TR&D #2: Software for Interactive Analysis of Large Molecular Assemblies

Research Strategy

Lead investigator: Thomas Goddard
Team members: Eric Pettersen, Greg Couch, Elaine Meng
Related driving biomedical projects:
DBP#1: Modelling Macromolecular Assemblies - Andrej Sali
DBP#4: HIV Accessory and Regulator Complexes – Yifan Cheng and Nevan Krogan
DBP#5: 3D Architecture of Hair Cell Stereocilia - Manfred Auer
DBP#6: Modelling Biological Assemblies from cryoEM Maps - Wah Chiu
DBP#7: Data Validation and Web Visualization at the EMDB and PDB – Helen Berman, Cathy Lawson and John Westbrook

Background and Rationale

This project develops software used by researchers who determine the architecture and operation of large molecular assemblies and those who study larger scale features such as cell, tissue and microbial community structures. The scope covers those structures observed by electron microscopy (EM), and primarily serves researchers analyzing EM data. Within this large range of length scales most of our developments concentrate on molecular assemblies composed of tens to thousands of macromolecules such as viruses and the thousands of cellular machines (e.g. ribosomes, polymerases for DNA/RNA transcription and copying, chaperones and proteasomes for protein folding and degradation) that orchestrate all functions of cells. Our software, an integral part of UCSF Chimera, provides interactive visualization and analysis of density map data and molecular models. The DBPs driving the technology development with this Core are CryoEM of Nanomachines (Wah Chiu), Tomography of Subcellular Structures (Manfred Auer), HIV Accessory and Regulatory Complexes (Yifan Cheng), EM Model Validation (Cathy Lawson), and Integrative Multi-Resolution Modeling (Andrej Sali).

This has been a core research and development project of the RBVI for the past 10 years, and the software is used by most of the labs world-wide that carry out structural biology research on molecular assemblies. (See letters of support from Michael Rossmann, Jack Johnson, Ken Downing, Pawel Penczek, Ed Egelman, David Mastronarde, and Chuck Sindelar). The primary experimental technique to determine structures of assemblies is to dock X-ray and NMR models of molecular components into electron microscopy density maps. Many other experimental and computational methods contribute, including small-angle x-ray scattering, genome sequencing, homology modeling, mass-spectrometry, and a variety of protein interaction assays (affinity purification, cross-linking, two-hybrid screening). It is well known that the majority of proteins and nucleic acids in cells and viruses act as parts of assemblies. Knowledge of genomic sequences and protein structures has become remarkably rich in the past decade but we are only beginning to understand the next functional level, the catalog and rules of operation of molecular assemblies. This is essential to understanding how cells function, misfunction (e.g. cancer, Alzheimer's disease), and are invaded by pathogens such as HIV and malaria.

We divide this project into four major software development areas: analysis and visualization tools for 3 types of data and technology for communicating the results:

1) Analyzing atomic models of molecular assemblies.
2) Analyzing single-particle EM maps.
3) Analyzing structurally heterogeneous assemblies.
4) Enabling researchers to share and archive analysis (models and maps).

These four categories are the same as in past descriptions of this core project, with the exception of 4) which formerly was limited to software development for producing animations and now encompasses the wider area of technology that enables data exchange. The four areas are tightly interlinked and cover analysis of molecular assemblies over 3 resolution ranges and presentation of the results (models, maps, segmentations,

images, and animations). A researcher studying a specific biological machine (e.g. the ribosome) uses Chimera tools covering all of these areas. The categories in this proposal are introduced as an aid to understanding the many software components we develop for researchers studying molecular assemblies. We will describe the key components of each area.

Atomic Models of Molecular Assemblies

Chimera provides a large collection of tools for analyzing atomic models of symmetric molecular assemblies such as icosahedral viruses and helical filaments (actin, myosin, microtubules, amyloids, flagella), and large asymmetrical complexes such as ribosomes and RNA polymerases. The assemblies typically consist of tens or hundreds of macromolecules. To visualize and analyze such systems methods for coarse-grain depiction of proteins and nucleic acids are essential. At the same time it should be possible to select small subsets of contacting molecules in order to study the interfaces at atomic resolution, for example, which residues are making contacts. Most assemblies accessible to current experimental methods are symmetric and are represented by atomic coordinates for just the asymmetric unit, together with a specification of symmetry. Analysis software must be able to apply the symmetries to allow studying interactions between neighboring asymmetric units. These three essential capabilities: coarse-grain depiction, multi-resolution representation, and applying symmetry, are functions of the central Chimera tool, called Multiscale Models (1). It is used to produce the illustrations of viruses at the Protein Data Bank (e.g. rotavirus 3IYU) and Virus Particle Explorer database (ViPERdb).

Density Maps from Single-Particle Electron Microscopy

The single-particle electron microscopy produces 3-dimensional density maps of molecular assemblies at resolutions ranging from 3 to 30 Angstroms. Only a very few maps have been solved at resolutions of 3 to 4 Angstroms where it is possible to trace protein backbones and build atomic models directly as is done in x-ray crystallography. At lower resolution, atomic models from crystallography or NMR, or homology models derived from those, are fit into the single-particle maps to obtain atomic models of the imaged molecular assembly. Chimera is used for displaying the maps, fitting atomic models into maps, and analyzing the resulting models.

Structurally Heterogeneous Assemblies

The molecular assemblies studied by single-particle electron microscopy and x-ray crystallography have rigid and highly reproducible shapes. Those techniques achieve high resolution and high signal to noise ratio by averaging many copies of identical molecular arrangements. At larger sizes, molecular assemblies invariably become heterogeneous in structure. For example, no two influenza virus or HIV virus particles are the same. They are surrounded by flexible lipid bilayers, with variable numbers of protruding spikes, not arranged with any set symmetry, and the interior contents of proteins and packaged RNA genome (8 segments for influenza and 2 for HIV) are not precisely ordered. Within cells some assemblies comprising hundreds of proteins such as nuclear pores, basal bodies, carboxysomes and clathrin cages lie at the boundary of having substantial structural order and a diversity of forms, while at larger scales mitochondria, transcription factories, chromosomes, a host of organelles are highly polymorphic. This area of our molecular assemblies proposal focuses on analysis and visualization of all such heterogeneous assemblies.

Communicating Analysis Results

Research results are communicated through journal publications using images and animations, and also via public databases where atomic models and electron microscopy density maps are available as starting points for future research. We have focused on developing software to enable researchers to create animations showing experimental data, models and how molecular assemblies function. Creating animations is a complex task with a difficult learning curve that few biology researchers have the time to master. Our developments have aimed to make it easy to create simple animations, and feasible to create complex animations. Articles, images and animations distill the insights of research. Equally valuable are the computational results: maps, atomic models, structural and sequence alignments, symmetry parameters, lists of interacting residues – computer-readable data that serves as building blocks for future research. The vast majority of published computational results on molecular assemblies using electron microscopy are not publically available. This is

a fundamental obstacle to progress in the field. While we plan to continue work on animation tools, our efforts in coming years will focus on improving technology for data sharing.

Subsequent sections of this core proposal will describe progress during the past funding cycle, and the significance and details of our future plans, and the relation to our driving biomedical projects.

Progress in Last Grant Period

The value of our Chimera software lies in the synergistic use of its many visualization, analysis and communication facilities. We'll describe several advances in these capabilities achieved in the last grant period for each of the four main areas.

Atomic Models of Molecular Assemblies

In the background section we described three fundamental software capabilities for studying large atomic models implemented in the Chimera Multiscale Model tool: coarse-grain depiction, multi-resolution representation, and symmetry handling. Beyond these prerequisites many additional capabilities support the analysis and enhance visualization. In the current funding period our effort has focused on building up additional capabilities.

Some of the added features support creating symmetric models from different types of symmetry specifications. We developed the sym command that handles cyclic, dihedral, helical, icosahedral and translational symmetries. It is useful in cases where the researcher has derived the symmetry from an electron microscopy reconstruction has fit an asymmetric unit into the experimental map, and then wishes to construct the symmetric assembly. Figure 2-1A shows another use where a chromatin filament was built from an x-ray model that contains just 4 stacked nucleosomes, the defining helical parameters being deduced by measuring the rotation and translation between nucleosomes. We also enhanced capabilities to apply crystallographic space group symmetries, to allow constructing several adjacent unit cells. Crystallized viruses often span multiple. Sometimes one unit cell is sufficient to show the interactions of interest, but the standard space-group symmetries do not pack the molecules tightly. We added the ability to apply crystal translation symmetries to pack the molecules so their centers all like in one unit cell box. Also the crystal symmetry can now be read from mmCIF and CIF file formats.

Many new capabilities provide analysis of assemblies. For crystallized viruses we added a command to find all contacting residues at all unique interfaces between viruses in the crystal. These are areas of possible distortion due to crystal packing forces. We implemented this for ViPERdb to allow them to report crystal contacts for all their archived structures, and also the PDB uses it to validate new submissions, checking for

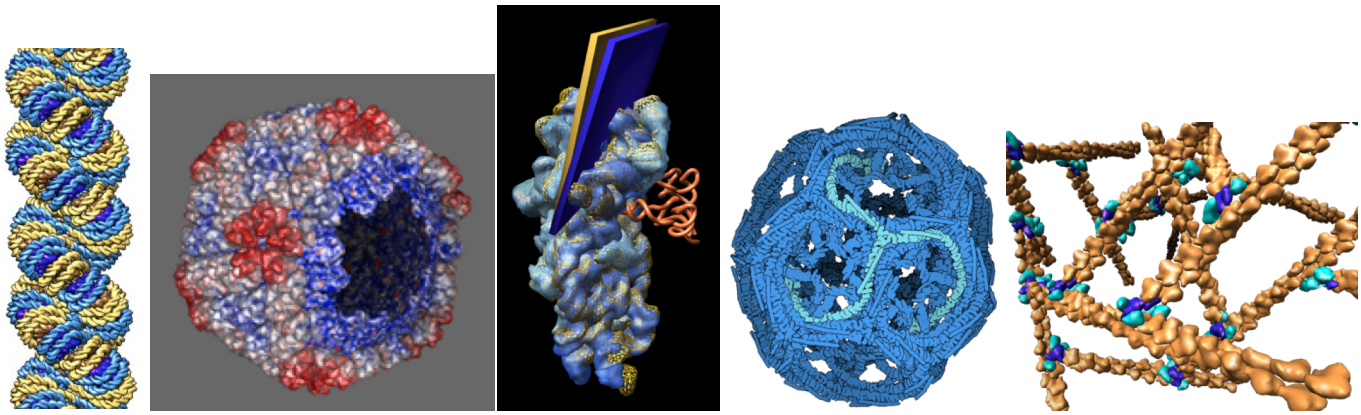


Figure 2-1: Illustrating advances during the last grant period. A) Filament made using sym command from nucleosome model 1zbb. B) Coloring by quantitative properties, here b-factor of x-ray structure asymmetric unit. C) Measuring and depicting domain rotation of small ribosomal RNA on tRNA binding. D) Flat shading for Goodsell-like images. E) A basic "composite model" made with a script from ARP2/3 junction complex and actin filament.

steric clashes. We implemented a more general tool called Find Clashes and Contacts that visually displays lines between interacting atoms and can optionally list the atoms and distances involved. This is valuable when fitting atomic models in electron microscopy maps to assess the collisions and complementarity between adjoining molecules. We added a command to measure the buried surface area between adjoining molecules, a basic indicator of the strength of interaction. We added a Coulombic electrostatic potential calculation useful for evaluating charge interactions at interfaces between molecules, and we can read more accurate solvent-screened potential calculations using Poisson-Boltzmann solvers such as APBS or DelPhi.

Several new developments improve visualization of assemblies. For attributes of atoms such as charge or b-factor we now support coloring symmetric assemblies using the attribute values associated with one asymmetric unit (Figure 2-1B). The symmetry is used not only to create surface depictions but also to propagate colors to those surfaces.. Many assemblies undergo large conformational changes. We've added a tool called Morph Conformations that interpolates between conformations. By using the time dimension this provides a clearer illustration of differences than a static superposition. It uses internal coordinates to preserve bond lengths and angles and reduce the incidence of parts of molecules crashing through each other. We added a schematic slab depiction of domain motions (Figure 2-1C). The standard 3-dimensional lighting with specular highlights works best at conveying depth when scenes have low complexity. With hundreds of surface bumps reflecting light, that method becomes much less effective, and flat shading with silhouette edging can be more effective (Figure 2-1D) as demonstrated by David Goodsell's illustrations (Figure 2-11). We added support for flat shading. For nucleic acids we created the ability to display bases as rods and rectangular slabs (2). This allows showing base pairing without requiring a full-atom depiction, especially useful in assemblies with large RNA molecules. We published a survey of ways of depicting molecular assemblies titled "Visualization software for molecular assemblies" (3).

Analysis of Density Maps from Single-Particle Electron Microscopy

We have added many features to the Chimera package to assist the analysis of EM maps. We described some of these in an article "Visualizing Density Maps with UCSF Chimera" (3). We'll describe below more recent features, including segmentation and fitting, measurement and visualization, and support for the complementary experimental approach of small-angle x-ray scattering (SAXS) which can help validate models derived from electron microscopy.

We added a segmentation and fitting method called Segger into Chimera (Figure 2-2A,B). It partitions a map into regions and allows fitting atomic models into chosen regions. Segmentation serves two important roles: it simplifies fitting atomic models by defining approximate boundaries between the molecules, and it can delineate regions of density for which no atomic model is available. The Segger method computes watershed regions. These are the neighborhoods of local density maxima, each region being the points that would reach the same maximum following a steepest-ascent walk. Successively smoothed copies of the map are used to coalesce the watershed regions into larger regions comparable to the expected size of the proteins composing the molecular assembly. This method is simple and fast, taking about 15 seconds on a desktop computer for the 255^3 Rous sarcoma virus map shown in Figure 2-2A. While the automated calculation of segmentation regions often comes close to correctly defining the molecular boundaries, an essential feature of Segger is the ability to manually split and combine the regions using mouse clicks. The regions can then be used to fit atomic models to the map. In the simplest case a region and an atomic model are selected and the fit is done by computing the principal axes of inertia of the density region and the atomic model and aligning those axes. There are 4 alignments considered allowing for flipping any 2 axes and each is used as a starting point for local correlation optimization and the best result is shown. This takes about one second in most cases allowing many fits to be explored. Fits can be made to a combined group of regions or fits to each of a number of regions can be done as a batch. When the correct molecule localization is very uncertain a powerful search capability can find all subsets of adjoining regions of approximately the size of the atomic model to be fit, and attempt fits to all such subsets reporting the best fits. In this usage the segmented regions are typically made to be half or one-third of the protein size to provide a good sampling of positions to fit. Segger was developed as a Chimera plug-in by Greg Pintilie at MIT, and subsequently included in the standard Chimera distribution by RBVI staff, much of it being rewritten to simplify the user interface and improve speed and reliability. We

published a detailed description of the method (4) and a video demonstrating how to use it is available at http://www.cgl.ucsf.edu/chimera/videodoc/segfit/.

We added a unique tool called Morph Map that can be used when maps are available for more than one conformation of a molecular assembly (Figure 2-2C). It does a simple linear interpolation between maps so that the differences are readily seen as an animation shown in real-time in Chimera. This kind of temporal depiction provides a remarkably clearer view of differences between maps than simply superimposing the two maps, where the maps invariably occlude each other. Morph Map was created by Wei Zhang with Pawel Penczek as a plug-in and later included in the standard Chimera distribution after being rewritten by RBVI staff to improve the user interface, speed and reliability.

We added a basic new capability to compute simulated maps from atomic models using a Gaussian density distribution for each atom (Figure 2-2D). The residual error between simulated and experimental maps expressed as a cross-correlation coefficient is the most commonly used measure of goodness-of-fit and Chimera now calculates and does rigid-body local optimization using that value.

A common map measurement problem is to determine positions of symmetry axes (Figure 2-2E), calculated by fitting a map to a rotated copy of itself. The axis is shown, and its direction and a point on the axis reported for use in subsequent operations, such as copying the placement of a monomer atomic model to generate the symmetrical arrangement. In cases where a map does not have exact symmetry, a 2-dimensional plot of self-correlation of the map as it is rotated around the symmetry axis can reveal the degree of symmetry (e.g. 6-fold). More commonly locating the center of symmetry is needed for single-particle maps where an exact symmetry was imposed, but due to limitations in existing map file formats, the center of symmetry is uncertain. This analysis can detect small errors in the center of symmetry of 1 or ½ grid spacing.

Some new and unusual ways to visualize maps have been added. A large proportion of single-particle EM studies examine icosahedral viruses, almost all of these forming lattices composed of 12 pentamers and a variable number of hexamers. The arrangement of the hexamers can be highlighted by drawing a cage

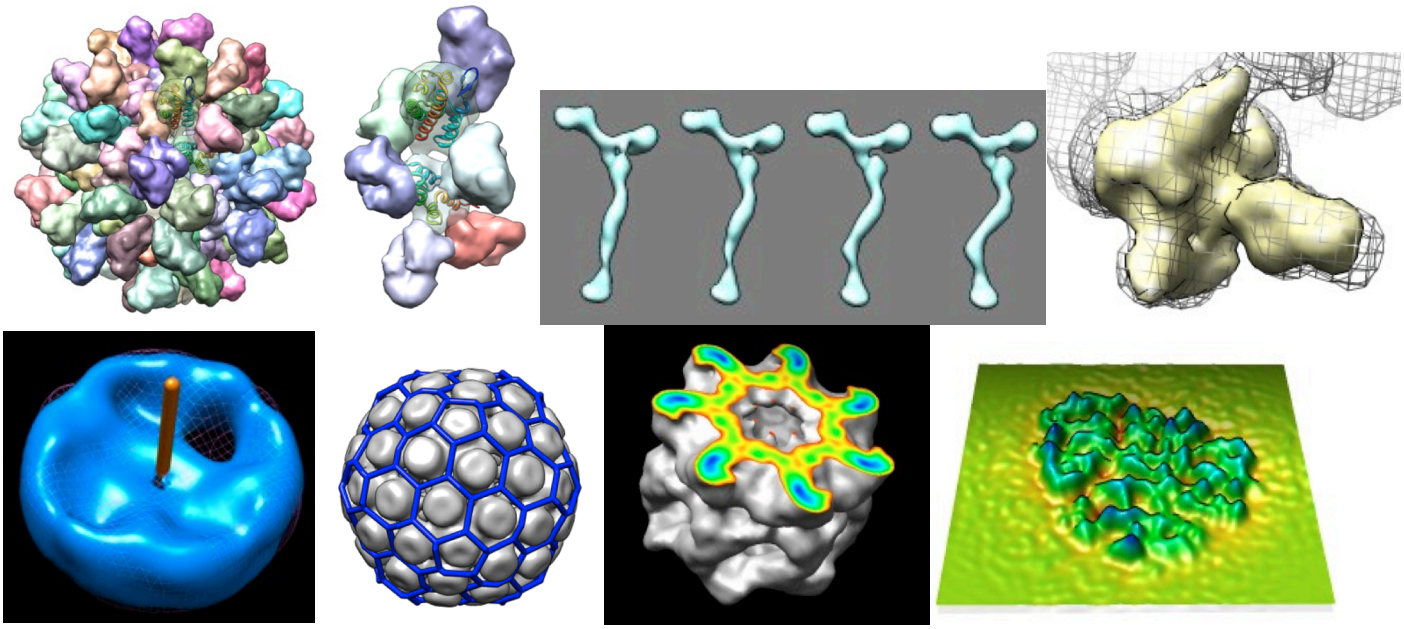
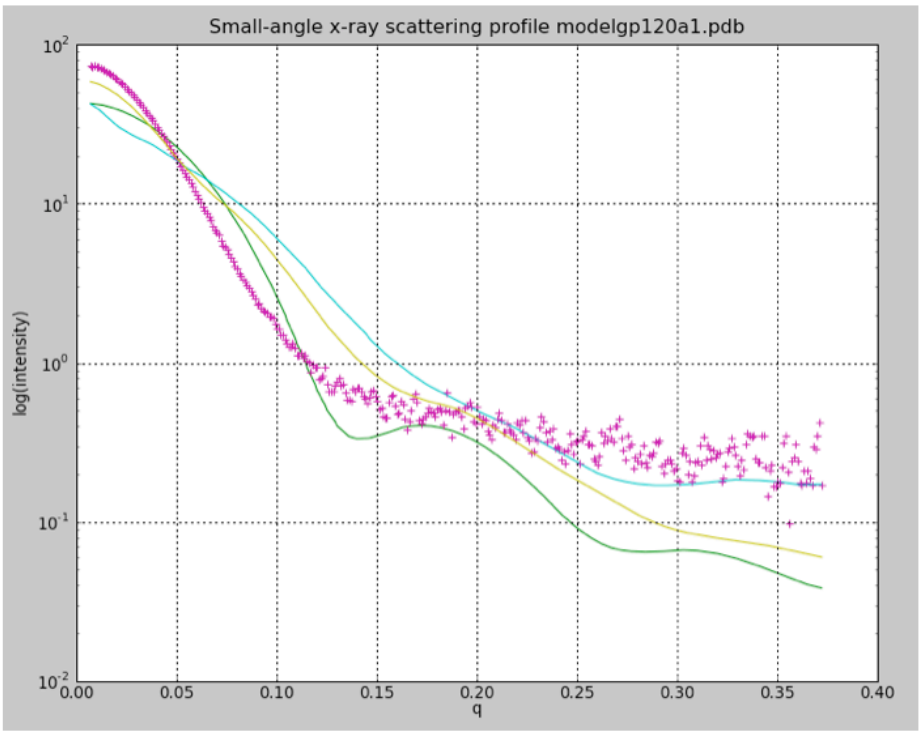
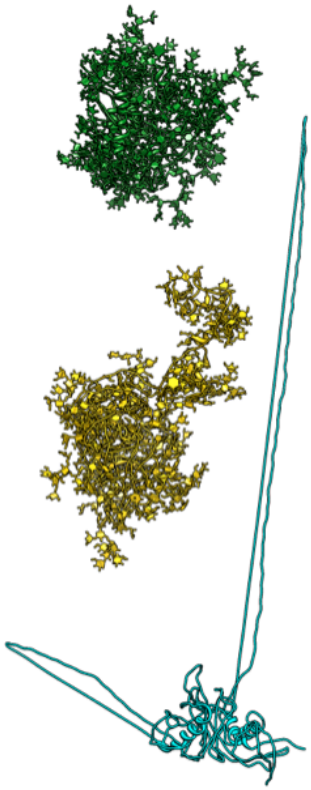


Figure 2-2: Some new Chimera map analysis tools. A) Segger (Rous Sarcoma virus, EMDB 1862) segmentation, B) fitting to segments, C) morph map (nup84, 6 protein nuclear pore subassembly in straight and kinked conformations, EMDB 1571), D) fitting using simulated maps from atomic models, E) symmetry axis calculation of 8-fold symmetric nuclear pore, F) icosahedral lattice cage for Simian Virus 40 illustrates arrangement of pentamers and hexamers, G) slices colored by density value (GroEL EMDB 1080), H) slice density shown as topography (ribosome EMDB 1007).



SAXS profiles. Experimental gp120 SAX data pink + marks, glycosylated model green, modbase homology model cyan, glycosylated model with moved v1/v2 loop yellow.



Alternate V1/V2 loop models. Colors match calculated curves in SAXS profile plot.

Figure 2-3: Small-angle x-ray scattering (SAXS) profiles calculated for 3 hypothetical HIV gp120 spike conformations, differing in placement of the large V1 and V2 variable loops, and compared to experimental gp120 SAX profile (pink + markers)

specified by two integer lattice parameters (5). This was developed for the Virus Particle Explorer database and is used to illustrate all of their EM maps. Two techniques for visualizing slices of maps were added, one that simply colors the slice to indicate density values (Figure 2-2G), and another that depicts density values as height above the slice in a topographic representation (Figure 2-2H). The latter style was developed for visualizing atomic force microscopy data where the image values actually do correlate with height of a scanned surface, but it is also useful and simple to apply to slices of maps from electron microscopy.

Validation of atomic models derived from single-particle reconstructions is a critical problem and a first meeting of an Electron Microscopy Validation Task Force (http://vtf.emdatabank.org) sponsored by the EM Databank was recently held. The central issue is how to assess the degree of confidence we have that map reconstructions and models derived from fitting electron microscopy data are correct. Basic requirements are that the analysis steps and standard quality metrics should be reported and archived. We discuss those requirements further in the section on communicating analysis results. Ultimately the limited resolution available in many electron microscopy studies places limits on how confident we can be about models derived from fitting into those maps. Complementary experimental information will be necessary to validate the models. One of the most promising methods for validating molecular assembly models is small-angle x-ray scattering (6). SAXS data is a one-dimensional curve quantifying rotationally averaged scattering. It provides a strong constraint on the shape of an assembly and is relatively simple to perform. The main current limitation is that a synchrotron radiation source is typically used. We have added to Chimera the ability to compute simulated SAXS profiles for atomic models and compare those to experimental SAXS profiles (Figure 2-3). This was a collaborative effort with Dina Schneidman in Andrej Sali's lab and uses their Fast X-Ray Scattering (FoXS) program (7) to compute the profiles. The RBVI developed a user interface to easily compare multiple

conformations of a molecule or assembly against experimental SAXS data. The computation is currently done by a web service hosted on RBVI servers.

Analysis of Structurally Heterogeneous Assemblies

Our software tools for analysis of heterogeneous structures focus on experimental data from electron tomography and new serial scanning electron microscopy (SEM) techniques: focused ion-beam scanning EM (FIBSEM) (8) and serial block-face scanning EM (9). Tomography reconstructs a 3-dimensional density map from a series of tilted images, typically over a range of tilt angles of +/- 60 degrees in 1 degree steps. Two important limitations are that the sample cannot be much thicker than half a micron because the electron beam must pass through the sample, and prominent anisotropic artifacts results from the fact the limited range of tilt angles because the flat wide sample cannot be imaged edge-on. The serial techniques work my milling away thin layers, typically 10 to 50 nm thick, and imaging the surface of the sample after each layer is removed. They can image whole cells tens of microns in thickness, but have lower resolution in the z-dimension than in x and y, and overall lower resolution than obtained with tomography. Advances in the serial sectioning techniques have the exciting potential of determining complete neural wiring in a brain at electron microscopy resolutions (10) (11). Other technologies such as X-ray tomography and super-resolution optical microscopy techniques also produce density maps that can be analyzed with our software, though our developments focus on the higher resolutions obtained with electron microscopy.

For most specimens the available imaging technologies do not resolve individual proteins without averaging equivalent structures. Consequently the goal is usually to build coarse-grain models, rather than atomic resolution assemblies. We have been able to leverage the Chimera capabilities for display of atomic models in creating and analyzing coarse-grain models. Atoms and bonds are redefined as markers and connectivity for tracing filaments, or representing whole assemblies such as vesicles or ribosomes seen in cellular tomography using a single pseudo-atom. The extensive surface display, coloring and measurement Chimera tools for single-particle analysis also are reused for surface representations of coarse-grain objects. All of the single-particle EM map tools work equally well on maps from tomography or serial SEM. A central reason to extend Chimera capabilities to lower than atomic resolutions is that much research spans multiple resolutions. A bridging technology we are especially interested in called subtomogram averaging (an equivalent term is single-particle tomography) extracts identical substructures such as HIV virus spikes from a structurally heterogeneous object and averages those to create molecular models.

In recent years we have added many tools to Chimera assist data exploration of tomography and SEM maps. A distinctive feature of these maps is that the signal to noise ratio is very low. We have added a collection of fast filtering methods: median filtering (preserves edges), Gaussian filtering, binning, density value inversion, Fourier transform, Laplacian filtering, and field flattening. A popular and unique method we introduced called Hide Dust hides noise without altering the map (Figure 2-4A). It hides connected specks of density if their size is less than a specified value. The value is adjustable with a slider and the display updates in real-time. Because of the high noise levels it is common to view these maps plane by plane with gray scale density levels. We implemented a highly optimized plane display to show the large map sizes (e.g. 4192 by 4192 by 200) that allows flipping through planes at many frames per second. Because objects are generally not aligned with image box axes, we allow extracting sub-boxes that needn't be aligned with the original data axes (using data resampling). For the same reason we incorporated an oblique plane slicing tool called TomoPlane developed by Karen Gross and Christoph Best at MPI Munich, and then later replaced it with a next-generation faster oblique slicing tool.

New coarse modeling tools allow placing markers on data planes and surfaces, tracing contours in planes and joining them to form surfaces (e.g. delineating membranes), placing geometric shapes such as ellipsoids, icosahedra, smooth splined tubes (Figure 2-4C,D). Surface models produced by the popular IMOD analysis software can be imported (Figure 2-4B). Measurement of distances between objects, contact surface areas, and principal inertia axes are now supported. Density within a surface bounding an object can be extracted into a separate map with the new mask command, and density in a slab centered on a membrane can also be extracted.

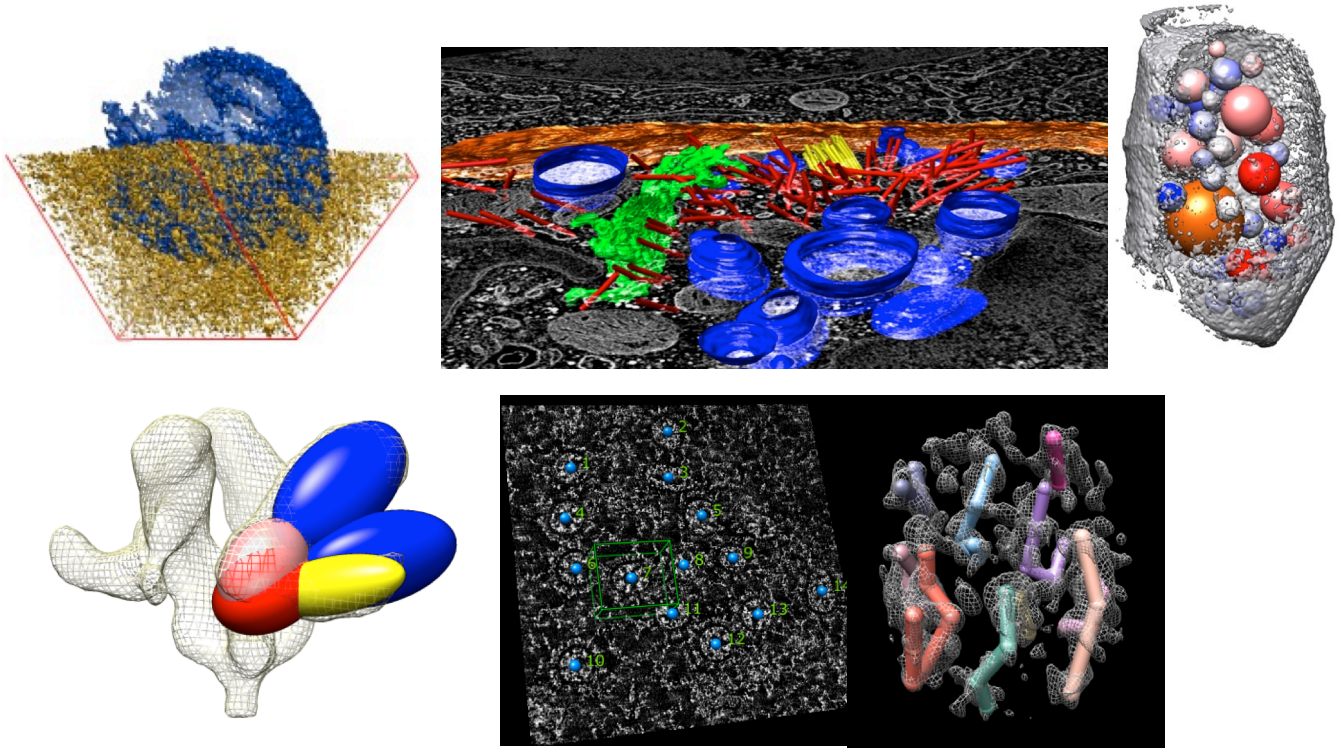


Figure 2-4: Representative data and coarse modeling for heterogeneous structures. A) An SIV virus particle from tomography with noise removed with the hide dust Chimera tool. B) Tomography of a human cytotoxic T-cell, single data slice shown with microtubules, lysosomes, golgi, centriole, and cell wall segmentations imported from IMOD. C) X-ray tomography of a Schizosaccharomyces *pombe* with nucleus (orange) and vesicles modeled as spheres with coloring (red to blue) indicating interior vesicle density. D) HIV spike from subtomogram averaging with ellipsoids for gp120 (red), CD4 (yellow) and a FAB (blue), orientations with and without CD4 bound compared. E) Tomography of influenza virus with traces of 8 RNA sausages in genome of one virus particle.

We've developed a number of more sophisticated capabilities as part of a collaborative project with Bernhard Knierim, Monica Lin and Manfed Auer studying the microbial community in termite hindgut for clues about which bacteria among the 200 species present might have value for producing biofuels. Several FIBSEM data sets, each showing about 500 bacteria and lignocellulose feedstock, were segmented using a newly developed Chimera segmentation method using watershed and flood-fill techniques (Figure 2-5). The segmentation is done by clicking on a bacterium and dragging the mouse to color it by flood-filling to adjacent watershed regions with threshold level controlled by the mouse drag. If strong contacts with a neighboring bacterium cause it to be colored, a higher threshold is used to avoid the incorrect connection, the neighbor bacterium is

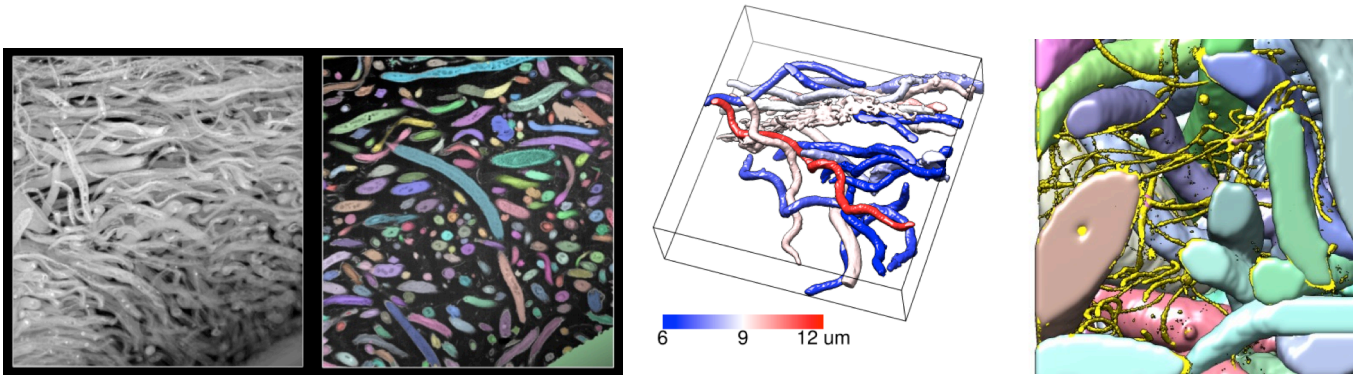


Figure 2-5: Analysis of termite hindgut microbial community. A) FIBSEM map, 9 by 9.6 by 2.4 microns, containing about 500 bacteria. B) Segmented bacteria, randomly colored, single plane shown. C) Several segmented bacteria colored by length. D) Flagella (yellow) in spaces between bacteria (multiple colors).

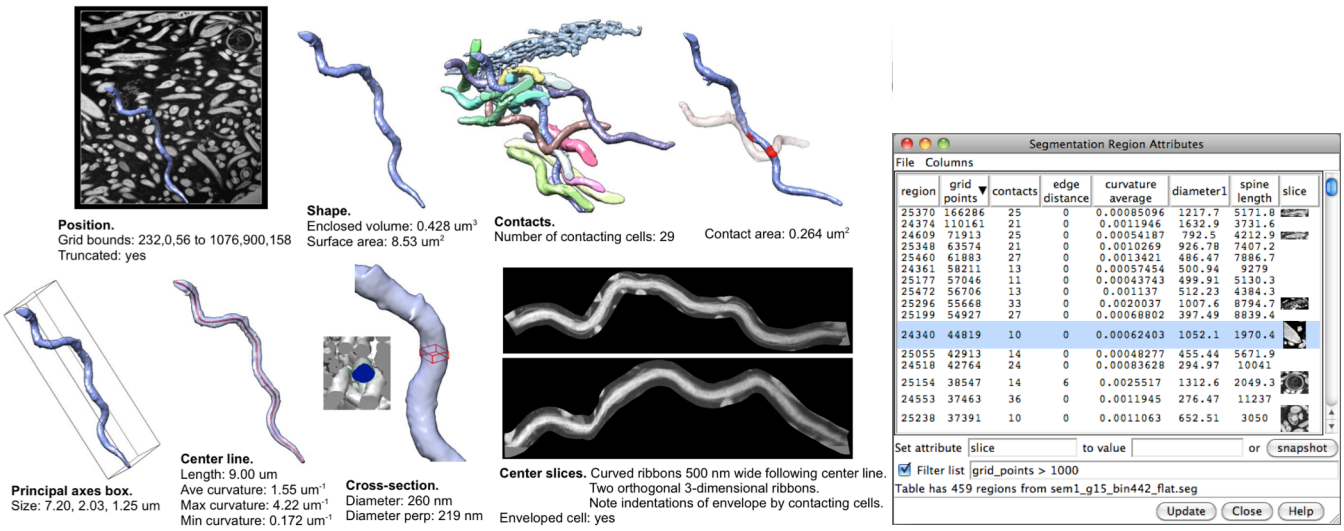


Figure 2-6: Measurements on a single bacterium segmented from FIBSEM data.

then colored, and another mouse click and drag on the original bacterium can extend it further without extending into the already colored neighbor. A few clicks and drags per bacterium is needed to produce the segmentation.

Many measurement tools were developed to provide a basic characterization of bacteria morphology (Figure 2-6). Enclosed volume, surface area, number of contacting cells, contact area with neighbors, and dimensions along principal axes can be computed. An automated modeling capability can trace the center-line of a long thin segmented bacterium, and that can be used to measure length, cross-sectional diameters and areas and to extract 2-dimensional ribbon slices following the 3-d helical shape of the bacterium with gray-scale density coloring to reveal internal morphological features (organelles, vesicles). An additional capability can use the centerline to computationally straighten the density producing a map representing an elongated bacterium, and a montage of slice images along the 3 perpendicular axes can be automatically created. All of the numerical attributes as well as images can be saved in a table with a row for each bacterium. The segmentation and tabulated properties can be saved to an HDF5 format file for later additional analysis and archiving. All of these capabilities are distributed with Chimera. More details about this work and tools is shown in a poster (http://www.cgl.ucsf.edu/chimera/termitegut.pdf) and a manuscript titled “Multiscale Three-Dimensional Organization of the Termite Hindgut Elucidated by FIB/SEM” is in preparation.

Communicating Analysis Results

This area of communicating analysis results replaces and broadens the scope of what we called animation tools in the molecular assemblies core technology proposal of the previous grant. We’ll review here some of the progress on tools to create animations, as well as progress related to communication of analysis results.

In the current grant period we added the two most valuable animation tools in the Chimera suite, Morph Conformations and Morph Map. The first interpolates between atomic models in different conformations using the Krebs and Gerstein morphing algorithm (12) to depict a smooth transformation between different states of single molecules or molecular assemblies. The depicted motion is not expected to be the actual transition pathway since no consideration of the energy landscape is used. But the animated motion gives a highly insightful view of the differences between the states, much clearer than a static superposition of the models. It uses internal coordinates that preserve reasonable bond angles and lengths and rarely causes unphysical motions where atoms passing through each other. The complementary tool Morph Map interpolates electron microscopy maps representing different conformations of an assembly. It uses simple point-wise linear interpolation and is especially valuable in cases where atomic models for the entire density are missing. Morphing maps is also useful for comparing maps with and without ligands bound, to clearly visualize where

the extra density associated with the ligand is located. Examples of these morphing capabilities are on the Chimera animation gallery web page (http://www.cgl.ucsf.edu/chimera/animations/animations.html).

A sustained multi-year effort has made commands for all Chimera tools that are useful for making animations. Many Chimera capabilities were available only through a graphical user interface dialog, or using Python programming, and hence were difficult to incorporate in an animation script. The improved commands allow per-frame calculations while showing molecular dynamics trajectories, for instance, depicting the changing hydrogen bond network. It is possible to cycle through stacks of tomogram planes, fade in and out surfaces, and rotate about axes specified in the coordinate system of any model. A very popular addition was the fly command that flies the camera through a scene using cubic spline interpolation (Figure 2-7) of a few chosen viewpoints. The command that records and encodes movies has been substantially enhanced to enable better quality capture using supersampling or raytracing (which adds shadows), and allows round-trip looping and cross-fading. For educational outreach the camera model now handles planetarium dome projection (modified fish-eye) and movie recording at larger than screen size (4096 by 4096 for commercial planetarium projectors), and background movie encoding, for example, on a server. The EM Navigator (http://www.pdbj.org/emnavi/) web site creates animations for all EM Databank maps using Chimera, as does the Virus Particle Explorer database (http://viperdb.scripps.edu/), the National Resource for Automated Molecular Microscopy (http://nramm.scripps.edu/), and the Conformational Dynamics Data Bank (http://www.cdyn.org/).

The extensive development of Chimera commands enables creating animations using the full range visualization capabilities using command scripts. The command script creates the motions, morphing, coloring changes, labels, and annotations (e.g. hydrogen bonds, clash highlighting) and uses the movie command to record the action and encode into a standard movie file (e.g. MPEG-4). All commands are extensively documented. The command script method of movie making is powerful, but requires learning many commands. We have given courses at UCSF on movie making and at an EMBO training workshop and have online tutorials demonstrating the process (see training section). During the recent 2011 EMAN single-particle reconstruction workshop we demonstrated movie making many times; it was the most common request. To simplify movie making we have developed a graphical user interface called the Animation tool based on key-frames. A key-frame is any scene that can be shown in Chimera. Pressing a button records that scene and adds a thumbnail image to a gallery. Any of the scene thumbnails can be dragged and dropped into a time-line sequence, and the tool will animate transitions from one scene to the next using standard interpolation methods (smooth motion, fades, morphs, coloring blending). Details of the transitions such as speed can be controlled. This allows making animations without understanding the Chimera command language.

While movies are usually made entirely within Chimera, more sophisticated Hollywood-style productions use commercial 3-d animation software such as Maya (http://usa.autodesk.com/maya/) and component models exported from Chimera. For example, animator Janet Iwasa created an illustration of clathrin cage endocytosis made from triskellion arms built from atomic models in Chimera

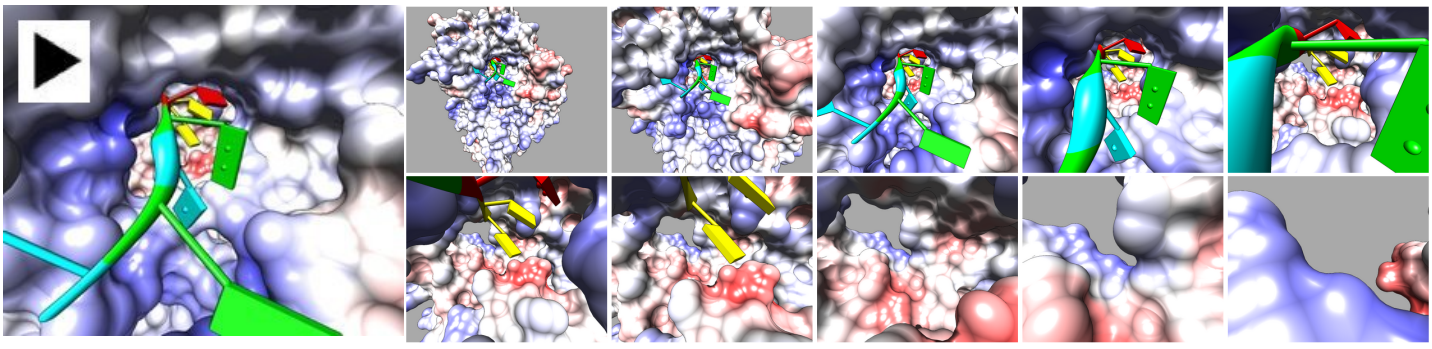


Figure 2-7: Fly through animation of rotavirus RNA polymerase (PDB 2r7r). Camera flies along a cubic splined path along the RNA being copied by a viral RNA polymerase with the polymerase surface colored by electrostatic potential. RNA backbone may interact with blue positively charged regions of polymerase. View the animation at the Chimera animation gallery http://www.cgl.ucsf.edu/chimera/animations/animations.html.

(https://iwasa.hms.harvard.edu/project_pages/endocytosis/endocytosis.html). We added a new export format Wavefront OBJ for exchanging such surface models. The Chimera multiscale model tool was reimplemented in Molecular Maya (http://www.moleculармovies.com/toolkit/) by Gael McGill, a successful technology transfer.

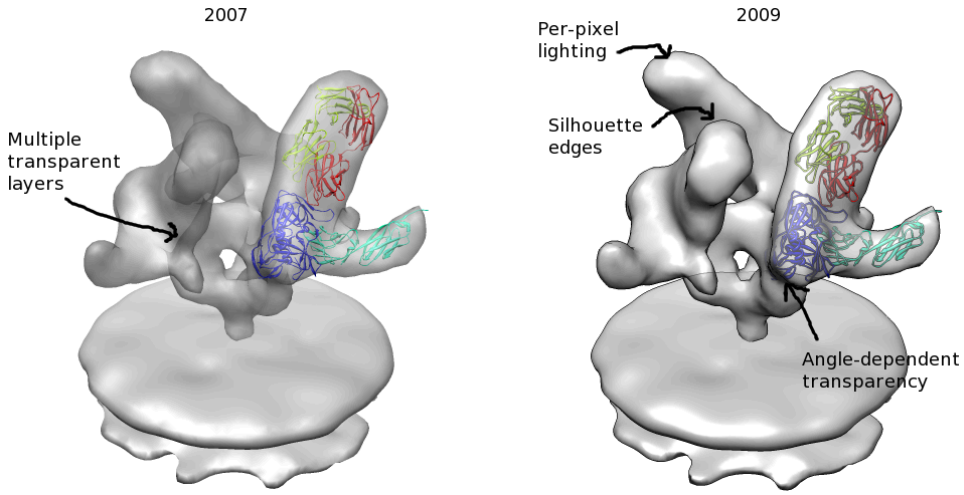


Figure 2-8: Improved visualization quality using GPU programming and other advanced OpenGL techniques.

Beyond the mechanisms of creating movies, we've made substantial improvement to the quality of images, significantly improving the perception of depth using programming graphics (Figure 2-8). Lighting highlights are now calculated at every pixel instead of more coarsely at each surface mesh vertex giving a much more realistic effect. Transparent surfaces can display only the front-most layer, hiding the complexity of many internal layers. Edge-on transparent surfaces appear more opaque and darker accounting for the increased thickness along the line of sight. And dark edging can be added to clearly distinguish overlaying features at different depths. Interactive shadows are also supported on some graphics hardware. Presets are new, offering several options for standard coloring, display styles and level of detail.

Data Sharing

We did not propose data sharing technology developments beyond animation in our past grant but we will describe some recent developments as it is a new focus area in this proposal. We have developed new file formats for electron microscopy maps and segmentations in collaboration with the National Center for Macromolecular Imaging (13) based on HDF5 (http://www.hdfgroup.org/HDF5/), a widely used high-performance extensible multidimensional array standard. The Chimera HDF5 map format resolves many technical limitations of existing formats (e.g. CCP4) used in electron microscopy that were original designed for X-ray crystallography, and also provides higher-performance. The new format contains the map symmetry. It can record rotated coordinate axes generated when extracting non-axis aligned parts of tomograms and also when aligning any two maps. It can represent unsigned 8-bit data commonly used in tomography, in addition to all other integer and floating point types. Subsampled copies of the map (e.g. sizes reduced by powers of 2 along each axis) can be included in the map file for quick display of large maps. This allows showing maps that are many gigabytes in size instantly instead of waiting tens of second or minutes to read the full resolution data. HDF5 stores the data in bricks allowing quick access to small regions without reading unneeded portions of the map. For large maps (e.g. tomograms and FIBSEM) xz and yz slices can be accessed rapidly. With other formats only xy planes can be read quickly unless the entire map is loaded into memory. The files can readily handle additional data such as precomputed smoothed versions of the map and fit atomic models, although we have not implemented those features. Our Chimera HDF5 segmentation format was developed because no community standard format exists. In addition to the attributes for maps, the segmentation format utilizes compression. The segmentation files compress extremely can be 100 times smaller than the maps, greatly facilitating transfer of files over the network, and allowing fast loading of segmentations in software. Our format also supports grouping of segmentation regions. Currently these formats are internal standards used only by Chimera.

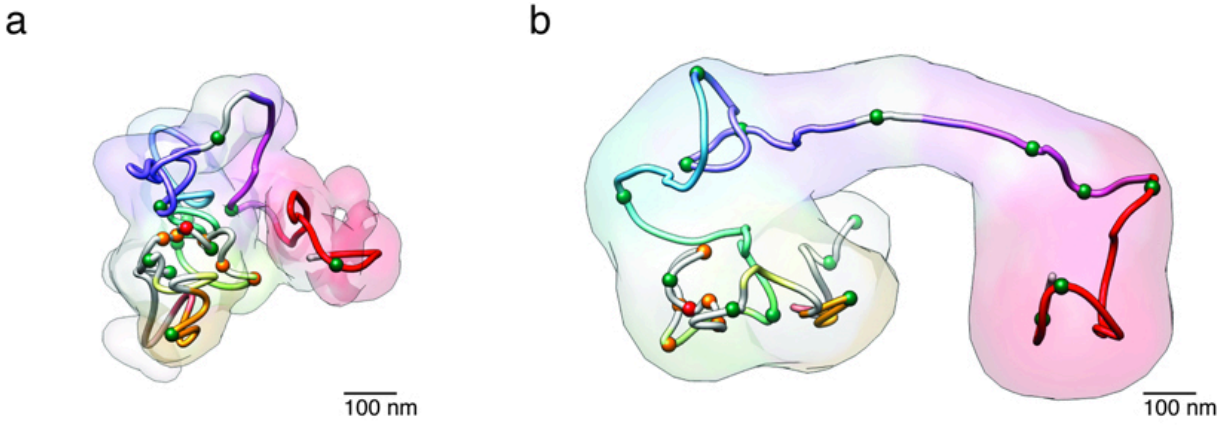


Figure 2-9: Coarse grain model of chromatin containing 50 Kbases of DNA, including the multi-protein alpha-globin locus (14). A) Alpha-globin expression silenced. B) Alpha-globin expression active shows a less compacted form.

We also have an internal XML format for coarse-grain models, representing spheres and connecting cylinders, and have begun development in collaboration with Andrej Sali an HDF5 coarse-grain model format for exchanging data between their Integrative Modeling Platform (IMP) (15) and Chimera. This will facilitate studies of large-scale assemblies, such as the fold of 50 Kbases of chromatin containing the multi-protein alpha-globin domain (14) that used IMP for computational modeling and Chimera for visualization (Figure 2-9).

Besides developing advanced data formats we support exchange of data with other software, reading for example 23 different density map file formats. Some of the latest added formats are EMAN HDF5 map format, IMOD surface segmentations, and image stacks (TIFF, PNG, and many common image standards). We also write a smaller set of formats, the most recent added being BRIX maps used by the popular crystallographic modeling package O. An important and popular Chimera feature is the ability to directly fetch data from the web. We added support to directly load maps from EMDB and the associate fit atomic models from the PDB, and directly search the EMDB (Figure 2-10). Atomic models of molecular assemblies, be directly fetched from the EBI Probably Quaternary Structure server (16), and we will soon update to their PISA server (17). Also direct download of homology models from MODBASE (18) has been added, commonly needed to build molecular assembly models when X-ray or NMR models of the components are not known.

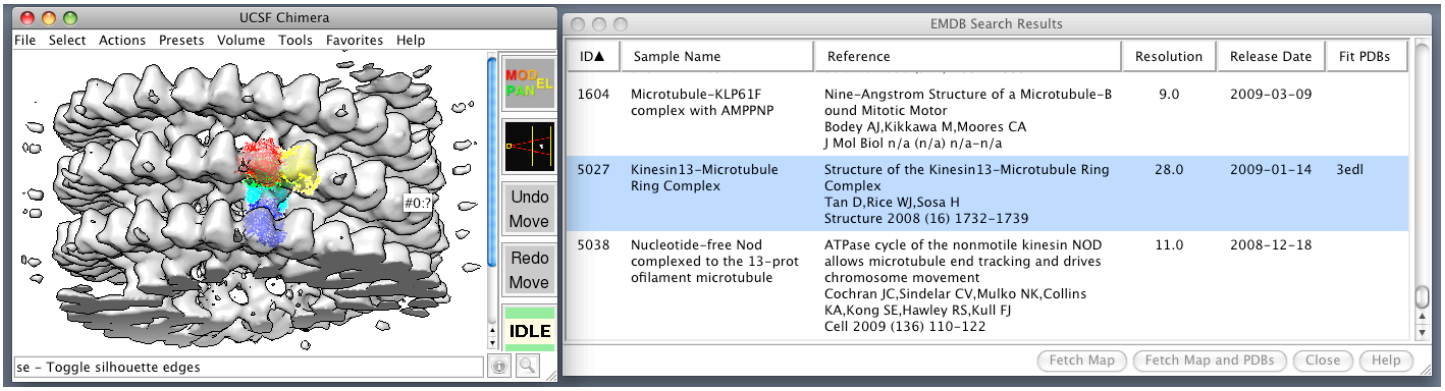


Figure 2-10: Chimera direct download from the web of EM Data Bank maps and fit atomic models.

Significance

Determining the catalog and functions of molecular machinery in cells is a frontier in biology that builds on the immensely successful developments of high-throughput protein structure determination, genomics, and protein-protein interaction mapping technologies. The RBVI develops interactive software to visualize, analyze

and model experimental data for molecular assemblies, and to communicate the results. A defining aspect of our software is that it is interactive. Most computations complete in less than one second. The software tries to enable visual and quantitative exploration of data as fast as the researcher can think, and provides over a hundred capabilities to view, dissect, compare, measure, and model to gather insights and guide analysis. Most software development in the field is of a different character, implementing specialized long-running algorithms to perform a specific in-depth analysis. Our interactive data exploration tools are an essential element to setting up those calculations and understanding the results, assess the failures and plan the next approach. This project proposes new, fast and flexible data exploration capabilities that will significantly improve the researcher's ability to understand data and achieve insights. A second and equally important aspect of this project is to enable new research to build on old results by enabling sharing of maps, analysis protocols, models, alignments – all types of computer-readable and operational results. The field of molecular assemblies research has developed few community data standards beyond those established decades ago for x-ray crystallography (e.g. PDB and CCP4 file formats) and we believe this currently impedes the build-up of knowledge more substantially than a lack of analysis algorithms. The RBVI has a central role in improving sharing and archiving of molecular assembly data.

Data Sharing

Decades of research on the structure and function of molecular assemblies has produced a substantial published literature. Those results are described in words and pictures. The data and models produced by the research are in most cases accessible only to the labs that did the work. While the journal articles are a valuable distillation of the insights found by the original researchers, the models of molecular assemblies and primary data those works describe are equally valuable as starting points for future research. Those computer readable results are sometimes available by personal request to the lab that produced them. But few electron microscopy maps, molecular assembly models, symmetry parameters, SAXS profiles, mutagenesis results, and lists of contacting residues have made it into public archives. This stymies the whole research community effort to build computational understanding of molecular machines and cells. The build-up of computer accessible knowledge from past decades of work is surprisingly small at EM resolutions compared to what has been achieved at the finer levels of proteins and sequences. It is hard to imagine how impaired structural biology research would be without comprehensive public archives of protein structures and genomic sequences, and yet that is the current state of molecular assemblies. The lack of standards, software, and databases needed to support the exchange between researchers and public accumulation of computer-readable models of molecular assemblies is natural given the early stage of development of this field of research. We believe we are at the right time to develop this infrastructure and it will increase the rate of progress in this research field.

The past decade has seen progress in archiving molecular assembly data with the creation of the Electron Microscopy Databank (EMDB) (19) in 2002 currently holding 1027 maps and the Virus Particle Explorer database (ViPERdb) (20) in 2004 currently holding 300 atomic models and 59 EM maps. The RBVI has worked with those databases since their founding and Chimera has been the primary visualization program used by both databases. These databases are capturing only a small proportion of maps and models, perhaps 1 in 5. Most published results modeling molecular assemblies use electron microscopy (EM) data, yet ViPERdb holds 267 xray models and just 33 EM models. The Protein Data Bank (PDB) holds 364 EM models. Of the 1027 maps at the EMDB only 221 are reported to have deposited models at the PDB while journal publications describe models for almost every map.

Capturing a higher proportion of models built from EM data, the maps and meta-data like map symmetry, for use in future research is the primary aim of this project. While this may seem to be primarily the responsibility of the public databases (e.g. EMDB, PDB, and ViPERdb) we believe the solution lies with the analysis software used to produce and view the models and maps, such as our Chimera program. The fundamental limitation is that software used in this field does not offer easy ways to exchange data between programs (and researchers), thus the databases can't simply adopt data exchange formats already in use by the community. Efforts of databases to establish new formats, rather than borrowing ones already in use in the research community, have not led to wide use.

Innovation

The primary focus of innovation for this project is to enable reliable, easy to use, integrated, real-time exploration of large data sets on desktop and laptop computers. This contrasts with most other computational biology research software where the focus is heavily on innovation in algorithms. While Chimera offers many unique capabilities, the algorithms behind them are in most cases trivial.

A few examples illustrate the innovative aspects of our software. Chimera has a "hide dust" capability used on noisy electron microscopy maps that hides all small specks of density, showing only the larger connected pieces. Adjusting the size limit of the specks or the density level that determines what is connected to what causes immediate display update, continuously updating many times per second for moderate data sizes. We have not seen any other software that offers this capability (although other data filtering methods can achieve a similar result, albeit with poorer interactive response). The primary innovation in this case is in the high-performance, robust, and easy to understand (one adjustable parameter) characteristics of the this filtering method, that can, for example, remove a blizzard of noise obscuring the view of an imaged virus particle.

A second example of how Chimera tools are innovative is the "morph map" capability which interpolates a sequence of two or more density maps, often related conformations of a molecular assembly, to show a real-time animation of the sequence. Utilizing the time dimension gives a remarkably clearer depiction of the changes between maps than a simple static superposition. Algorithmically it is trivial -- a linear interpolation between density values at each grid point. But high-performance real-time rendering at many frames per second (optionally preserving enclosed volume) is challenging requiring innovative engineering approaches. We have also not seen 3-dimensional map morphing available in any other software. This capability was developed as a Chimera plug-in by Wei Zhang and Pawel Penczek and then the RBVI staff recoded a high performance version with simplified user interface that is included in the Chimera distributions.

The current Chimera molecular assembly analysis tools are an integrated collection of a hundred data exploration capabilities. By "integrated" we mean that the tools can be used in combination, for example hiding dust and morphing at the same time. Combined use greatly enhances the power to obtain insights from real-time data exploration. The capabilities solve specialized problems often encountered in studying molecular assemblies such as fitting atomic-resolution protein models in electron microscopy maps, or using point-group symmetry in modeling icosahedral viruses. They are used by most labs in the world (about two thousand labs) that study molecular assemblies by electron microscopy. These capabilities are the product of 10 years of development. From discussions with more than 100 different researchers each year we have ideas for many broadly useful additional features, more than our combined 10-year effort has so far produced. This proposal details some of the new capabilities that we believe (gauged by high interest expressed by experienced researchers) will substantially advance the field and are technologically achievable.

Approach

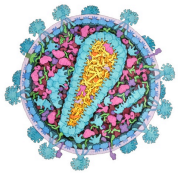
We describe specific aims in each of the 4 project areas. It is not possible to forecast all of the developments over the next 5 years. In the past 10 year period the development of new software capabilities for molecular assemblies has been strongly collaboration driven. Comparing our progress to past proposals suggests that we can at most forecast half of the developments that we will pursue. We believe that the specific aims described here are a representative subset of the problems we will tackle and that they span the major problem areas. This has been born out by molecular assembly RBVI proposals for the past two funding cycles.

Specific Aims

Specific Aim 1: Develop software to analyze atomic models of molecular assemblies.

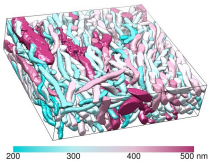
Aim 1A: Develop tools for researchers to layout 3-dimensional "composite models" of complete biological systems that depict structural components in a prototypical arrangement.

Aim 1B: Provide direct access to all database structures, multiple sequence alignments and EM maps associated with each molecule in a composite model.

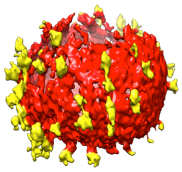


Aim 1C: Create a demonstration composite model of a full HIV virus particle. This is a 3-d version of a recent David Goodsell's HIV painting:
http://www.rcsb.org/pdb/static.do?p=education_discussion/educational_resources/hiv-animation.html

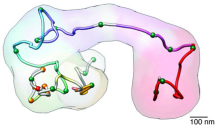
Specific Aim 2: Develop software to analyze single-particle EM maps.
Aim 2A: Develop real-time fitting methods for EM maps handling symmetry and packing constraints, flexible turns, ensemble fitting, and incremental global searches.
Aim 2B: Develop approaches for efficient handling of large maps, such as asymmetric unit display, hidden surface removal, fast transparency, and Fourier space representations.
Aim 2C: Develop EM map and model validation tools, for example, computing small-angle x-ray scattering profiles directly from EM maps to use SAXS as complementary validating data.



Specific Aim 3: Develop software to analyze structurally heterogeneous assemblies.
Aim 3A: Develop interactive segmentation and feature extraction methods, focused on multi-resolution approaches and annotation of simple features such as filaments and spheres.
Aim 3B: Develop tools to support single-particle tomography, including alignment algorithms with missing wedge correction and filtering to minimize missing wedge artifacts.



Specific Aim 4: Develop software enabling researchers to share and archive analysis results.
Aim 4A: Automatically create an HTML log of Chimera analysis (Figure 2-15) including images, links to data (PDB files, maps, sequence alignments, residue lists), Chimera session files, movies, quantitative measurements (correlation, buried area). Organizes hundreds of files of a research project.
Aim 4B: Document HDF5 map, segmentation, and coarse-grain model standards and promote use by other software packages.
Aim 4C: Develop animation creation tools to illustrate hypotheses about molecular assembly architecture and function. Support 3-dimensional movie recording.



New Directions

There are two new directions in this core project aimed at enabling researchers to organize and disseminate their biological results. First we will provide tools so that researchers can build 3-dimensional models composed of proteins, nucleic acids and lipids to represent large functional entities, and we will demonstrate the new capabilities by making a whole HIV virus or muscle filament sarcomere model. Essentially we are aiming at a true 3-dimensional representation of a recent David Goodsell depiction of an entire HIV particle (Figure 2-11). We will use our existing multi-scale display capability to depict large molecules or complexes in cartoon surface style and there will be direct access to the many experimentally determined structures, sequences and EM maps known for each of the components allowing easy exploration and analysis. This effort is related to our Biological Context core project.

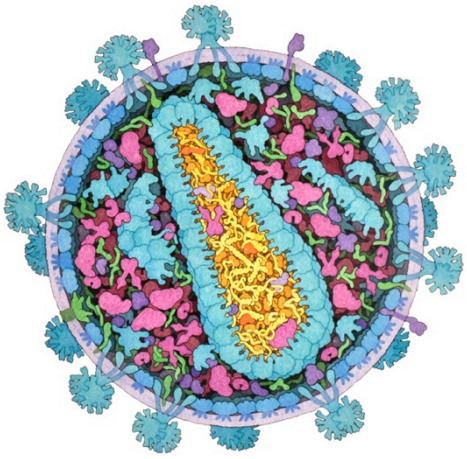


Figure 2-11: HIV virus particle including all structural components illustrated by David Goodsell. We will develop software enabling researchers to build 3-dimensional prototypical models that link to underlying structure, sequence, and EM map data.

The second major new direction aims to help researchers organize, share, and archive their results, by creating a web page including images, links to Chimera sessions and data files, animations, measurements like distances, angles, RMSD, and user notes. The researcher using Chimera will be able to record any of this data at any time during their analysis and HTML multi-media log of the results will have the requested links and information appended. Links to Chimera sessions will allow returning at a later date to any saved scene. The

immediate benefit of this tool is to organize the many data files and results associated with a research project. But our primary motivation is to attack what we see as the biggest challenge facing the molecular assembly research field, that most (80%) models and EM maps are lost. These computer readable results are as valuable to long-term research as the journal articles that describe the insights gained. A key part of our proposal is that the data put into our web page logs will conform to completely documented standards that we will provide (e.g. matrices placing molecules in maps, map symmetry specification, segmentation files, …).

Technology Themes

This proposal emphasizes the biological analysis problems we are solving. Another viewpoint useful in understanding the directions of the work is from an engineering and technology perspective. Major facets of this project from a technology perspective are listed here.

1) High-performance computing: enabling real-time analysis and visualization of larger data sets and exploiting advances in commodity computer hardware (multiple CPUs, programmable graphics).

2) Multi-resolution capabilities, from atomic resolution to tissues or microbial communities.

3) Multi-modal data integration: combining all available data for the biological complex under study to produce a consistent model, e.g. EM, x-ray, sequence, homology, covalent linkage constraints, SAXS, ...

4) Direct data interaction: real-time morphing, filtering, fitting, slicing to explore data with continuous hand-eye interaction.

Continuous hand/eye interaction using mouse dragging is highly valuable in analysis. Most obvious example is rotating a model using a mouse drag. The advantage 30 frame/sec continuous hand control becomes very apparent when compared to only being able to change view direction with a typed command (as in some older software). Translating, zooming, volume contour level adjustment, rotamer bond rotation, clip plane positioning, hand fitting, volume cropping, molecular dynamics playback, volume morphing are all powerful data exploration methods in Chimera. Many more are not available in Chimera.

5) Comparing large numbers of objects, e.g. dozens of similar protein structures (various mutations, truncations, species, ligands, pH/salt), all protein-protein interfaces in a virus, dozens of fitting alternatives of an x-ray protein model in an EM map, ...

Tools to compare large numbers of objects: e.g. conformations from BLAST pdb, interfaces between virus proteins, bacteria in termite gut, enzymes binding sites (SFLD), alternative fits of molecules in maps. Researchers often compare dozens of homologous structures, alternate map fits, segmented volume regions, binding interfaces. I commonly see Chimera user's with 10 - 30 open models. As biology research accumulates more models, analysis of many models becomes as important as the one-at-a-time analysis that Chimera, designed in a more data poor era, focuses on. Working with many models becomes too time consuming and tedious to be feasible without multi-model analysis tools. The Chimera View Dock tool is a successful example of multi-model analysis.

6) Technologies to allow researchers to communicate their results, keep records of their analysis, exchange and archive computational models.

7) Enabling prototyping new analysis methods by many labs, and widely distributing successful tools.

Related Dissemination Projects

Two dissemination projects will significantly increase the value of the tools developed by this core technology project. First we plan to make a hundred or more video tutorials demonstrating all common uses of our software. We have made and posted on the web 25 such videos in the past year where an expert user demonstrates standard analyses using Chimera. These are highly effective and appreciated training materials.

We find in face-to-face training that even the most experienced Chimera users know only a fraction of the software capabilities important to their work. Secondly we propose to create online Chimera programming documentation for all useful functions and classes. Small Python scripts utilizing the many libraries in Chimera immensely extend the scope of analysis that is feasible. We provide about 40 such scripts per year and lack the resources to address a hundred more requests to do analysis slightly out-of-reach of the built-in commands and dialogs. These two projects are described in the dissemination section of the proposal.

Methods

Specific Aim 1: Software for Analyzing Atomic Models of Molecular Assemblies

Aim 1A: Software for Lay Out of Typical Configurations for Pleiomorphic Molecular Assemblies.

The suite of tools we have developed for studying atomic models of molecular assemblies has targeted relatively rigid assemblies. This is a natural limitation because the techniques used to produce such high-resolution models, single-particle electron microscopy and x-ray crystallography, work by averaging data for many instances of the same biological complex. Imaging requires many identical copies of an assembly. Unfortunately this is not possible for many interesting assemblies, for example, HIV and influenza virus particles are very heterogeneous in shape. Although there is no unique configuration for such molecular assemblies, we frequently have extensive knowledge about typical configurations, and often know structures of many of the constituent proteins at atomic resolution. We propose to develop software tools to allow researchers to create models of typical configurations. We will call these “composite models” because the atomic resolution pieces of the models come from separate imaging data, and the combined model represents the “composite” features rather than any one observed complex. It is also possible that a composite model could be based on a unique assembly imaged by electron tomography.

Two beautiful depictions of composite models of HIV virus were recently published (Figures 2-11, 2-12), one a two-dimensional watercolor by David Goodsell and the other a three-dimensional depiction, the 2010 first place winner of the Science magazine image contest made by Ivan Konstantinov (http://www.sciencemag.org/content/331/6019/848.full#F1). Both images try to depict all the structural components of HIV. A version of the Goodsell image on the web (http://www.rcsb.org/pdb/static.do?p=education_discussion/educational_resources/hiv-animation.html) can be used to find out more information. Clicking on a protein in the virus image presents a detailed image of that protein, text describing its function and a link to an atomic model at the PDB. Our aim is to allow researchers to build similar models in 3-dimensions.

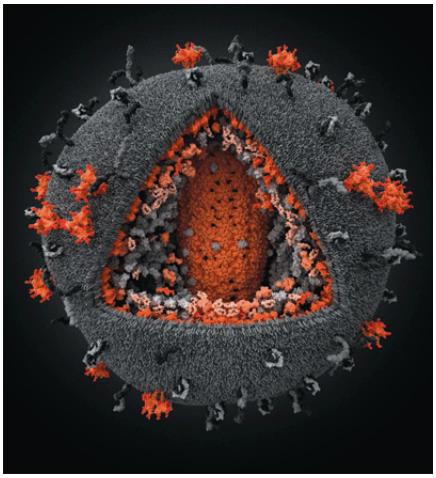


Figure 2-12: Science magazine image contest HIV depiction with virus-encoded molecules in orange and cell-derived components in gray by Ivan Konstantinov.

Aim 1B: Database Access for all Components of a Composite Model.

In addition to providing a visual synthesis of many structural results, we envision composite models being an index to the available structure and sequence data. By selecting a component molecule of say HIV capsid protein you will be able to access all structures of that protein from the PDB (obtained for example by BLAST). Sequence variability is also a very rich source of information about the functionally significant residues in proteins and Chimera has sophisticated multiple sequence analysis tools. Clicking on the capsid protein will also give the option of view multiple sequence alignment data, in the case of HIV from the Los Alamos National Laboratory HIV sequence database. HIV capsid protein has also been studied by electron microscopy with maps at the Electron Microscopy Databank and this too would be available. The objective is to make the composite model provide streamlined access to the available structural data, with only a few clicks to obtain atomic

models, sequences and maps. The available data could be fetched from the web databases (Chimera currently fetches from 10 databases), or for higher performance could be collected in a single directory to be distributed with the composite model.

Aim 1C: Create a Demonstration Composite Model of a Complete HIV Virus.

Our efforts will focus on creating the software to enable researchers to make such models for the biological systems they study. But an important aspect will be demonstrating these tools by building an HIV composite model that we will undertake with links to several hundred Protein Databank structures of the more than 15 distinct proteins comprising the virion. Also we will enable fetching multiple sequence alignments from the Los Alamos National Laboratory HIV sequence database to depict known conservation and mutation between viral strains using the powerful Chimera MultAlign Viewer tool.

A simple example of a composite model is shown in Figure 2-1e which shows an actin network. Actin networks inside cells are used to dynamically reshape cells, extending the cell in some direction enabling movement. The network shown consists of just two assemblies, filamentous actin (F-actin) and a branching complex Arp2/3. It was built with a Python script that made random branch points and filament lengths, and used an experimentally observed branching angle. Our aim is to develop capabilities in Chimera will allow researchers to build composite models without requiring them to write code. How would a researcher build the actin network? The script we used relied on two structural alignments, how Arp2/3 sits on the base filament, and how the nucleated branch filament extends from Arp2/3. Both of these positions were derived interactively in Chimera using existing alignment tools. Existing Chimera facilities can also extend an actin filament to any length using known helical parameters. With these building blocks, the new tool would allow making many copies of filaments and Arp2/3 and snapping them together. Snapping them together could consist of clicking an actin monomer and an Arp2/3 to join them using the appropriate alignment.

The user interface to allow building the actin network is a simple case and a number of other building techniques would be needed for the more challenging HIV composite model. Already Chimera allows click and drag placement at arbitrary positions, for example, to place HIV spikes in the lipid envelope. The lipid envelope itself would be most easily modeled as a spherical slab (already supported in Chimera). To build the conical capsid from 12 pentamers and a few hundred hexamers of capsid protein (both have been experimentally determined) will require a more sophisticated joining where adding a hexamer might snap to the best fit with two neighboring hexamers based on the idealized alignments of hexamers from the 2-dimensional crystallographic arrays. Layouts for HIV capsid pentamers and hexamers have been reported in the literature (21) and it will be useful to allow a simple file format for placement of subunits, so such externally computed placements can be used. Another tool for doing this placement would use electron tomography data to define the surface, and allow dropping whole pentamers or hexamers on that surface. For the two identical strands of RNA a tube model based on a cubic spline of hand-placed markers (currently in Chimera) could be used. It would be valuable to have clicking the RNA tube provide the full secondary structure of all base paired sections of the RNA which has been experimentally determined (22),

In addition to creating simple molecule and subassembly placement tools, this project will drive a number of lower level enhancements to Chimera capabilities associated with the large numbers of objects in these models and to support analysis of large numbers of structures and sequences that are parts of the models. It will be important to support hierarchical grouping. For example an HIV spike is a trimer composed of gp120 and gp41 and these 6 molecules would probably best be represented as a single mushroom surface when viewing the entire virus. It will be possible to group and ungroup components. For fast and memory efficient display of many copies of the same shape we will use just a single copy of that shape with positioning matrices to render it in many places (“graphical instancing”). Because we envision opening many atomic structures of a given molecular component for comparative analysis we will need to improve memory efficiency of atomic models. Our current memory use of approximately 2 Kbytes per atom could be reduced by up to a factor of 10 by carefully avoiding creating Python language copies of the atoms that consume most of that memory. Chimera sessions already save models such as the actin network (Figure 2-1e) and the composite models will also be saved as sessions.

Driving projects

Our effort to create tools for composite model building will benefit from strong connections to four of our driving biomedical projects: HIV Accessory and Regulatory Complexes (Nevan Krogan, Yifan Cheng), coarse grain modeling (Andrej Sali), modeling EM tomography (Manfred Auer), EMDB and PDB databases (Cathy Lawson).

Specific Aim 2: Software for Analysis of Single-Particle Density Maps

The experimental methodology of single-particle electron microscopy has advanced substantially in recent years, enabling faster structure determination and at higher resolutions. An example of almost high-throughput reconstruction was recently published by the Glaeser lab (23), where 16 of the most abundant multi-protein complexes found in prokaryote Desulfovibrio vulgaris Hildenborough were isolated and 8 solved by single-particle reconstruction using negative-stain at resolutions from 15 to 29 Angstroms (0.5 Fourier shell correlation). Technical advances have also allowed high resolutions, with 9 structures determined in the past two years at 3 to 4.5 Angstroms resolution where some residues of proteins become visible, and direct building of atomic models is feasible (24). All of the highest resolution maps are of viruses and utilize icosahedral 60-fold symmetry to reconstruct density from more than a million asymmetric units. Matt Baker at the National Center from Macromolecular Imaging is developing a program called Gorgon (25) to perform the complex task of do direct model building in high resolution maps. Chimera continues to focus on simpler fitting scenarios in lower resolution maps, 5 to 10 Angstroms where helices and sheets are the finest features, and greater than 10 Angstroms where domains or whole molecule shapes are distinguishable. Almost all maps from electron microscopy are at those lower resolutions.

Aim 2A: Fitting Enhancements

The Chimera map analysis tool most frequently cited in the literature is Fit in Map, which rigidly aligns a molecule or part of a molecule (domain or helix) with density to maximize correlation. The optimization is local, steepest descent and generally requires an initial placement of the molecule to within 60 degrees and half a diameter of the optimal position to obtain convergence. The method is fast, executing in less than a second, allowing interactive testing of many possible molecule placements. In many cases a convincing optimal fit can be obtained. We plan to extend this interactive fitting capability in five ways to handle common more difficult scenarios, and to improve quality and confidence in the resulting fits.

The first enhancement is to include symmetry in the local optimization. Most assemblies being studied have symmetry and it imposes a useful constraint when fitting: if a monomer is fit to the density but its symmetric neighbors clash with it then the fit is unreasonable. A simple method to account for these clashes is to optimize the placement of one monomer based on correlation coefficient of the full symmetric atomic model (26). The abnormally high density where two copies of a molecule overlap contribute to unfavorable correlation with the experimental density. While this will be a slower calculation than the case without symmetry it is only necessary to consider overlaps with the nearest symmetric neighbors so we believe it will still take only seconds, permitting rapid exploration of different starting orientations.

Clashes between neighbor molecules fit into a map where the molecules are not related by symmetry represents another useful constraint. Because molecules pack together tightly with interface atom spacing comparable to intra-molecular atom spacing, the divides between contacting molecules are seldom visible. This often results in a scenario where fitting either of two contacting molecules into a density map without consideration of its neighbors merely pushes the molecule to the center of the density containing those two molecules. This is especially common at lower resolutions where the deeply buried regions often have higher density than peripheral ones. A simple solution is to fix one molecule, subtract its simulated density from the map, and fit the other molecule. Then exchange roles, subtracting the previous fit from the map, and fitting the previously fixed molecule. Iterate this procedure for several rounds. This method was successfully performed by hand with Rous sarcoma virus (27). We will automate it.

At 5 to 10 Angstrom map resolutions helices and beta sheets are usually visible and frequently available atomic models from crystallography, NMR, or homology represent different conformations. It is necessary to

move a domain or helix to achieve a good fit to the density. This is currently possible using a rigid fit of the piece of the molecule, but it distorts bond lengths where the piece connects to the rest of the molecule. We plan to use the existing energy minimization capability in Chimera based on MMTK and the Amber force field to restore reasonable geometry to the turns that connected a fit piece of a molecule to the rest of the molecule. At these resolutions there are several flexible fitting approaches combining molecular mechanics and density matching (28) and we do not intend to duplicate that work. Those methods typically require a good starting model since molecular dynamics is not able to sample large-scale changes such as different pairings of model helices with helical density. Our modeling with turn energy minimization can provide appropriate starting models.

Two fitting improvements are aimed at increasing confidence that hand-exploration with local fits did not miss good configurations and choosing the best model to fit. We will add the ability to sample rotational orientations of a molecule, locally optimizing each orientation, and producing a list of all unique fits found. The sampling can be uniform or random, random having the advantage that one does not need to choose how fine the angular step should be. The fits will be computed and the list updated continuously and it will be possible to inspect the best fits obtained so far while others are being computed. Similarly the center of the map can be moved to random positions in the map or a uniform scan can be done. A test of this idea fitting a GroEL monomer into a high resolution 5.4 Angstrom map (EMDB 1457) performed 1000 fit optimizations with random rotations and translations in just 90 seconds. Surprisingly it produced only about 50 unique fits (not counting symmetry equivalents). Only a few had high correlations. Another enhancement will allow running fits against all members of an ensemble of atomic models and reporting the best. Common cases where this has been used include NMR ensembles (27) and ensembles of homology models (29).

These fitting improvements build on the existing Chimera fit optimization capability. Where multiple independent fits are being computed, we will compute the fits in parallel on computers with multicores. Typical new desktop computers have 4 or 8 cores. Current fitting code uses only a single core. We will also be making improvements to the user interface, allowing lists of fits sorted by correlation or other metrics (e.g. atoms within density contour). It will be useful to quickly inspect a large number of fits with some of the new methods. We envision being able to use the mouse scroll wheel or arrow keys to go through the list and the molecule will automatically to show that fit. It will move by a smooth rotation and translation to allow visual indication of how near it is to the previously viewed fits. The correlation can be reported in the graphics window so attention can be focused there when quickly surveying a large number of candidate fits.

Aim 2B: Efficient handling of large maps

High-resolution maps pose performance challenges for interactive software. A recent 3.8 Angstrom rotavirus map (30) has a 1.2 Angstrom voxel size with grid size 1000^3 resulting in a file size of 4 Gbytes if floating point density values are used. Because of the large size the EM Databank provides only one octant of the map, still a very time-consuming download. Displaying such a large map also can make redrawing graphics too slow for interactive use. We will explore a range of techniques to make it easier to work with large maps.

First we will provide the ability to produce symmetrized maps from asymmetric units, or larger partial maps such as virus octants. Another use of this capability requested multiple times at the 2011 EMAN single-particle reconstruction workshop was the ability to make symmetric starting density models for reconstructions. We will also create a command to extract asymmetric units from full maps. In principle this would reduce the size of an icosahedral virus map by a factor of 60 making reading such maps and transferring them over the internet much faster. In practice, map formats are only able to handle rectangular grids, so we will take a practical solution of choosing a box just big enough to enclose a wedge representing the asymmetric unit. The capabilities to convert back and forth between asymmetric units and a full density will work for all common symmetries: icoshahedral, cyclic, dihedral, helical, dodecahedral, crystal symmetries. Using our HDF5 density map file format we can easily record the symmetry in the map header so that the user will not have to separately enter the symmetry group whenever an operation involving symmetry is used.

For improving interactive display speed the choices are to display only part of the map, or display less detail. Our current rendering code uses state-of-the-art OpenGL hardware accelerated rendering achieving speeds

close to the limits of the hardware. We have experimented with advanced mesh decimation techniques using MeshLab (http://meshlab.sourceforge.net/) and found reductions in surface vertices by a factor of 2 were possible with only small loss in visual quality. Some practical solutions are to allow display of just a small number of joining asymmetric units. This would be a display option and not require making a reduced map. Another effective technique is to not display uninteresting regions of density. Much of the rendering complexity for virus maps is inside the capsid layer often showing many concentric layers where the viral genome is packaged. That density is rarely studied because the genome does not have the symmetry that was imposed in the map calculation. Removing it from the display could increase rendering speed by a factor of several. Even when it is completely hidden in the interior, the graphics hardware spends much time processing it, especially if all that extra unseen data does not fit in memory on the graphics card. Another important performance enhancement will increase the speed of rendering transparent surfaces. These currently render many times slower because the computer processor is used rather than the graphics processor. A relative simple optimization of caching depth for several view angles can make it as fast as drawing opaque surfaces. A final performance enhancement for large map display is to use Fourier space map representation. This is used by UROX map visualization and fitting software (Siebert 2009, PMID 19564685) and has the key advantage that the rendered resolution can be changed continuously very efficiently allowing the researcher to choose level of detail interactively.. With high-resolution maps it is frequently useful to fit and visualize smoother versions showing less detail.

Aim 2C: EM map validation with SAXS

We plan to extend our current small-angle x-ray scattering (SAXS) profile calculation to compute predicted profiles from electron microscopy density maps. Currently Chimera is able to compute profiles for atomic models and compare them to experimental SAXS curves to validate such models. Part of the uncertainty in models from single-particle reconstruction is simply that the map may be wrong. Being able to directly compare the expected SAXS profile for a map to experimental SAXS data would allow a direct assessment without requiring fit atomic models. We have discussed using the em2dam program developed by Alex Shkumatov in the Svergun lab, a leader in the development of SAXS analysis software to make such calculations, to create a dummy atom model from EM maps that would allow computing the profiles.

Driving projects

Three of our driving biomedical projects focus on single particle reconstruction and will contribute to the above developments and guide other unforeseen developments. These collaborations are with the National Center for Macromolecular Imaging (Wah Chiu, Greg Pintilie, Matthew Baker, Steve Ludtke), the HIV Accessory and Regulatory Complexes (HARC) structural biology center (Yifan Cheng), and the Electron Microscopy Databank (Cathy Lawson).

Specific Aim 3: Software for Analysis of Heterogeneous Assemblies

We propose to develop more methods for segmentation and feature extraction and also more tools for filtering, visualization and analysis to support subtomogram averaging.

Aim 3A: Interactive Segmentation and Feature Extraction

Segmentation of crowded cellular environments is an extremely challenging problem. Many segmentation algorithms for 3-dimensional data, especially for medical imaging, are available, for example, the Insight Toolkit (ITK) developed originally for the visible human project (31). Our goal is to create easy to use interfaces combined with the simplest algorithms. Almost all our development effort is centered on the user interface rather than incorporating sophisticated algorithms. The most common method of segmenting EM maps is by manually tracing contour plane-by-plane and joining them to create bounding surfaces, for example, as is done in IMOD software (32). The methods we are exploring perform segmentations directly in 3 dimensions.

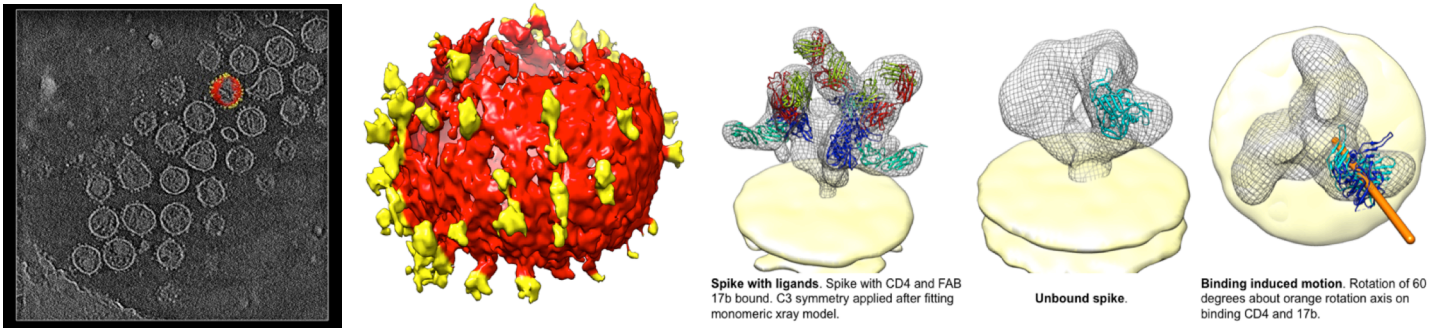


Figure 2-13: Subtomogram averaging to determine SIV virus spike structure. A) Tomogram containing many irregular shaped SIV virus particles. B) One virus particle from tomogram, filtered to reduce noise and radially colored to highlight spikes (yellow). C) Map obtained from aligning and averaging many spikes with CD4 and a fragment anti-body bound, fit with X-ray atomic models. D) Spike with no ligands bound from different tomogram. E) Comparison of spike protein orientations in bound and unbound states, rotation axis shown (orange).

We intend to solve a major limitation of the watershed and flood-fill segmentation method used successfully on the termite gut project. While the existing method worked well for segmenting whole cells it was not possible to segment finer features such as individual flagella, organelles and vesicles (Figure 2-5C). Our aim is to make multi-resolution segmentations feasible. The current limitations exist because the code is based on the single-particle segmentation Segger method which only requires segmenting at one length scale, for example at the size of proteins or domains. We will generalize this to allow segmenting different map features at different levels of map smoothing or binning. The degree of smoothing will be interactively controlled and only computed in the region near the object being segmented so that the researcher can quickly adjust this parameter through experimentation to an optimal value.

We will implement two feature extraction methods, one to trace paths in maps and another to delineate spherical objects such as vesicles and viruses. To trace a path two markers can be placed in the density using the mouse. The new code will compute the path traversing the highest density that joins those markers, optionally making a cubic spline through equispaced control points. It will be possible to segment the density nearby using the traced path. We have worked with several data sets containing filaments where this would be valuable, for example, microfibrils in plant cell walls, bacterial flagella, tip-links in auditory cilia bundles, and arms of clathrin cages. A second feature extract capability will place spheres to match vesicles and viruses in density maps, automatically locating the center and radius to achieve an optimal fit. This also will work by a simple mouse click inside the particle of interest and compute and display results immediately.

Aim 3B: Analysis and Visualization for Single-Particle Tomography (also known as "subtomogram averaging").

We will develop two techniques to assist studies using subtomogram averaging, where many equivalent objects are extracted from a tomogram, aligned and averaged to produce resolution and signal to noise suitable for fitting atomic models (Figure 2-13). The limited range of tilt angles at which the sample was imaged causes severe artifacts in the tomogram, a problem called the "missing wedge" because a whole wedge of data is Fourier space is not observed. This accounts for the hole in the top of the virus shown in Figure 2-13B. The missing wedge causes problems choosing spikes and aligning them. The alignments become highly biased because standard correlation maximizing methods simply end-up aligning the missing wedges. We will implement a known method to correctly align different objects by treating the missing data as unknown, and hence not to be used when computing the quality of alignment (33).

The other subtomogram averaging tool will attempt to correct the missing wedge reconstruction artifact using a novel algorithm we have developed based on solvent flattening (Figure 2-14). It has so far only been evaluated on test data with simulated noise, so an important aspect of this work will be algorithm testing and development with data from real tomograms. The algorithm is simple and fast and utilizes an extra constraint that much of the map is water and should have nearly uniform average density. Standard tilt-series

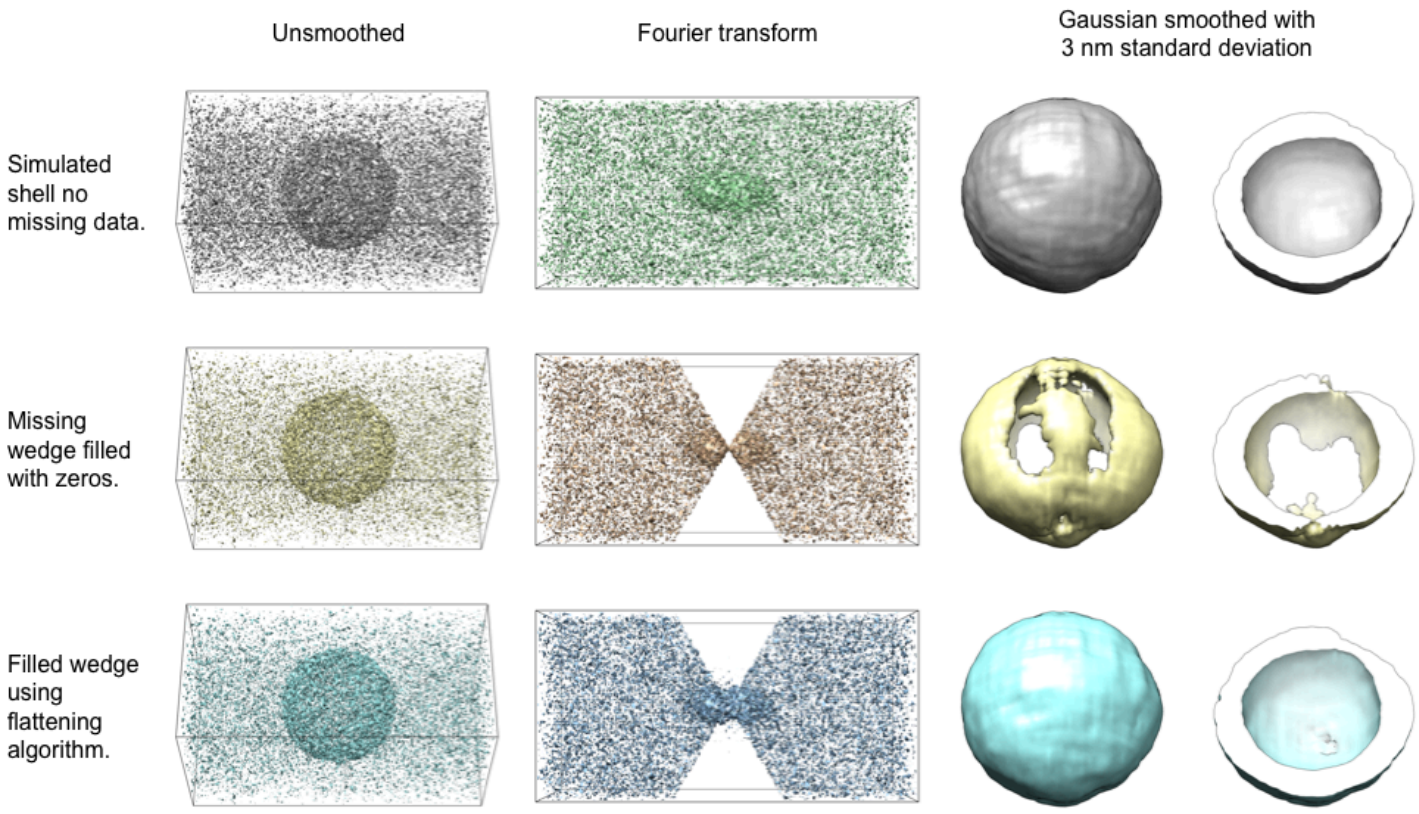


Figure 2-14: Solvent flattening to correct missing wedge artifact on a spherical shell test data, inner radius 20 nm, outer radius 25 nm, shell density 1 and Gaussian noise with standard deviation 1. Top row is without missing wedge, middle row with missing wedge, bottom row with flattening correction. Map size is 128 by 64 by 64 and correction using 50 flattening iterations and 2 nm Gaussian smoothing took 18 seconds on a laptop computer using a single core.

reconstruction methods that produce 3-d maps from the series of 2-d images place zeros in the missing wedge in Fourier space that results in ripples in the solvent density. Our algorithm computes non-zero values in the missing wedge to flatten to remove the large-scale ripples, and that redistributes density that went into solvent ripples back to its correct location in for instance the virus particle shell. An exactly analogous technique is used to determine phases in x-ray crystallography, where the phases play the same role as the missing wedge. The EM tomography solvent flattening condition is a linear constraint while the X-ray phasing problem is non-linear, so we might expect our case to be simpler to solve. We solve it by an iterative procedure, each cycle doing one correction in real space and one correction in Fourier space. In real space we increase all solvent density values that are less than the mean solvent value to the mean value. Then Fourier transform and restore all Fourier coefficients outside the missing wedge to the experimentally determined values (those of the original tomogram). Leave the non-zero missing wedge values alone. Then transform back to real-space for another iteration. Gradually energy builds up in the missing wedge that flattens the solvent and corrects associated artifacts in the structure.

Driving projects

FIBSEM and tomography studies, Manfred Auer. Subtomogram averaging methods, National Center for Macromolecular Imaging, Michael Schmid, Wah Chiu. Coarse-grain model representations, Andrej Sali.

Specific Aim 4: Software for Communicating Analysis Results

Aim 4A: Software to Create Web Page Logs of Analysis Results.

We plan to develop a Chimera tool to record analysis results in a local web page (an HTML file) for organizing a research project and for sharing computer-readable results with other researchers. It will create a web page

(an HTML file on the local computer) that contains a chronological log of analysis done in Chimera with what to log controlled by the researcher (Figure 2-15). The web page will contain images, links to Chimera session files, links to PDB models and density maps, placements of models in maps (Euler angles, quaternions, matrices), precise descriptions of map symmetry (exact center for point symmetries and reference frame), measurements (distances, angles, contacting residues), precisely defined measures goodness of fit measures, RMSD values, links and images to computed radial density, Fourier shell correlation and other 1-dimensional plots, sequence alignments, embedded spin movies, 3d stereoscopic movies, and researcher notes. The aim is to provide a record of useful analysis results obtained which is both visual and functional in the sense that links to the actual quantitative data in standard public formats. Links to session files or links to primary data in this web page will allow a researcher to quickly restart software and revisit analysis. This is an organizational tool for the researcher to keep track of the tens or hundreds of diverse data files associated with a project, but also a way to promote data exchange. A lab could make these organized project results available to a collaborator by simply providing the web page and directory of linked files. The detailed analysis and data could easily be made public on the research lab's web site.

Aim 4B: Develop Formats for Data Exchange.

Our long-term mission is to promote community standards for data exchange. As part of our web log facility we will make our internal data representations (maps, symmetry, segmentations, model placement, coarse grain models) documented specifications and refine them when needed to support uses outside our software. Also we will promote and assist use of these representations by other developers, many of whom we have worked with (EMAN, IMOD, Situs, Spider, BSOFT). Can Chimera formats be useful for data exchange between software package? Many of the scientific data formats in use are inflexible (fixed length headers), poorly documented and designed for other specialized needs (e.g. representing crystal symmetry). We believe the formats we plan to promote are better designed and better suited to address the problems in molecular assemblies research. Our formats for maps, segmentations, and coarse grain models use HDF5 and XML, both easily extensible. These formats contain version numbers so software can choose to handle only specific format versions, and new versions can correct inadequacies of old versions. Half of our effort to promote data exchange involves only definitions of the data, and does not propose any specific file format. For example, our public specification of Chimera symmetry operators or placements of X-ray models in EM maps will explain precisely what Euler angle system, or quaternion representation or 3 by 3 rotation matrix is used and what coordinate frames are used, but will not specify how those values should be saved in a file. These values can be embedded just as numbers in our web page log with a link to the public specification defining how those numbers are interpreted. Copy and paste may be needed to use those values in other software, or for submitting data to a public archive like EMDB in cases where they are not already embedded in map and model files.

In addition to the HDF5 map and segmentation and XML coarse-grain model primary formats used by the Chimera molecular assembly tools we will develop and HDF5 coarse-grain model format. This is a collaborative project to develop a format that can handle larger ensemble data sets than our current XML format and will be used for exchanging computed models from the Integrated Modeling Platform (IMP) software (Andrej Sali lab) with Chimera for visualization and analysis. IMP generates large ensembles of models to characterize the degree of variability consistent with experimental restraints, for example, 50000 models were computed in the chromatin analysis shown in Figure 2-9 and 10000 having the fewest constraint violations were analyzed to determine statistical properties. The format will handle geometrical models: spheres, ellipsoids, tubes, connections, coloring, and grouping with representations suitable for making quantitative calculations (distances, angles, path lengths, positional variation).

Rous sarcoma virus capsid compared to HIV

Thu Apr 7 17:48 PDT 2011

1 Opened emd_1862.map (#0), size 255^3, grid spacing 1.27 Å
contour level -363, minimum -17270, maximum 14818
mean -6649.5, sd 3176.6, rms 7369.3, 16-bit integer

2 Opened 1em9.pdb (#1), 2 chains A B, 2257 atoms
rous sarcoma virus capsid protein: n-terminal domain
weight 30.04 KDa, x-ray 2.05 Angstroms

3 Deleted 1091 atoms (chain B), 1em9.pdb (#1)

4 Opened 1eoq.pdb (#2), 1 chain A, 1209 atoms, NMR
rous sarcoma virus capsid protein: c-terminal domain
weight 8.497 KDa

5 Molecular weight 24.04 KDa
2375 atoms 1em9.pdb (#1), 1eoq.pdb (#2).

6 Contour level 1239, emd_1862.map (#0)

7 Enclosed volume 1.78e6 Å^3, area 454.6e3 Å^3, emd_1862.map (#0)

8 **Note**: Set contour level to enclose volume 60 * 24 KDa * 1.2 Å^3/Da

9 Fit 1em9.pdb (#1) in emd_1862.map (#0)
correlation 0.7883, average density 4850
173 of 1166 atoms outside contour
simulated map 10.4 Å
envelope 0.95 mass enclosed, optimized overlap

rock movie
session9.py

10 Segmented emd_1862.map (#0)
regions 360, smoothing steps 3, step size 1 voxel

11 Segmentation grouping emd_1862.map (#0)
regions 120, smoothing level 4 voxels

12 Fit to segment 1eoq.pdb (#2)
in region 1966 of emd_1862.seg (#3)
correlation 0.8737, masked to region
simulated map 10.4 Å
envelope 0.95 mass enclosed, optimized overlap

13 Spin movie 1em9.pdb (#1), 1eoq.pdb (#2),
emd_1862.seg (#3)

spin
session13.py

14 Combined 1em9.pdb (#1), 1eoq.pdb (#2)
producing combination (#4)

15 Saved ca_fit.pdb (#4) relative to emd_1862.map (#0).

16 Symmetry copies ca_fit.pdb (#4) producing (#5.1-59)
icosahedral 222r, coordinate system emd_1862.map (#0).

17 Simulated map ca_fit.pdb (#4), (#5.1-59)
producing molmap res 10.4 (#6)
resolution 10.4, grid spacing 1.27 Å

session17.py

18 Saved map ca_fit_60_r10.4.mrc (#6)

19 Measure correlation ca_fit_60_r10.4.mrc (#6), emd_1862.map (#0).
corr 0.9403, corr about mean 0.6705, within contour level 0.19 (#6).

20 Resampled map ca_fit_60_r10.4.mrc (#6)
on grid emd_1862.map (#0) producing (#7).

21 Saved map ca_fit.mrc (#7).

22 Shifted map emd_1862.map (#0) by 6649 producing (#9).

23 Scaled map (#9) by 2.63e-5 producing map (#10).

24 Saved map emd_1862_normal.mrc (#10)

25 Morph maps emd_1862_normal.mrc (#10)
to ca_fit.mrc (#7) producing map (#11).

morph25.mov

26 Difference map
ca_fit.mrc (#7), emd_1862_normal.mrc (#10).
Minimum RMS scaling.
Contours -0.19 blue, 0.19 red.

spin26.mov

27 **Note**: 2 pentamers

session26.py

28 Molecular surfaces ca_fit.pdb (#4), (#5.1)

rock28.mov

29 Buried area ca_fit.pdb (#4), (#5.1)
solvent accessible area 486.4 Å^2, solvent excluded area 138.0 Å^2.

30 Buried residues ca_fit.pdb (#4), (#5.1)
11 residues, solvent accessible area cutoff 10 Å^2.
Total buried SAS area 303.0 Å^2, average 27.5 Å^2.

residue list

31 Opened 3p05.pdb (#12), x-ray 2.50 Angstroms
5 chains A B C D E, 7918 atoms, 1043 residues
pentameric hiv-1 ca
weight 105 KDa

32 Fit 3p05.pdb (#12) in emd_1862.map (#0)
correlation 0.53, average density 3548
2160 of 7918 atoms outside contour
simulated map 10.4 Å, envelope 0.95 mass enclosed
optimized overlap

33 **Note**: Superposition of HIV and RSV pentamers.
HIV 3p05.pdb (light blue and pink)
RSV ca_fit.pdb (blue and red)
Helices align. HIV squeezed toward center.

spin32.mov
session32.py

Fri Apr 8 00:15 PDT 2011

Figure 2-15: Illustration of web log file fitting Rous sarcoma virus (RSV) map and comparing HIV virus capsid pentamer. Records analysis done in Chimera fitting of N-terminal and C-terminal capsid protein domains using segmentation, applying symmetry, visualizing residual density, measuring C-terminal dimerization domain buried area, finding dimer interface residues, and aligning and comparing HIV pentamer. Links to maps, PDB files, the segmentation, session snapshots, lists of residues, spin, rock and morph movies, and larger images are provided. Quantitative results are recorded. The proposed Chimera logging tool will create this type of web page, semi-automatically, to record interactive analysis sessions. This example was mocked up by hand

Aim 4C: Software for Creating Animations conveying Hypotheses about Molecular Assembly Function.

While our primary focus is on sharing data and models useful as starting points for analysis by other researchers, we will also continue to make improvements for communicating ideas with images and animations. Molecular assemblies frequently undergo large-scale conformational rearrangements. For example, dengue virus undergoes a maturation process activated by reduced pH and proteolytic cleavage that completely rearranges the protein capsid architecture (34), a process required for infectivity. Immature and mature conformations are know from electron microscopy studies, and one hypothesis is that the transition involving rearrangement of 180 proteins is nucleated at one location on the icosahedral capsid and then spreads as a wave across the entire capsid. To illustrate such a transition we will allow morphing between conformations that propagates from one location in a large assembly to nearest neighbour molecules, continuing until the final experimental conformation. This will utilize developments we make for composite model building (Aim 1A) where relative positions and orientations of pairs molecules are used to make assemblies. Also we will explore using interface energies computed by the Virus Particle Explorer Database (ViPERdb) to allow the researcher to modulate the order in which conformational rearrangements occur. This is related to one of our goals in our collaborative project with ViPERdb that will color-code interfaces in molecular assemblies according to computed binding energies. An additional development to enhance animations of transitions will be the addition of simple Brownian to convey the stochastic nature of events. These capabilities will be applicable to both conformational rearrangements and assembly and disassembly processes, such as the formation of the conical HIV capsid that assembles from hundreds of capsid protein hexamers and 12 pentamers (21). These animation developments aim to convey conceptual information and hypotheses and are not intended as realistic depictions of the molecular events.

We plan to enable Chimera directly produce stereo movies for display on the latest generation of 3-dimensional computer displays (including a few currently available laptops) and televisions. The popular video archiving web site YouTube already supports stereo movies. Also the BluRay DVD movie standard supports 3-d content. And many entertainment movies are being shown in theaters with polarized glasses stereo viewing. Chimera already is able to capture left and right eye movies and the work involved in this effort is to evaluate software that can encode in standard 3-d formats. Simple improvements in our animation tools will allow dynamically changing depth of field in the same way as other camera parameters like field of view and zoom level are smoothly varied. Changes in perceived depth are a standard element of 3-d Hollywood movies.

Driving projects

Standardizing map, segmentation, symmetry and fit placements, EM Data Bank, Cathy Lawson – EMDB/PDB. Formats for coarse-grain models, IMP, Andrej Sali. Application of coarse-grain models to cellular structues, Manfred Auer. Stereo animations, National Center for Macromolecular Imaging, Matt Dougherty and Wah Chiu.

Timetable

Estimated level of effort reflects the full cost of development including research, collaboration, implementation, documentation, dissemination via talks, tutorials and publications, and refinements, enhancements and maintenance over the funding period. Effort to produce a prototype implementation is generally a small fraction (roughly one tenth) of this total effort required to produce a simple, reliable and widely used software tool. These specific aims cover only about one half of the projects that we will undertake in the area of molecular assemblies software development over the funding period. Additional projects in the target areas will be driven by collaborator research that cannot be foreseen over the long 5 years time frame.

Specific Aim	Task	Estimated Effort (man-months)	Potential Impact
Aim 1A	Composite model building	4	Unique next-generation modeling of pleiomorphic molecular assemblies.
Aim 1B	Composite model database access	4	Unified data access for a biological system.
Aim 1C	HIV demonstration composite model	4	Synthesizes rich HIV structural data.
Aim 2A	Map fitting methods	6	Higher accuracy of molecular interfaces in models from electron microscopy.
Aim 2B	Efficient handling of large maps	6	Analysis tools for highest-resolution cryoEM for assemblies that cannot be crystallized.
Aim 2C	EM validation using SAXS	2	Verify correctness of EM maps using simple experimental technique.
Aim 3A	Segmentation and feature extraction	6	Enables analysis of cellular imaging data sets that is currently too time-consuming.
Aim 3B	Single-particle tomography analysis	6	Advances nascent technology for in-vivo data averaging to obtain high resolution.
Aim 4A	Web page analysis log	12	Simplifies sharing computational models for future studies and database archiving.
Aim 4B	Exchange data formats	6	Develops community data standards to allow build-up of operational models of molecular machinery.
Aim 4C	Animation creation tools	4	Enables researchers to communicate hypotheses on molecular assembly function.
Total		60	

References for TR&D #2

1. Goddard TD, Huang CC, Ferrin TE. Software extensions to UCSF chimera for interactive visualization of large molecular assemblies. Structure. 2005;13(3):473-82.
2. Couch GS, Hendrix DK, Ferrin TE. Nucleic acid visualization with UCSF Chimera. Nucleic Acids Res. 2006;34(4):e29.
3. Goddard TD, Huang CC, Ferrin TE. Visualizing density maps with UCSF Chimera. J Struct Biol. 2007;157(1):281-7.
4. Pintilie GD, Zhang J, Goddard TD, Chiu W, Gossard DC. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. Journal of structural biology. 2010;170(3):427-38. PMCID: 2874196.
5. Johnson JE, Speir JA. Quasi-equivalent viruses: a paradigm for protein assemblies. Journal of Molecular Biology. 1997;269(5):665-75.
6. Krukenberg KA, Street TO, Lavery LA, Agard DA. Conformational dynamics of the molecular chaperone Hsp90. Q Rev Biophys. 2011:1-27.
7. Schneidman-Duhovny D, Hammel M, Sali A. FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic acids research. 2010;38(Web Server issue):W540-4. PMCID: 2896111.
8. Marko M, Hsieh C, Schalek R, Frank J, Mannella C. Focused-ion-beam thinning of frozen-hydrated biological specimens for cryo-electron microscopy. Nature methods. 2007;4(3):215-7.
9. Denk W, Horstmann H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. PLoS biology. 2004;2(11):e329. PMCID: 524270.
10. Helmstaedter M, Briggman KL, Denk W. 3D structural imaging of the brain with photons and electrons. Curr Opin Neurobiol. 2008;18(6):633-41.
11. Li A, Gong H, Zhang B, Wang Q, Yan C, Wu J, et al. Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. Science. 2010;330(6009):1404-8.
12. Krebs WG, Gerstein M. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. Nucleic Acids Res. 2000;28(8):1665-75.
13. Dougherty MT, Folk MJ, Zadok E, Bernstein HJ, Bernstein FC, Eliceiri KW, et al. Unifying Biological Image Formats with HDF5. Communications of the ACM. 2009;52(10):42-7. PMCID: 3016045.
14. Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, et al. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol. 2011;18(1):107-14. PMCID: 3056208.
15. Lasker K, Phillips JL, Russel D, Velazquez-Muriel J, Schneidman-Duhovny D, Tjioe E, et al. Integrative structure modeling of macromolecular assemblies from proteomics data. Molecular & cellular proteomics : MCP. 2010;9(8):1689-702. PMCID: 2938050.
16. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends in Biochemical Sciences. 1998;23(9):358-61.
17. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. Journal of Molecular Biology. 2007;372(3):774-97.
18. Pieper U, Eswar N, Webb BM, Eramian D, Kelly L, Barkan DT, et al. MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic acids research. 2009;37(Database issue):D347-54. PMCID: 2686492.
19. Lawson CL. Unified data resource for cryo-EM. Methods in enzymology. 2010;483:73-90. PMCID: 2966391.
20. Carrillo-Tripp M, Shepherd CM, Borelli IA, Venkataraman S, Lander G, Natarajan P, et al. VIPERdb2: an enhanced and web API enabled relational database for structural virology. Nucleic acids research. 2009;37(Database issue):D436-42. PMCID: 2686430.
21. Pornillos O, Ganser-Pornillos BK, Yeager M. Atomic-level modelling of the HIV capsid. Nature. 2011;469(7330):424-7.

22. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr., Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature. 2009;460(7256):711-6. PMCID: 2724670.
23. Han BG, Dong M, Liu H, Camp L, Geller J, Singer M, et al. Survey of large protein complexes in D. vulgaris reveals great structural diversity. Proc Natl Acad Sci U S A. 2009;106(39):16580-5. PMCID: 2742403.
24. Grigorieff N, Harrison SC. Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy. Curr Opin Struct Biol. 2011;21(2):265-73.
25. Baker ML, Abeysinghe SS, Schuh S, Coleman RA, Abrams A, Marsh MP, et al. Modeling protein structure at near atomic resolutions with Gorgon. Journal of structural biology. 2011.
26. Siebert X, Navaza J. UROX 2.0: an interactive tool for fitting atomic models into electron-microscopy reconstructions. Acta crystallographica Section D, Biological crystallography. 2009;65(Pt 7):651-8. PMCID: 2703571.
27. Cardone G, Purdy JG, Cheng N, Craven RC, Steven AC. Visualization of a missing link in retrovirus capsid assembly. Nature. 2009;457(7230):694-8. PMCID: 2721793.
28. Trabuco LG, Schreiner E, Gumbart J, Hsin J, Villa E, Schulten K. Applications of the molecular dynamics flexible fitting method. Journal of structural biology. 2011;173(3):420-7. PMCID: 3032011.
29. Topf M, Baker ML, John B, Chiu W, Sali A. Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. Journal of structural biology. 2005;149(2):191-203.
30. Settembre EC, Chen JZ, Dormitzer PR, Grigorieff N, Harrison SC. Atomic model of an infectious rotavirus particle. Embo J. 2011;30(2):408-16. PMCID: 3025467.
31. Yoo TS, Ackerman MJ, Lorensen WE, Schroeder W, Chalana V, Aylward S, et al. Engineering and algorithm design for an image processing Api: a technical report on ITK--the Insight Toolkit. Stud Health Technol Inform. 2002;85:586-92.
32. Kremer JR, Mastronarde DN, McIntosh JR. Computer visualization of three-dimensional image data using IMOD. Journal of structural biology. 1996;116(1):71-6.
33. Schmid MF, Booth CR. Methods for aligning and for averaging 3D volumes with missing data. Journal of structural biology. 2008;161(3):243-8. PMCID: 2680136.
34. Yu IM, Zhang W, Holdaway HA, Li L, Kostyuchenko VA, Chipman PR, et al. Structure of the immature dengue virus at low pH primes proteolytic maturation. Science. 2008;319(5871):1834-7.