase in resolution as a function of voltage demonstrated the Ewald sphere curvature problem.

Analogous reconstructions of the 15 kV data set were then performed with Bsoft, IMIRS, and EMAN implementations of *Prec*. All three programs completely overcame the effects of Ewald sphere curvature. Because in this context the exact wave aberration values χ used to simulate the images in Bsoft could only be estimated by interpolation in the IMIRS coordinate system, the *Prec* in IMIRS reconstruction failed to reach all the way to Nyquist frequency, but instead was eventually limited by the rate of change of χ to ~ 3 Å resolution. In practice, where voltages much higher than 15 kV are used, this behavior of χ will not be limiting for either program.

Next the effects of smaller numbers of images and noise were explored using multi-slice images. Five-thousand FMDV images were again calculated, this time using Peach (Leong, Heymann et al. 2005), a distributed computation system, to meet the heavier computational demands of the multi-slice algorithm. A voltage of 15 kV was again assumed to ensure that the Ewald sphere curvature limitations would be manifest well before Nyquist frequency. Multiple sets of images with different signal-to-noise-ratios (SNRs) were then produced by first calculating the standard deviation of the raw multi-

slice image (σ_{image}), and then adding random Gaussian noise with standard deviation σ_{noise} to each pixel such that $(\frac{\sigma_{image}}{\sigma_{noise}})^2$ was equal to the desired SNR.

To confirm the presence of the Ewald sphere curvature problem in the multi-slice images, conventional reconstructions were produced from 25, 50, 100, 250, 500, 1000, 2500, and 5000 images, respectively, all with a SNR of 0.1, using the conventional *reconstruct* program in IMIRS. The reconstructions were again limited to ~ 4.2 Å, regardless of how many images were included (data not shown, except for the 5000-image reconstruction curve, which is part of the set described next). Application of the Bsoft, IMIRS, and EMAN versions of *Prec* removed the limitation (Figure 3-3, IMIRS results only), although the IMIRS reconstructions were again limited to ~ 3 Å resolution by the CTF-correction interpolation problem explained earlier.

In order to test how robust *Prec*'s refinement algorithm is to the presence of noise, similar reconstructions were calculated from 5000-image data sets with SNR ratios of 0.05, 0.01, and 0.001. While the resolutions of the corresponding reconstructions progressively decreased with increasing noise, in every case *Prec* clearly overcame the basic problem of Ewald sphere curvature (Figure 3-3). Further improvements were not realized by second or third iterations of *Prec* (the cylindrical-coordinate-based IMIRS version), probably for the reason described in Section 2.4.

3.3.6 Application to the CPV, ε15, and DLP reconstructions — Although several groups have proposed solutions to the Ewald sphere limitation in the context of complex wavefront recovery (Saxton 1994), none to our knowledge have shown a successful correction in a 3D reconstruction from experimental data. The recent publication of three near-atomic resolution (~ 4 Å) reconstructions of large (~ 700-Å diameter) viruses presents an opportunity to do so. According to Jensen and Kornberg's envelope function (Jensen and Kornberg 2000), half of the signal in a conventional reconstruction of such a large virus at 300 kV would be lost due to curvature of the Ewald sphere by 3.5 Å resolution. Likewise, DeRosier's formula (DeRosier 2000) predicts that the curvature problem in this same situation would become significantly limiting by 3.3 Å resolution. Thus as a further test of *Prec*, it was next used to refine the experimental reconstructions of CPV, ε15, and DLP.

CPV is a 750 Å diameter dsRNA virus in the *Reoviridae* family. Using the same cryo-EM images, two different 3D reconstructions were obtained using *Prec* and, for comparison, the conventional *IMIRS* reconstruction program. While in the previous tests of *Prec* with simulated images, only the cCTF-correction was used, in order to optimize this experimental reconstruction of CPV at all spatial frequencies, the low frequency Fourier coefficients of the (cCTF-corrected) *Prec* reconstruction were replaced with those from the conventional (rCTF-corrected) *reconstruct* version, as discussed above in conjunction with equation 10. The transition point was chosen as the spatial frequency where the two reconstructions matched best (~ 17 Å, Figure 3-4).

Disappointingly, the *Prec* reconstruction of CPV was not significantly better than the conventional. By visual inspection, the *Prec* reconstruction looked just slightly higher resolution in several locations, but not conclusively so (Figure 3-4a–i). Because the same images and particle parameters (defocus, origin, orientation) were used in these reconstructions, all the differences were due to *Prec*'s correction for Ewald sphere curvature. To compare the resolutions of the two maps quantitatively, the CPV cryo-EM image dataset was split into halves and independent "half-maps" were generated by *Prec* and then again by the conventional IMIRS *reconstruct* program. After all four maps were normalized and a soft spherical mask was imposed to remove noise inside the capsid shell, FSC curves were calculated (Ludtke, Baldwin et al. 1999)(Figure 3-4j). While the large spaces around the turrets and within the capsid shell devoid of protein reduce correlation and make it difficult to relate these FSC curves to the actual interpretability of the map, again these curves suggested that the *Prec* map might have had just slightly higher resolution at frequencies where the signal seemed reliable (i.e., $< \sim 1/6$ Å^{-1}), but not significantly. Similarly, the experimental reconstructions of the 700 and 710 Å diameter ε15 and DLP viruses calculated with the EMAN and IMIRS programs, respectively, were also only marginally if at all improved by refinement with *Prec* (data not shown).

In order to explore why more significant improvements were not realized, a single set of images were simulated of the equally large (754 Å diameter) reovirus core (Reinisch, Nibert et al. 2000) at the same voltage used in the experiments (300 kV). In this case applying the multi-slice algorithm was not computationally practical, so only Ewald

projections were used. After a conventional (EMAN) reconstruction was calculated from these simulated images, FSC analysis indicated a resolution of ~ 2.5 Å, much better than predicted by either Jensen and Kornberg's envelope or DeRosier's formula. Because the experimental reconstructions of CPV, ε15, and DLP had significantly worse resolution, we conclude that other resolution limitations besides the Ewald curvature problem must still be experimentally dominant. *Prec* in either Bsoft or EMAN again alleviated the problem in this simulated context as expected (Figure 3-5, EMAN results only).

3.4 Discussion

Here we have described *Prec*, an iterative algorithm based on the oft-described PM that corrects for curvature of the Ewald sphere in 3D reconstructions. Three versions were implemented: two Cartesian-coordinate-based versions in the software packages Bsoft (where multi-threading was also introduced) and EMAN, and a cylindrical-coordinate-based version in IMIRS. To test *Prec*, numerous images of a moderately sized virus were simulated in two different ways, namely simple Ewald projection and the more sophisticated multi-slice method. All three versions of *Prec* corrected for the curvature problem in reconstructions from both types of simulated images, even in the presence of noise greater than that found in typical experimental images. *Prec* was then used to refine the experimental reconstructions of CPV, ε15, and DLP from cryo-EM images. Disappointingly, none of these experimental reconstructions were significantly improved. To explain this result, a single set of images were simulated of a similarly large virus with the same imaging parameters used in the experimental reconstructions. Reconstructions from these simulated images showed that, contrary to expectations, the

Ewald curvature did not become severely limiting until ~ 2.5 Å resolution. Thus other factors besides Ewald sphere curvature are still the predominant resolution limitation even in these high-resolution experimental reconstructions. As the size of reconstructed viruses, the number and quality of images that are included in reconstructions, and the precision to which those images can be mutually aligned continue to increase, Ewald curvature correction will nevertheless eventually become essential.

During the course of this project, Wolf et al. implemented a Cartesian-coordinate-based version of the PM similar to ours but in the FREALIGN package and with minor differences in the weighting factors involved in CTF correction (Wolf, DeRosier et al. 2006). These differences allowed a single CTF correction strategy to be used throughout the resolution range rather than the combination of real and complex CTF corrections used by *Prec* at low and high spatial frequencies, respectively. Wolf et al. further proposed an iterative, "reference-based insertion" method similar to our iterative algorithm, and tested it on simulated images, but reported that under conditions of low signal, iteration decreased FSCs. Here the Cartesian-coordinate- but not the cylindrical-coordinate-based version of *Prec* realized slight gains through iteration, even in the presence of noise, but the specific reasons for the difference remain unclear.

The cylindrical-coordinate-based version of *Prec* has two major advantages in comparison to the Cartesian implementations (Bsoft, EMAN, and FREALIGN). First, the cylindrical expansions allow all the measurements on a particular ring to be used to sample specific Fourier coefficients (Crowther, Derosier et al. 1970). Second, the

cylindrical-coordinate-based *Prec* program is much faster and requires less memory. Our CPV reconstructions from over twelve thousand 1k x 1k images required less than a day on a single-processor personal PC and used less than 2 Gbytes of memory. In contrast, even the multi-threaded and distributed versions of the Cartesian-based *Prec* in Bsoft and EMAN would have required a prohibitive ~ 20 and ~ 16 Gbytes of memory, respectively, and approximately 10 times more computing power to match the computation times of IMIRS. Likewise, Equation 5 of (Grigorieff 2007) suggests that FREALIGN would need 30 Gbytes of memory for such images.

The programs created for this project are freely available at www.jensenlab.caltech.edu.

3.5 Acknowledgements

We thank Andy Rawlinson, Bernard Heymann, Bill Tivol, Dylan Morris, Yuyao Liang, Wong Hoi Hui, Xiaokang Zhang, Weimin Wu, Wen Jiang and Nikolaus Grigorieff for helpful discussions about Ewald sphere curvature and the manuscript as well as the Bsoft, IMIRS, and EMAN packages and for providing experimental data. This work was supported in part by NIH grants R01 AI067548 and P50 GM082545 to GJJ and R01 GM071940, CA094809 and AI069015 to ZHZ; DOE grant DE-FG02-04ER63785 to GJJ; a Searle Scholar Award to GJJ; the Beckman Institute at Caltech; and gifts to Caltech from the Parsons Foundation and Agouron Institute. Access to the 4-node and 8-node Sun Fire X4600 computers, located at the California Institute of Technology, was provided by the Center for Advanced Computing Research.

3.6 References

Cowley, J. M. and A. F. Moodie (1957). "The Scattering of Electrons by Atoms and Crystals .1. a New Theoretical Approach." Acta Crystallographica **10**(10): 609−619.

Crowther, R. A., D. J. Derosier, et al. (1970). "Reconstruction of 3 Dimensional Structure from Projections and Its Application to Electron Microscopy." Proceedings of the Royal Society of London Series A-Mathematical and Physical Sciences **317**(1530): 319−340.

DeRosier, D. J. (2000). "Correction of high-resolution data for curvature of the Ewald sphere." Ultramicroscopy **81**(2): 83−98.

Erickson, H. P. and A. Klug (1971). "Measurement and Compensation of Defocusing and Aberrations by Fourier Processing of Electron Micrographs." Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **261**(837): 105−118.

Frank, J. (2002). "Single-particle imaging of macromolecules by cryo-electron microscopy." Annual Review of Biophysics and Biomolecular Structure **31**: 303−319.

Fry, E., R. Acharya, et al. (1993). "Methods Used in the Structure Determination of Foot-and-Mouth-Disease Virus." Acta Crystallographica Section A **49**: 45−55.

Grigorieff, N. (2007). "FREALIGN: High-resolution refinement of single particle structures." Journal of Structural Biology **157**(1): 117−125.

Henderson, R., J. M. Baldwin, et al. (1990). "Model for the Structure of Bacteriorhodopsin Based on High-Resolution Electron Cryomicroscopy." Journal of Molecular Biology **213**(4): 899−929.

Heymann, J. B. (2001). "Bsoft: Image and molecular processing in electron microscopy." Journal of Structural Biology **133**(2−3): 156−169.

Jensen, G. J. and R. D. Kornberg (2000). "Defocus-gradient corrected back-projection." Ultramicroscopy **84**(1−2): 57−64.

Jiang, W., M. L. Baker, et al. (2008). "Backbone structure of the infectious epsilon 15 virus capsid revealed by electron cryomicroscopy." Nature **451**(7182): 1130−1134.

Klug, A., F. H. C. Crick, et al. (1958). "Diffraction by Helical Structures." Acta Crystallographica **11**(3): 199−213.

Kuhlbrandt, W., D. N. Wang, et al. (1994). "Atomic Model of Plant Light-Harvesting Complex by Electron Crystallography." Nature **367**(6464): 614−621.

Leong, P. A., J. B. Heymann, et al. (2005). "Peach: A simple perl-based system for distributed computation and its application to cryo-EM data processing - Ways & means." Structure **13**(4): 505−511.

Liang, Y. Y., E. Y. Ke, et al. (2002). "IMIRS: a high-resolution 3D reconstruction package integrated with a relational image database." Journal of Structural Biology **137**(3): 292−304.

Ludtke, S. J., P. R. Baldwin, et al. (1999). "EMAN: Semiautomated software for high-resolution single-particle reconstructions." Journal of Structural Biology **128**(1): 82−97.

Pettersen, E. F., T. D. Goddard, et al. (2004). "UCSF chimera - A visualization system for exploratory research and analysis." Journal Of Computational Chemistry **25**(13): 1605−1612.

Reinisch, K. M., M. Nibert, et al. (2000). "Structure of the reovirus core at 3.6 angstrom resolution." Nature **404**(6781): 960−967.

Saxton, W. O. (1994). "What Is the Focus Variation Method - Is It New - Is It Direct." Ultramicroscopy **55**(2): 171−181.

Schiske, P. (1968). "Zur Frage der Bildrekonstruktion durch Fokusreihen." Proc. 4th Eur. Conf. on Electron Microscopy Rome.

Shannon, C. E. (1949). "Communication in the Presence of Noise." Proceedings of the Institute of Radio Engineers **37**(1): 10−21.

Shepherd, C. M., I. A. Borelli, et al. (2006). "VIPERdb: a relational database for structural virology." Nucleic Acids Research **34**: D386−D389.

Unwin, N. (2005). "Refined structure of the nicotinic acetylcholine receptor at 4 angstrom resolution." Journal of Molecular Biology **346**(4): 967−989.

Van Dyck, D. and M. Op de Beeck (1990). "New direct methods for phase and structure retrieval by HREM." Proc. 12th Int. Congr. on Electron Microscopy Seattle.

Wan, Y., W. Chiu, et al. (2004). "Full contrast transfer function correction in 3D cryo-EM reconstruction." IEEE Proceedings of ICCCAS 2004 Chengdu, Sichuan, China.

Whittaker, E. T. (1915). "On the Functions which are Represented by the Expansion of Interpolation Theory." Proceedings of the Royal Society of Edinburgh **35**: 181−194.

Wolf, M., D. J. DeRosier, et al. (2006). "Ewald sphere correction for single-particle electron microscopy." Ultramicroscopy **106**(4−5): 376−382.

Yu, X. K., L. Jin, et al. (2008). "3.88 angstrom structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy." Nature **453**(7193): 415−419.

Zhang, X., E. Settembre, et al. (2008). "Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction." Proceedings of the National Academy of Sciences of the United States of America **105**(6): 1867−1872.

Zhou, Z. H. and W. Chiu (1993). "Prospects for using an IVEM with a FEG for imaging macromolecules towards atomic resolution." Ultramicroscopy **49**(1−4): 407−416.

3.7 Figures

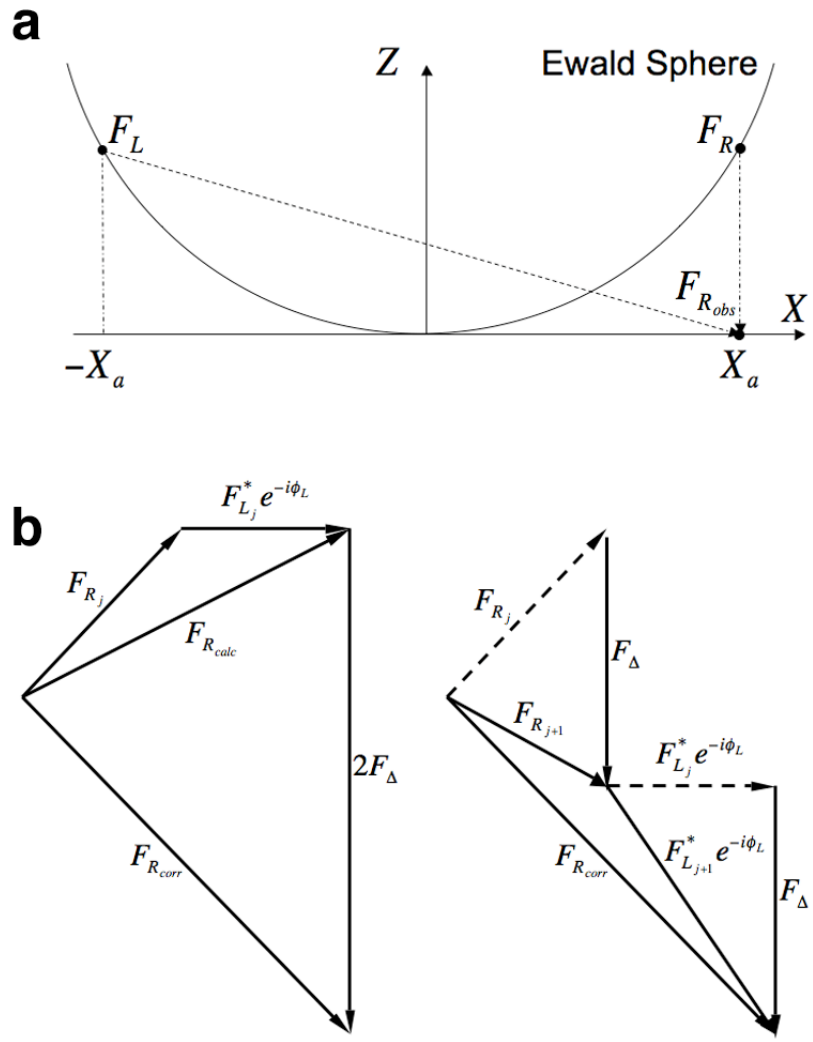


Figure 3-1. **The Ewald sphere and *Prec* algorithm.** (a) Fourier coefficients in the transforms of electron microscope images ($F_{R_{obs}}$) are actually combinations of coefficients (F_L and F_R) that lie on a spherical surface through the 3D transform of the specimen called the Ewald sphere. (b) *Prec* iteratively recovers the independent values of these coefficients by comparing CTF-corrected observations ($F_{R_{corr}}$) with the calculated sum ($F_{R_{calc}}$) that would have been expected from the right (F_{R_j}) and left (F_{L_j}) terms of some previous reconstruction, with appropriate phase factors $e^{i\phi_L} = (\alpha + i\beta)^2 e^{i2\chi}$. Half the difference ($F_\Delta$) is then added to F_{R_j} and F_{L_j} to produce the next iteration ($F_{R_{j+1}}$ and $F_{L_{j+1}}$).

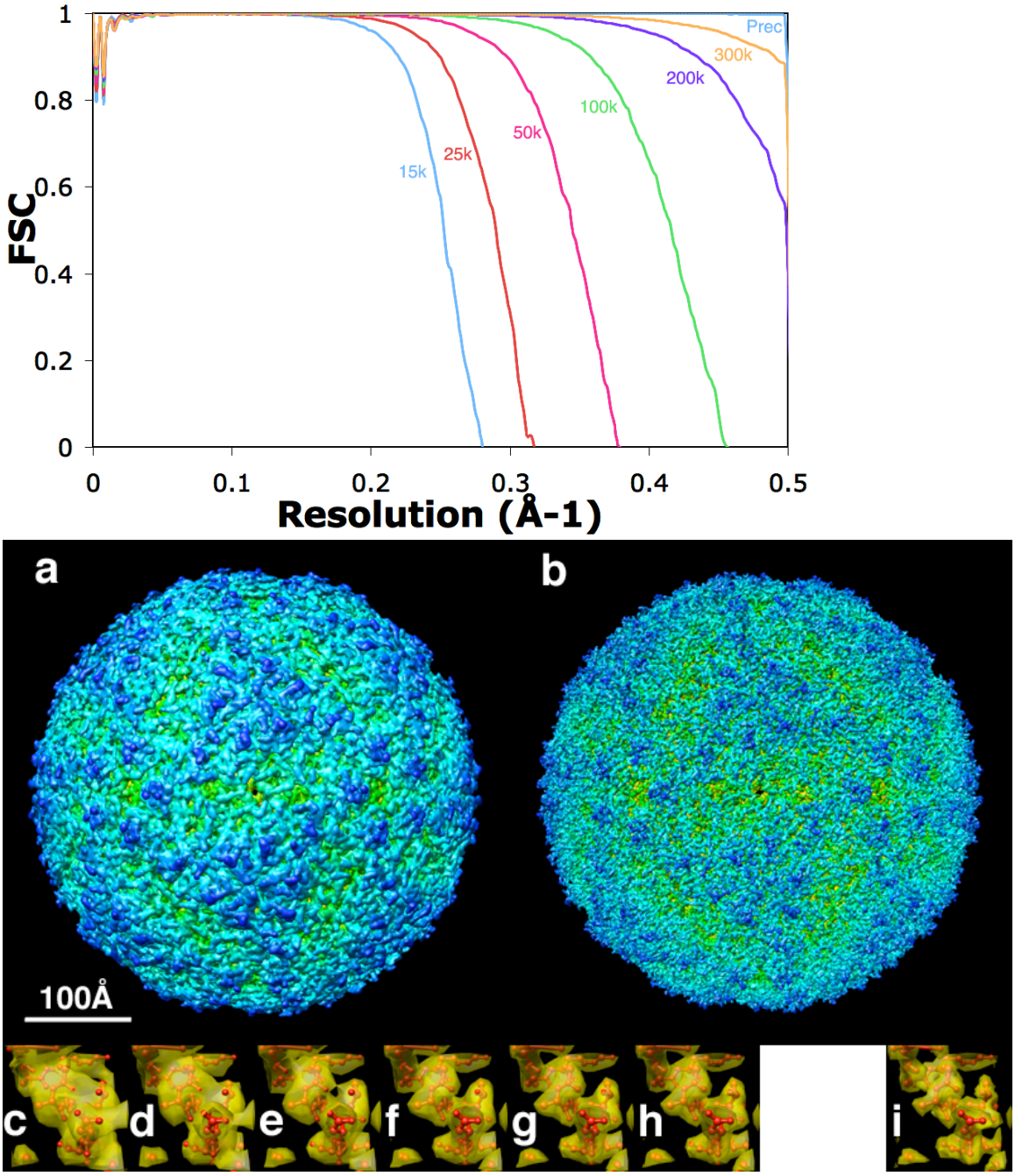


Figure 3-2. ***Prec* overcomes the curvature problem in Ewald projections.** (top) FSC curves for conventional Bsoft reconstructions of the foot and mouth virus from 5000 "Ewald projection" images simulated with the voltages shown, plus a reconstruction from the 15 kV images calculated by the *Prec* program, which completely corrects for the curvature problem. (a and b) Isosurface renderings of the conventional and *Prec* 15 kV reconstructions, respectively. (c, d, e, f, g, h) Transparent isosurfaces of a single α-helix from the 15, 25, 50, 100, 200, and 300 kV reconstructions, respectively, surrounding the atomic model used to simulate the images. (i) The same helix from the *Prec* 15kV reconstruction. FSC curves were calculated with *bresolve* (Heymann 2001) and isosurfaces were rendered with *Chimera* (Pettersen, Goddard et al. 2004)

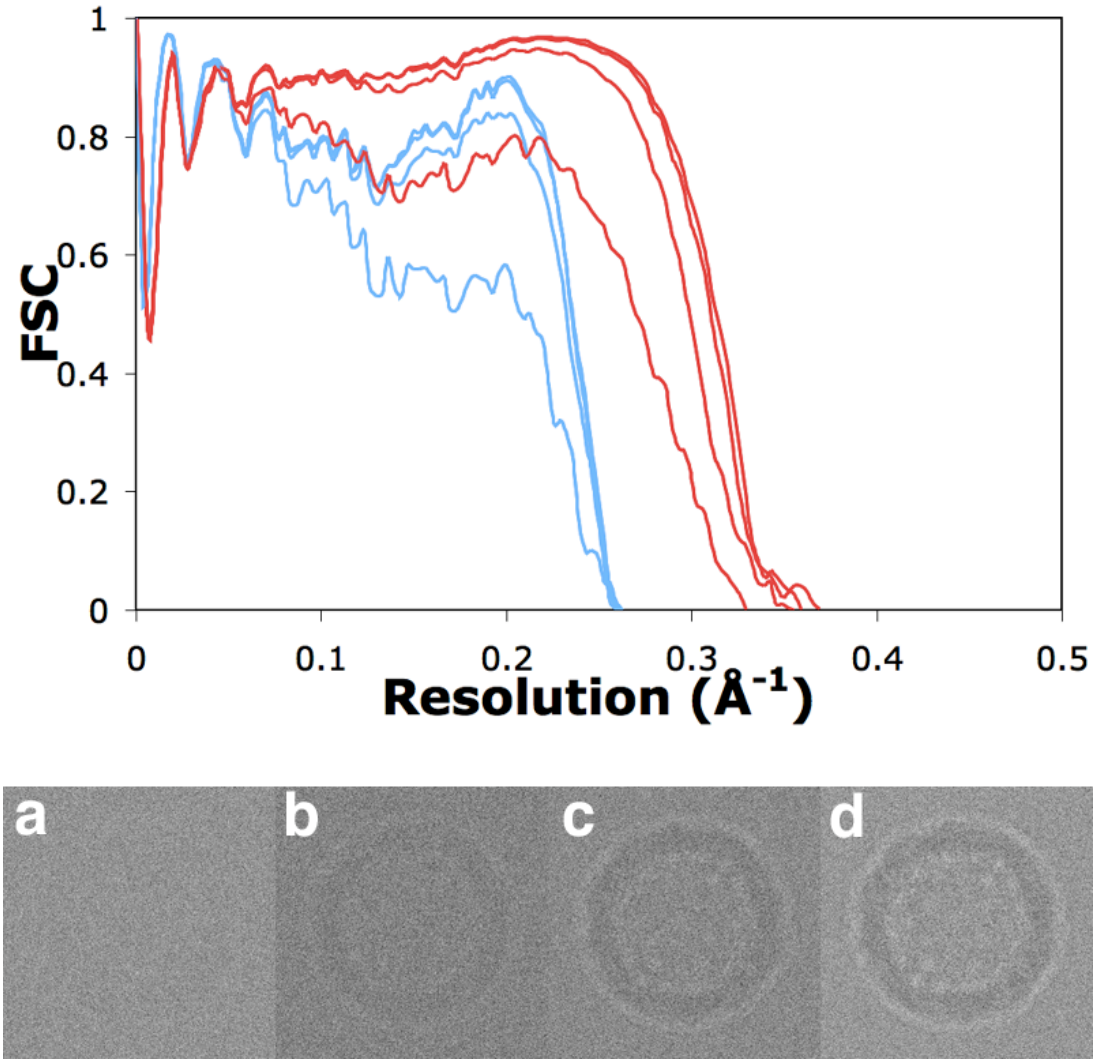


Figure 3-3. ***Prec*** **overcomes the curvature problem in multi-slice images and in the presence of noise**. (top) FSC curves reporting the resolution of reconstructions calculated using conventional methods (the IMIRS *reconstruct* program, blue) and *Prec* (IMIRS implementation, red) from 5000 fifteen-kV multi-slice images with SNRs of 0.001, 0.01, 0.05, and 0.1 (progressively with higher resolution). (a–d) One example multi-slice image for each noise level

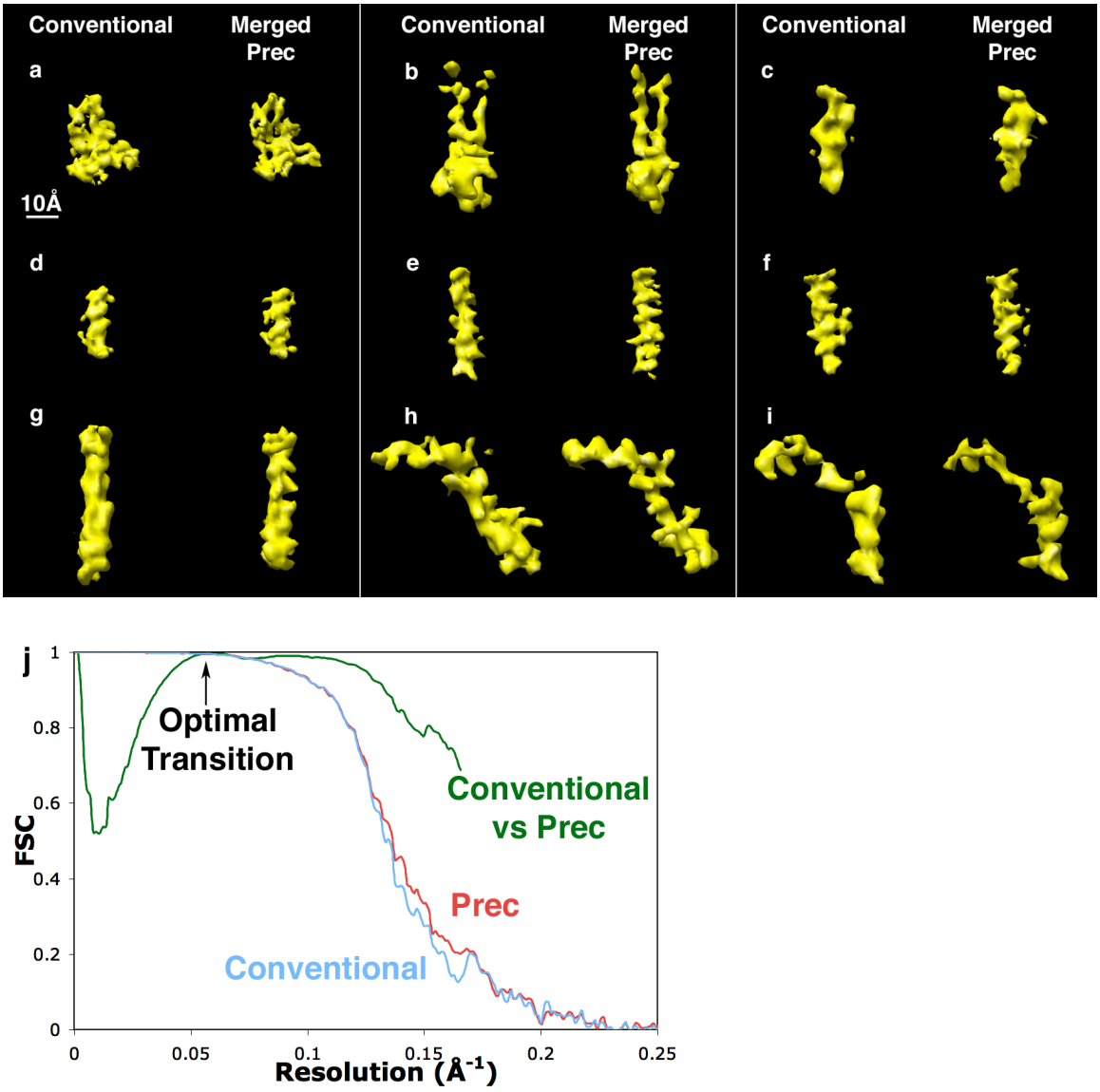


Figure 3-4. **Application of *Prec* to experimental images: 3D reconstruction of CPV.** (a–i) Isosurfaces of selected β-sheets (a, b) and α-helices (c–i) from the conventional and *Prec* reconstructions, respectively, do not clearly show improved interpretability of the *Prec* map. (j) FSC curves for the *Prec* (blue) and conventional (red) reconstructions of CPV, plus a third FSC curve (green) comparing the two that identifies the resolution at which *Prec*'s complex CTF-correction method becomes more appropriate than the conventional real CTF-correction

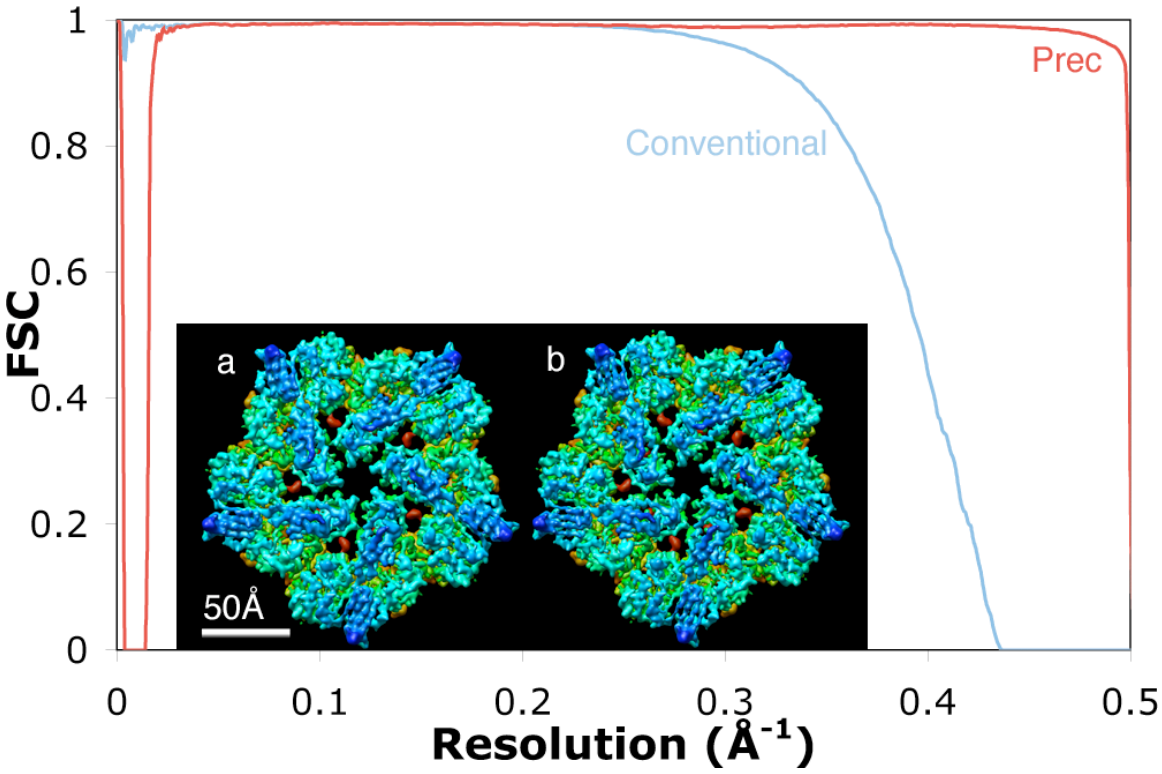


Figure 3-5. **Reconstructions of the 754 Å diameter Reovirus from 300 kV simulated images.** Curvature of the Ewald sphere does not limit the resolution of the conventional reconstruction (blue curve) until ~ 2.5 Å, showing that other factors must still be dominant in the recent experimental reconstructions of similarly sized viruses. *Prec* in EMAN eliminates the limitation, recovering the full resolution present in the simulated images. (a and b) Turrets from conventional and *Prec* (EMAN) reconstructions, respectively

Chapter 4

Conclusion

4.1 **Progression of Single Particle Analysis**

To avoid the challenges of crystallization and the size limitations of nuclear magnetic resonance spectroscopy, it has long been hoped that single-particle cryo-electron microscopy would eventually produce atomically interpretable maps. Steady progress towards this goal has been made (Frank 2002), led by reconstructions of large icosahedral viruses, whose 60-fold symmetry, size, and rigid architecture all facilitate precise image alignment. 3D single-particle reconstructions of virus particles from electron micrographs were first accomplished by Fourier synthesis in 1970 (Crowther, Amos et al. 1970). By the turn of the 21st Century, single particle techniques had already achieved sub-nanometer resolutions (Bottcher, Wynne et al. 1997; Conway, Cheng et al. 1997; Trus, Roden et al. 1997) but were still limited in resolution by various factors (Baker, Olson et al. 1999; van Heel, Gowen et al. 2000). The difficulty in modeling some of these factors led to the lack of accurate predictions about the severity of each of these limits and it was unclear which was the most dominant limit. Thus, when I began my thesis work in 2002, I chose to address two of these problems, namely the lack of computing power for high-resolution reconstructions and the depth of field or, equivalently, the Ewald sphere curvature problem (DeRosier 2000), as they were best suited to my interests and abilities.

4.2 Hybrid Approach to Address Lack of Computational Power

I have addressed the lack of computational power using a hybrid computational approach (parallel computation used in conjunction with a distributed computation system), which utilizes untapped resources to effectively increase computational power. This approach consisted of (1) Parallel implementations of conventional and paraboloid reconstruction algorithms (Chapter 3), which are also compatible with distributed computation systems and (2) Development of a distributed computation system (Chapter 2) designed specifically for (but not limited to) large scale image processing.

Thus, the hybrid approach, when applied to single particle reconstructions, allows for the utilization of all cores on each computer and all available computers participating in the distributed computation system. This leads to a massive computational speedup and is necessary for high-resolution reconstructions of large virus particles.

4.3 Paraboloid Reconstruction Algorithm to Address Ewald Sphere Curvature

I have addressed the Ewald sphere curvature problem, or equivalently the depth of field problem, by development of the *Prec* algorithm. The algorithm, unlike conventional reconstruction algorithms that are based on the projection theorem, takes into account the curvature of the Ewald sphere and is able to correct for this resolution limitation completely (Chapter 3).

The *Prec* algorithm was applied to simulated images and recent experimental data sets of three 700–750 Å diameter viruses, which had been reconstructed to ~ 4 Å resolution

(Jiang, Baker et al. 2008; Yu, Jin et al. 2008; Zhang, Settembre et al. 2008) by conventional methods. Two main conclusions could be drawn from the results: (1) The Ewald sphere curvature problem has been completely solved and (2) The curvature of the Ewald Sphere is currently not the dominant resolution limit.

Thus, in order for the effects of the Ewald sphere curvature correction to be significant, higher resolution reconstructions of larger virus particles have to be achieved. It would seem that with the rapid improvements in single-particle reconstruction resolutions over the past decade, it is just a matter of time before these resolutions become sufficiently high. When this occurs, the application of the Prec algorithm will be necessary for high-resolution reconstructions of large virus particles.

4.4 References

Baker, T. S., N. H. Olson, et al. (1999). "Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs." Microbiology and Molecular Biology Reviews **63**(4): 862−922.

Bottcher, B., S. A. Wynne, et al. (1997). "Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy." Nature **386**(6620): 88−91.

Conway, J. F., N. Cheng, et al. (1997). "Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy." Nature **386**(6620): 91−94.

Crowther, R. A., L. A. Amos, et al. (1970). "3 Dimensional Reconstructions of Spherical Viruses by Fourier Synthesis from Electron Micrographs." Nature **226**(5244): 421−425.

DeRosier, D. J. (2000). "Correction of high-resolution data for curvature of the Ewald sphere." Ultramicroscopy **81**(2): 83−98.

Frank, J. (2002). "Single-particle imaging of macromolecules by cryo-electron microscopy." Annual Review of Biophysics and Biomolecular Structure **31**: 303−319.

Jiang, W., M. L. Baker, et al. (2008). "Backbone structure of the infectious epsilon 15 virus capsid revealed by electron cryomicroscopy." Nature **451**(7182): 1130−1134.

Trus, B. L., R. B. S. Roden, et al. (1997). "Novel structural features of bovine papillomavirus capsid revealed by a three-dimensional reconstruction to 9 angstrom resolution." Nature Structural Biology **4**(5): 413−420.

van Heel, M., B. Gowen, et al. (2000). "Single-particle electron cryo-microscopy: towards atomic resolution." Quarterly Reviews of Biophysics **33**(4): 307−369.

Yu, X. K., L. Jin, et al. (2008). "3.88 angstrom structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy." Nature **453**(7193): 415−419.

Zhang, X., E. Settembre, et al. (2008). "Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction." Proceedings of the National Academy of Sciences of the United States of America **105**(6): 1867−1872.

Appendix

Supplementary Information

A.1 Introduction

The appendix provides further details about the results of *Prec* refinement cycles, the effect of additional images in conventional reconstructions on the Ewald sphere resolution limit, and a comparison of Ewald sphere resolution limit predictions with reconstructions from simulated data in the first three sections. This information could not be included with Chapter 3 due to the brevity required of academic papers. In addition, the orientation conventions of software packages used during the testing of CPV, ε15, and DLP are described. Lastly, a list of all the important programs used in Chapters 2 and 3 is provided.

A.2 *Prec* Refinement in Practice

As described in Chapter 3, *Prec* possesses an iterative capability, which allows for errors due to the Ewald sphere curvature in the 3D Fourier transform (FT) of the reconstruction to be reduced by successive applications of the *Prec* algorithm.

In order to determine the significance of an additional refinement cycle, the improvement in the FSC curves of reconstructions of one application of the refinement algorithm versus an additional refinement were calculated for simulated data sets of 25, 50, 100, 250, 500, 1000, 2500, and 5000 images using *Prec* in Bsoft on Ewald projections of

FMDV at 15 kV. The results of the tests indicated that the additional refinement produced an insignificant improvement in the FSC curves and this improvement decreased as the number of images used increased (Figure A-1).

To understand these results, we observe the form of the Fourier values after the first iteration as described in Chapter 3:

$$F_{R_0} \approx \overline{F}_R + \varepsilon \qquad (1)$$

where $\overline{F}_R$ is the average F_{R_k} and ε is the residual error which consists of the average of the $F_{L_k}(\alpha - i\beta)^2 e^{-i2\chi_k}$ terms, which is a random walk with step size of approximately $\overline{|F_{L_k}|}$. The residual error after the first iteration falls off as ~ $\frac{1}{\sqrt{N}}$, thus the error is small for large numbers of images and only small improvements can be expected from additional iterations.

In practice, large numbers of images, on the order of 10^4 (Jiang, Baker et al. 2008; Yu, Jin et al. 2008; Zhang, Settembre et al. 2008), are used for reconstructions that achieve high resolution, thus no additional refinement is necessary.

A.3 The Effect of the Ewald Sphere Resolution Limit on Conventional Algorithms

In Chapter 3, the Ewald sphere curvature problem was characterized by observing the resolution achieved by conventional algorithms as the number of images with significant Ewald sphere curvature increased. Reconstructions were generated using sets of 25, 50, 100, 250, 500, 1000, 2500, and 5000 multi-slice images at 15 kV. The results of these

tests (Figure A-2) indicated that, regardless of how many images were used, the Ewald sphere curvature problem could not be overcome by additional images.

In contrast, current state-of-the-art high-resolution reconstructions of large particles (~ 700–750 Å in diameter) (Jiang, Baker et al. 2008; Yu, Jin et al. 2008; Zhang, Settembre et al. 2008) have not reached the Ewald sphere resolution limit of ~ 2.5 Å as predicted by our simulations of a 754 Å diameter virus particle at 300 kV, despite the large number of images being used. Once these limits are approached, significant improvements in resolution should be observed without an increase in the number of images when the *Prec* algorithm is applied.

A.4 Comparison of Ewald Sphere Resolution Limit Predictions with Simulations

Currently there are two formulas for predicting the resolution limits imposed by the curvature of the Ewald sphere. The first is an envelope function by Jensen and Kornberg (Jensen and Kornberg 2000), which indicates the percentage of information content remaining as resolution increases. The second formula by DeRosier (DeRosier 2000) indicates a resolution limit. A third approach to predicting the resolution limit is through simulations where Ewald projections of a model generated from pdb files are used to produce a reconstruction, which is subsequently compared with a reference model using an FSC curve (Chapter 3).

According to the resolution of the reconstructions of simulated data sets (Figure A-3), the simulation method predicts the highest resolution limits due to the Ewald sphere

curvature, indicating that both formula predictions may be too strict. When taking into account only the Ewald sphere curvature effect, the simulation method is the most accurate as it produces resolution limits without making any other assumptions about the information content and also simulates entirely the reconstruction process. Its drawback is that it requires a large amount of computation time in the generation of simulated images and the reconstruction process.

A.5 Icosahedral Symmetry Conventions

During the testing of *Prec*, four image-processing packages were used. These were Bsoft (Heymann 2001), IMIRS (Liang, Ke et al. 2002), EMAN (Ludtke, Baldwin et al. 1999) and FREALIGN (Grigorieff 2007). *Prec* was implemented in all the packages except FREALIGN, as it possessed its own Ewald sphere correction functionality. The packages were chosen because the three highest resolution reconstructions by cryo-EM to date, CPV, DLP, and ε15 were achieved using IMIRS, FREALIGN, and EMAN, respectively, while Bsoft was used for testing and generating simulated images.

For images to be used in the reconstruction process, their orientations have to be specified. While a standardization of these conventions has been proposed (Heymann, Chagoyen et al. 2005), each of the packages still possessed their own orientation conventions. There are multiple ways that orientations can be specified, one way is by defining three Euler angles (used in IMIRS, EMAN, and FREALIGN) that correspond to rotation matrices such as

$$R_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix} \quad (2)$$

$$R_y(\alpha) = \begin{pmatrix} \cos\alpha & 0 & -\sin\alpha \\ 0 & 1 & 0 \\ \sin\alpha & 0 & \cos\alpha \end{pmatrix} \quad (3)$$

$$R_z(\alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (4)$$

where $R_x(\alpha)$, $R_y(\alpha)$, and $R_z(\alpha)$ represent right-handed rotations of the axes by angle α around the x, y, and z axes, respectively.

The alternative approach, which is used in Bsoft, is to define an axis of rotation by a normalized 3D vector known as a "view vector" and an angle of rotation.

The confusion which surrounds the use of Euler angles is due to the numerous possible combinations of rotation axes and angles directions that are possible. Through careful examination of the software code, the Euler angle conventions (Table A-1), as well as their order of listing in the various orientation file formats (Table A-2), were determined. In addition to the different Euler conventions, each of the packages had different reference orientations (Table A-3), i.e., when all three Euler angles are equal to zero.

Once the correct conventions had been determined for each of the packages, the conversion of angles between packages was straightforward for Bsoft (I90), IMIRS, and FREALIGN (I2). However, conversions from EMAN to Bsoft (I90) required additional

rotations of $R_z(-90°)$, $R_x(31.7175°)$, and $R_z(90°)$ before the application of EMAN Euler angles, due to a different reference orientation.

A.6 List of Important Programs in Peach and Prec

Peach — Distributed computation system

Pjobd	Job daemon
Pserv	Job server
Pview	Interactive client
Psubmit	Client for submission of jobs

Prec in Bsoft

Brec	Multi-threaded version of Breconstruct
Prec	Multi-threaded implementation of Prec
Pref	Multi-threaded implementation of Prec refinement loops
Ewald_proj	Generates Ewald projections

Prec in IMIRS

Prec	Implementation of Prec
Pref	Implementation of Prec refinement loops
Reconstruct_ast	Modified version of reconstruct with astigmatism correction
Prec_ast	Modified version of Prec with astigmatism correction

Prec in EMAN

Make3d (modified version)	Contains implementation of Prec compatible with multi-threaded and distributed computation capabilities of EMAN

Euler Conversion Programs

Eman_to_bsoft	Converts EMAN Euler angles to Bsoft view vector and angle
Bsoft_to_eman	Converts Bsoft view vector and angle to EMAN Euler angles

A.7 References

DeRosier, D. J. (2000). "Correction of high-resolution data for curvature of the Ewald sphere." Ultramicroscopy **81**(2): 83−98.

Grigorieff, N. (2007). "FREALIGN: High-resolution refinement of single particle structures." Journal of Structural Biology **157**(1): 117−125.

Heymann, J. B. (2001). "Bsoft: Image and molecular processing in electron microscopy." Journal of Structural Biology **133**(2−3): 156−169.

Heymann, J. B., M. Chagoyen, et al. (2005). "Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology." Journal of Structural Biology **151**(2): 196−207.

Jensen, G. J. and R. D. Kornberg (2000). "Defocus-gradient corrected back-projection." Ultramicroscopy **84**(1−2): 57−64.

Jiang, W., M. L. Baker, et al. (2008). "Backbone structure of the infectious epsilon 15 virus capsid revealed by electron cryomicroscopy." Nature **451**(7182): 1130−1134.

Liang, Y. Y., E. Y. Ke, et al. (2002). "IMIRS: a high-resolution 3D reconstruction package integrated with a relational image database." Journal of Structural Biology **137**(3): 292−304.

Ludtke, S. J., P. R. Baldwin, et al. (1999). "EMAN: Semiautomated software for high-resolution single-particle reconstructions." Journal of Structural Biology **128**(1): 82−97.

Yu, X. K., L. Jin, et al. (2008). "3.88 angstrom structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy." Nature **453**(7193): 415−419.

Zhang, X., E. Settembre, et al. (2008). "Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction." Proceedings of the National Academy of Sciences of the United States of America **105**(6): 1867−1872.

A.8 Figures and Tables

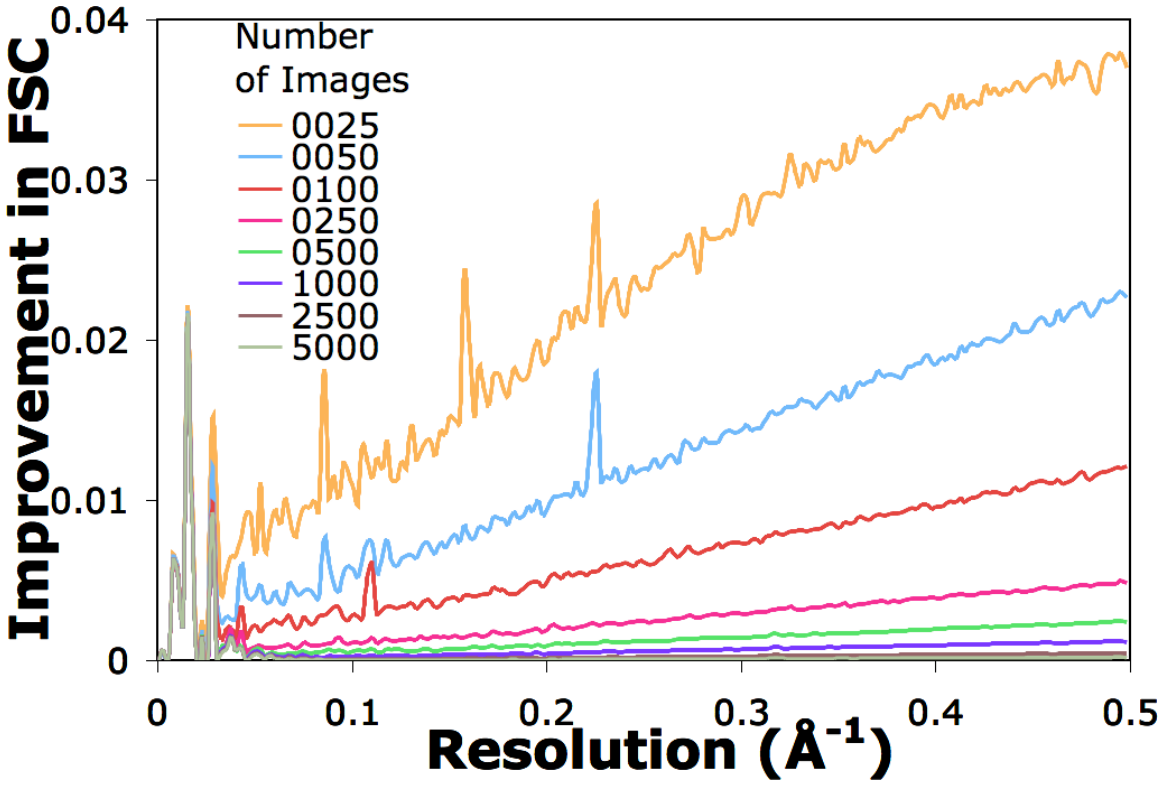


Figure A-1. **Effect of additional refinement loop**. Improvements in FSC for reconstructions using 25, 50, 100, 250, 500, 1000, 2500, and 5000 Ewald projections at an acceleration voltage of 15 kV using *Prec* in Bsoft demonstrate the decreasing significance of the improvement between the first and second cycles of refinement. When a large number of images are used in the reconstruction, the additional refinement has no significant effect.

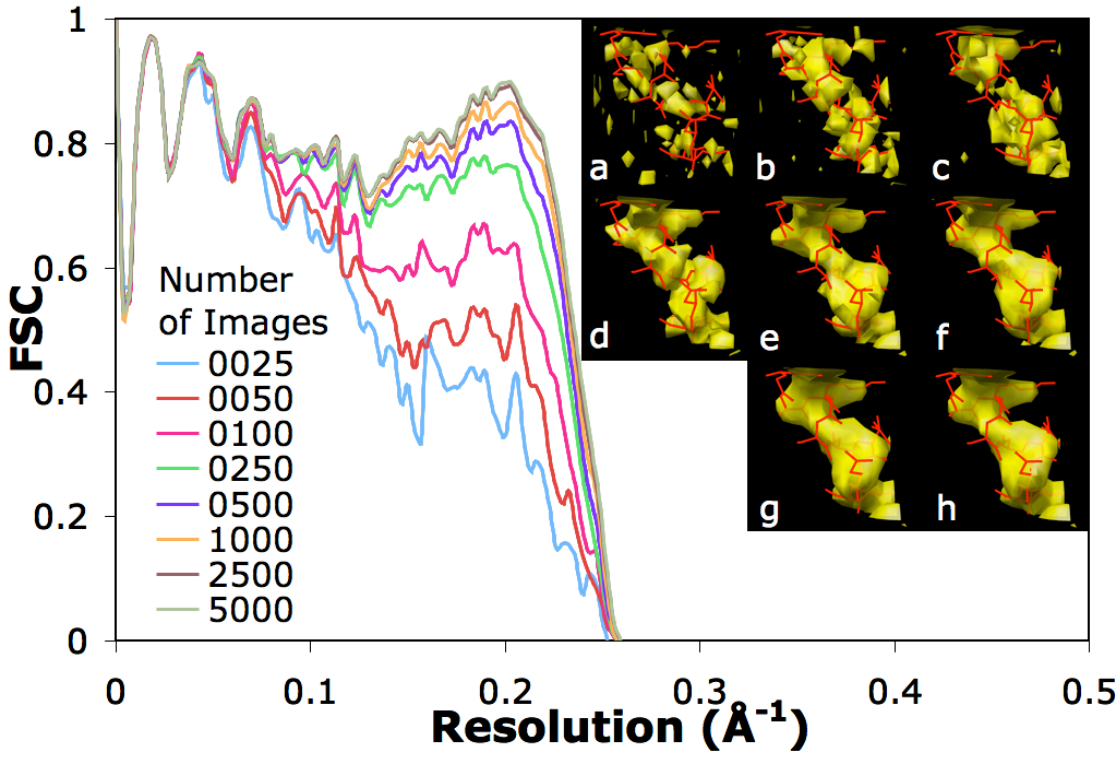


Figure A-2. **Effect of number of images on Ewald sphere curvature resolution limit**. FSC curves of conventional reconstructions performed using *reconstruct* of the IMIRS package, using 15 kV multi-slice images, demonstrate that the maximum resolution imposed by the curvature of the Ewald sphere cannot be overcome by increasing the number of images. Insets a–h show volume extracts of a single α-helix from the reconstruction from 25, 50, 100, 250, 500, 1000, 2500, and 5000 images, respectively.

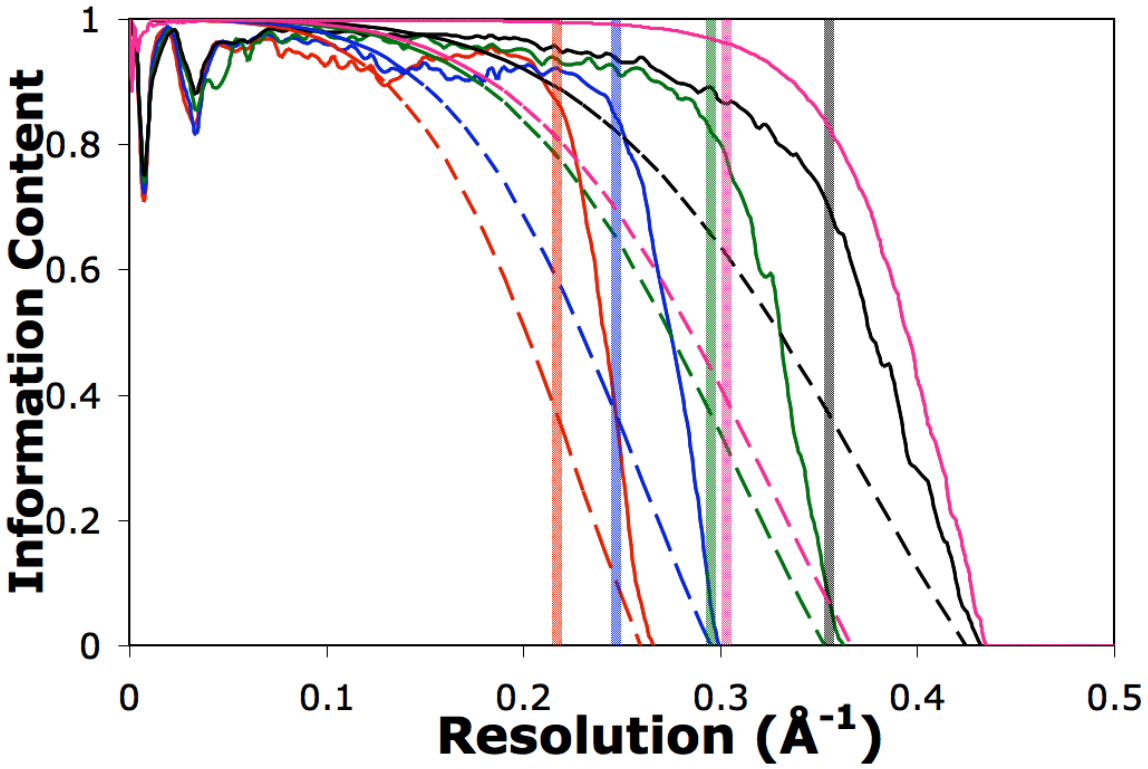


Figure A-3. **Comparison of Ewald sphere resolution limitations**. The comparison of the maximum achievable resolutions at acceleration voltages of 15 (red), 25 (blue), 50 (green), and 100 kV (black) for the foot and mouth virus and 300 kV (pink) for the reovirus core using the FSCs of the reconstructions (solid curves) from Ewald projections, the sinc envelopes by Jensen and Kornberg (dashed curves), and the empirical threshold by DeRosier (vertical lines), where the dimensionless constant p is 0.7. The envelopes and the limit formula both predict limits significantly lower than the resolutions achieved by reconstructions from simulated images.

Software Package	1st		2nd		3rd	
	angle	axis	Angle	axis	Angle	axis
Bsoft	Phi	Z	Theta	Y	Psi	Z
IMIRS	Phi	-Z	Theta	Y	Omega	Z[1]
EMAN	Az	Z	Alt	X	Phi	Z
FREALIGN	Phi	Z	Theta	Y	Psi	Z

Table A-1. **Table of Euler angle conventions**

[1]IMIRS Omega angle requires an addition of 180° for it to be correct.

Software Package	Orientation File Format	Orientation File Type
Bsoft	View vector and angle (degrees)	Star
IMIRS	Theta, Phi, Omega (degrees)	Dat
EMAN	Alt, Az, Phi (degrees)	Lst
FREALIGN	Psi, Theta, Phi (degrees)	Par

Table A-2. **Table of orientation file formats**

Software Package	Symmetry Axis along Z-Axis	Additional Symmetry Axis for Orientation Clarification
Bsoft (I)	2-fold	5-fold axis along (0, 1, ϕ) vector
Bsoft[1] (I90)	2-fold	5-fold axis along (1, 0, ϕ) vector
IMIRS[2]	2-fold	5-fold axis along (1, 0, ϕ) vector
EMAN[3]	5-fold	2-fold axis along (0, -1, ϕ) vector
FREALIGN (I)	2-fold	5-fold axis along (0, 1, ϕ) vector
FREALIGN[4] (I2)	2-fold	5-fold axis along (1, 0, ϕ) vector

Table A-3. **Table of reference orientations**

[1]Bsoft (I90) was used for all simulations and for compatibility with other packages

[2]IMIRS uses a 5-fold axis along z internally during the reconstruction process

[3]EMAN uses a 5-fold axis along z but 2-fold axis along (1, 0, golden ratio) internally

[4]FREALIGN (I2) was used during rotavirus reconstruction

ϕ is the golden ratio which is defined as $\frac{1+\sqrt{5}}{2}$