Concurrent Computations and Data Visualization for Structure Determination of Spherical Viruses

Ioana M. Martin
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
ioana@us.ibm.com

Dan C. Marinescu
Computer Science Department
Purdue University
West Lafayette, IN 47907
dcm@cs.purdue.edu

Abstract

In this paper we address the problem of combining concurrent computations with analysis and visualization of structural biology data. We provide an overview of two methods for structure determination of spherical viruses, X-ray crystallography and electron microscopy, and discuss our efforts to design concurrent algorithms and programs for structure determination using these methods. We also present two interactive software systems we have developed which support processing of large data sets produced in structural biology experiments.

1 Introduction

Biological molecules differ in their complexity, small proteins have several thousands of atoms, whereas large macromolecules like viruses could have millions of atoms. For example a protein like streptavidin has some 7,000 atoms while the monkey tumor virus has 900,000 non-hydrogen atoms [4]. Structure determination of small molecules has become a routine process nowadays. In contrast, the determination of macromolecular structures remains a lengthy and difficult task.

To gain insight into biological processes scientists need to know where the atoms are located in biological molecules and how they interact during biochemical reactions. The atomic model of a macromolecule can be built from high-resolution (2–3 Å) electron density maps. If all computations are accurate a polypeptide chain can be traced at about 3.5 Å and individual atoms can be isolated at about 1.2 Å [4]. Structural biology uses nuclear magnetic resonance (NMR), X-ray crystallography, and electron microscopy (EM) methods to gather information about the three-dimensional (3D) atomic structure of macromolecules like proteins and viruses. NMR methods can be used to obtain 3D models of small proteins but cannot be used to obtain detailed information about the arrangement of atoms in large macromolecules. X-ray crystallography is the only method to obtain secondary and tertiary structures and processing of electron microscope images is crucial for the analysis of high resolution structures of biological molecules [4].

The 3D model of a specimen is normally represented as an electron density function sampled at the points of a regular grid. Intricate computations, often involving parallel computers, are used to refine the experimental data and produce high-resolution electron density maps. In the model building phase high-resolution electron density maps, information gathered through EM studies, and chemical information allow the structural biologist to determine the 3D atomic structure, in other words, to place atoms and groups of atoms in the electron density "clouds". Two and three-dimensional visual representations of the data are invaluable at this step.

The remaining sections of this paper are organized as follows: in §2 we describe parallel algorithms and methods for the determination of the 3D atomic structure of spherical viruses; in §3 we present two interactive tools we have developed and discuss as a case study the fitting of electron density maps of related 3D virus structures; our conclusions are summarized in §4.

2 Concurrent Computations for Structure Determination of Spherical Viruses

.

The structural biology groups we are collaborating with, led by Michael G. Rossmann and Timothy S. Baker, are interested in the analysis of viruses, virus-receptor, and virus-antibody complexes, and the study of new anti-viral compounds that interact with viral capsids, interfere with viral-receptor interaction, or inhibit uncoating. Their studies include rhinoviruses (HRV16), enteroviruses, human and animal paraviruses, and the human immunodeficiency virus (HIV) [5], [14]. Such studies require high-performance computing to analyze large amounts of experimental data and to produce atomic-level models of the viruses.

Icosahedral symmetry governs the arrangement of protein subunits within the shells of spherical viruses. An icosahedron is shown in Figure 2(a) in standard orientation: three of its two-fold axes are aligned with the axes of coordinates. Three angles define the orientation of the icosahedron: θ – rotation in the xz plane, from z towards x, ϕ – rotation in the xy plane, from x towards y, and ω (not depicted) defines the orientation about the (θ, ϕ) direction. An icosahedral asymmetric unit is shown shaded.

Structure determination aims to obtain high-resolution electron density maps by combining experimental information with a model of the virus particle. The more accurate the experimental data collected, the higher the resolution of the electron density maps that can be potentially obtained. Modern sources of X-rays, electron microscopes, and data acquisition devices such as Charge Coupled Device (CCD) detectors have increased the data acquisition rates. It is not unrealistic to expect thousands of large images (e.g., $2,000 \times 2,000$ to $4,000 \times 4,000$ pixels with 16 to 24 bit/pixel) to be collected in a few days time. Processing very large volumes of data at speeds comparable to the data acquisition rates poses distinctive challenges and can only be done by exploiting concurrency.

In this section we focus on the calculations required by the Molecular Replacement Method for phase refinement and extension using X-ray diffraction data and on particle orientation determination and 3D structure reconstruction from cryo-electron microscopy data, as shown in Figures 2(b) and (c), respectively.

2.1 Parallel Algorithms and Programs for Phase Refinement and Extension

Since the mid 50's X-ray crystallography has had a profound impact on structural biology. The discovery of the double helical structure of the DNA by James Watson and Francis Crick was influenced by the knowledge of the DNA diffraction patterns obtained by Rosalind Franklin and Maurice Wilkins. Max Perutz's studies of the hemoglobin, as well as Michael Rossmann's discovery of the atomic structure of Rhinovirus 14, were based on X-ray diffraction methods. Nowadays, increasingly more complex macromolecules and biological assemblies are investigated using new sources of intense X-radiation and modern CCD detectors.

The X-rays scattered by biological molecules infer both constructively and destructively, producing *diffraction patterns* that can be recorded using photographic emulsions or, more recently, CCD detectors. The first step in structure determination requires measuring the *structure factors*. Macromolecules produce millions of reflections while small proteins typically produce tens of thousands of reflections. Each reflection is characterized by a structure factor consisting of an *amplitude*, determined by the strength of interference at that point, and a *phase* determined by the relative time of arrival of the scattered radiation to the recording medium. The information about the phases is lost when the diffraction pattern is recorded and therefore, phases cannot be measured directly from X-ray diffraction images. Several methods to solve the so-called *phase problem* are used. In the *Heavy Atom* method an assumption is made that the phases of the diffracted X-rays are close to the phases which would be observed if only the heavy atoms were present. This approach is generally not applicable to proteins, since the heavy atom contribution to scattering is small with respect to the protein. In the *Isomorphous Replacement* and related methods phase information is retrieved by making isomorphous structural modifications to the native protein,

usually by including a heavy atom or changing the scattering strength of a heavy atom already present and then measuring the diffraction amplitudes for the native protein and each of the modified cases. If the position of the additional heavy atom or the change in its scattering strength is known then the phase of each diffracted X-ray can be determined by solving a set of simultaneous phase equations. Methods which use such a strategy are Single Isomorphous Replacement (SIR), Multiple Isomorphous Replacement (MIR), Single Isomorphous Replacement with Anomalous Scattering (SIRAS) and within the last 15 years, the Multiple-wavelength Anomalous Diffraction method (MAD).

We have developed a suite of parallel algorithms and programs [7] for phase refinment and extension based on the *Molecular Replacement* (MR) method [18] introduced by Michael G. Rossmann and David Blow in 1963. The objective of the MR method is to determine rotation and translation transformations to position a model structure into the unit cell so that the experimental diffraction data correlate best with calculated data. A low-resolution model of a virus, for instance a hollow sphere or a related virus with a known structure, provides the starting point for the MR method. The initial model is refined by taking into account the symmetry of the virus. Two types of symmetry are of interest for structure determination by molecular replacement. The *crystallographic symmetry* property implies that an operator applies throughout the whole, infinite crystal. The *crystallographic asymmetric unit* is the smallest unit from which the crystal can be generated by symmetry operations of its group. The *non-crystallographic symmetry* is related only to a localized volume within the crystal.

The MR method utilizes the similarity (or identity) of structure in different parts of the crystallographic asymmetric unit, caused by the repetition of the same subunit structure in the formation of a whole molecule [18]. Let N be the *non-crystallographic redundancy*, i.e., the number of identical replicas of the asymmetric unit that make up the whole structure. For example, in the case of icosahedral viruses, N is a multiple of 60. Solving the phase problem reduces to solving a set of equations which represent the condition that the electron density distribution within the volume of the unit cell is identical within all subunits related by crystallographic and non-crystallographic symmetry and it is constant outside these volumes. Solving these equations for the unknown phases is an iterative process, known as *phase refinement and extension*.

The basic computational procedure for phase refinement and extension is depicted in Figures 1 and 2 (c). The most intensive computations are required by the calculation of the molecular envelope of the virus. In this process the electron density value at a given grid point of the electron density map situated inside the protein shell is replaced by the average of the electron density values of all points related to the given point by non-crystallographic symmetry. Let n_x, n_y, and n_z be the dimensions in grid points of the electron density map. The total number of operations required by the electron density averaging process is $k \times N \times n_x \times n_y \times n_z$, where k is a constant. For a high-resolution map, each of the n_x, n_y, n_z values may be in the 500-800 range. The constant k may also be quite large, reflecting the fact that, once a symmetry transformation is applied to a point, the resulting coordinates may not correspond to an integral grid point and therefore some interpolation is necessary to determine the value of the electron density at that point.

The main issues in the parallelization of the electron density averaging algorithm are data distribution to minimize communication and load balancing. The entire electron density map is partitioned into 3D volumes called *bricks* which are distributed to processors to preserve locality of reference. Load balancing is non-trivial due to the fact that averaging is carried out only for points within the protein shell, whereas the density of points inside the solvent or the nucleic acid is flattened, i.e., assigned a constant value. Therefore distributing equal numbers of bricks to the processors does not guarantee load balance.

After an averaged electron density map is obtained, a 3D Fourier transformation from real to reciprocal space is carried out. Experimental and calculated structure factors are combined: experimental structure factor amplitudes replace calculated ones, while calculated phases are preserved. Next, the structure factor data is transformed back to real space using another 3D Fourier transformation to complete a cycle of refinement. When phase improvements from one iteration to the next are no longer observed, the

resolution is increased and the process continues.

For large structures, at high resolution, a single cycle of refinment requires tens of CPU-hours on the fastest sequential machines. Running on a large partition of a distributed memory MIMD system like the Intel Paragon or the IBM SP2 reduces the execution time to less than one hour. For a typical structure hundreds of such cycles are needed and parallel computing opens entirely new possibilities. Details are provided in [15] and [7] where we document our effort to reduce the computing time required for phase refinement and extension by two to three orders of magnitude.

2.2 Processing of Cryo-Electron Microscopy Data

In correlation with X-ray diffraction, biochemical, genetic, immunological, and model building studies, image processing of electron micrographs is a powerful tool for investigating the basis of molecular events in living systems. In general, this method gives structural information at low resolution, usually enough to reveal the shape of individual subunits, but rarely enough to determine the path of the polypeptide chain within a protein molecule. Efforts are currently being made to bridge the resolution gap between X-ray crystallography and electron microscopy by increasing the resolution of electron microscopy methods.

Using electron microscopy methods, scientists are routinely able to produce electron density maps at 20 Å resolution. Maps at higher resolution have been obtained for th first time in 1997. The structure of the core protein of the hepatitis B virus at 7.4 Å resolution [3] and the 4-helix bundle of the same virus at 9 Å resolution [6] were obtained by cryo-electron microscopy. Low-resolution electron density maps use a few hundred virus particle projections selected from micrographs, whereas the high-resolution maps require thousands of such projections. Clearly, a ten to fifty-fold increase in the amount of data used for 3D structure reconstruction from electron micrographs requires new, possibly parallel and distributed computational methods.

The images of individual particles in electron micrographs are approximate projections of the specimen in the direction of the electron beam. The problem of determining the specimen structure from the micrographs is equivalent to the problem of reconstructing a spatial density distribution from its projections. Fourier theory provides a simple approach to finding the 3D structure of an object from its projections. The *Projection Theorem* [8] connects the Fourier transform of the object with the transforms of its projections. It states that *the Fourier transform of the projected structure of a 3D object is equivalent to a 2D central section of the 3D Fourier transform of the object, normal to the direction of projection.* Indeed, the Fourier transform of the function $\rho(x, y, z)$ is:

$$F(X, Y, Z) = \int\int\int \rho(x, y, z) e^{2\pi i(xX+yY+zZ)} dxdydz$$

.

The central section $Z = 0$ of the transform is given by:

$$F(X, Y, 0) = \int\int \sigma(x, y) e^{2\pi i(xX+yY)} dxdy,$$
$$\text{where } \sigma(x, y) = \int \rho(x, y, z) dz.$$

Figures 1 and 2(b) illustrate the basic steps of the 3D reconstruction process. The first step is the *selection* (boxing) of individual particle images from one or several digitized micrographs. These images provide the different views needed to fill in the 3D Fourier transform of one virus particle. The number of views required at this step depends on the desired resolution of the final structure and on the size of the virus. The next step is to determine for each image its center and the orientation of the particle that produced it. Best results are usually obtained in the case of highly symmetrical particles such as icosahedral viruses because the high symmetry (see Figure 2 (a)) leads to redundancies in the Fourier transform data which in turn aids the orientation search process. The 3D Fourier transform of a particle is calculated from 2D Fourier transforms of the projected views which are equivalent to central sections of the 3D Fourier transform perpendicular to the direction of projection. Interpolation methods are required to obtain the values of the 3D transform on a regular grid. Finally, the last step is to compute the electron density function from the 3D Fourier transform by an inverse Fourier transformation.

The computationally demanding steps are (a) the determination and the refinement of the orientations of the virus particle projections and (b) the calculation of the 3D Fourier transform from values on central sections. Parallel algorithms for (a) have been implemented by Johnson & al. [11] who report that, in addition to an increase in resolution, the signal-to-noise ratio of the resulting maps is improved. We are developing a new parallel algorithm [2] for orientation determination, based on an improved sequential version of of the Polar Fourier Transform (PFT) method [1]. The goal of our algorithm is to reduce the computing time of the PFT method by as much as three orders of magnitude by taking advantage of the speed, storage capacity, I/O bandwidth, and latency of high-performance parallel systems. The basic idea of the algorithm is briefly described next.

The input to the algorithm consists of an Image Pool (IP) containing n 2D particle images whose orientations have to be determined. In addition, the algorithm uses as input a 3D electron density map which serves as a high signal-to-noise model. Such a model may be a computer-generated model, a low-resolution map previously computed, or a map corresponding to a similar, already known, structure. From this model, a Reference Database (RDB) consisting of m different views is generated. For instance, for an icosahedral particle, these views cover one half of the asymmetric unit of the structure (e.g., $1/120th$ of an icosahedron, from $\theta = 69$ to 90° and $\phi = 0$ to 32°), as shown in Figure 2 (a). The images in RDB are correlated against those in the IP to determine for each image in IP a best match in RDB. The angular coordinates of the best match thus found provide a solution for the unknown orientation of the corresponding IP image.

The execution time T_0 for one iteration of the search process is $T_0 = I \times n \times m$, where I is the time required to compare an image in IP with one in RDB. The algorithm discussed in [2] (i) uses compressed data, (ii) performs a multi-phase, multi-resolution search of the database, and (iii) supports concurrent processing of projections. Each of these elements are discussed below. (i) Compression: I can be reduced by storing and comparing compressed images instead of the original ones. The speedup due to compression is $S_C = C$, where C is the compression factor, since the time to compare two images is proportional to their size. In addition to a reduction of I, the space necessary to store the images is also considerably reduced. (ii) Multi-phase, multi-resolution search: instead of building a high-resolution RDB and trying to find in it a best match for each image in IP, we propose a three-phase scheme. In the first phase a low-resolution RDB consisting of m_1 elements is built (e.g., at 3° angular increments). The search time in this RDB is $T_1 = I \times n \times m_1$. In the second phase, a medium-resolution RDB of size m_2 is generated (e.g., at 1° angular increments). The search in this new database is restricted to a small subset of size $r_2 \times m_2$, with $r_2 \ll 1$, around the best match obtained in the previous phase. The search time in this phase is $T_2 = I \times n \times r_2 \times m_2$. In the third phase, a high-resolution RDB (e.g., at 0.3° angular increments) is generated and the search is restricted to a subset of size $r_3 \times m_3$, with $r_3 \ll 1$. The total search time for the new scheme is: $T = I \times n \times (m_1 + r_2 \times m_2 + r_3 \times m_3)$. The speedup due to multi-resolution search is:

$$S_M = \frac{T_0}{T} = \frac{1}{(m_1/m_3 + r_2 \times m_2/m_3 + r_3)} \approx \frac{1}{r_3}$$

(assuming $m = m_3$ in the calculation of T_0). (iii) Parallel search: given P processors, each could process a fraction n/P of images in IP for a speedup of $S_P = P$. The implementation of the parallel search requires solutions to two problems: work allocation and data management and distribution. Solutions to these problems are discussed in detail in [2]. The speedup due to the cumulative effect of all improvements previously enumerated is:

$$S = S_C \times S_M \times S_P = \frac{C \times P}{r_3}.$$

For instance, for $C = 4, P = 20$, and $r_3 = 0.01$, the total speedup of the search process obtained with our method is $8,000$.

For problem (b) a sequential algorithm using Fourier-Bessel inversion was proposed in [8]. However, as previously pointed out, increasing the resolution of the 3D reconstruction requires several thousands individual virus particle projections. Time and memory constraints make sequential methods unsuitable for this task. We are currently designing concurrent programs for 3D reconstruction based on a new method proposed by R. E. Lynch at al. [13]. The

input for this method consists of n particle projections and their orientations given by the angles θ, ϕ, and ω (see Figure 2 (a)). A 2D Fourier transform is applied to each individual projection. The orientation of each projection is used to determine the corresponding central section in the 3D Fourier domain (according to the Projection Theorem). Next, an interpolation procedure is used to determine the values of the electron density Fourier transform at regular grid points. In addition to parallelism, this method has a number of advantages over the method described in [8]. All computations are performed in cartesian coordinates, thus reducing their complexity. Also, the symmetry of the structure to be reconstructed is not built-in the algorithm, which means that, in principle, this method could be used in the future to investigate structures with no symmetry (e.g., nucleic acid).

3 Data Visualization and Computation Steering

Graphics plays an important role in the three stages of the structure determination: data collection, data analysis, and model building. Numerical simulations and scientific experiments produce information hard to comprehend. Image processing and data visualization help convey this information to the scientist in a form which can be better exploited by the human analytic capabilities and then further employed to steer computations. Image processing refers to any technique which alters and displays, in more tangible form, the information contained in images. In the case of structural biology, it extends the scientist's ability to study imaged biological structure because details that may be invisible to the naked eye can be clearly revealed. An obvious benefit of clearer images and structural information is an enhanced understanding of biological structure-function relationships. Data visualization is the transformation of numerical information into visual representations. The result is a simple and effective medium for analyzing complex information. The main challenge in the visualization of biological data is the sheer volume of data. A high-resolution electron density map of a large virus may contain $500 \times 500 \times 500$ grid points. Any real-time transformation of such a volume requires efficient computational algorithms and powerful graphics engines.

In this section we describe two software tools we have developed for structural biology. The first one, Emma, is an image processing tool developed around the Crosspoint method for automatic selection of spherical virus particle images from low-contrast cryo-electron micrographs. The second one, Tonitza, supports visual data exploration and various computations required during the data analysis phase. As a case study, we discuss the fitting of electron density maps of related virus structures.

3.1 Processing of Electron Micrographs

The first step in the 3D reconstruction process by means of electron microscopy is the selection of individual particle projections from electron micrographs. This step is generally performed manually and because such a task can be very tedious, mostly low-resolution reconstructions (e.g., 20 Å) of relatively small virus particles have been computed from fewer than 100 particle images.

It has been estimated (and recent results at 7–9Å resolution with Hepatitis B virus capsids [3], [6] have confirmed this estimate) that approximately 2,000 particle images are necessary for the reconstruction of a virus with a diameter of 1,000Å at 10Å resolution [19]. Hence, manual selection methods are becoming impractical. The need for computer-aided particle detection methods provided the motivation for developing Emma [16].

At high magnification, noise in electron micrographs of unstained, frozen-hydrated macromolecules is unavoidable and makes automatic detection of particle positions a challenging task. For high-resolution reconstruction work it is necessary to analyze large numbers of micrographs at speeds comparable to the data acquisition rates. An ideal automatic particle selection method must produce a reliable solution and be computationally efficient.

Template matching methods have been proposed by several groups [9], [17], [19]. They produce reasonable results only when applied to images with a good signal-to-noise ratio, i.e., formed with medium to high electron dose, and after background variations are minimized or removed. However, it is commonly

agreed that it is difficult to identify peaks in the cross-correlation maps computed from such low-dose micrographs, and peak discrimination is extremely sensitive to fluctuations of the average intensity value throughout the image.

The Crosspoint method we have developed combines traditional image processing techniques with heuristics and a new algorithm for the detection of particle centers. The time complexity of various algorithms used by this method depends linearly on the number of pixels in the digitized micrograph. The method is described in detail in [16]. Its main steps are summarized below and illustrated in Figure 3.

(a) Image Enhancement. The digitized micrograph is enhanced by histogram equalization, followed by image averaging to smooth out local fluctuations of pixel intensities. High-resolution 3D reconstructions usually include close-to-focus, i.e., low-contrast images in which the high-resolution details are not destroyed by the electron beam. Histogram equalization helps improve image contrast by redistributing the gray levels in the image more uniformly over the gray-scale range (Figures 3 (a) – (c)). The role of neighborhood averaging is to smooth out high-intensity fluctuations which tend to be sharp and scattered throughout the entire area of projected particle images (Figure 3 (d)).

(b) Particle Identification. Our particle identification algorithm consists of two phases: *marking* and *clustering*. The image is scanned horizontally in a first pass and then vertically in a second pass. A local contrast value is computed at each pixel based on information about the radius of the particles. The result of the *marking phase* (see Figure 3 (e)) is a binary image obtained by a thresholding operation. A pixel with a high local contrast value is considered to belong to a particle projection (in which case it is marked, i.e. set to 1), whereas a pixel with a low local contrast value is assumed to belong to the background (therefore left unmarked, i.e., set to 0).

In the *clustering* phase, the connected components in the binary image resulting from the marking phase are detected. These are subsequently filtered based on their size. For those components that are not rejected in the filtering process a center of mass is computed and assumed to correspond to the center of a virus particle image.

(c) Postprocessing. A particle identification method is affected by two types of errors: missed particles and false hits. The role of the post-processing phase is to attempt to improve the quality of the final solution by reducing the number of missed particles and false hits. Figure 3 (f) depicts the final solution produced by the Crosspoint method in the case of the micrograph shown in Figure 3 (a).

Emma is interactive software environment built around the Crosspoint method. In addition to automatic particle selection and refinement, it includes capabilities to decompose large images and to display sub-images, to perform various traditional image processing transforms on digitized micrographs, to select, unselect, and extract individual particles interactively, and to store particles into files. It allows for an easy composition of such transforms in random order.

3.2 Tonitza

Tonitza is an interactive package for visualization, analysis, and data manipulation for computational structural biology. It has a modular structure, consisting of several components: input/output, visualization, and computation modules accessible to the user via a unified Motif interface. The main interface style is a direct-manipulation one: operations are invoked by actions performed on the visual representations of the objects.

Objects displayed can be manipulated in various ways using the mouse or the dials. Depending on the representation, objects may be rotated, translated, scaled, and/or clipped with planes. The appearance of objects may be customized via editors.

The input/output (I/O) module is responsible for: reading/writing data files from/to disc, automatic file format recognition, and handling of I/O errors. The program accepts as input structured data produced by scientific and engineering software. It provides support for a variety of data formats used in X-ray crystallography and electron microscopy. In addition, it accepts as input image and movie files. Tonitza also facilitates the generation of new data sets, images, and movies.

Specific computations support data analysis. They may be used independently or in conjunction with the visualization module to obtain information about the contents of the data. Some of the most frequently

used computational functions supported by Tonitza are described next.

Data rotation allows computation of rotated maps by resampling the original data. This feature is useful, for example, in the case of three-dimensional electron density maps representing virus particles reconstructed by electron microscopy methods. Such particles usually exhibit a high degree of symmetry and it is important to be able to select sub-volumes for visualization based on the symmetry elements relative to the original data set (usually given in standard orientation as shown in Figure 2 (a)). This feature also allows clipping the data volume with planes in arbitrary orientations.

The *correlation and scaling of two data sets* allows the user to compare the two sets and to adjust one relative to the other by re-sampling based on the variation of a *scaling factor*. Determination of the scaling factor requires a sequence of computational steps as described in the case study presented in §3.3.

The *composite map* feature allows one to compute a linear combination of two data sets. The new map can be subsequently displayed in various representations and may give some insight about the relationship between the two maps. A special case is the *difference* of two data sets. Figure 4 (e) shows a shaded surface representation of the Ross-River virus [5] in its native form together with Fab antibody fragments attached to it. The surface of the antibodies corresponds to a map computed as the difference between data representing the complex form of the virus with antibodies attached and the native virus structure.

Tonitza supports 2D and 3D visualization of multivariate data. Examples of some of the representations available and how they may be used to investigate biological structures follow.

Planar sections allow the display of data in sections parallel to the planes of coordinates. Using the data rotation feature previously described, arbitrary sections through the data volume may also be obtained. The data in planes may be represented as a set of *contour lines* at user-defined levels, as a *continuous scale* map, by mapping data values to colors, or using both representations superimposed. The transfer function that maps a given scalar data value to a color can be interactively edited using a *colormap editor*. *Stacks* of contours plots can be created for a set of sections to allow interpretation from a 3D perspective. Figure 5 (a) shows a contour map for the Ross-River virus, with electron density contoured at three levels. The map illustrates the overall organization within the multi-layered virus structure [5]. The positions of the icosahedral two-, three-, and five-fold axes are shown. Figure 5 (b) depicts an equatorial section (through the same virus structure) as a continuous scale map. The map shows regions of membrane pinching which are suggested to be the regions of transmembrane connections [5]. The arrows in Figure 5 (b) indicate two such regions. A stack of contours for a Coxsackievirus B3 mask map is shown in Figure 5 (c). The map reveals the spatial arrangement of particles within an asymmetric unit.

Spherical sections are similar to continuous scale maps, except that in this case data values are interpolated on spheres instead of planes. Such a representation is useful when one is interested in visualizing the distribution of scalar values at various radiae within the data volume. Sweeping such sections through the entire volume may reveal particular properties which cannot be inferred from planar representations. For instance, in the case of the Ross-River virus spherical sections reveal the glycoprotein spikes as flower-like structures that project outward from the virus structure and have a hollow base [5]. Figure 5 (d) shows one such section viewed along a five-fold symmetry axis.

Isosurfaces are the 3D analog of 2D contour lines. They provide information about sets of points within a data volume that have associated a particular scalar value. In Tonitza, isosurfaces can be displayed as *wireframe* or *shaded*, for the entire data volume or for selected sub-volumes. Analysis of high-resolution features requires a fine polygonalization of the surface under study. To enable real-time spatial manipulation of such a surface we have combined several techniques. Starting with the classical Marching Cubes method [12], we have added a preprocessing step for speeding up the computation of the isosurfaces [10]: gradient vectors used for the interpolation of the surface normals are calculated once and then reused for reconturing. Figures 4(e) and (f) show a combination of shaded isosurfaces for a Ross-River virus particle. The surfaces in Figure 4(f) reveal the three-fold nature of the virus spikes and the bilobal nature of each of the spike petals [5].

A few lessons learned during the design and the im-

plementation of Tonitza are discussed next. Tonitza started out as a general interactive scientific visualization tool, running on virtually any workstation. It included commonly used 2D and 3D representations, such as contours, continuous scale maps, isosurfaces, histograms, etc. The first challenge occurred when we attempted to visualize high-resolution maps with 20 to 250 million grid points. Using memory mapping of large data files proved to be a significant step to improve the I/O performance. Next, there was the need for interactive manipulation of the objects displayed, for viewing selected volumes from arbitrary positions, and for animating individual frames. Efficient algorithms were not enough to achieve the real-time effect, given the large data sets. This motivated a major change in the design of Tonitza: portability was traded for interactivity and we redesigned the package to take full advantage of graphics hardware by using the OpenGL library. Finally, generality was traded for usefulness and ease of use in molecular modeling, which led us to the current version of Tonitza. A considerable fraction of the effort to design and implement this package as it is now has gone into the user interface, the I/O, and the computation modules which are specific to molecular modeling.

3.3 Case Study: Fitting Electron Density Maps

The study of new anti-viral compounds that interact with viral capsids is a component of the design of anti-viral drugs and requires the ability to display and manipulate in real-time native and complex 3D structures with antibodies attached. For example, cryo-electron microscopy studies showed that the structure of the Ross-River virus inferred from several crystallographic experiments was incorrect [5]. The problem was detected by comparing surface features of the Ross-River virus from two data sets with and without bound antibody fragments.

Usually, 3D electron density maps for various structures are computed separately and before a native structure and its complex form can be displayed together, one of the maps has to be "fitted" to the other. In other words, a scaling factor must be determined and applied to one of the maps to bring it to the same scale as the other map. Finding the optimal scaling factor to be used for resizing requires computing the correlation of the two data sets. The correlation procedure is iterative and consists of: defining the correlation regions within the protein (there is little interest in correlating densities within the nucleic acid core) and the scaling range, determining an optimal scaling factor that maximizes the correlation coefficient, refining it, and scaling one map relative to the other. In Tonitza this sequence of computation and visualization steps benefits from the integration of data transformations and data visualization into a single, specialized tool.

The correlation procedure, as implemented in Tonitza is described next. Let Map 1 denote an electron density map representing a native virus structure, and Map 2 an electron density map representing the same virus with antibodies attached. The centers of the two maps must coincide. Map 1 serves as a reference and Map 2 is scaled to fit Map 1. The correlation procedure used for fitting consists of three steps.

(a) Define the correlation volume. Computing the correlation of two data sets is an expensive operation when each volume consists of ten to a hundred million grid points. Due to the special structure and high-symmetry of spherical viruses, the correlation volume may be considerably reduced by taking into account only the protein regions. The average density inside the protein shell is larger than inside the nucleic acid core. This step allows the user to specify pairs of radiae which define shells in which the correlation is to be performed based on a plot of the average electron density as a function of the particle radius. Such a plot may be a 2D graph or a surface. Figure 4 (a) illustrates the the average electron density as a function of the radius for a Ross River virus reconstruction. Based on this graph, the user may interactively select two spherical annuli where the correlation is to be performed.

(b) Compute an optimal scaling factor, that is, a number for which Map 1 and Map 2 scaled correlate the best. For a quick, rough estimate, the correlation volume may be restricted even further to a few slabs (groups of planes or sections) within the previously defined shell. Figure 4 (b) shows a plot of the correlation coefficient as a function of the scaling factor. The optimal value of the scaling factor is one which maximizes the correlation coefficient. In this exam-

ple the best correlation coefficient is 0.85 for a scaling factor of approximately 0.94.

(c) Display the correlation coefficient as a function of the particle radius. Such a display allows to check the correctness of the initial correlation regions and to redefine them if needed (Figures 4 (c) and (d)). If such a refinement takes place, steps *(b)* and *(c)* are repeated iteratively until the desired accuracy is reached.

In practice, the spatial fitting of two maps is usually followed by a linear scaling of the electron density values such that both maps have the same average electron density. After the two maps are fitted, they can be displayed together and manipulated in real-time.

The result of fitting two data sets representing Ross-River electron density data with and without antibody fragments attached are illustrated in Figures 4 (e) and (f). The blue surface was computed using the native virus structure data set. This was the reference set and the other set was scaled to fit it. The orange surface represents the surface of the antibodies and was computed as a difference between the data set containing antibody information and the native data set after a scaling of the former to fit the latter. Figure 4 (f) shows a close-up view of one of the virus spikes which reveals the binding sites of the antibody fragments.

4 Conclusions

Computational structural biology relies heavily on high-performance computing and graphics to produce atomic level models of biological macromolecules such as viruses. Recent developments of data acquisition devices and experimental methods pose new challenges: processing of very large amounts of data at speeds comparable to the data collection rates, maintaining large databases of experimental data, creating new methods to extract more information from the raw data, providing integrated environments, and allowing the scientist to steer the computations efficiently.

Designing concurrent algorithms and programs is the logical choice to deal with large amounts of experimental data and intricate models in structure determination. Interactive environments which integrate data visualization and computations are needed.

The two packages presented in this paper were developed in collaboration with structural biologists at Purdue University and other centers. Emma is currently used by researchers in the Biology Department at Purdue University and Karolinska Institute in Sweden. Tonitza is used by researchers in the Biology Department at Purdue University and the National Institutes of Health.

Acknowledgments

Timothy Baker's enthusiasm and willingness to share his knowledge with others contributed significantly to our understanding of some of the computational problems in structural biology. Michael Rossmann's constructive criticism and advice were invaluable to us. We would like to thank Timothy Baker and Jodi Muckelbauer for providing the data sets for the Ross River and the Coxsackievirus B3 viruses presented in this paper. We also thank Andrea Hadfield, Holland Cheng, and David Belnap for their critical observations, and Zhongyun Zhang for helping with the data conversion. The authors express their gratitude to the anonymous reviewers for their constructive criticism.

This research has been partially supported by the National Science Foundation grants BIR-9301210 and MCB-9527131, by the Scalable I/O Initiative, by grants from the Intel Corporation, the Computational Science Alliance, Purdue Research Foundation, and a Grant-In-Aid of Research from Indiana University.

References

[1] T.S. Baker and R. H. Cheng, "A Model-Based Approach for Determining Orientations of Biological Macromolecules Imaged by Cryoelectron Microscopy", *Journal of Structural Biology*, Vol 116, No. 1, 1996, pp. 120–130.

[2] T.S. Baker, I.M. Boier Martin, and D.C. Marinescu, "A Parallel Algorithm for Determining Orientations of Biological Macromolecules Imaged by Electron Microscopy", Technical Report, Department of Computer Sciences, Purdue University, 1997, CSD-TR 97-055.

[3] B. Bottcher, S.A. Waynne, and R.A. Crowther, "Determination of the Fold of the Core Protein of Hepatitis B virus by Electron Microscopy", *Nature*, Vol. 386, 1997, pp. 88–91.

[4] C. Branden and J. Tooze, *Introduction to Protein Structure*, Garland Publishing, 1991.

[5] R. H. Cheng, R. J. Kuhn, N. H. Olson, M. G. Rossmann, H. K. Choi, T.J.Smith, and T.S. Baker, "Nucleocapsid and Glycoprotein Organization in an Enveloped Virus", *Cell*, Vol. 80, 1995, pp. 621–630.

[6] J.F. Connway, N. Cheng, A. Zlotnick, P.T. Wingfield, S.J. Stahl, and A.C. Steven, "Visualization of a 4-Helix Bundle in the Hepatitis B Virus Capsid by Cryo-Electron Microscopy", *Nature*, Vol. 386, 1997, pp. 91–94.

[7] M.A. Cornea-Hasegan, Z. Zhang, R.E. Lynch, D.C. Marinescu, A. Hadfield, J. K. Muckelbauer, S. Munshi, L. Tong, and M. G. Rossmann, "Phase Refinment and Extension by Means of Non-crystallographic Symmetry Averaging Using Parallel Computers", *Acta Crystallographica*, Vol. D51, 1995, pp. 749–759.

[8] R.A. Crowther, D.J. DeRosier, and A. Klug, "The Reconstruction of a Three-Dimensional Structure from Projections and Its Applications to Electron Microscopy", *Proceedings of the Royal Society London*, Vol. A317, 1970, pp. 319–340.

[9] M. van Heel, "Detection of Objects in Quantum-Noise-Limited Images", *Ultramicroscopy*, Vol. 8, 1982, pp. 331–342.

[10] C. Henn, M. Teschner, A. Engel, and U. Aebi, "Real-Time Isocontouring and Texture Mapping Meet New Challenges in Interactive Molecular Graphics Applications", *Journal of Structural Biology*, Vol. 116, No. 1, 1996, pp. 86–92.

[11] C.A. Jhonson, N.I. Weisenfeld, B.L. Trus, J.F. Conway, R.L. Martino, and A.C. Steven, "Orientation Determination in the 3D Reconstruction of Icosahedral Viruses Using a Parallel Computer", *Proceedings of Supercomputing '94 Conference*, 1994, pp. 550–559.

[12] W.E. Lorensen and H.E. Cline, "Marching Cubes: A High Resolution 3D Surface Algorithm", *Computer Graphics*, Vol. 21, 1987, pp. 163–169.

[13] R.E. Lynch and D.C. Marinescu, "Parallel Reconstruction of Spherical Virus Particles from Digitized Images of Entire Electron Micrographs Using Cartesian Coordinates and Fourier Analysis", Technical Report, CSD-TR 97-042, Department of Computer Sciences, Purdue University, 1997.

[14] J.K. Muckelbauer, M. Kremer, I. Minor, G. Diana, F.J. Dutko, J. Groarke, D.C. Pevear, and M.G. Rossmann, "The Structure of Coxsackievirus B3 at 3.5ÅResolution", **Structure**, Vol. 3, 1995, pp. 653–667.

[15] D.C. Marinescu, J.R. Rice, M.A. Cornea-Hasegan, R.E. Lynch, and M.G. Rossmann, "Macromolecular Electron Density Averaging on Distributed Memory MIMD Systems" *Concurrency: Practice and Experience*, Vol. 5, No. 8, 1993, pp. 635–657.

[16] I.M. Boier Martin, D.C. Marinescu, T.S. Baker, and R.E. Lynch, "Identification of Spherical Virus Particles in Digitized Images of Entire Electron Micrographs", *Journal of Structural Biology*, in press.

[17] N.H. Olson and T.S. Baker, Magnification "Calibration and the Determination of Spherical Virus Diameters Using Cryo-Microscopy", *Ultramicroscopy*, Vol. 30, 1989, pp. 281–298.

[18] M. G. Rossmann, *The Molecular Replacement Method*, Gordon & Breach, 1972.

[19] P. Thuman-Commike and W. Chiu, "Automatic Detection of Spherical Particles from Spot-Scan Electron Microscopy Images", *Journal of the Microscopy Society of America*, Vol. 1, 1995, pp. 191–201.

GLOSSARY

Structure determination: finding the spatial coordinates of all atoms in a macromolecule.

Primary structure: the sequence of amino acids in a protein.

Secondary structure: different regions of the amino acid sequence form local regular secondary structures such as alpha helices or beta strands.

Tertiary structure: is formed by packing secondary structures into compact globular units called domains.

Crystal: regular arrangement of atoms, ions, or molecules.

Unit cell: structural pattern from which a crystal is conceptually built up by its continuing translational repetition.

Asymmetric unit: part of a symmetric object from which the object can be generated by symmetry operations.

Structure factor: complex number associated with each reflection point of a diffraction pattern. It is characterized by its amplitude and phase.

Real/Reciprocal Space: the electron density domain and its Fourier transform domain.

Electron microscope: instrument that converts electron radiation scattered by a specimen into recorded images.

Electron micrograph: image obtained with the electron microscope.

Charge Coupled Device (CCD): imaging device used to collect X-ray diffraction images and micrographs.

Cryo Fixation: high-resolution electron microscopy technique in which biological specimens are rapidly frozen in a thin layer of glassy ice so as to avoid structural changes.

TONITZA

Tonitza is a tool for analysis and visualization of large structural biology data sets.

The main features of Tonitza are:
• Computations and visualization are combined under a common framework to facilitate computational steering and effective analysis of the data at various stages of the structure determination process.
• It has an open-ended design, new representations and algorithms can be easily incorporated.
• It is written in C, is portable, it runs on Unix workstations that support X-Windows, Motif, and OpenGL.
• The code is available at no cost.

EMMA

Emma is an image processing tool for electron microscopy.

The main features of Emma are:
• It implements the Crosspoint method for automatic selection of spherical virus particle images from low-contrast electron micrographs.
• The Crosspoint method [16] has a high success rate and rarely produces false hits.
• Supports for processing of large images by operating on sub-images or on compressed images.
• It is written in C, it is extensible and portable. It runs on platforms that support X-Windows and Motif.
• The code is available at no cost.

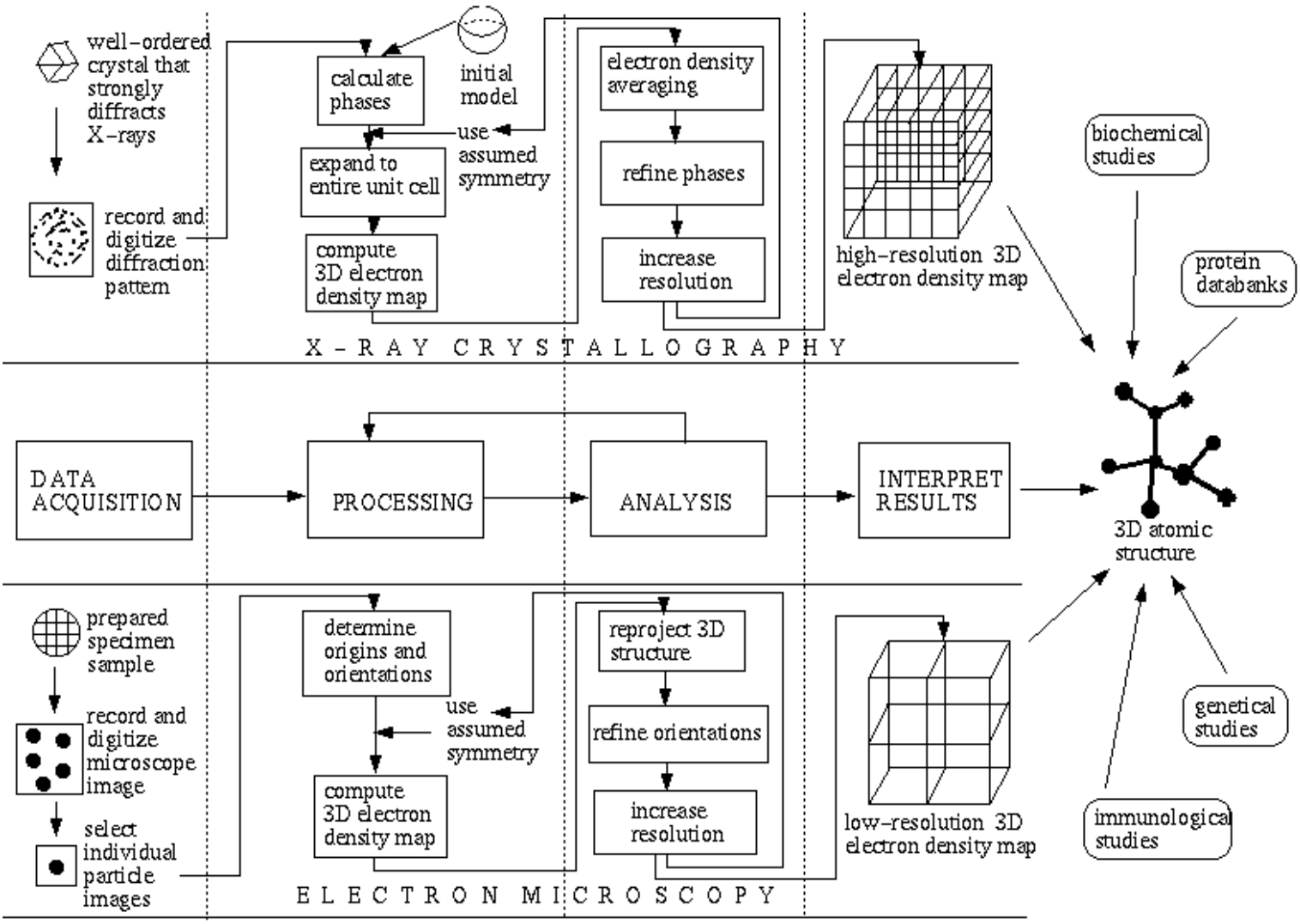

Figure 1: The main steps in 3D atomic structure determination in X-ray crystallography and electron microscopy.

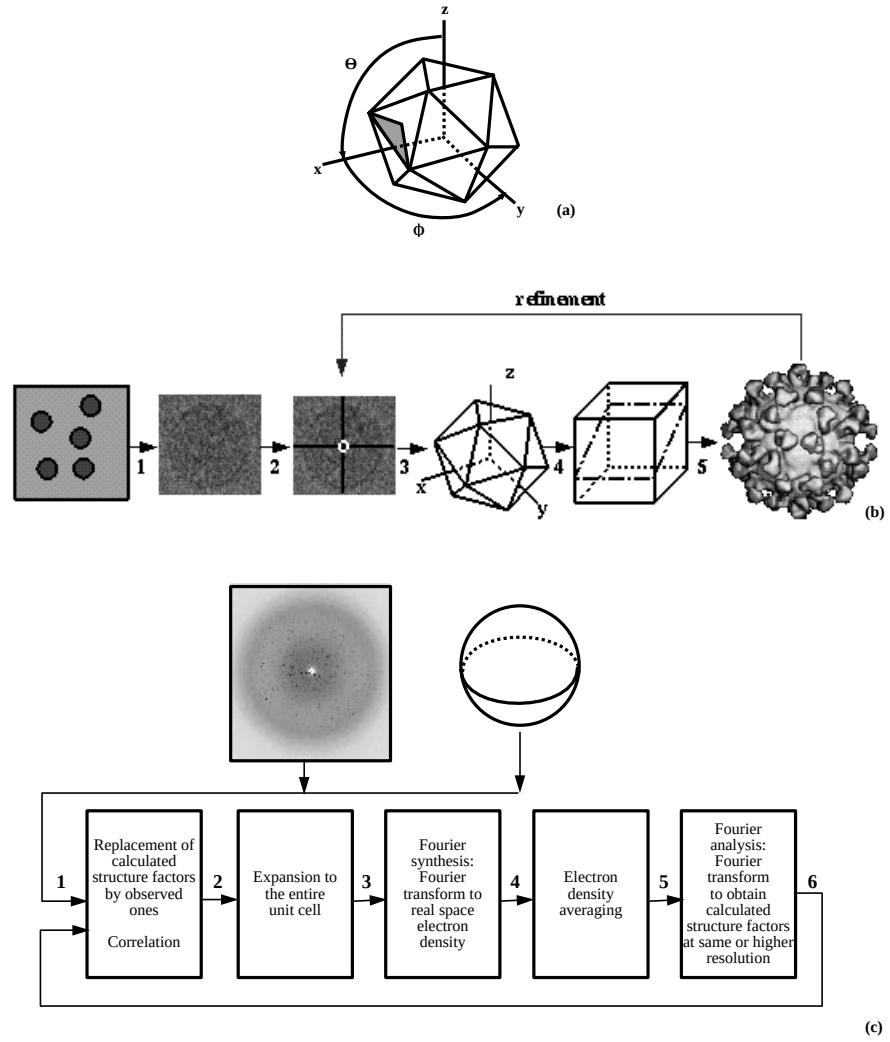

Figure 2: Structure determination of spherical viruses: (a) Icosahedral symmetry governs the arrangement of protein subunits within a spherical virus. An icosahedron is shown in standard orientation with an icosahedral asymmetric unit shaded. (b) Schematic representation of the processing steps from 2D images to the 3D structure of a virus particle by means of electron microscopy. (b1) Extract individual particle images from electron micrograph(s). (b2) Determine the location of each particle center. (b3) Determine the orientation of each particle view. (b4) Carry out 3D reconstruction in the Fourier domain. (b5) Compute 3D electron density map. (c) Schematic representation of the processing steps required by the MR method. (c1) Use X-ray diffraction data and an initial low resolution model to derive a first set of structure factors for an asymmetric unit. (c2) Apply symmetry operators to derive structure factors for the entire unit cell. (c3) Compute an electron density map from structure factor data by applying an inverse Fourier transform. (c4) Average the computed electron density among all asymmetric units to obtain a more accurate map. (c5) Obtain a new structure factor data set from the electron density map obtained in (c4) by applying a Fourier transform. (c6) Combine calculated and observed structure factors and repeat (c2)–(c6) until convergence of phases is achieved.

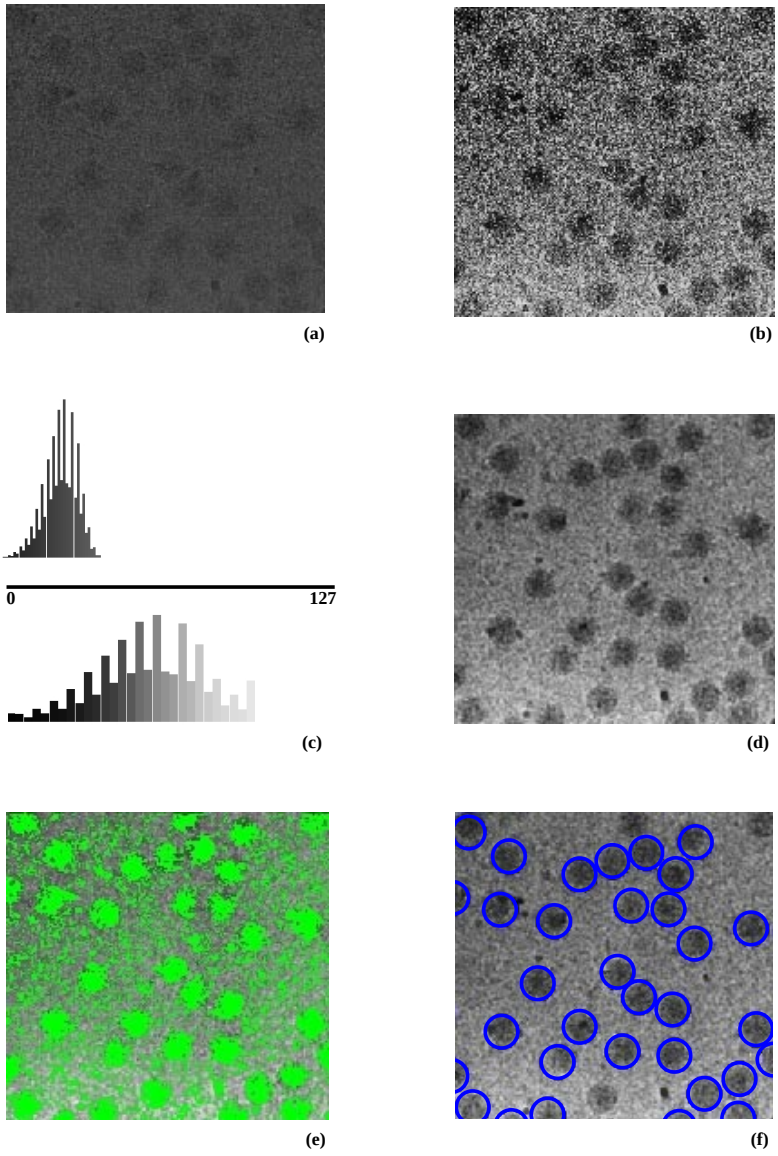

Figure 3: Automatic virus particle identification in electron micrographs using the Crosspoint method [16] (a) Portion of low-contrast micrograph of frozen-hydrated sample of reovirus cores. (b) The micrograph after histogram equalization. (c) Gray level histograms before (top) and after (bottom) histogram equalization. (d) The micrograph in (b) after neighborhood averaging with a 10×10 filter. (e) Contents of the binary image after pixel marking (green) superimposed on the micrograph in (d). (f) The result of the Crosspoint method.

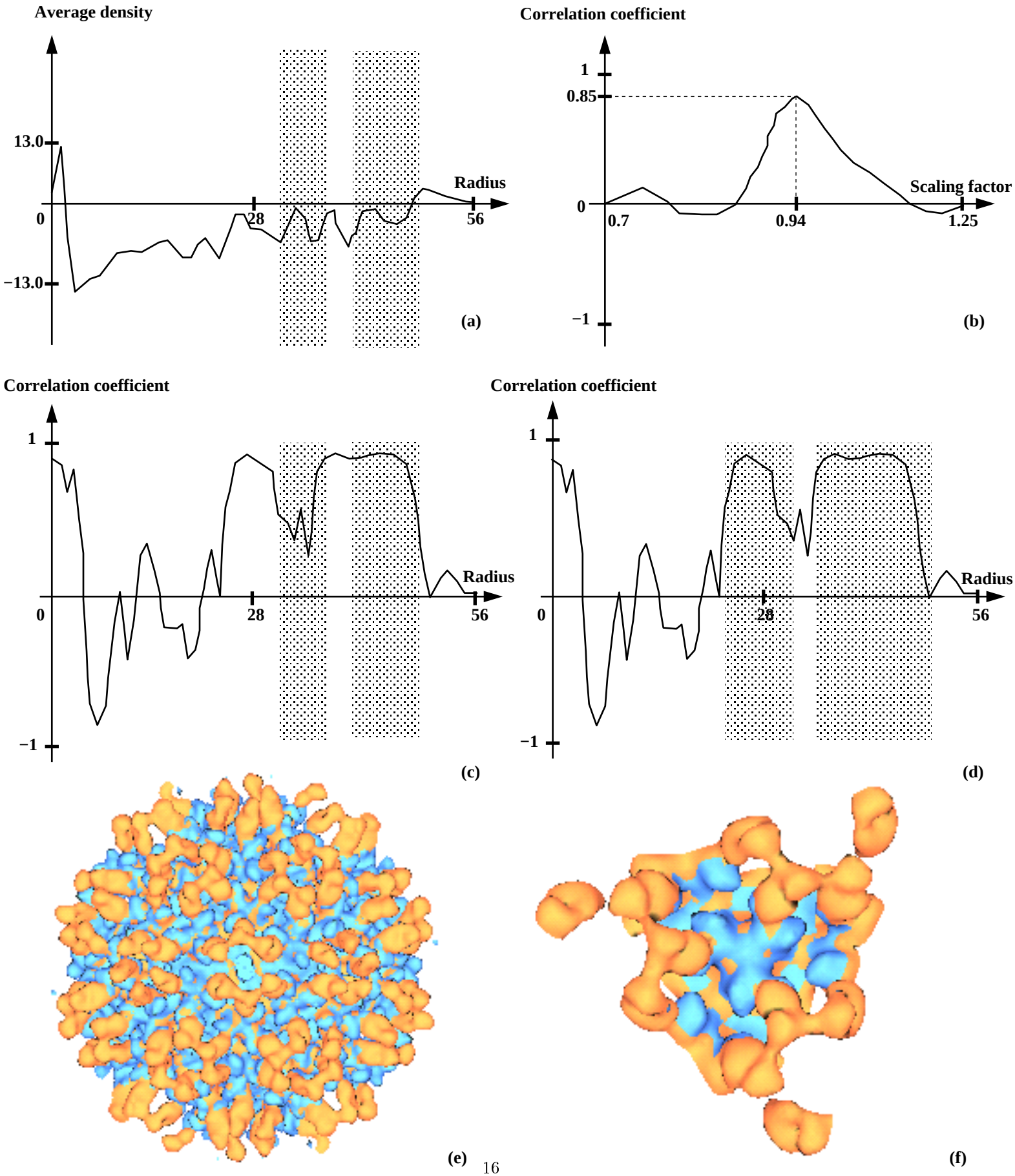

Figure 4. Application: fitting an electron density map corresponding to a native Ross−River virus structure (blue) to a separately computed map representing the virus with antibodies attached (orange). (a) Plot of the average electron density of the native virus particle as a function of its radius. Annuli where correlation is to be performed are selected at this step. (b) Plot of the correlation coefficient as a function of the scaling factor. The optimal scaling factor yields the highest correlation. (c) Using the optimal scaling factor obtained in (b), the correlation coefficient is computed and plotted against the particle radius. Original correlation annuli are shown. (d) Correlation annuli are redefined to match the regions where correlation was highest. For refinement of the optimal scaling factor, step (b) is repeated for the new annuli. (e) Fitted maps after scaling. (f) Close−up view of one of the virus spikes reveals the binding sites of the antibody fragments. Ross−River virus data courtesy of Prof. Timothy S. Baker of Purdue University.

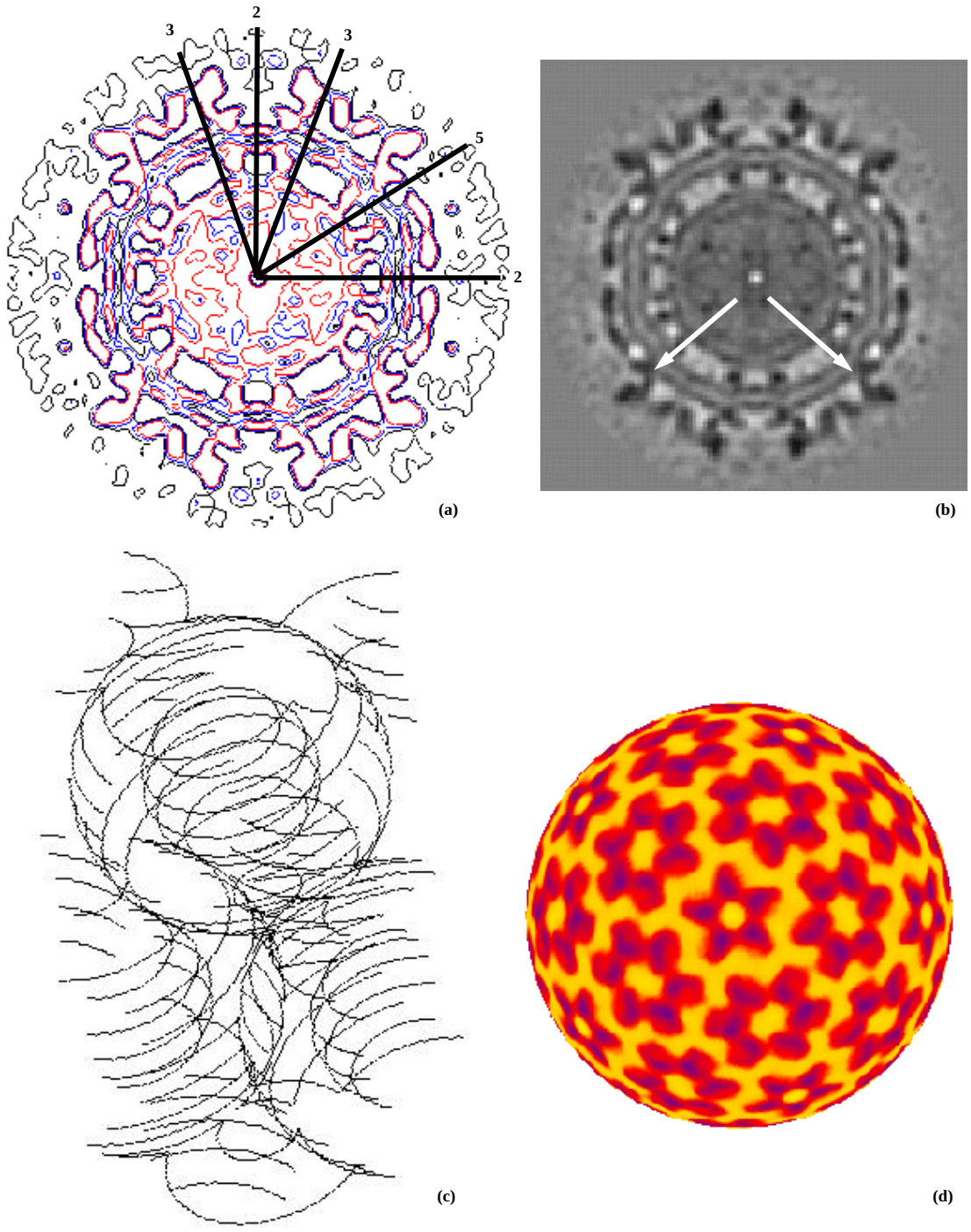

Figure 5: Using Tonitza. (a) Equatorial contour map for the Ross-River virus. Lines indicate the positions of the two-, three-, and five-fold symmetry axes. (b) Equatorial continuous scale map for the Ross-River virus. Arrows indicate regions of membrane pinching. (c) Stack of mask contours for a Coxsackievirus B3 asymmetric unit. (d) Spherical section of Ross-River virus viewed along five-fold axis.