Segmentation and Registration of Molecular Components in 3-Dimensional Density Maps from Cryo-Electron Microscopy

by

Grigore D. Pintilie

B.Sc. York University, Toronto (1999)
M.Sc. University of Toronto (2001)

SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL ENGINEERING & COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2010

Signature of Author: ____
Department of Electrical Engineering & Computer Science
December 17, 2009

Certified by: ____
David C. Gossard
Professor of Mechanical Engineering
Thesis Supervisor

Accepted by: ____
Terry P. Orlando
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Segmentation and Registration of Molecular Components in 3-Dimensional Density Maps from Cryo-Electron Microscopy

by

Grigore D. Pintilie

Submitted to the Department of Electrical Engineering & Computer Science on December 17, 2009,
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Electrical Engineering and Computer Science

ABSTRACT

Cryo-electron microscopy is a method that produces 3D density maps of macromolecular complexes. Segmentation and registration methods are heavily used to extract structural information from such density maps.

Segmentation aims to identify regions in a density map corresponding to individual molecular components, so as to allow us to understand their complex arrangements and the relation of these arrangements to the function of the complex. Currently used segmentation methods rely to a large degree upon user interaction and thus are tedious and yield subjective results. We present a multi-scale segmentation method requiring very little interaction and guidance from the user. The segmentation accuracy of this method is quantified for simulated density maps, using a shape-based metric. The method is applied to several density maps of various sizes and complexity, producing accurate results.

Registration of molecular structures with density maps helps to relate the vast structural information from X-ray crystallography with the structural information contained in cryo-electron microscopy density maps. The most reliable registration methods to date depend on exhaustive search, which is time-intensive and scales poorly with map and structure size. Two methods are presented that achieve direct registration of structures with density maps, based on the alignment of the structures to segmented regions. The registrations are refined using a gradient-based method, which locally optimizes the density cross-correlation score. A search algorithm is presented for automatically finding groups of regions that produce correct registrations. The accuracy of these registration methods is measured using simulated density maps, and then the methods are used to register structures of individual proteins and subunits with density maps obtained by cryo-electron microscopy.

Thesis Supervisor: David C. Gossard
Title: Professor of Mechanical Engineering

Acknowledgements

I'd like to thank my advisors David Gossard and Jonathan King for their guidance and inspiration through the past four years. I'd also like to thank professors Srini Devadas and Berthold Horn, members of my doctoral committee, for their insights into this work, and for their help in seeing it become a thesis. Also key in this work has been Wah Chiu from the Baylor College of Medicine, who has provided much motivation along the way.

I'd also like to thank all my past advisors, who have inspired me to pursue a path in academia. They are my undergraduate research advisors James Elder, Michael Jenkins, and Wolfgang Stuerzlinger at York University, M.Sc. thesis advisors Tim McInerney and Demetri Terzopoulos at University of Toronto, and Christopher Hogue now at National University of Singapore. They all played key roles in helping me find my passion for research, and I strive to make them proud.

Lastly and most importantly, I'd like to thank my wife Kelly Slate for putting much joy in the years during this journey. I also owe much gratitude to my family for their support; thank you.

Table of Contents

List of Figures

Chapter 1. Introduction

1.1 Motivation

Macromolecular complexes are the building blocks and workhorses of biological organisms. A macromolecular complex is composed of components such as proteins and ribonucleic acids (RNA). Obtaining the structures of such complexes is critical for better understanding how they function, and also for understanding why they sometimes fail to function properly, which is a common cause of many diseases.

A well-established method used to determine the structure of macromolecular complexes is X-ray crystallography. With this method, 3-dimensional electron density maps are reconstructed from crystal diffraction patterns, typically to high-enough resolution such that the position of individual atoms can be determined [1]. However the complexes must first crystallize for this method to be applicable. Thus this method cannot be applied universally, for example to very large complexes, complexes that are structurally dynamic, or complexes that are embedded in cellular membranes.

Cryo-electron microscopy (cryo-EM) doesn't require crystallization and hence can be applied to a wider variety of complexes. On the other hand, it involves a large amount of computation and typically produces lower-resolution maps from which atomic positions cannot be directly determined. However the reconstruction methods are improving, and cryo-EM is increasingly being used to uncover the structure of an increasing number of complexes. As a result, methods for analyzing the resulting density maps are also increasing in importance.

1.2 Problem statement

The relatively low resolution of cryo-EM density maps makes the extraction of structural information a challenge. The tools used to build atomic models in an electron density map from X-ray crystallography have not yet been applied successfully, since those tools require higher resolutions. Hence, segmentation and registration methods are commonly used to extract structural information from cryo-EM density maps.

The segmentation and registration methods currently used in the analysis of cryo-EM density maps have serious limitations, as will be described below. These limitations will only get worse as the size and complexity of the density maps obtained continue to grow. In this thesis, the aim is to develop fast, efficient, accurate, and objective methods for segmentation and registration. Achieving such methods will mean that the analysis of density maps will be easier and faster, and it will allow us to more quickly and accurately extract structural information from density maps obtained using the cryo-EM method.

1.2.1 Segmentation

During the segmentation process, the goal is to identify regions in a density map that belong to individual molecular components. This helps us to learn what components make up the complex, and how its composition may be related to its function. Current segmentation methods applied to density maps from cryo-EM require a lot of guidance from the user, and hence are tedious and labor intensive. The results depend to a large degree on the skill and knowledge of the user, and hence they can be highly subjective.

Segmentation of cryo-EM density maps is a hard problem for several reasons. Firstly, the regions to be segmented are 3D in nature, and interacting with 3D structures on 2D display devices is not an easy task to start with. Secondly, molecular shapes are very complex, and thus it is not easy for users with little prior skills or knowledge to identify these shapes. Thirdly, the boundaries between molecular components don't clearly stand out when visualizing a density map, and thus are hard to identify. To solve these problems, the segmentation method must be able to efficiently deal with 3D data, and be mostly automated, so that the users are not relied upon to identify the 3D shapes themselves.

1.2.2 Registration

The process of registration involves taking a known structure of a molecular component, for example obtained using X-ray crystallography, and placing it such that it best overlaps an analogous component in the density map. The registration of a structure to a density map is a useful analysis tool. Since cryo-EM maps are of lower resolution, they cannot be used to directly determine atomic positions of each component. The registration of structures with the map can be used as a way to build atomic-level structural models of the complexes captured in the cryo-EM density map. Moreover, the process of registration can be used to uncover the relationships between the many presently known structures and the density maps being produced by cryo-EM.

In this work, the rigid-body registration problem is considered, which assumes that the structures to be registered are similar to the structures in the cryo-EM density map. Such a registration involves only 6 degrees of freedom: a 3-dimensional position and a 3-dimensional orientation. Current methods for rigid registration are problematic. For example, the structure can be positioned and oriented manually by the user, however this is not an easy task since 2D displays and input devices are ill suited for 3D manipulation. Exhaustive search can solve the problem, however it requires long computation times, and the computation times scale poorly with the sizes of the density map and the structures being registered. Methods based on aligning feature points in the structure and density map are very fast. However to reliably identify feature points is itself a hard problem, and automatic identification of feature points does not always work well, especially in the presence of noise in the density map. To solve these problems, the registration method must therefore not rely on user-interaction, feature points, or exhaustive search. A more direct way to register the structure is needed.

Chapter 2. Background

2.1 Atomic structures

An atomic structure is defined by specifying a 3D position for each atom that it contains, along with a list of covalent bonds between atoms. In a molecule, each atom is bonded to at least one other atom that is also in the molecule. A structure can consist of a single molecule or it can consist of multiple molecules that are held together by van der Waals and electrostatic forces [2]. Structures that are composed of two or more molecules are commonly called *molecular complexes*. Macro-molecular complexes are composed of two or more large molecules such as proteins or ribo nucleic acid molecules (RNAs).

2.1.1 Proteins

A protein is a chain of covalently bonded amino-acid molecules [3]. An amino-acid molecule has *backbone* atoms, which connects it to other amino acids, and *side-chain* atoms. The atoms that make up the side-chain branch out from one of the atoms in the backbone. The sequence of amino acids along in the chain is commonly referred to as the *primary structure* of the protein. A chain of amino-acid molecules commonly folds into two types of *secondary structures*, namely *alpha helices* or *beta strands*. These secondary structures tightly pack against one another when the protein folds. The fold is also referred to as the *tertiary structure* of a protein. Finally, the *quaternary structure* describes how different proteins bind to one another to form complexes, through hydrophobic or electrodynamic interactions. These four levels of protein structure are illustrated in Figure 2.1.

In every molecule such as a protein, an atom is covalently bonded with at least one other atom that is part of the same molecule. Atoms that are covalently bonded to each other tend to be close together. On the other hand, non-covalently bonded atoms tend to be further away from one another. Hence, interfaces between different molecules tend to be less dense, since atoms are further apart, whereas the spaces within a molecule, in which atoms are closer together due to covalent bonding, are denser.

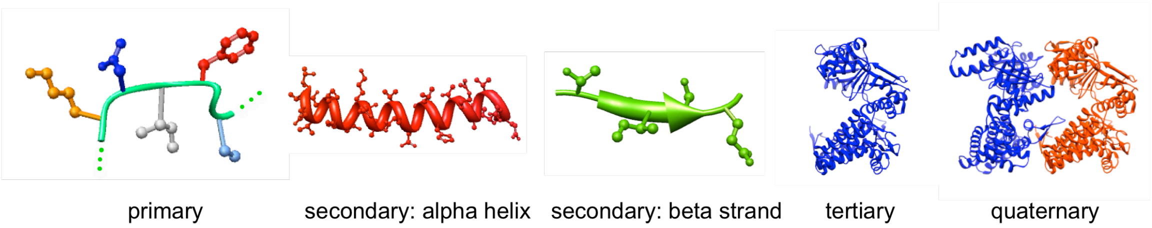

Figure 2.1. Protein structure. The primary structure is the sequence of amino acids connected together to form a long chain. In the image on the left, a chain of 4 amino acids is drawn. Secondary structures include alpha helical and beta strand segments. The tertiary structure is the overall arrangement of the entire chain. Quaternary structure captures how multiple proteins are bound to one another. The structures here are show in the ribbon representation, in which a tube interpolating the positions of the backbone atoms is drawn. Side chains in the first three figures are shown with ball-and-stick model representing atoms and covalent bonds respectively. They are not drawn in larger structures at the tertiary and quaternary level for clarity. All the images are created from the crystal structure available in the protein data bank (PDB:1xck).

2.1.2 Ribonucleic acid (RNA) molecules

Ribonucleic acid (RNA) molecules are long chains of covalently bonded molecules called *nucleotides* or *bases*. RNA molecules are usually single stranded, a strand consisting of a single chain of covalently bonded bases. In RNA molecules, bases from different points along the chain can form base pairs, which physically holds those points together. RNA strands, much like proteins, have specific 3D structures. RNA molecules also often bind to other RNA molecules or proteins. An example of such an RNA-protein complex is the ribosome. The ribosome consists of two subunits, where each subunit is a RNA-protein complex. These 4 levels of RNA structure are illustrated in Figure 2.2.

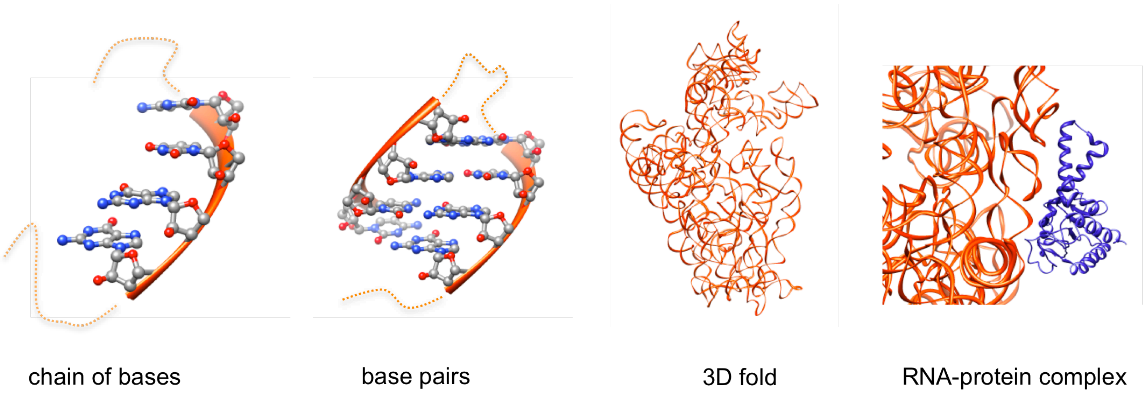

Figure 2.2. RNA structure. On the left, the RNA chain structure is shown, showing the atoms in 4 of the bases along the chain. The backbone atoms in the 4 bases are simplified and drawn as an orange ribbon, for clarity. Second from left, bases from different positions in the RNA chain forming base pairs are shown. Second from right, the 3D structure that a chain can take is shown. On the right, an RNA (orange ribbon) and protein (blue ribbon) complex is shown. All images are created from the crystal structure of the E-Coli ribosome small subunit, available from the protein data bank (PDB:2avy).

2.2 X-ray crystallography

The atomic positions in a molecular complex can be determined using X-ray crystallography [1]. An X-ray beam is sent through a crystal composed of many copies of the same complex, all arranged regularly in a crystal lattice, producing a diffraction pattern. A 3D electron density map is computed from this diffraction pattern, from which the atomic positions are determined. The structures of many proteins and complexes have been obtained to date with this method. All these structures can be accessed in a publicly accessible database, the protein data bank (PDB) [4]. The main limitation of X-ray crystallography is that the proteins or protein complexes must first crystallize. Many proteins and protein complexes, which are flexible in nature or are typically embedded in cellular membranes, do not crystallize, and hence heir structures may never be found using this method.

2.3 Cryo-electron microscopy (cryo-EM)

In cryo-EM, the crystallization of a molecular complex is not required. Instead, a purified solution containing many copies of the same complex is frozen, and then imaged with an electron microscope. This produces many 2D images of the complex in varying orientations. First, the images that correspond to the same orientation are found and averaged to improve the signal-to-noise ratio. Then the orientation of each average image is found with respect to the 3D volume, and back-projection is used to reconstruct a 3D density map. This back-projection methodology has been used in other domains, for example radio astronomy and medical imaging, even with arbitrary ray-sampling schemes [5].

The process described above for the reconstruction of 3D density maps of molecular complexes is commonly called *single particle reconstruction*, and is illustrated in Figures 2.3 and 2.4. Several software tools have been developed implementing this method, for example EMAN [6] and SPIDER [7]. These tools are constantly being improved, and thus are they are becoming more efficient and automated. Many 2D images (>100,000) and computational resources are typically required to produce an accurate density map.

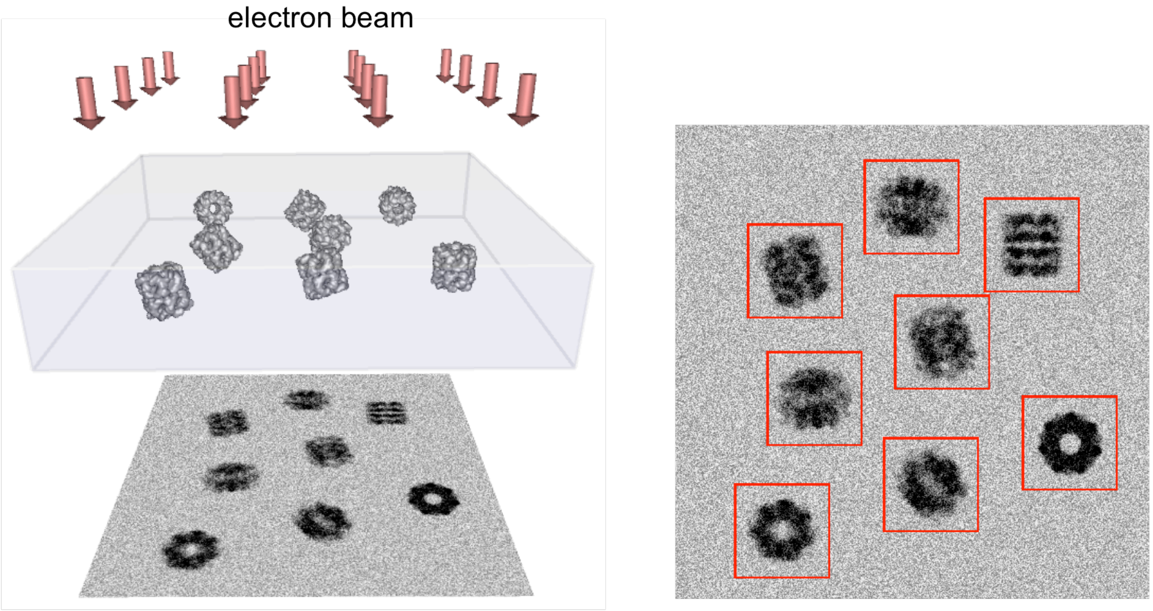

Figure 2.3. Illustration of imaging and boxing steps in cryo-EM. Many copies of the same complex are placed in a thin film, which is then cryogenically frozen to drastically reduce Brownian motion. The film is imaged using an electron beam, producing many 2D images of the complex in different orientations (left). Each image is a radon projection of the complex [8]. The resulting 2D image is shown on the right, where individual images of the complex have been identified with red boxes. These illustrations contain only simulated data.

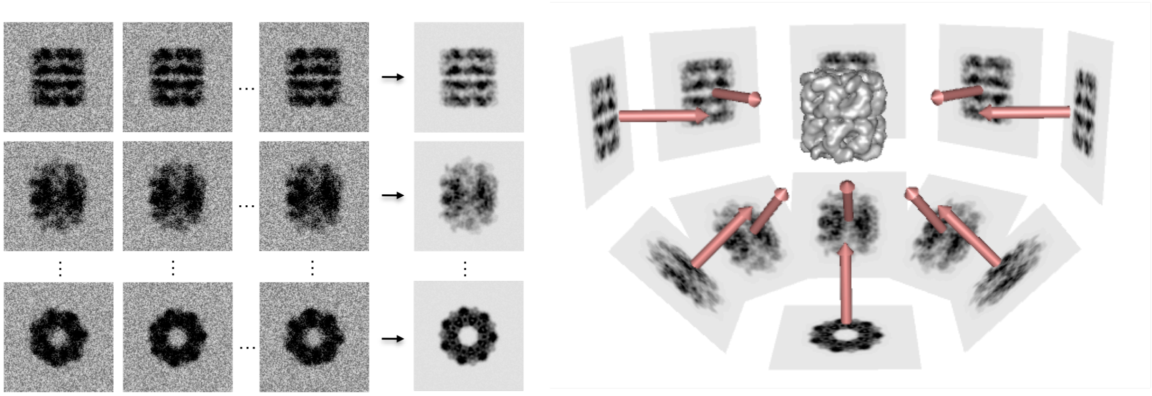

Figure 2.4. Single-particle reconstruction process for cryo-EM. Boxed 2D images are first clustered, based on similarity, to find representatives of different orientations. On the left, images from 3 clusters are shown. Images in the same cluster are averaged to create a 'cluster-average image' in which the signal-to-noise ratio is improved. The 3D orientations of the resulting average images are then found, with respect to the 3D volume. Using these orientations, the images are back-projected to yield a 3D density map (right). The last two steps are iterative – first a guess of the 3D volume is made, which is used to compute the orientations of each average image with respect to this volume. The average images are then used to recreate the volume by back-projection. Then the first step is repeated, but using the newly computed volume. The process stops when covergence is reached. These illustrations contain only simulated data.

2.3.1 Visualization of 3D density maps

A density map is defined as a 3D grid, with each grid point having a density value that reflects the electron density at the corresponding point in space. To visualize a density map, this 3D grid must somehow be projected to form a 2D image. A common way of doing this is to create a 3D iso-surface through the grid, which is projected to create the 2D image. The iso-surface is a collection of points that have the same density value; the latter is also often called a *threshold*. The surface points, along with triangles between them, which can be projected and drawn, can be obtained using the marching cubes algorithm [9].

Another way to visualize a density map is to show a 2D slice through the 3D grid. The slice shows the density values using varying intensities, e.g. darker values representing higher densities. Both iso-surface and slice representations are illustrated in Figure 2.5. Figure 2.6 shows the iso-surfaces resulting at different thresholds in the same density map.

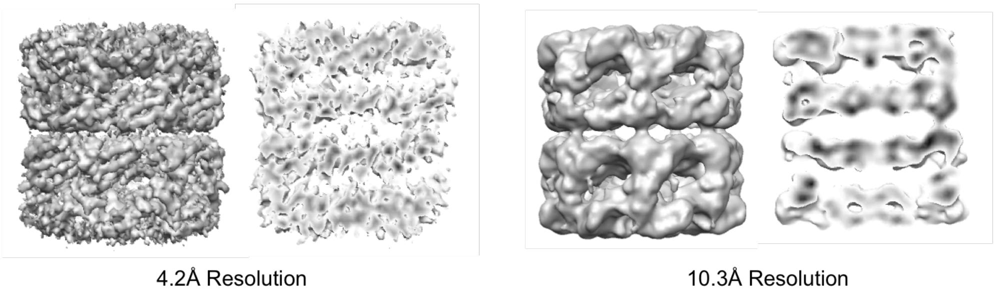

Figure 2.5. Cryo-EM density maps of the GroEL complex at two different resolutions. The maps were obtained from the EMDB (EMDB:5001 and 1042 respectively). Iso-surface (left) and slice (right) representations are shown for maps at two resolutions. Higher resolution density maps (lower number) have a greater amount of detail, while lower resolution (higher number) are smother.

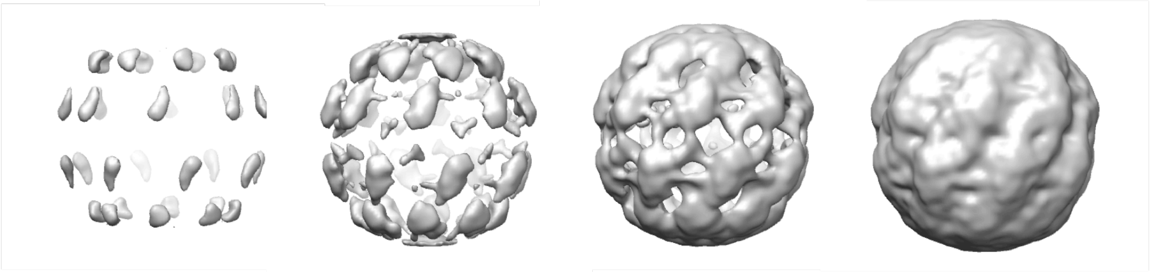

Figure 2.6. Cryo-EM density map of Mm-cpn, visualized with iso-surfaces at 4 different density thresholds. The surfaces shown are drawn at decreasing threshold values, with the surface on the left having the highest threshold. At higher threshold, the inner and denser parts of the complex are seen, while at lower thresholds a larger outer envelope of the complex can be seen.

2.3.2 The Cryo-Electron Microscopy Data Bank (EMDB)

The EMDB is a publicly-accessible repository where cryo-EM density maps are deposited [10]. This database is relatively new, with maps having been deposited only as far back as 2002. In figure 2.7, a bar-char shows how the number of cryo-EM density maps in the EMDB has been increasing. Figure 2.8 shows 5 cryo-EM maps taken from the EMDB, illustrating the wide range in size and complexity of maps in the EMDB.

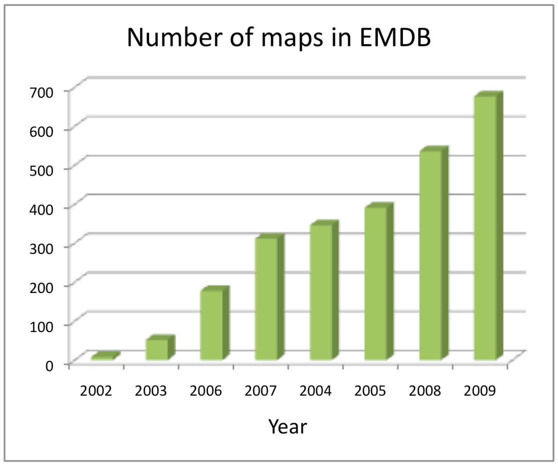

Figure 2.7. Chart of number of total cryo-EM maps in EMDB vs. year since its conception in 2002. The number is increasing, a sign of increasing adoption of the method.

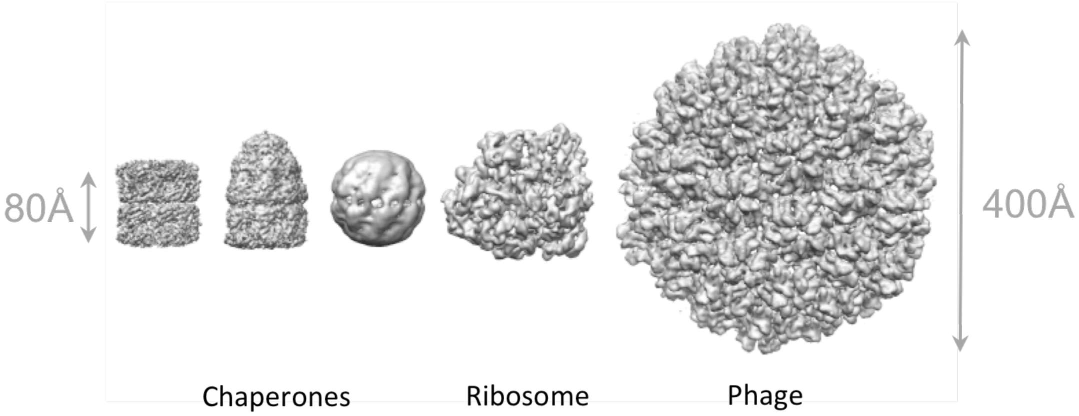

Figure 2.8. Five cryo-EM density maps of complexes with varying sizes and complexity. All density maps are visualized using iso-surfaces, all drawn at the same scale, thus showing their relative sizes.

2.3.3 Resolution of a cryo-EM density map

The resolution of a cryo-EM density map is calculated using the Fourier shell correlation criterion [11], and reflects the level of detail in the density map. Resolutions of up to ~4Å have been obtained to date [12-15]. These resolutions are not high enough for the positions of individual atoms to be easily determined. By comparison, X-ray crystallography is capable of obtaining density maps with higher resolutions, from which the positions of the individual atoms can be determined.

2.4 Simulation of density maps from atomic structures

A simulated density map aims to capture the varying electron-density due to atoms in a structure at discrete points in space. The electron density in due to a single atom is complicated and requires quantum mechanics for an accurate description. A simplified approximation is based on placing Gaussians functions at the coordinates of each atom [16]. The Gaussian function captures both the exponentially decaying radial functions of electron density in an atom as well as the resolution limitations inherent in the cryo-EM reconstruction process. Given a structure with *N* atoms at positions $\vec{r}_i$ and atomic number a_i, the approximate electron density function ρ at every point in space $\vec{p}$, is expressed as:

$$\rho(\vec{p}) = \sum_{i=1}^{N} a_i \exp\left(-\frac{\|\vec{p}-\vec{r}_i\|^2}{2\sigma^2}\right) \quad (2.1)$$

To simulate a density map, this function is discretized on a 3-dimensional grid, with grid points evenly spaced in all dimensions. The typical step sizes vary from 1Å to 4Å. The maps can vary in size from 100x100x100 voxels to as high as 500x500x500 voxels. The latter, even at such small number of grid points per dimension, can already push the limits of current computers in terms of memory storage and visualization.

To discretize the density function, a 3D grid is first created around the structure, and the atomic mass of each atom is extrapolated to the 8 grid points nearest to its position. This map is then convolved using a Gaussian filter. The width of the filter is proportional to the resolution of the density map – the larger the width, the lower the resolution. Based on the same principle as the Fourier shell correlation criterion, which is used to determine the resolution of an experimental cryo-EM density map [11], the width of the Gaussian kernel is set to $0.187r$ where *r* is the resolution of the resulting density map. This makes the Fourier transform of the Gaussian filtering function fall to half its maximum magnitude at the frequency *1/r*. Two density maps of the same complex, simulated at different resolutions, are illustrated in Figure 2.9.

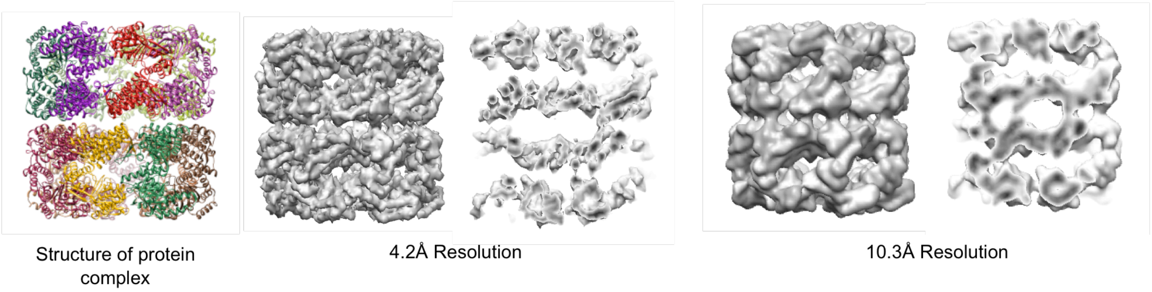

Figure 2.9. The crystal structure of a protein complex, and resulting simulated maps at two different resolutions. The structure is of the chaperone GroEL, and contains 14 proteins, which are arranged circularly in a barrel-like shape. Each protein in the image on the left is drawn as a ribbon with a different color. The resulting density maps are shown using iso-surface and slice representations. These simulated maps can be visually compared to cryo-EM density maps of the same complex, shown in Figure 2.5. By comparison they are very similar, however the cryo-EM maps can be seen to have a slightly more jagged appearance due to noise.

Chapter 3. Prior work

3.1 Segmentation

Segmentation has been a widely studied subject, for example in computer vision [17] and medical image analysis [18]. Approaches to segmentation include finding contours around objects by edge detection [19,20], active contours [21], or level sets [22], and partitioning regions in an image based on graph cuts [23,24], random walks [25], or topological methods such as mean-shift [26,27] and watershed [28-30]. Multi-scale analysis has been used along with some of these methods; it involves smoothing of the input image, which reduces the number of segmented contours or regions while retaining salient features [31-36].

For the segmentation of cryo-EM density maps into regions corresponding to individual molecular components, several of these methods have been used, for example the level set and watershed methods. The level-set method can produce good results [37], but it heavily depends on prior placement of seed points in each region to be segmented, which cannot be done automatically in a reliable way. The watershed method is very effective in lower-resolution density maps, and requires little if any user interaction [38,39]. However its main limitation is that it typically produces too many regions in maps with high resolution or a lot of detail, an effect typically referred to as over-segmentation.

Methods for dealing with over-segmentation include grouping of regions based on metrics such as topological height [40] or topological persistence [28]. They generally do not produce accurate segmentations because the metrics they use are based on local information, which is unreliable in the presence of noise and discretization error. User-guided grouping of regions is not feasible, given that the number of regions produced in a high-resolution density map can be on the order of thousands or more.

Due to these difficulties, segmentation of cryo-EM maps is presently performed mainly using interactive methods, in which users manually select out regions belonging to each molecular component. Several software tools implement such approaches [41-43]. This is a labor-intensive task that can take many hours to accomplish, and requires prior knowledge and skill. A faster, more automatic and less subjective method has thus far remained elusive.

3.1.1 Filtering methods

The application of a filter to an image is usually an important step before the application of a segmentation method [17]. Filters can be broadly divided into linear filters and non-linear filters. Linear filters include low-pass filters, which keep only the low frequencies in an image and thus have a smoothing effect, and high-pass filters, which keep only higher frequencies, typically emphasizing contours around objects. A filter based on the Gaussian function, for example, is a low-pass filter, while a filter that computes gradient magnitudes is a high-pass filter.

High-pass filters are standard in image processing for computer vision and medical image analysis, since they bring out the contours around objects, as shown in Figure 3.1. Low-pass filters smooth the image, thus reducing high-frequency noise, and in doing so they also tend to blur the contours around objects. This effect is also illustrated in Figure 3.1. To reduce the blurring of object contours, anisotropic low-pass filters can be used, which attempt to smooth in directions perpendicular to contours [32], thus reducing the effect of blurring on the contour.

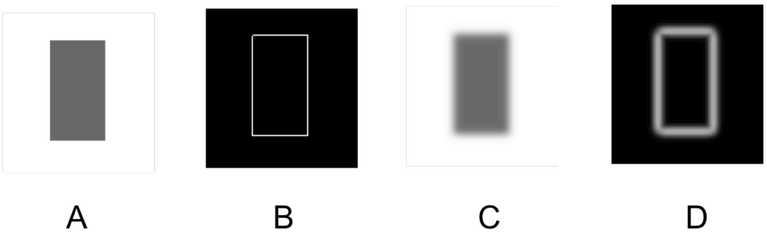

Figure 3.1. Effect of low-pass and high-pass filters applied to the image of a grey rectangle. A is the initial image. B is the result of the application of a high-pass filter to A – it shows the contour of the rectangular object. C is the result of the application of a low-pass filter (Gaussian) to A – the result is that the image of the rectangle is blurred. D is the result of the application of a high-pass filter to image C – again the contour is brought out, however the contour is now blurry.

3.1.2 Scale-space

Low-pass filters, in both isotropic [31] and anisotropic forms [32], are used for scale-space analysis of an image. The scale space for an image is created by the application of a low-pass filter, which progressively blurs the image. Coarser scales have a larger degree of smoothing. The features that persist through scale space tend to be the ones that are more salient [32]. The use of an anisotropic filter avoids the blurring of contours at coarser scales [32].

3.1.3 The Gaussian filter and scale-space

The Gaussian filter is the most commonly used filter in scale-space analysis, and it has been shown to be ideal in this process [44]. Its effect is similar to the application of a diffusion or heat transfer equation [32]. A provable property is non-enhancement of local extrema in any dimension. The latter is important, because it implies that as a result of the application of the filter, the result is bounded, and no spurious detail or noise will be introduced into the image.

3.2 Registration

Registration is also extensively used in computer vision and medical image analysis. For example, it can be used to find out which images of objects taken from a database appear in a picture of a scene containing many objects, where exactly they appear, and in what orientation. In the study of cryo-EM, very similarly, a database of structures – the protein data bank (PDB) [4], is available, and structures from it can be registered with density maps obtained by cryo-EM. This process helps to build more detailed models of the structures seen in cryo-EM maps, since structures from the PDB contain accurate atomic coordinates of most if not all of the atoms in a structure.

Several approaches are typically taken for registration [45]. They include interactive placement by the user, alignment of corresponding feature points in the template and reference image [46], exhaustive search [17], and the use of moment-based shape-descriptors [47]. Generally these methods are driven by a registration metric, which evaluates how well the image being registered matches the corresponding sub-part that it overlaps in the image it is being registered with.

Registration based on matching of two sets of feature points, given an correspondence between points from each set, can be done very efficiently in closed form [48]. In such methods, the mean distance between feature points can serve as the registration metric, with the goal of the registration being to reduce the mean distance between corresponding feature points. However, corresponding feature points must first be identified before such a procedure can be used. An example of a method that automatically identifies and registers images based on feature points is SIFT [46].

When feature points are not used, the metrics are typically based on cross-correlation [49] or mutual information [50] between corresponding intensities or color values in the template and reference images. Computation of the cross-correlation in Fourier space is also particularly useful for speeding up registration methods based on exhaustive search [51,52].

In the cryo-EM literature, previously reported registration methods include manual placement [41,53,54], exhaustive search, e.g. EMFIT [55], DOCKEM [56], SITUS [57], URO [58], Foldhunter [59], FRM [60], and ADP_EM [61], and matching of feature points [62], or surface features, e.g. 3SOM [63,64]. Manual placement is tedious and prone to error, exhaustive search is time-intensive and scales poorly with map size, and feature-matching methods depend on reliably identifying the same features in the map and structure being registered. The difficulties in the latter in particular mean that manual placement or exhaustive search methods are the predominantly used methods, and thus the process remains laborious and very time-consuming.

3.2.1 Density cross-correlation for density maps

The density cross-correlation metric can be computed between two density maps, but not between a structure and a map. Thus, to be able to compute this score, a density map is simulated for the structure being registered. The simulated density map is generated at the same resolution as the reference density map, and using the same grid spacing. The simulated density map of the structure being registered will be referred to as the *template* density map, and the density map with which it is being registered will be referred to as the *reference* density map.

The cross-correlation score is computed between density values taken at points in the template density map, and density values at the same positions in the reference density map. The points in the template density map are translated and rotated by transforms $\vec{T}$ and $\vec{R}$ respectively, which define the registration parameters. The cross-correlation is computed as follows:

$$cc(\vec{T},\vec{R}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^{\ 2}}\sqrt{\sum_{i=1}^{N} b_i^{\ 2}}} \quad (3.1)$$

where

$$\vec{a} = [a_1, a_2, \ldots, a_N]$$

and

$$\vec{b} = [b_1, b_2, \ldots, b_N]$$

In the above, $\vec{a}$ and $\vec{b}$ are vectors containing *N* scalar values. The vector $\vec{a}$ contains density values above a given threshold, at grid points in the template density map. The vector $\vec{b}$ contains density values from the reference density map, calculated by trilinear interpolation at locations corresponding to the grid points from which the density values in $\vec{a}$ are taken. An example registration between two 2D density maps is illustrated in Figure 3.2.

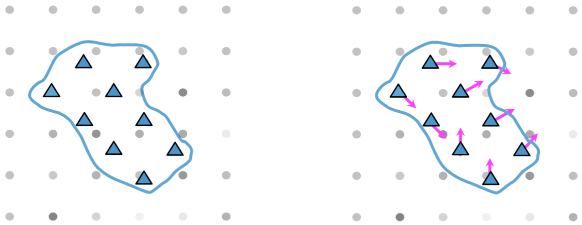

Figure 3.2 Example registration of 2D density maps. Grid points in the reference density map are drawn using solid circles, and grid points in the template density map are drawn using triangles. An isocontour is shown in the template density map and only the grid points within this contour are shown. The density values taken from the reference map, to compute the cross-correlation score, are computed by interpolation, at the positions of the grid points in the template map. Gradients, which are used in the local refinement of a registration, are computed at these same positions, also by interpolation, as shown by arrows in the image on the right.

3.2.2 Local refinement of a registration

Local refinement aims to improve the density cross-correlation score of the registration, by changing the translation and rotation ($\vec{T}$ and $\vec{R}$) registration parameters. This process can be accomplished through optimization methods, randomized Monte-Carlo search [65,66], or gradient-based search [67]. The gradient-based approach was adopted here, mainly because it is faster and more efficient compared to the other methods. It has previously been implemented in UCSF Chimera [68], a software which we extended in order to implement the segmentation and registration methods.

The gradient-based local refinement method makes use of density gradients computed in the reference map, at the positions of the grid points in the template density map (see Figure 3.2). First, the gradients are computed at the grid points in the reference map, using a second order discretization scheme for the derivatives of the density function. Then the gradients at the grid points in the template map are found by trilinear interpolation, using gradients from the nearest 8 grid points in the reference map. The movement of the structure follows the average gradient direction in the reference map. These moves, since they're in the direction of the average density gradient, tend to increase the cross-correlation score, however the steps must be small so as to not overshoot the local maximum. When the local maximum is reached, the average density gradient becomes 0, and hence movement ceases.

Two types of moves are used during the refinement process: translation and rotation steps. These steps are alternated until movement of the template due to both types of steps becomes insignificant. In the applications of this method, convergence is typically obtained quickly, typically after less than 100 steps.

3.2.2.1 Translation step

In the translation step, the gradients taken from the reference map at the positions of the voxels in the template map are averaged together. This produces a displacement vector, $\vec{D}$, which an be expressed as:

$$\vec{D} = \frac{1}{N}\sum_{i=1}^{N}\vec{g}_i, \text{ (3.2)}$$

where $\vec{g}_i$ is the gradient vector computed at the position of a voxel *i*, for *i=1..N*, in the template density map. This displacement vector is scaled so that its length is not greater than the distance equal to the spacing between two grid points in the reference map. The new position for the structure, after the displacement, is $\vec{T} + \vec{D}$.

3.2.2.2 Rotation step

In the rotation step, the gradients taken from the reference map, at the positions of the voxels in the template map, are used to compute a torque on the structure being registered. The torque is computed with respect to the center of rotation for the structure, which is simply its center of mass. The total mass is:

$$m = \sum_{i=1}^{N} d, \text{ (3.3)}$$

and the center of mass is:

$$\vec{c} = \frac{1}{m}\sum_{i=1}^{N} d \cdot \vec{r}_i. \text{ (3.4)}$$

In the above, $\vec{r}_i$ is the position of the i^{th} voxel in the template density map, and d is the density value at that voxel. The torque due to a single gradient $\vec{g}_i$ from the reference map, computed at a voxel position $\vec{r}_i$ in the template density map, is:

$$\vec{t}_i = (\vec{r}_i - \vec{c}) \times \vec{g}_i \quad (3.5)$$

The average torque used to rotate the template density map is computed by averaging the torque at every voxel being considered in the template density map:

$$\vec{\tau} = \frac{1}{m}\sum_{i=1}^{N} d\vec{t}_i \quad (3.6)$$

The average torque $\vec{\tau}$ is used to rotate the template density map about its center of mass. The rotation axis is the normalized direction of $\vec{\tau}$ itself, and the degree of rotation is proportional to the magnitude of $\vec{\tau}$. The degree of rotation is scaled so that the largest displacement for any voxel in the template density map does not exceed the distance between two grid points in the reference density map.

3.2.2.3 Limitations of local refinement

Local refinement of an initial registration only locally optimizes the cross-correlation score, converging to a local maximum rather than a global maximum. Thus to find a good registration, local refinement of a registration has to start with a good initial placement of the structure which is close to the global maximum.

Chapter 4. Contributions

4.1 Watershed segmentation of cryo-EM density maps

The immersion-inspired watershed method, presented in [29], was adapted for application to 3D density maps. The algorithm is illustrated in Figure 4.1. The algorithm is as follows:

- All density values in the density map are first sorted and then considered in descending order.
- For each density value, if the corresponding grid point is:
    - not adjacent (26-connected) to any voxels in any existing regions, it is assigned to a new region.
    - adjacent to voxels in a single region, it is assigned to that region.
    - adjacent to voxels in two or more regions, the regions are sorted in descending order of number of adjacent voxels, and the voxel is assigned to the first region in the list

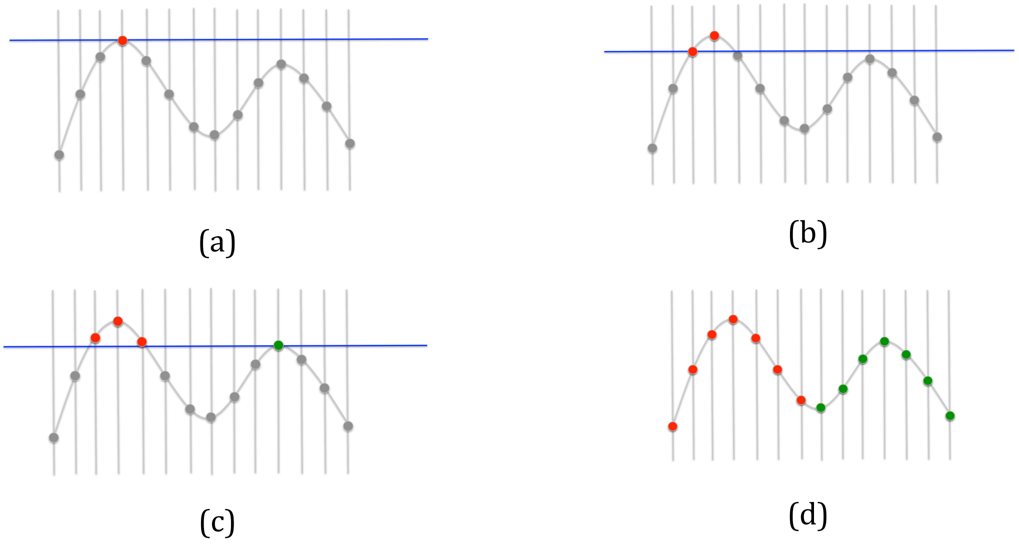

Figure 4.1. Illustration of the immersion-inspired watershed segmentation algorithm applied to a 1D map. Each grid point is drawn at an elevation which is proportional to its density value. Initially, all values are sorted from highest to lowest, and then considered one at a time in decreasing order. For a point being considered, if it is adjacent to an existing point that is already labeled, it takes the same label, otherwise it is assigned a new label. (a) to (c) illustrate this process. (d) illustrates the labels given to each point, red and green, which specify two segmented 'regions'.

Each resulting region thus contains a number of adjacent voxels. The boundaries between regions are the points with the lowest densities between local maxima. The watershed segmentation of a 2D map, simulated from a slice of atoms taken from a complex of two proteins, is illustrated in Figure 4.2. The pictures show that for such a density map, many more regions than proteins result. However, the region boundaries closely follow protein boundaries, and so if it were possible to join regions corresponding to each protein, a segmentation that contained only two regions and accurately captured the boundaries between the proteins could be found. Such grouping methods, as previously discussed, are typically not accurate. A multi-scale approach will be described in the next section, which attempts to achieve such a grouping.

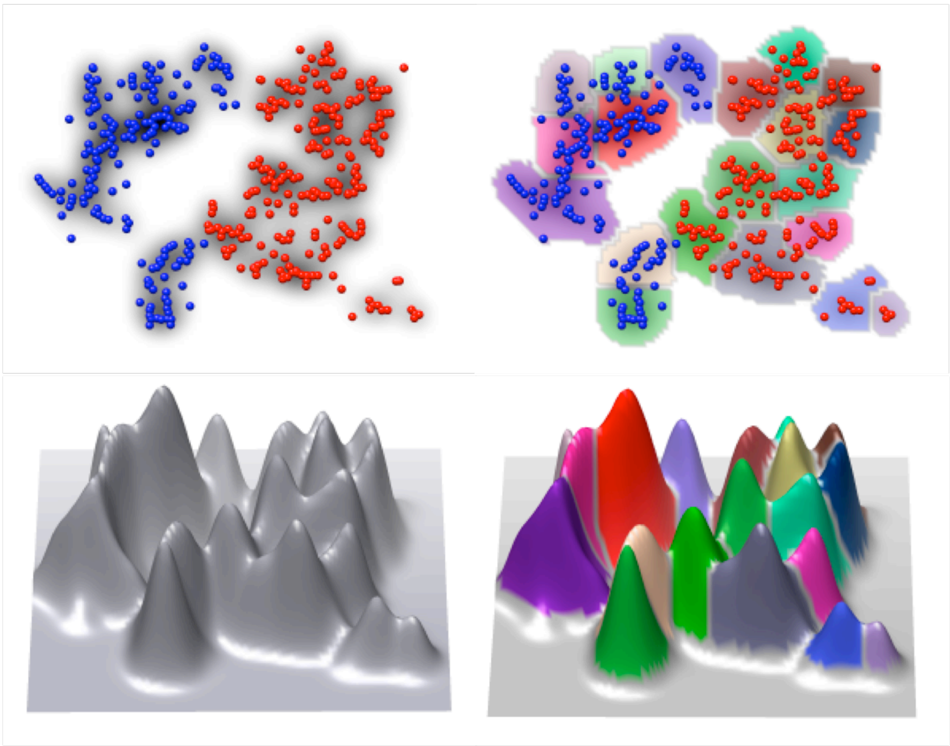

Figure 4.2. Topology and watershed segmentation of a 2D density map, simulated from a slice of atoms through two proteins. The atoms are drawn as spheres, colored blue if they are from one protein and red if from the other. In the top-left image, the density function and atoms are shown. Darker pixels represent denser regions, which coincide with dense clusters of atoms. In the top-right image, the regions resulting from watershed segmentation of this density function are shown, each region having a random color. More regions than proteins result, however the region boundaries coincide well with protein boundaries. In the bottom two images, the topological representation of the density function (left) and the watershed segmentation (right) are shown.

The watershed segmentation method can be applied to either density values or to gradient magnitudes. For computer vision, the later is typically done, since object contours tend to have higher gradient magnitudes. For cryo-EM density maps, Figure 4.3 illustrates that applying it to density values is a better choice, since it produces fewer regions, and the region boundaries tend to fall closer to molecular boundaries.

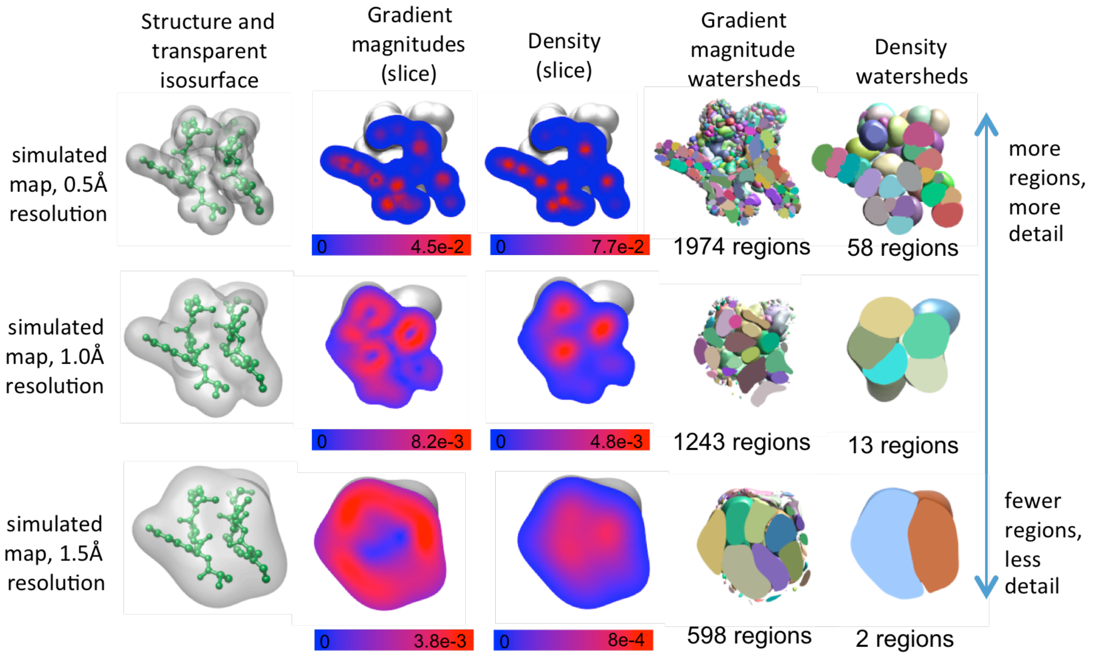
Structure and transparent isosurface
Gradient magnitudes (slice)
Density (slice)
Gradient magnitude watersheds
Density watersheds
simulated map, 0.5Å resolution
0
4.5e-2
0
7.7e-2
1974 regions
58 regions
more regions, more detail
simulated map, 1.0Å resolution
0
8.2e-3
0
4.8e-3
1243 regions
13 regions
simulated map, 1.5Å resolution
0
3.8e-3
0
8e-4
598 regions
2 regions
fewer regions, less detail

Figure 4.3. Watershed segmentation of a small density map, using density values and gradient magnitudes. A small complex of 2 molecules is shown in the left-most images, along with iso-surfaces from simulated density maps at three different resolutions. Slices through gradient magnitudes, density values, and segmentation regions are shown across the horizontal axis. The segmentation regions are shown as smoothed surfaces encapsulating each region. As indicated with text above each column, the slices are colored by density gradient magnitude, density value, and a random color for each region.

4.2 Multi-scale segmentation and sharpening

The topological watershed segmentation of a density map, as previously shown, produces numerous small regions, since many local density maxima are present. Here, a new, multi-scale approach is presented, which groups these numerous, small regions, into fewer, larger regions that correspond to single proteins or subunits.

The multi-scale approach uses the watershed segmentations obtained from the map at progressively smoothed levels. In more-smoothed maps, which correspond to coarser scales, the finer features in a density map are blurred out, and larger components such as proteins appear as single regions. The boundaries between regions in smoother maps tend to follow protein or subunit boundaries, which are normally less dense. However, the regions in smoothed maps lose the finer detail of the original non-smoothed density map. The sharpening process reintroduces this detail by joining regions from less smooth maps using a simple overlap test. The sharpening achieves the grouping of regions in less-smoothed maps based on which regions from more smoothed maps they overlap the most. The process is illustrated in Figure 4.4.

To describe the multi-scale segmentation process more precisely, let M_i represent a density map in scale space, with $i=0..n$. M_0 is the initial non-smoothed density map, and M_n is the most smoothed map after n smoothing steps. The map M_i, with $i=1..n$, is obtained by smoothing M_{i-1} with a Gaussian filter of a user-specified standard deviation, or step size. Each map M_i is segmented using the watershed algorithm, to produce a set of regions R_i, $i=0..n$.

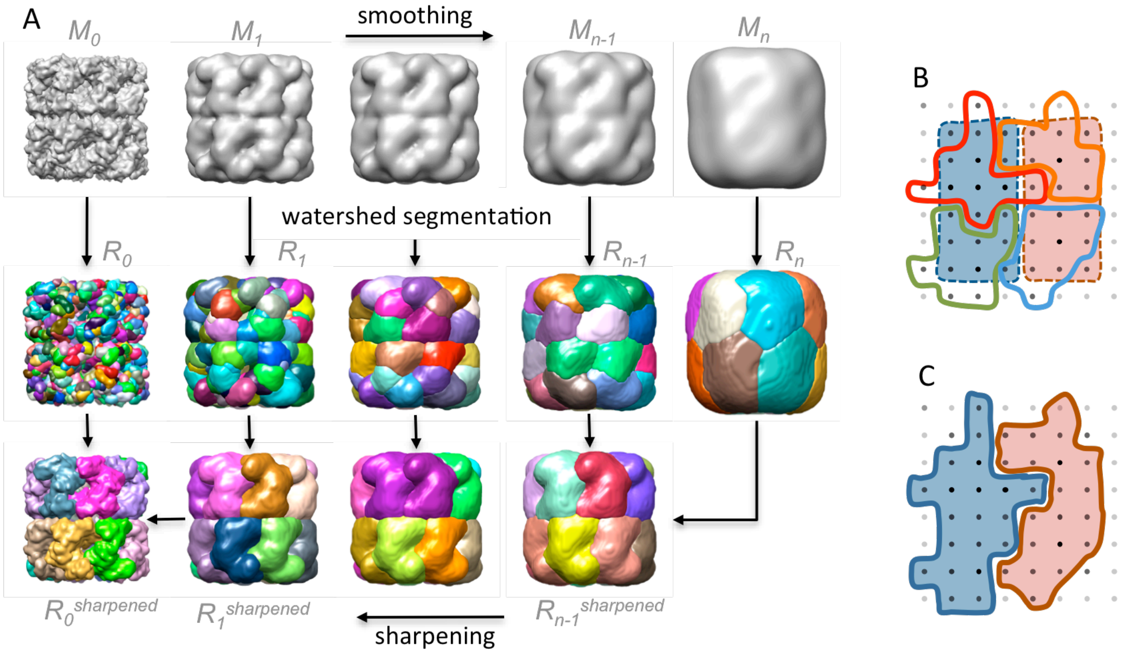

Figure 4.4. Illustration of the multi-scale segmentation and sharpening methods for a cryo-EM density map of GroEL. (A) The top row shows the original density map (left), which is progressively smoothed (left to right). The middle row shows the watershed segmentation of each of the maps; each region is shown using a surface enclosing the voxels it contains. The bottom row shows the regions from the most smoothed map successively sharpened using regions from less smoothed maps. (B) Sharpening of 2 regions from a smoothed map (dashed contours) by grouping regions from a less-smoothed map (solid contours). The latter are grouped based on which of the 2 regions from the more-smoothed map they overlap the most. (C) The 2 sharpened regions have more detail than the corresponding regions from the smoothed map.

In the first step of the sharpening process, the overlap between every region in R_{n-1} and every region in R_n is computed. Regions are defined using voxels from the same grid, and thus the overlap between two regions is simply the number of voxels that the two regions have in common. For every region in R_{n-1}, the region it overlaps the most in R_n is recorded. Regions in R_{n-1} that overlap the same region in R_n the most are joined. The resulting region is assigned the voxels from every region being joined. The set of resulting regions becomes $R_{n-1}^{sharpened}$. In the next step, regions in R_{n-2} are grouped based on which regions in $R_{n-1}^{sharpened}$ they overlap the most, producing $R_{n-2}^{sharpened}$. This process is repeated, grouping regions in R_i based on their overlap with regions in $R_{i+1}^{sharpened}$, to produce $R_i^{sharpened}$, with $i=n-3,n-4,\ldots,0$.

The regions $R_0^{sharpened}$ are the final result. It should be noted that while regions in $R_0^{sharpened}$ directly correspond to regions in R_n, not every region in R_n produces a region in $R_0^{sharpened}$. This happens when for some region in R_n, no region in R_{n-1} overlaps it more then it overlaps any other region in R_n.

Figure 4.5 illustrates this method on a density map generated from a small two-molecule complex. As shown in Figure 4.6, the multi-scale process is able to group the regions obtained in the non-smoothed map into two regions that very closely match the map of each individual protein.

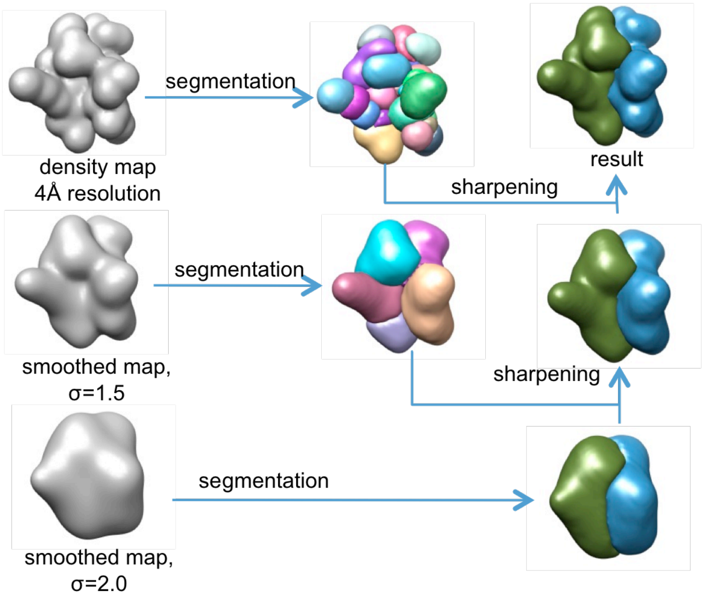

Figure 4.5 Multi-scale segmentation method applied to a simulated density map generated from two small molecules.

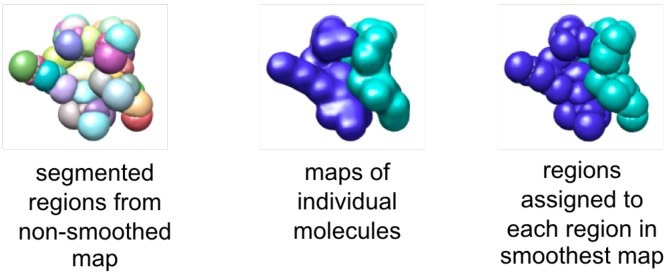

Figure 4.6 Results of multi-scale method on regions from the simulated density map of a small two-molecule complex. The regions grouped by the multi-scale method (right) closely resemble the density maps of each molecule individually.

4.2.1 Region hierarchies

During the sharpening process, a hierarchical grouping of regions is created. This hierarchical grouping is ideal for allowing the user to modify the segmented regions by subdividing them into smaller regions. The creation of a hierarchy during the sharpening process is illustrated in Figure 4.7.

A hierarchy is defined as an arrangement of items, or nodes, in which a node can have a number of descendant or children nodes, and a single parent node. All nodes that have no parents are the root nodes, and all nodes that have no children are leaf nodes. A hierarchy has multiple levels, and nodes are either at the same level, or “above” or “below” other nodes. Root nodes are at the top level, while leaf nodes are at the bottom level.

A hierarchy of regions as created during the sharpening process meets the following conditions:

1. Every sharpened region in $R_0^{sharpened}$ corresponds to a root node in the hierarchy.
2. Every region in R_0 corresponds to a leaf node.
3. Regions in R_i, for $i=1..n-1$ correspond to nodes at level i of the hierarchy.
4. Every region in R_i, for $i=0..n-1$, has exactly one parent, and the parent can only be a node that corresponds to a region from R_{i+1}.

These conditions aim to make the user-editing process simple and intuitive, and to maintain consistency when dividing a region into smaller regions. More specifically:

1. The regions first presented to the user are the sharpened regions. Hence these regions are at the top of the hierarchy.

2. The regions in the unsmoothed-map, which correspond to leaf nodes, are the ideal ones to present to the user, since they contain the most detail. These regions, being from the non-smoothed map, represent the highest level of detail.
3. The ungrouping of a single region at any level should result in a small number of sub-regions. By grouping regions in a hierarchy with multiple levels, the ungrouping of a region makes use of descendants only one level down into the hierarchy, which are fewer than descendents at levels further down.
4. Since every region has only one parent, this means that the ungrouping is unambiguous. If this condition is not met, then the same region could be seen in the subdivisions of different ancestor regions.

At the first sharpening step, the root nodes are added to the hierarchy. These root nodes correspond to every region in R_n. At subsequent sharpening steps, another level is added to the hierarchy as follows:

- In the second sharpening step, the regions in R_{n-1} which overlap the same region *r* in R_n more than they overlap any other region in R_n, are assigned to nodes whose parent is the root node corresponding to region *r*.

- At subsequent steps, the following process is repeated, which is described here for the third sharpening step (so the indexes can be compared to those shown in Figure 4.7). All subsequent steps are the same, with the indexes decreased by one. The process for the third sharpening step is:
  - The regions in R_{n-2} that overlap the same region in $R_{n-1}^{sharpened}$ the most are grouped together.
  - The regions in R_{n-2} that overlapped a region *r* in $R_{n-1}^{sharpened}$ the most, which are part of a group g_{n-2}, are considered one at a time, computing their overlap with every region in the group of regions from R_{n-1}, g_{n-1}, which also overlapped the region *r* from R_n the most.

- The regions in group g_{n-2} which overlap the same region from g_{n-1} the most are assigned to nodes whose parent is the node corresponding to the region in g_{n-1} which they overlapped the most.

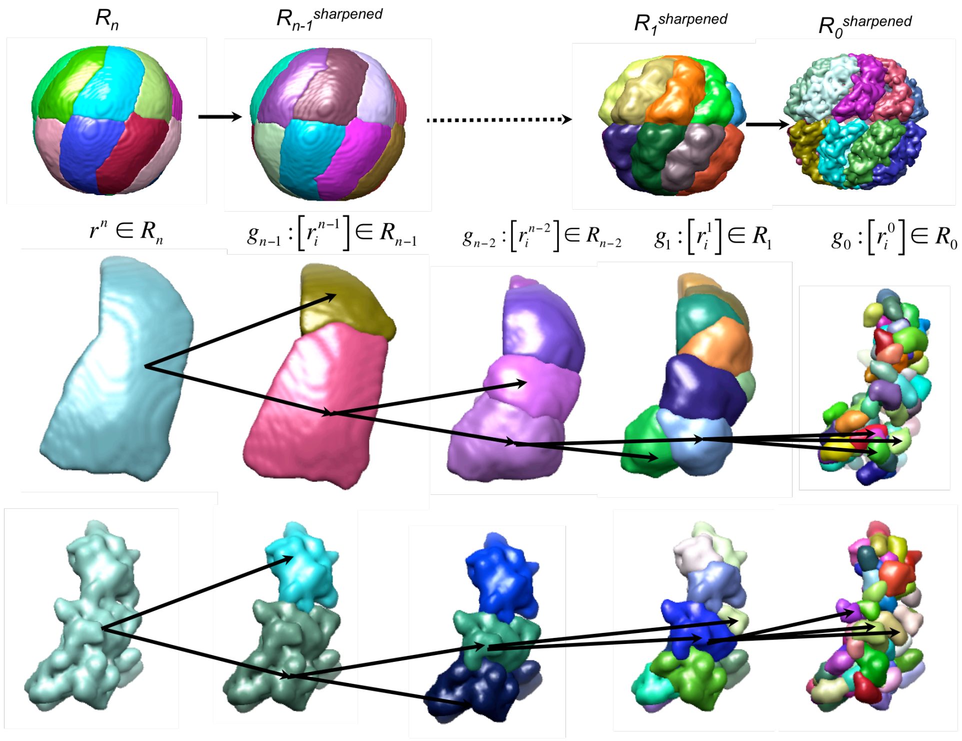
R_n
$R_{n-1}^{sharpened}$
$R_1^{sharpened}$
$R_0^{sharpened}$
$r^n \in R_n$
$g_{n-1} : \left[r_i^{n-1} \right] \in R_{n-1}$
$g_{n-2} : \left[r_i^{n-2} \right] \in R_{n-2}$
$g_1 : \left[r_i^1 \right] \in R_1$
$g_0 : \left[r_i^0 \right] \in R_0$

Figure 4.7. Hierarchical grouping of regions in the Mm-cpn cryo-EM density map. The top row shows the complete set of sharpened regions at each step during the sharpening process. The middle row shows a subset of regions [r_i] from each complete set of sharpened regions. Each region corresponds to a node in the hierarchy, and the parent-child relationships for one region, *r* from R_n, and its descendants are illustrated with arrows. The bottom row shows the sharpened versions for the same region *r* and its descendants. For any region at any level in the hierarchy, the sharpened version is constructed by combining all regions from R_0 that are descendants of that node.

4.2.2 Segmentation procedure and parameters

The user first selects a desired threshold for the map to be segmented, and all voxels with density value above this threshold are segmented using the watershed method. This threshold affects the resulting regions much like it affects the iso-surface visualization of the density map:

- At higher thresholds, the inner parts of a component are segmented, since they are denser. In particular, in high resolutions maps, at high threshold values, the backbone and secondary structures are seen.
- At lower thresholds, the outer surface of each protein is segmented, and thus, regions tend to be larger.

The user then chooses a smoothing step size, which specifies the standard deviation of the Gaussian kernel used to smooth the map. This step size determines how much smoothing is performed at each step. For example, a step size of ~2A produces a small decrease in the number of regions and a small change in the boundaries between regions. Larger step sizes cause a larger drop in the number of regions at each step, and also more drastic changes to the boundaries. Thus, smaller step sizes are preferred, since they produce more gradual changes. Larger step sizes may however be used if the density map is particularly large and memory is an issue.

The multi-scale segmentation procedure can be performed either interactively, as directed by the user, or automatically, based on a number of regions to be segmented.

4.2.2.1 Interactive multi-scale segmentation

In the interactive approach, the user triggers every smoothing step. At each step, the smoothest map so far is further smoothed using the specified step size. The resulting map is then segmented using the watershed algorithm (the segmentation threshold is automatically chosen so that the resulting regions cover every region in the map from the previous step). The user then inspects the resulting regions. In the ideal case, the process is repeated until the obtained regions correspond to individual proteins or subunits. If instead a point is reached where single regions span more than one protein or subunit, the user can backtrack to a previous point in the process, where small groups of regions appear to correspond to single proteins or subunits. These regions are then sharpened, which does not require any further parameters.

4.2.2.2 Automatic multi-scale segmentation

In the automatic approach, it is assumed that after a number of smoothing steps, every segmented region will correspond to a single protein or subunit. Thus the user can simply enter the number of proteins or subunits expected. This can be based on prior knowledge about how many proteins or subunits are expected in the density map (e.g. from a crystal structure or from biochemical experiments). The input density map is then repeatedly smoothed and segmented, until the number of segmented regions matches this number.

4.2.3 Dependence of segmentation time on map size

Typically, simulated or cryo-EM density maps can range in size from approximately 80x80x80 voxels to 500x500x500 voxels, using a grid spacings of 1Å-4Å. There is no upper limit on the map size that can be segmented with this method. The running time for watershed algorithm is O($n$log$n$), where n is the total number of voxels to be segmented, and thus the method scales favorably with map size. The smoothing operation is performed in Fourier space after transformation of the map and

Gaussian kernel using FFT, and thus its running time also scales well with *n*. However *n* itself scales poorly with map dimension *d*, by O(d^3).

4.3 Segmentation accuracy

Simulated maps are used to measure the accuracy of the multi-scale segmentation method. To do this, the segmented regions produced by the method are compared to *protein-masked* or *subunit-masked* regions. Protein/subunit-masked regions are generated by masking the density map with structures of the individual proteins or subunits in the structure used to simulate the map. Hence, these regions give the ideal segmentation. The process for measuring the segmentation accuracy, including simulation and segmentation of a density map, and generation of protein-masked regions is diagrammed in Figure 4.8.

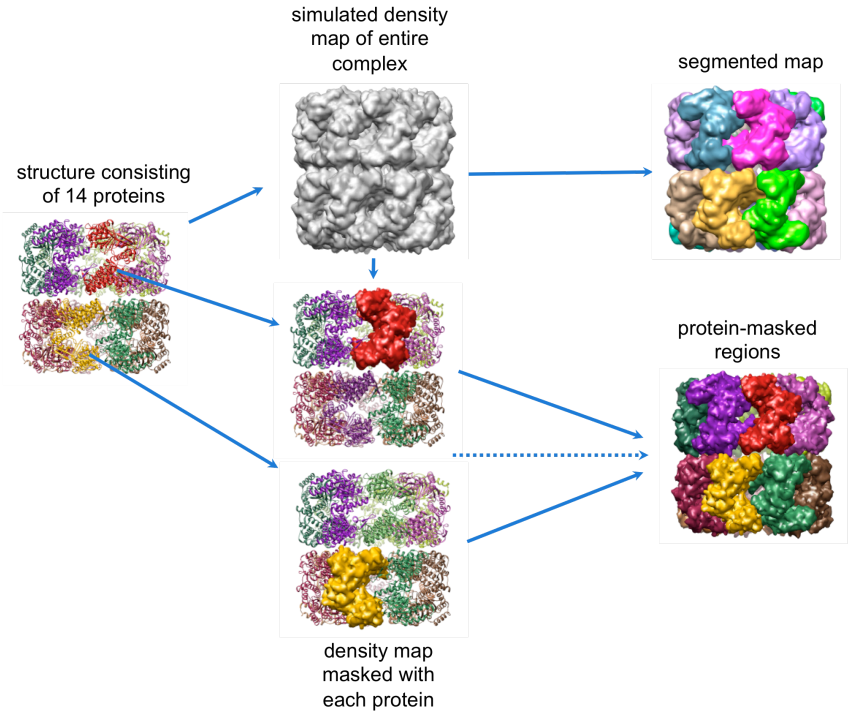
simulated density map of entire complex
segmented map
structure consisting of 14 proteins
protein-masked regions
density map masked with each protein

Figure 4.8. Illustration of how segmented and protein-masked regions are generated from a simulated density map of GroEL (PDB:1xck). A comparison of the segmented and ground-truth regions can then be done to validate and test how accurate the method is.

4.3.1 Protein/subunit-masked regions

To generate a protein-masked region or subunit-masked region, all voxels in the density map that are closer than 2.0Å to any atom in the protein or subunit keep their density values, and all others are given density values of 0. The value of 2.0Å is chosen because it is close to what the radius of an atom is when drawing a molecular surface for a structure. The voxels in the density map with density value lower than the threshold used to segment the map are also given density values of 0, so as to eliminate voxels that weren't included in the segmentation of the density map. The remaining voxels with non-zero density value are taken to belong to the protein-masked or subunit-masked region.

4.3.2 The shape-match score

Segmented regions are compared to protein/subunit-masked regions using a shape-match score. This score is defined as follows:

$$sm = \frac{volume(R \cap G)}{volume(R \cup G)} \qquad (4.1)$$

In the above equation, $volume(R \cap G)$ is the volume of the intersection of regions R and G, and $volume(R \cup G)$ is the volume of the union of the two regions. The shape-match score will be 0 if the two regions do not match at all (the intersection will have 0 volume), and it will be 1 if they match exactly (the volumes of the intersection and the union will be the same). Both regions being compared are defined by voxels on the same grid, so the intersection and union operations are performed directly on these sets of voxels.

Figure 4.9 illustrates this metric for 2D shapes. On the right, assuming the red region is the segmented region, and the blue region is the protein-masked region, ‘wrong’ segments are present in the segmented region but not the protein-masked region or are not present in the segmented region but are present in the protein-masked region. The ‘right’ segments are in both the segmented and protein-masked regions. The shape match score captures the proportion of ‘right’ segments to ‘right’ + ‘wrong’ segments. In the most accurate segmentation, the volume of ‘wrong’ segments would be zero, and hence the score would be a maximum of 1.

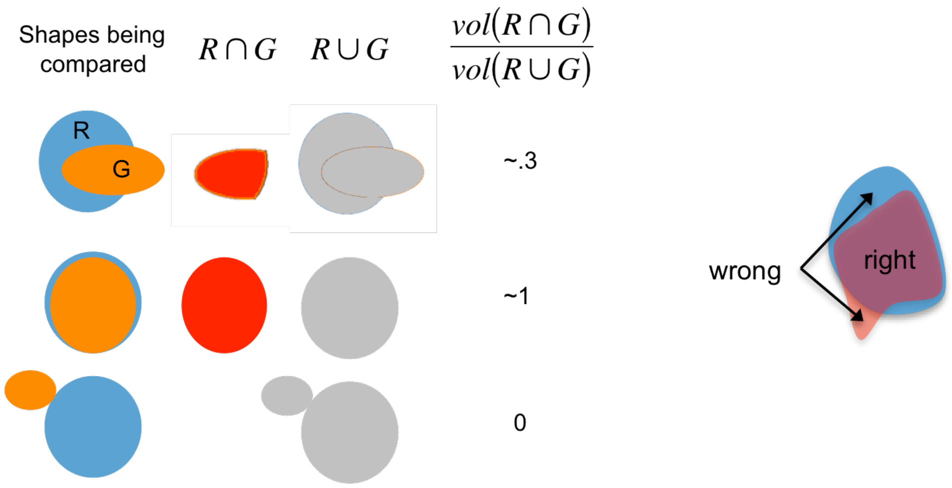

Figure 4.9: Illustration of the shape-match score in 2D. The score is illustrated for 3 examples (left). ‘Right’ and ‘wrong’ regions are illustrated on the right. The shape match score captures the ratio of the ‘right’ to ‘right+wrong’.

4.3.3 Maximum segmentation accuracy by grouping watershed regions

We also measure what is the best that the multi-scale grouping could do, i.e. what is most accurate segmentation attainable by grouping watershed regions in R_0, obtained from the non-smoothed map M_0. To obtain this *maximum watershed segmentation accuracy*, the regions in R_0 are joined based on which protein-masked or subunit-masked region they overlap the most. This process is illustrated in Figure 4.10.

The resulting regions are compared to the protein-masked or subunit-masked regions using the shape-match score. This score will tell us how accurate a segmentation could be obtained using the watershed segmentation method followed by perfect grouping of the resulting regions, and more importantly, how the multi-scale and sharpening methods perform in comparison.

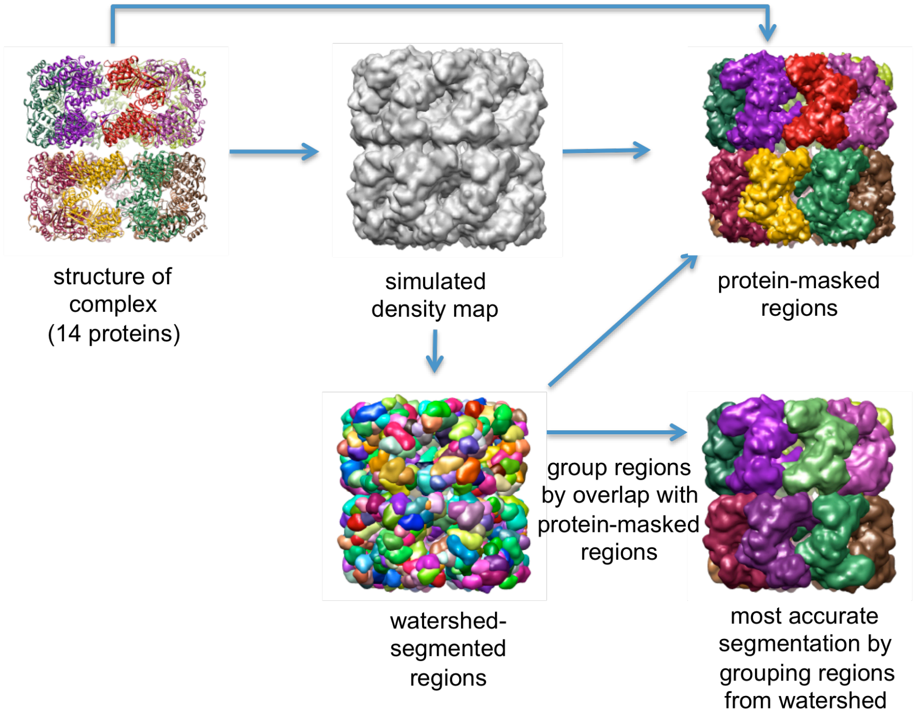
structure of complex (14 proteins)
simulated density map
protein-masked regions
watershed-segmented regions
group regions by overlap with protein-masked regions
most accurate segmentation by grouping regions from watershed

Figure 4.10. Optimal grouping of regions generated by the watershed method in a simulated density map. The regions generated by the watershed method are grouped based on which protein-masked region they overlap the most.

4.4 Registration of structures by alignment with regions

The results of the segmentation method are single regions, or small groups of regions, which correspond to each individual molecular component such as a protein or subunit. These regions are used to generate registrations of structures of individual components with the density maps. The registrations are created by aligning the structures with a single region or a small group of regions. The resulting alignments are then locally refined using the gradient-based method that optimizes the cross-correlation score.

Two methods for creating alignments of structures to regions are presented. The first is based on the principal-axes transform, which aligns the centers of mass and principal axes of the structure and regions. This works well for structures that are not spherical, cubical, or rod-like. When the principal axes transform does not produce a good registration, this can be detected by visual inspection, and it is also reflected by a low cross-correlation score. For such cases, a second method has been implemented, which is based on the alignment of centers of mass, followed by exhaustive search through the 3 degrees of freedom in rotational space.

4.4.1 Principal-axes transform

The principal axes of a structure are computed directly from its atomic positions. The principal axes of the region (or regions) that the structure is being aligned with are also computed in the same way, but using the positions of all the constituent voxels. The principal axes are coarse shape descriptors and are not affected greatly by noise or small differences in the shapes.

The principal axes are obtained from the second-moment tensor of a structure or region. The second-moment tensor is computed with respect to the center of mass of the structure or region (Eqn. 3.3). It is a 3x3 matrix defined as:

$$M_{jk} = \frac{\sum_{i=1}^{N} m_i \left(\vec{r}_i^{\,j} - \vec{c}^{\,j}\right)\left(\vec{r}_i^{\,k} - \vec{c}^{\,k}\right)}{m}, \text{ (4.2)}$$

where $j,k=0,1,2$, $\vec{r}_i$ is the position of the i^{th} atom or voxel in the structure/region, and $\vec{c}$ is the center of the structure/region. The indexes 0,1,2 for *j* and *k* refer to the *x,y* and *z* components of the vectors $\vec{r}_i$ and $\vec{c}$. The total mass of the structure or region, m, is the sum of the masses of each atom or voxel in the region, m_i. The mass of a voxel is the density value at the corresponding grid point, multiplied by the volume of the voxel.

The tensor matrix M is symmetric and can be diagonalized using the Jacobi transformation [69]. The resulting three eigenvectors are the directions of the principal axes, and the corresponding eigenvalues represent the relative lengths of these axes. The eigenvectors and the corresponding principal axes are sorted in order of decreasing eigenvalues.

The principal axes transform is illustrated for 2-dimensional shapes in Figure 4.11. The signs of the principal axes are ambiguous, so 2 possible alignments are possible. In the first alignment, the principal axes of the structure are pointing in the same direction as those of the region. In the second alignment, the two axes of the structure are flipped. The transform resulting from flipping a single axis would involve a reflection, which would not be a valid registration.

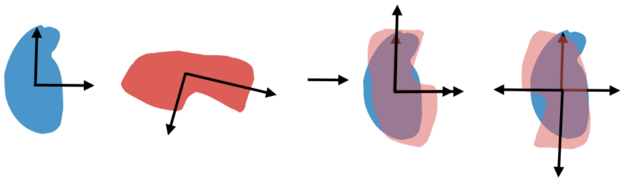

Figure 4.11. Illustration of the principal axes transform for 2D shapes. The transform aligns centers of mass and principal axes. The signs of the principal axes are ambiguous, so two alignments are possible in this 2D scenario.

In the 3D case, a total of 4 alignments are possible, as illustrated in Figure 4.12. After each alignment, the registration is first refined using the gradient-based method, and then the registration with the highest cross-correlation score is kept.

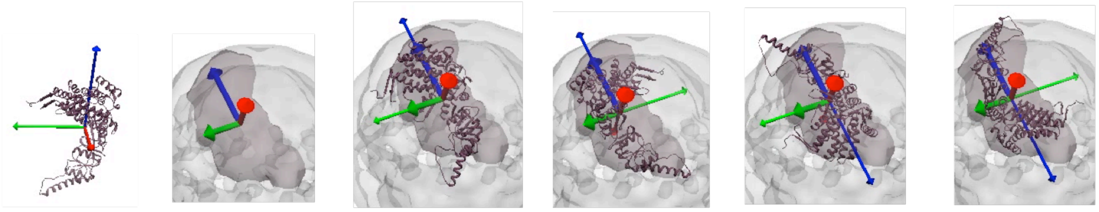

Figure 4.12. Illustration of the principal axes transform for a structure and segmented region from a 3D density map. The structure, region and their respective principal axes are shown in the two images to the left. The remaining 4 images show the 4 possible alignments in which centers of mass and principal axes are matched. In the first alignment, the principal axes are unmodified. In the other three alignments, two of the three axes have their signs flipped. Alignments where one or three axes are flipped at a time are not considered, since that would result in a reflection of the structure.

4.4.2 Alignment of centers and rotational search

When the principal-axis registration method doesn't produce a good registration, as indicated by a low cross-correlations score or by visual inspection, an alternate registration method is used. This method first aligns the centers of mass of the structure and region(s), and then evenly samples rotational space. Each resulting registration is first locally refined to optimize the cross-correlation score, and the registration with the highest cross-correlation score is kept.

To evenly sample rotational space, 3 degrees of freedom are required. The 3 degrees of freedom include 2 degrees of freedom specifying an axis of rotation, and one degree of freedom specifying the amount of rotation. The axes of rotation are obtained by evenly sampling points on a sphere, and taking the axis to be the direction from the origin to each point on the sphere. The amount of rotation is specified by a scalar which is varied between 0° and 360°.

Since only 3 degrees of freedom are required to specify all possible orientations, this process is also relatively fast when compared to exhaustive search, which must search through 6 degrees of freedom. For each rotation considered, local refinement is also performed using the gradient-based method to optimize the cross-correlation score.

4.4.3 Interactive specification of regions for alignment

The segmentation method produces single regions or small groups of regions corresponding to each protein or subunit. To register the structure of a single protein or subunit, the regions with which it is to be aligned so as to create the correct registration have to be determined. One way in which this can be accomplished is for the user to interactively select the region or small groups of regions to align the structure with. The alignment is then performed using the principal-axes transform first, since it is faster. If the resulting registration does not look right, or if the cross-correlation score is low, rotational search can be used to see if a better registration is found.

4.4.4 Automated alignment of structures to regions

The structure can also be aligned to groups of regions that are automatically generated from all segmented regions, as described below. Again, the principal-axes alignment method is used first, followed by rotational search if the resulting fits do not appear to be correct or produce low cross-correlation scores.

After aligning a structure to all groups, the resulting fits are sorted in order of decreasing cross-correlation score, and the first *N* fits are kept, where *N* is the number of times the structure is expected to appear in the density map. *N* can be determined by inspecting the cross-correlation scores, since cross-correlation scores of incorrect fits tend to be much lower than cross-correlation scores of correct fits. For the fits kept, the regions overlapping the fitted structure are joined, to create single regions corresponding to the fitted structure. This process is illustrated in Figure 4.13.

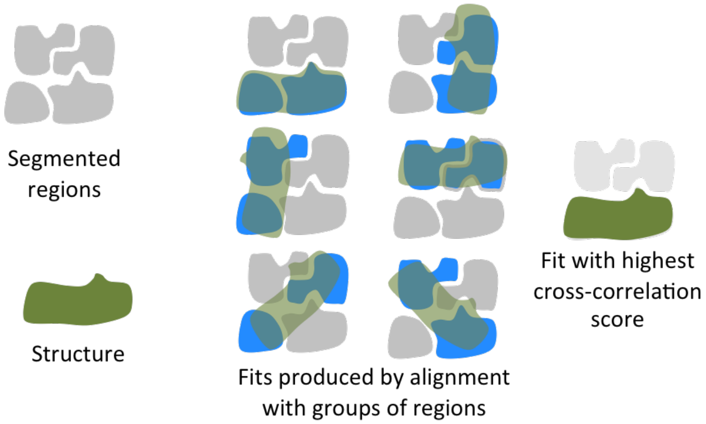
Segmented regions
Structure
Fits produced by alignment with groups of regions
Fit with highest cross-correlation score

Figure 4.13. Registration a structures by alignment with groups of segmented regions. The best registration is found after aligning the structure to automatically generated groups of regions.

4.4.4.1 Generation of groups of adjacent regions

The goal in this process is to consider all possible groups of adjacent regions, so that when the structure is aligned with one or more of these groups, the correct registration is found. An exhaustive enumeration of all possible combinations of regions could generate a very large number of groups. However, by requiring that the groups contain adjacent regions, since each region is adjacent to only a small number of other regions, the number of possible groups is drastically reduced.

To automatically generate groups of adjacent regions, a recursive algorithm was implemented, which uses a queue. A queue is a list of elements that are yet to be processed. The elements in the queue are groups of adjacent regions. The queue is initialized with the same number of groups as regions, with each group containing a different region. At each step, a group is removed from the front of the queue and is processed. The algorithm stops when the queue becomes empty.

In parallel, a set of groups is maintained, which is the resulting list of groups. This set of groups is initially empty, and groups are added to it during the recursive algorithm. The list of resulting groups is maintained such that every group in it is different from any other group in the list. Two groups are different if the set of regions they contain are not the same.

The processing of each group removed from the queue is as follows. If the group is the same as a group already added to the list of resulting groups, it is ignored. Otherwise, it is added to the list of resulting groups, and it is further considered as follows. If the volume of the group is smaller than the volume of any of the structures to be fit, then further regions are considered for addition to the group. First, all regions that are adjacent to at least one region in the group are listed. All possible combinations of these adjacent regions are added to the group to create new groups, which are all added to the queue.

4.4.4.2 Filtering of groups

When considering a structure for alignment to groups of regions, the groups are first filtered to remove groups that are too dissimilar from the structure, and thus which would not create correct registrations. Considering fewer groups for each structure reduces the number of alignments considered, and thus makes the automated process faster. The groups are filtered using two metrics: the ratio of volumes and ratio of bounding radii.

The bounding radius of a structure is the largest distance from its center to any of the atoms it contains. The bounding radius of a group of regions is the largest distance to any of the voxels in any of the regions, from the center of the voxels from every region in the group. The volume of a group of regions is the number of combined voxels from all the regions in the group, multiplied by the volume of each voxel. The volume of a structure is computed from its simulated density map: it is the number of voxels with density values above a threshold, multiplied by the volume of each voxel. The threshold can be adjusted by the user, to get a volume that is close to the volume of the group of regions that it correctly aligns with. We do this interactively, varying the threshold until the iso-surface of the simulated map looks similar to the segmented regions.

To compute the volume ratio, the difference between the volume of the structure and the volume of the combined regions in a group is computed. The absolute value of this difference is divided by the volume of the structure to get the ratio. If this ratio is greater than a cut-off value, for which we use 0.5, the group is ignored. The above process is the same for the bounding radius, with a cut-off of 0.1. These values were determined by starting with small values, and increasing them until we found that all correct registrations were found for the structures considered here. They can be set to different values by the user if necessary. Decreasing them speeds up the process but the correct registrations may not be found, while increasing them will make the process take more time but increases the chances that the correct registrations will be found.

4.5 The *Segger* software

A tool has been developed that allows a user to perform the multi-scale segmentation and registration procedures described above, which we call *Segger*. It has been developed as a plug-in to Chimera [70,71], which is an extensible platform based on Python and C++ for molecular visualization. The plug-in is written mostly in Python, making extensive use of functionality already implemented in Chimera. Computation-intensive functions such as the watershed segmentation and sharpening procedures were compiled in C++ for speed.

The reasons for developing the software as a plug-in to Chimera are that firstly, Chimera already implements tools for visualization and manipulation of 3D density maps, and hence development time was greatly reduced. The Segger plug-in simply adds further functionality rather than recreating the functionality already there. Secondly, many researchers already use Chimera, and hence it would be much easier for them to use Segger than if it was a stand-alone software.

Chapter 5. Results

5.1 Effects of parameters on segmentation accuracy

The effect of parameters used in the multi-scale segmentation process on segmentation accuracy was measured. The two parameters that were varied are the initial segmentation threshold and the step size. The total number of steps depends on how many segmented regions are desired in the segmentation, and thus it automatically detected for the simulated maps used.

5.1.1 Initial segmentation threshold

The initial segmentation threshold mostly affects the visualization of each segmented region. A simulated density map of GroEL, segmented at various thresholds is shown in Figure 5.1. As illustrated, at high threshold values, the denser inner regions of each protein are segmented, and at low thresholds, the segmented regions are larger and capture the outer surface of each protein. Despite the threshold used, the same number of regions as proteins are produced.

The plots in Figure 5.1 show that the segmentation accuracies are not greatly affected by the threshold value, although at higher thresholds they increase slightly. The protein-masked regions are thresholded using the same threshold used to segment the density map, and so they also contain only the higher density values which were segmented. The accuracies increase slightly at higher thresholds because the ‘wrong’ segments become slightly smaller in proportion to ‘right’ segments.

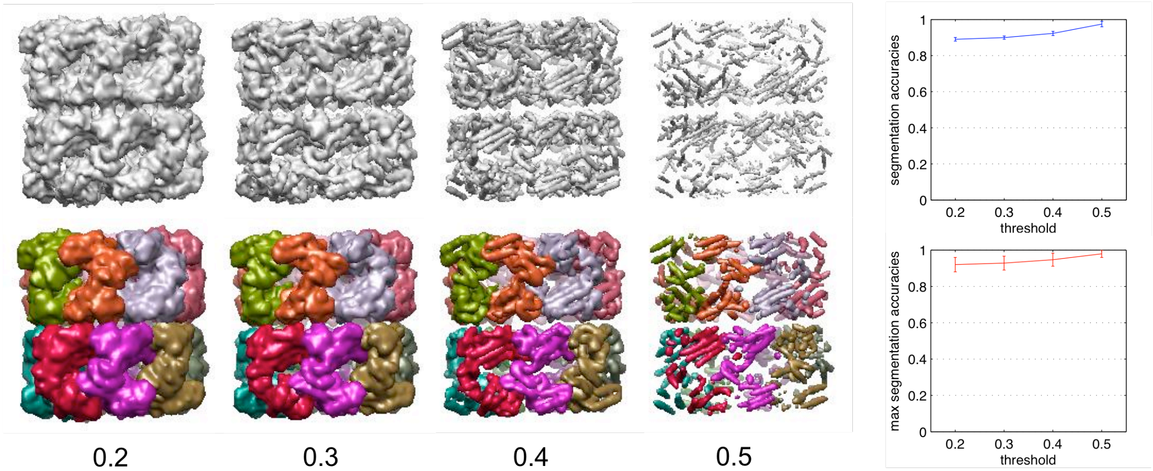

Figure 5.1. Effect of threshold on segmentation of a simulated map of GroEL. The simulated density maps (top) and resulting segmentations with the multi-scale procedure (bottom) are shown at various thresholds. The segmentation accuracies and maximum watershed segmentation accuracies are plotted using error bars indicating the lowest and highest shape-match scores between the segmented regions in each complex and the protein-masked regions.

5.1.2 Smoothing step size

The smoothing steps size determines how much smoothing occurs at each smoothing step. The multi-scale segmentation procedure was applied to 3 complexes, with step sizes varying between 2Å and 12Å. Regardless of the step size, the number of resulting regions is the same as the number of proteins in the complex. The segmentation accuracies were measured for the resulting regions. The plots of the segmentation accuracies at various step sizes, shown in Figure 5.2, show that the segmentation accuracies tend to be similar at different step size, however for the thermosome they are higher when smaller step sizes are used.

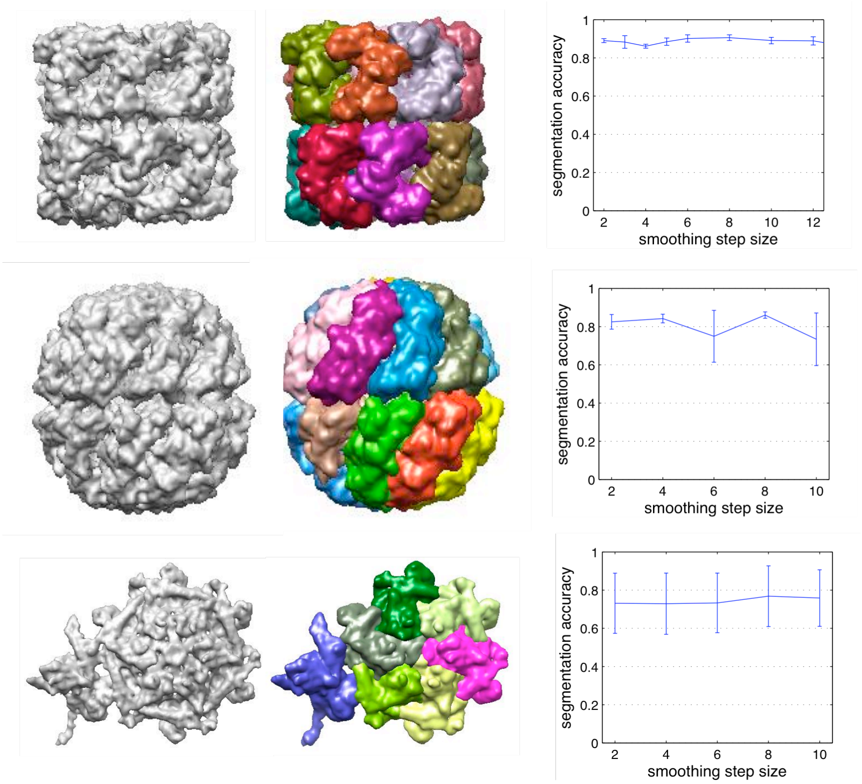

Figure 5.2. Effect of step size on segmentation accuracies of 3 simulated density maps. The maps are of GroEL (PDB:1xck, top), the thermosome (PDB:1aon, middle), and HK97 capsid (PDB:1ohg, bottom). For each complex, the density map (left) and the segmented regions for a step size of 2.0Å (middle) are shown. The step sizes were varied between 2Å and 12Å. The multi-scale procedure was applied to each density map, producing the same number of regions as proteins regardless of step size. The segmentation accuracies are plotted for each step size (right), as error bars indicating the lowest and highest segmentation accuracies amongst the regions in each complex. The plots show that the segmentation accuracies are somewhat similar regardless of step size, although for the thermosome complex smaller step sizes yield better accuracies.

5.2 Segmentation of 5 simulated density maps

Density maps of 5 molecular machines were simulated at 10Å resolution, using the Chimera *molmap* command, *sigmaFactor* 0.187, *grid spacing* 2.0Å. For the multi-scale segmentation procedure, only a segmentation threshold of 0.2, smoothing step size of 2.0Å, and target number of regions were specified. The target number of regions was set to the number of proteins or subunits: GroEL - 14 proteins, thermosome - 16 proteins, E-coli ribosome - 2 subunits, and HK97 - 7 proteins.

The numbers of smoothing steps taken were automatically determined based on the number of target regions. The number of steps taken for each density map was: GroEL - 48, thermsome - 40, ribosome - 115, HK97 pro-capsid - 11, HK97 mature capsid - 21. In all cases, the final number of regions matched the number of proteins or subunits. The process took only several minutes for each complex, and thus is extremely fast given that interactive segmentation can typically take many hours and require much input from the user. The results are shown in Figure 5.3.

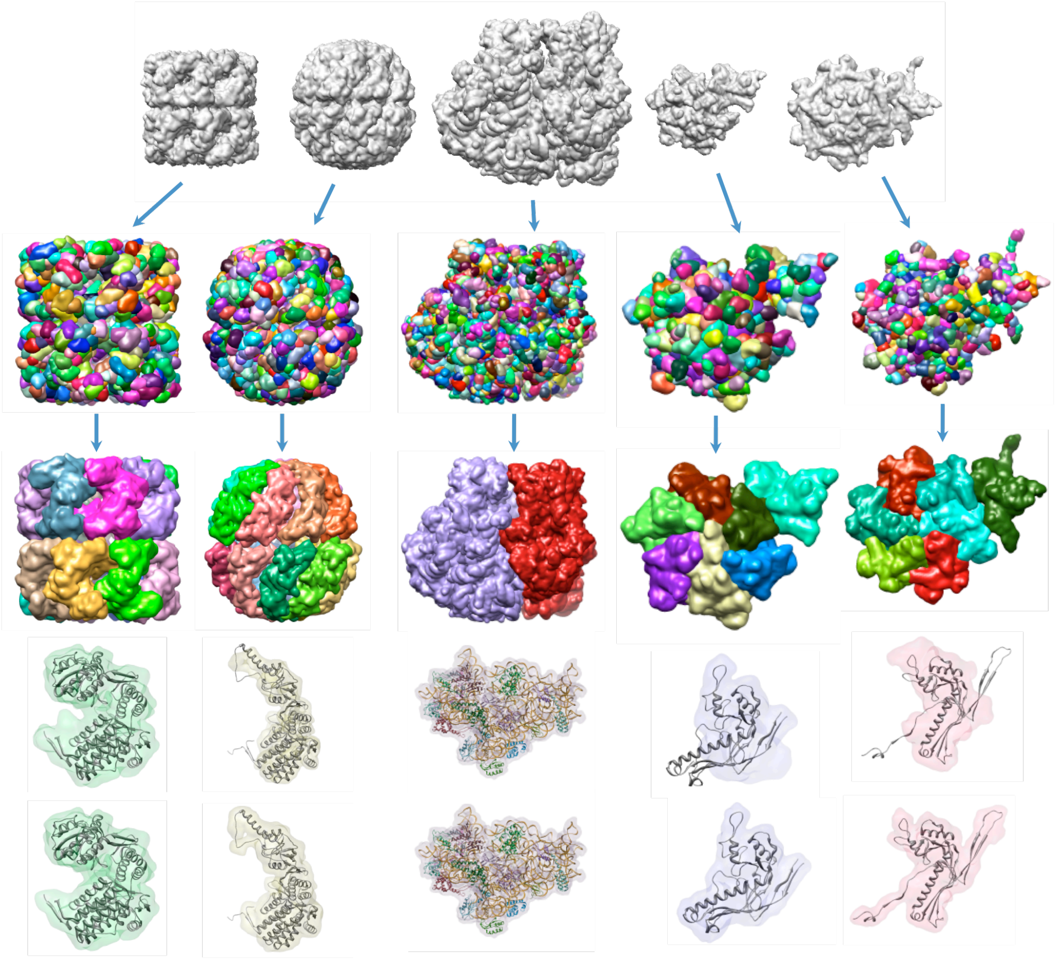

Figure 5.3. Segmentation results for simulated density maps of GroEL, thermosome, ribosome, HK97 procapsid and mature capsid asymmetric units. The first row shows the simulated density maps, all equally scaled. The second row shows the numerous regions resulting from watershed segmentation of these maps. The third row shows the regions produced by the multi-scale process. The fourth row shows single regions using transparent surfaces, along with the structure of the corresponding protein or subunit. The fifth row shows the same protein structure, along with a region that was generated based on which regions in the watershed segmentation (second row) overlap the protein-masked or subunit-masked region they overlap the most, thus yielding the maximum segmentation accuracy by grouping of watershed regions.

5.2.1 Segmentation accuracies

Segmentation accuracies for each component were measured by computing the shape-match scores between segmented regions produced by the multi-scale method and protein/subunit-masked regions. The scores are plotted in Figure 5.4. Good segmentation accuracies were obtained for GroEL (0.859-0.886), thermosome (0.812-0.880), and the ribosome large and small subunits (0.973, 0.983), but lower accuracies (0.501-0.886) for HK97. For the components segmented with high accuracy, the segmented regions closely match the corresponding structure of each protein or subunit (Figure 5).

Figure 5.4 also plots the maximum watershed segmentation accuracies. These are the best segmentation accuracies that can be obtained by joining regions obtained using the watershed method in the non-smoothed map, M_0. All the maximum watershed accuracies for each component are high, indicating that the watershed method could be used to produce very accurate segmentations. These accuracies however are not 1, because the protein-masked regions approximate the molecular surface of each protein, whereas the regions resulting from grouping watershed region are limited by the watershed method and the resolution of the simulated density map.

The segmentations accuracies produced by the multi-scale method are lower than the maximum watershed accuracies. Despite this, the results of this method are still close for the GroeL, thermosome, and ribosome complexes, however they are much lower for the HK97 asymmetric units. Even in the latter cases, the multi-scale method still produces a single region for each protein, a good result given that minimal user-interaction was required.

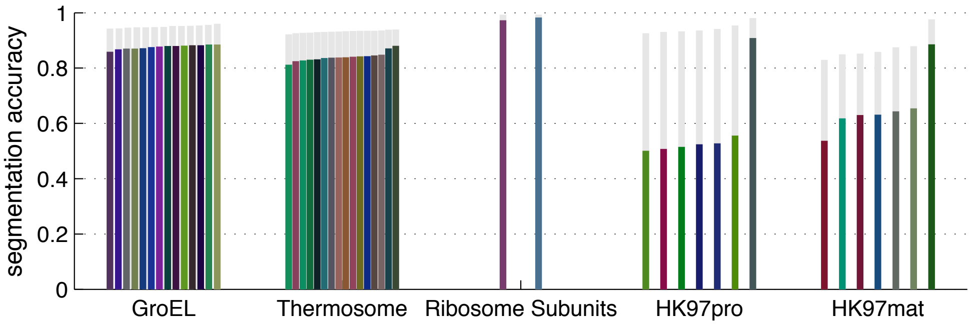

Figure 5.4. Segmentation accuracies for regions in the 5 simulated density maps shown in Figure 5.3. Each bar with a random color represents a single protein or subunit. The maximum watershed segmentation accuracies are also plotted for each component using light gray bars. The multi-scale and sharpening method do very well comparatively, except for proteins in the HK97 asymmetric units. Despite the lower accuracies for the multi-scale method in the latter, the same number of regions as proteins is produced.

5.2.2 Cause of low accuracies

Figure 5.3 shows that narrow segments in the proteins of HK97 were not captured correctly in the regions produced by the multi-scale method, and hence the segmentation accuracies for these components were low (0.5-0.6). This happens because the regions corresponding to these protruding segments are joined with regions corresponding to the nearby proteins they interact with. They appear as separate regions in less-smoothed maps, however the sharpening process does not join them with the correct regions corresponding to the protein they belong to, since they mostly overlap the smoother regions corresponding to the nearby proteins they interact with. We tried improving the sharpening process by taking into account local metrics such as density values between regions, however this doesn't work in general, most likely because the local metrics are easily influenced by noise and discretization error.

The segmentation accuracies plotted in Figure 5.4 also show that one of the proteins in HK97 has substantially higher segmentation accuracy than the other six. The units with lower accuracies are part of a 6-fold ring-like symmetric arrangement where each protein interacts with two others. The protein with the higher accuracy is part of a 5-fold symmetric arrangement that is formed with proteins from 4 other asymmetric units. The segmentation of this protein was more accurate because the two neighbors it has in adjacent asymmetric units are not present in the asymmetric unit.

5.3 Segmentation accuracy at various resolutions

An important question in the analysis of cryo-EM density maps is how accurately molecular components can be identified at different resolutions. This is an important question because high-resolution density maps cannot always be obtained, so the question has to do with how valuable maps with lower resolution might be. We try to answer this question using density maps simulated at a range of resolutions (6Å - 30Å, in steps of 2Å) for GRoEL, GroEL+GroES, Ribosome, HK97 procapsid and HK97 mature capsid. Each simulated map at every resolution was segmented using the multi-scale method, specifying only an initial threshold for each map, the number of proteins or subunits to be segmented, and smoothing step size of 2.0Å.

In Figure 5.5, the highest segmentation accuracy (blue lines) produced by the multi-scale method and highest maximum watershed segmentation accuracy (dashed red lines) in each density map is plotted vs. resolution. The plots show that both accuracies are higher for the high-resolution density maps, and decrease with resolution, but stay above 0.6 even at the lowest resolution of 30Å. At lower resolutions, the multi-scale segmentation method yields the same accuracy as the maximum accuracy possible with the watershed method, since the two lines coincide. This is because at lower resolutions, there are fewer watershed regions to join, and hence it becomes somewhat easier to join the correct regions.

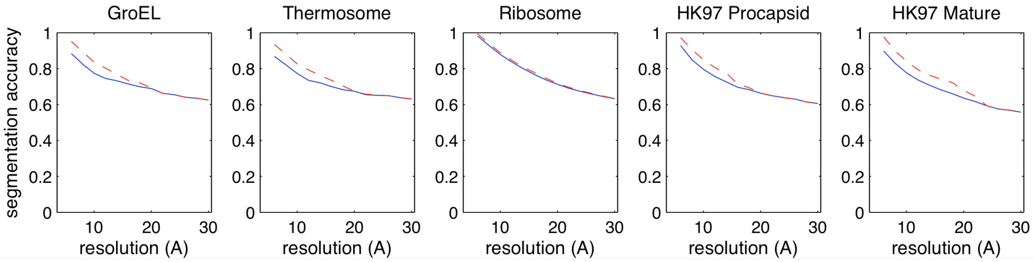

Figure 5.5. Segmentation accuracies for 5 density maps simulated at various resolutions (6Å-30Å, every 2Å). The highest segmentation accuracy (blue lines) and highest maximum watershed segmentation accuracy (dashed red lines) for a component in each density map is plotted vs. resolution. The plots show that segmentation accuracies drop as the resolution increases.

To illustrate the effect of resolution on segmented regions, a single protein from GroEL is shown in Figure 5.6. The regions shown are the protein-masked region and regions produced by the multi-scale method applied to simulated maps at different resolutions. At high resolution, the segmentation closely resembles the ground-truth region. At lower resolutions, the segmented region has a smoother surface compared to the protein-masked region. However, even at low resolutions, the segmented region still closely, if roughly, captures the shape of the protein.

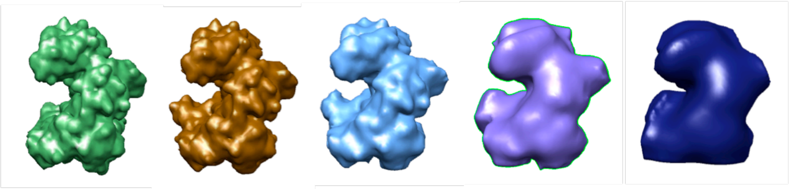

Figure 5.6. Protein-masked region and segmented regions corresponding to a single protein in simulated maps of GroEL at different resolutions. The protein-masked is the first from the left. The remaining segmented regions are from maps with resolutions of (left to right) 6Å, 10Å, 20Å, and 30Å.

5.4 User-edits of segmented regions

The user has no control over the regions generated from the multi-scale method other than the initial segmentation threshold, the number of smoothing steps and the step size. In some cases a user may desire to be able to modify and fine-tune the resulting regions. We use the hierarchical grouping of regions to make this process easy and intuitive. To do so, the user can perform two types of operations:

- Ungrouping: the user can select a region and split it up into smaller sub-regions. Ideally, only a small number of sub-regions result, so that the number of regions the user has to deal with does not become overwhelming.
- Group regions: once a larger region has been split into smaller regions, the user may decide to join one of the smaller regions with a different region.

The ungrouping and grouping processes are illustrated in Figure 5.7. This process relies on the user's knowledge about the structure of each protein or subunit being segmented, and it is can be used to try to improve the segmentation accuracy. For this example, which is performed on segmented regions from a simulated density map, comparison of the resulting regions with protein-masked regions shows that the segmentation accuracy can be increased during such a procedure (Figure 5.8).

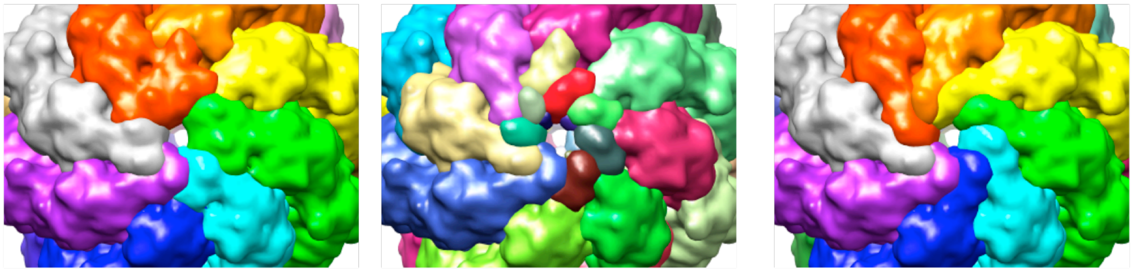

Figure 5.7. User editing of the simulated map of the thermosome by hierarchical ungrouping and regrouping of regions. Large segmented regions (left) are ungrouped, resulting in smaller regions (middle) around the area that needs to be modified. The small regions are then regrouped, reproducing the large starting regions but with small modifications in the center area (right).

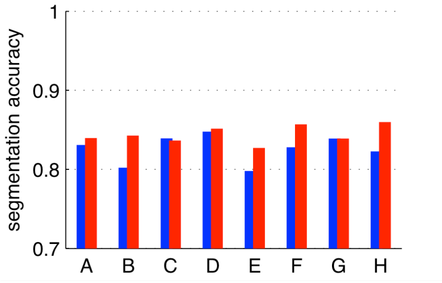

Figure 5.8. Segmentation accuracies of 8 segmented regions before and after user-edits. The blue bars are the accuracies for 8 of the segmented regions before the user-edits, and the red bars are the segmentation accuracies after the user-edits. Most of the accuracies are higher after the user-edits.

5.5 Segmentation of cryo-EM density maps

A total of five cryo-EM density maps were segmented using the multi-scale watershed method. The results are shown in Figure 5.9. The segmentation of each complex took only several minutes. For the GroEL, Mm-cpn, and ribosome density maps, the resulting segmentations contained the same number of regions as proteins or subunits being segmented. For the maps of GroEL+GroES and bacteriophage lambda, groups of at most two regions corresponded to single proteins. In the latter maps, segmentation of further smoothed maps produced regions spanning more than one protein, and so in these cases, a smoothing level where every region corresponded to a single protein could not be reached. However due to the small number of regions, it was very easy to interactively select out groups of regions belonging to individual proteins.

5.5.1 GroEL

GroEL is a barrel-like protein complex, with 7 proteins arranged in a symmetric fashion to form a ring with a cavity in the middle. Two rings are stacked on top of one another. This complex is also commonly referred to as a *chaperone*. Its function is to bind unfolded proteins in its large central cavity. With the help of the lid-like co-chaperone GroES, the unfolded protein is isolated from the environment and is helped to fold to its native state. The density map for GroEL at 4.2Å resolution [13] (EMDB:5001) was segmented using the watershed method, producing 2936 regions at a threshold of 0.597. The map was smoothed with 4 steps of size 7.5Å. The most smoothed map yielded 14 regions, each region corresponding to a single protein, which were then sharpened.

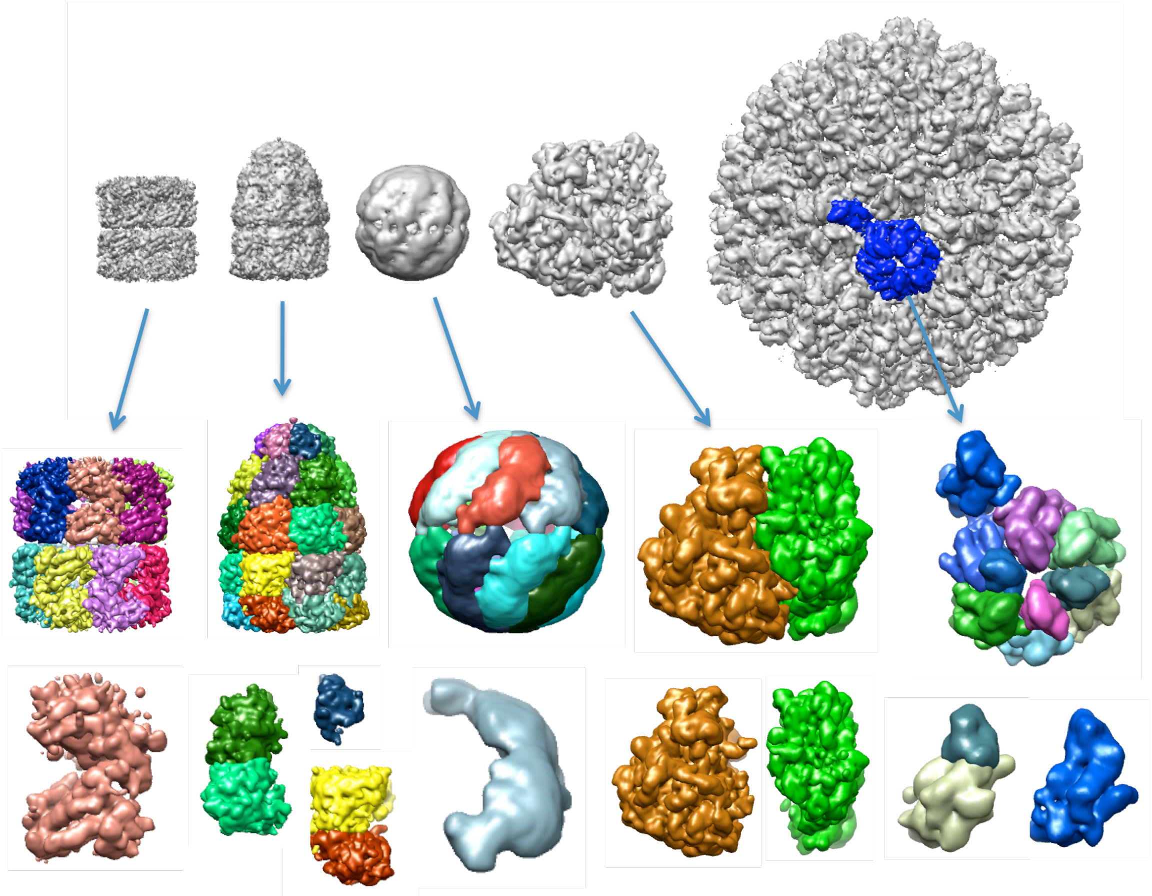

Figure 5.9. Five cryo-EM density maps segmented using the multi-scale watershed method. From left to right the maps are GroEL, GroEL+GroES, Mm-cpn, ribosome, and bacteriophage lambda. The top row shows the cryo-EM density maps (all scaled equally), middle row shows the resulting segmented regions, and the bottom row show single or groups of 2 regions corresponding to individual proteins or subunits.

5.5.2 GroEL+GroES

The GroEL+GroES complex consists of the barrel-like GroEL complex and a lid-like GroES complex attached to one end of the barrel. The GroES complex is made of 7 proteins, arranged in a symmetric ring-like fashion with no cavity in the middle. Its function is to close off one side of the barrel-like GroEL. The density map for GroEL+GroES at 7.7Å resolution [72] (EMDB:1180) was segmented into 2684 regions using the watershed method at a threshold of 0.608. The map was then smoothed with 3 steps of size 5Å. The smoothest map segmented into 42 regions, which were then sharpened. In the resulting segmentation, groups of two regions correspond to single proteins in the barrel-like GroEL complex, while single regions correspond to single proteins in the lid-like GroES complex.

5.5.3 Mm-cpn

Mm-cpn is also a barrel-like complex, consisting of two symmetric rings, each ring being made up of 8 proteins. This complex does not require a lid to close off the internal cavity; instead the top (apical) parts of the proteins bind to each other in an iris-like form in the closed state. (This iris-like arrangement can be seen in Figure 5.7). The density map at 10Å resolution [73] was segmented at a threshold of 1.25, producing 192 regions. The map was smoothed with 8 steps of size 5.0Å. The most smoothed map produced 16 regions, with each region corresponding to a single protein.

5.5.4 Ribosome

The ribosome is a large complex that consists of both proteins and RNA. It consists of two subunits, commonly reffered to as the *large* and *small* subunits. The cryo-EM density map of the E-coli ribosome at 9Å resolution [74] (EMDB:1056) was segmented at a threshold of 43.4, producing 897 regions. Smoothing was done with 32 steps of size 5.0Å. Segmentation of the most smoothed map produced only two regions corresponding to the large and small subunits, which were then sharpened.

5.5.5 Bacteriophage lambda

Bacteriophage lambda is a large capsid, which consists of 60 asymmetric units symmetrically arranged in an icosahedron-like shape. An asymmetric unit is composed of 7 proteins. This capsid encloses DNA, protecting it while the phage is outside of a cell. The capsid usually also has a portal complex which drives the DNA inside during assembly and pushes the DNA out during infection of a cell. The portal complex is not usually required for the capsid to assemble [75], which is why a portal is not present in this density map.

The density map of bacteriophage lambda at 14.5Å resolution [76] (EMDB:1507) was segmented at a threshold of 2.57, resulting in 12,580 regions. It was then smoothed with 5 steps of size 4.0Å. In the most smoothed map, 308 regions resulted, which were then sharpened. In this segmentation, 6-fold and 5-fold symmetric arrangements of proteins in the capsid are clearly visible, with single or groups of 2 regions corresponding to individual proteins. The regions making up an asymmetric unit, which includes 6 proteins in the 6-fold arrangements and 1 of the proteins from the 5-fold arrangement, were interactively selected and extracted from the rest of the regions. The entire asymmetric unit is shown in blue in the top row of Figure 5.9 superimposed on the entire density map. The segmented regions that make up a single asymmetric unit are also shown separately in the middle row.

5.6 Registration of structures with simulated density maps

To test the accuracy of the registration method, it was used to fit structures of individual proteins or subunits into simulated density maps. When the transformed positions, ($\vec{r}_i^{\,r}$), of the atoms in the registered structure, match the corresponding atomic positions, ($\vec{r}_i$), of a single component from the entire structure that was used to simulate the density map, the registration is accurate. The root-mean-square-deviation (RMSD) between these corresponding positions is computed as follows:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} \left\| \vec{r}_i^{\,r} - \vec{r}_i \right\|^2}{N}}$$

A low RMSD score indicates the corresponding atomic positions are close by, which means the registration is accurate. The entire process is illustrated in Figure 5.10.

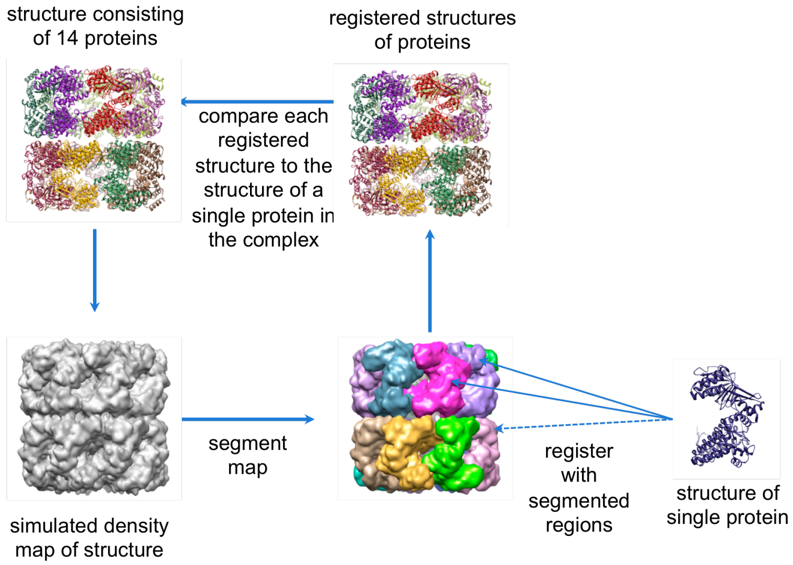
structure consisting of 14 proteins
registered structures of proteins
compare each registered structure to the structure of a single protein in the complex
segment map
register with segmented regions
structure of single protein
simulated density map of structure

Figure 5.10. Procedure for testing accuracy of registration. The structure consisting of 14 proteins (top left) is used to simulate a density map (bottom left), which is then segmented. The structure of a single protein (bottom right) is registered with each segmented region. Each resulting registered structures is compared with the structure of a single protein from the entire complex, using the RMSD score.

5.6.1 GroE, thermosome, ribosome, and HK97 asymmetric units

In the 5 simulated maps of GroE, thermosome, ribosome, and HK97 asymmetric units, as discussed in section 5.2, each segmented region corresponded to a single protein or subunit. The automated registration procedure was used, and since each region corresponded to a single structure being registered, all automatically generated groups contained only single regions. The RMSD between the atoms in each registered structure and the corresponding structure in the complex from which density maps was simulated were all less than 1Å, indicating that the registrations for all the structures were very accurate. The principal-axes registration method gave the correct registrations for all the structures, and rotational search was not required at all. The registration method was also tested with simulated maps at lower resolutions. Correct registrations were obtained for density maps simulated at up to 30Å resolution.

5.6.2 GroEL+GroES

We simulated a map of the structure (PDB:1aon) at 10Å resolution, with grid spacing of 2.0Å. Watershed segmentation of this map produces 1474 regions. As this map is smoothed, regions corresponding to individual proteins are not eventually obtained. Instead, regions spanning more than one protein result. However, stopping the smoothing process before this happens, the result is a small number of regions (48), with groups of 1-3 regions corresponding to single proteins (Figure 5.11). In total, we smoothed the map with 10 smoothing steps, with step size of 2Å. After sharpening, single regions are obtained for proteins in the lid (GroES) section, and groups of 2-3 regions for proteins in the barrel (GroEL) section.

The automated registration method was used to register the 3 different protein structures (chains A, H, and O) with the segmented regions. The principal-axes transform produced correct registrations for proteins in the barrel section, but not in the lid section, where rotational search was used. The RMSD between atoms in the registered structures and the corresponding atoms in the structure used to simulate the density map were all below 1.0Å, indicating that the correct registrations were produced. After the registration of all the structures, the regions overlapping the same structure were joined, to produce single regions corresponding to each structure. Segmentation accuracies for the resulting regions are plotted in Figure 5.13. Good accuracies are obtained for regions in the barrel section (0.81-0.90), but lower accuracies for regions in the lid section (0.64-0.71).

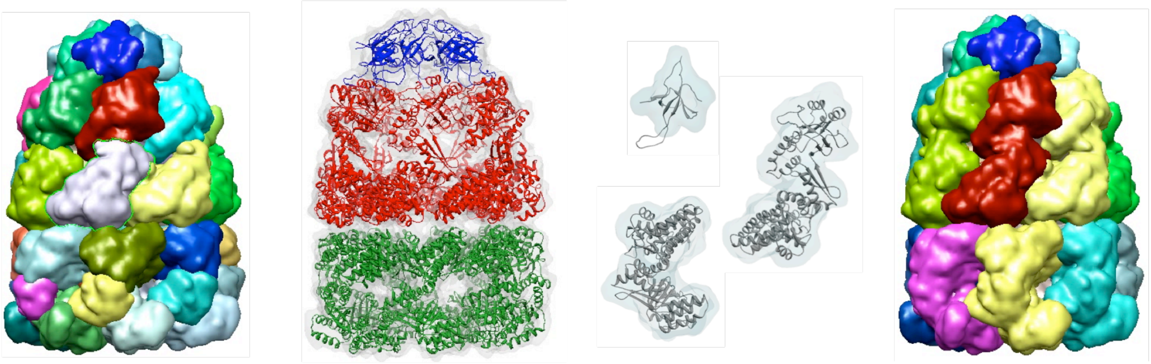

Figure 5.11. Registration of three structures with groups of regions from a simulated density map of GroEL+GroES. The regions resulting from the multi-scale method are shown at left. The structures were registered with groups of these regions, producing the registrations shown in the middle. Each structure is drawn as a ribbon, and each of the three different structures registered is drawn with a different color. The structures are shown individually along with the regions they were registered with as transparent surfaces. On the right, the regions joined based on which protein they join the most are shown.

5.6.3 Ribosome

For the simulated map of the E-coli ribosome (PDB:2avy,2aw4), as described in section 5.2, multi-scale grouping was able to produce single regions for the large and small subunits after a large degree of smoothing. In less-smoothed maps (5 steps of size 2.0Å), groups of 2-4 regions were found to correspond to single proteins.

The automated registration process was used to register all 49 protein structures to these regions. Most structures were registered correctly using the principal axes transform, however a few required rotational search. Regions corresponding to single proteins were produced; a few of these regions and the structures of the proteins that were registered with them are shown in Figure 5.12.

All correct registrations gave RMSD scores lower than 1Å between the registered structure and the corresponding structure in the complex, signifying correct registrations. Other regions not joined in this process were taken to belong to RNA components, shown with a gray surface in Figure 5.12. The segmentation accuracies computed by shape-match score between the segmented regions and ground-truth regions ranged between 0.322 and 0.933, and are plotted in Figure 5.13. Despite the lower segmentation accuracies for some of the regions (e.g. 0.322), the correct registration were still found, mainly because the centers of the segmented regions and the correct fit of the corresponding structure are still close enough to allow a good initial alignment.

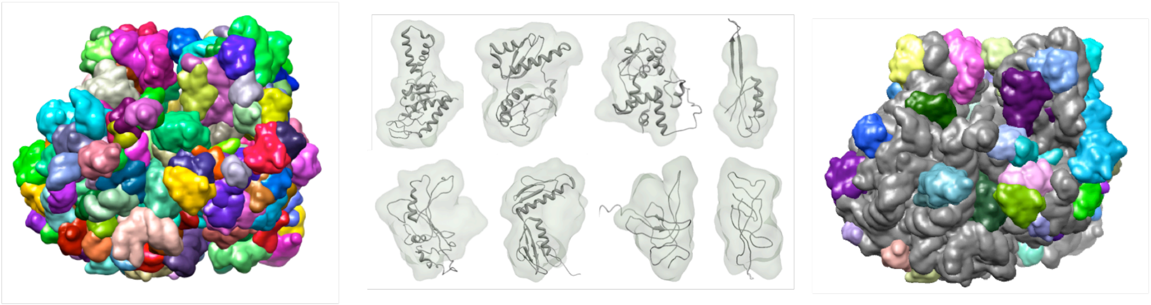

Figure 5.12. Segmented regions and registration of structures in simulated maps of the ribosome. On the left, the segmented regions after the multi-scale method was applied are shown. In the middle 8 of the 49 protein structures which were registered with groups of regions are shown, along with the corresponding region as a transparent surface. On the right, the regions were joined based on which structure they overlap. All remaining regions were joined to produce the grey region which corresponds to RNA.

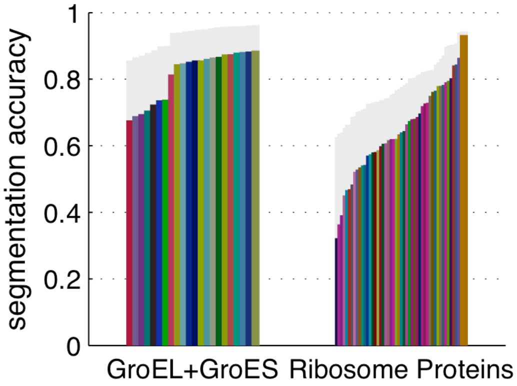

Figure 5.13. Segmentation accuracies in simulated maps of GroEL+GroES and ribosome. Segmentation accuracies of regions produced by the multi-scale method followed by registration of protein structures are plotted using randomly colored bars, and maximum watershed segmentation accuracies are plotted for each component using grey bars.

5.7 Registration of structures with cryo-EM density maps

Structures of individual proteins or subunits were registered to segmented regions in 5 cryo-EM density maps. For use in the registration process, maps for each structure were simulated at the same resolution as the experimentally reported resolution and grid spacing of the cryo-EM map. The segmentation and registration results are shown in Figure 5.14. The registered structures of each component were used to generate protein or subunit-masked regions. The shape match-score was used to measure how similar the segmented regions are to these regions. These scores, plotted in Figure 5.15, reflect segmentation accuracy, and also how similar cryo-EM and crystal structures of individual components in these structures are.

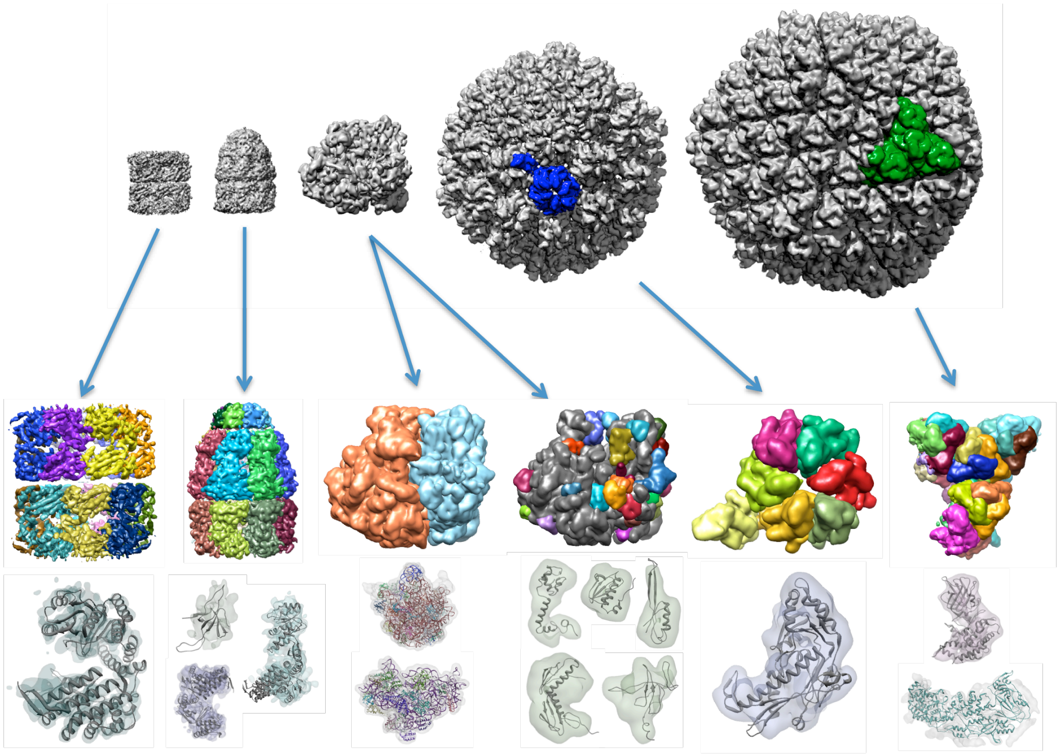

Figure 5.14. Segmented regions for 5 cryo-EM density maps and structures of proteins or subunits registered with them. The density maps are, from left to right, GroEL (EMDB:5001), GroEL+GroES (EMDB:1180), ribosome (EMDB:1056) large/small subunits and RNA/proteins, rice dwarf virus (EMDB:1060), and bacteriophage lambda (EMDB:1507). The top row shows regions after segmentation and registration, and the bottom row shows single regions as transparent surfaces and corresponding registered structures as ribbons. The structures are, from left to right, PDB:1xck chain A, 1aon chain A, 2avy all chains, 2avy chains M,I,J (top) and 2aw4 chains G,P (bottom), 1uf2 chain C, and 3bqw.

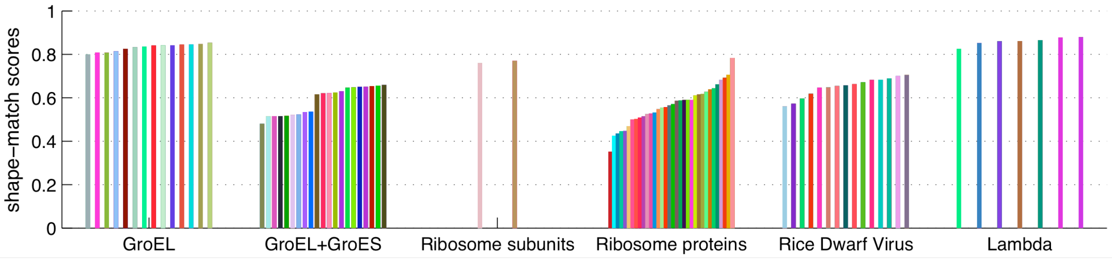

Figure 5.15. Shape-match scores between simulated density maps of registered structures and corresponding segmented regions in experimental density maps.

5.7.1 GroEL

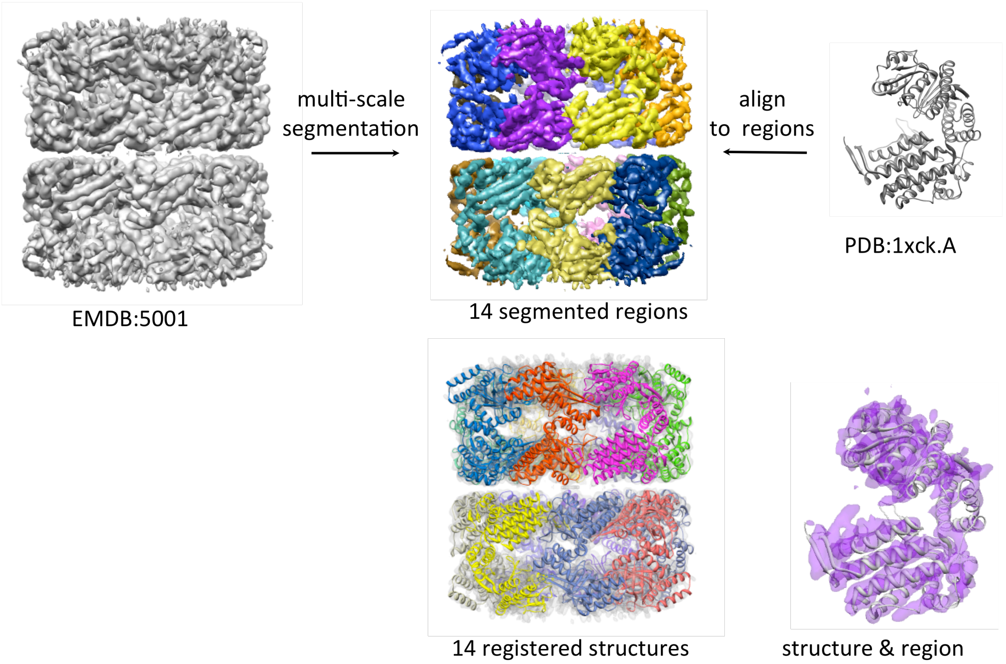

Figure 5.16 Segmentation and registration results for the density map of the GroEL chaperone.

The structure of a single protein in the GreEL complex (PDB:1xck, chain A), was used to simulate a density map which was aligned to the segmented regions, using the automated procedure. Because each region corresponded to a single structure, only 14 groups (each group containing a single region) were generated for alignment. The shape-match scores between each of the 14 segmented regions and protein-masked regions, generated from the registered structures, ranged between 0.799 and 0.854. The scores are slightly lower than for the analogous simulated density map, signifying lower segmentation accuracy (perhaps due to noise), and/or slight difference between crystal and cryo-EM structures

5.7.2 GroEL+GroES

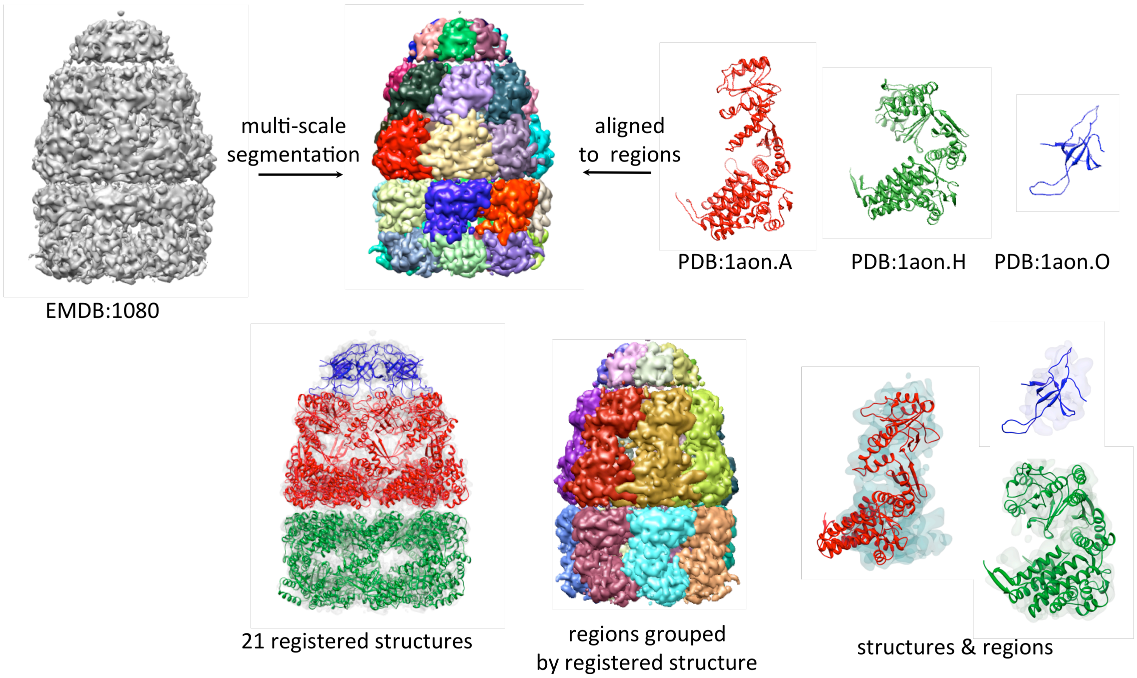

Figure 5.17 Segmentation and registration results for the density map of the GroEL+GroES chaperone.

Simulated maps of the 3 different proteins in the GroEL+GroES complex (PDB:1aon, chains A,H,O) were aligned to groups generated by the automatic procedure. For each structure, respectively, 58, 57, and 21 groups of regions were automatically generated, and alignment of the structures with these groups yielded the correct registrations.

Chain A registered correctly using rotational search with groups of 2 regions each. The resulting joined regions, compared to protein-masked regions, gave shape-match scores between 0.457 and 0.543. Chain H registered correctly using the principal-axis transform with 7 groups of 2 regions each in the lower barrel section with shape-match scores between 0.615 and 0.625. Chain O registered correctly also using the principal-axis transform with 7 regions in the lid section, with shape-match scores ranging between 0.412 and 0.558. All these scores are quite low, and by visual inspection, the cause appears to be a great deal of noise in the density map. Despite this noise, the segmentation and fitting methods still produced results consistent with the structure of the analogous simulated density map.

5.7.3 Ribosome

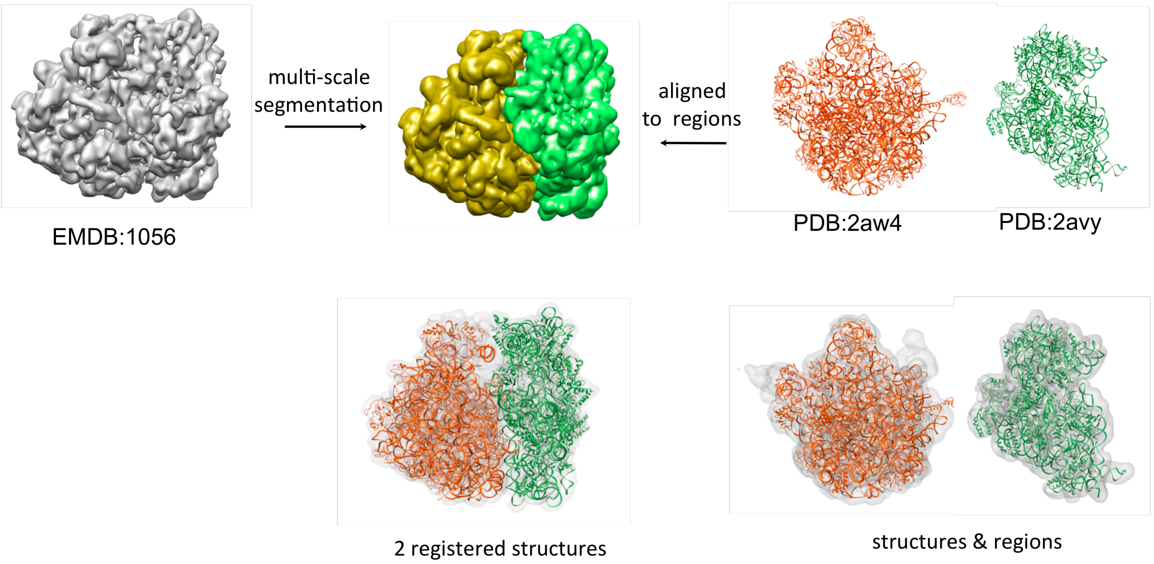

Figure 5.18 Segmentation of the ribosome density map into the large and small subunits, and registration of structures with the resulting regions.

Simulated maps from the structure of the large (PDB:2aw4) and small (PDB:2avy) subunits were registered correctly to the corresponding regions of the larger and small subunit, using the principal-axes transform, giving cross-correlations of 0.618 and 0.597 respectively. The shape-match scores between the segmented regions and protein-masked regions were 0.770 and 0.761.

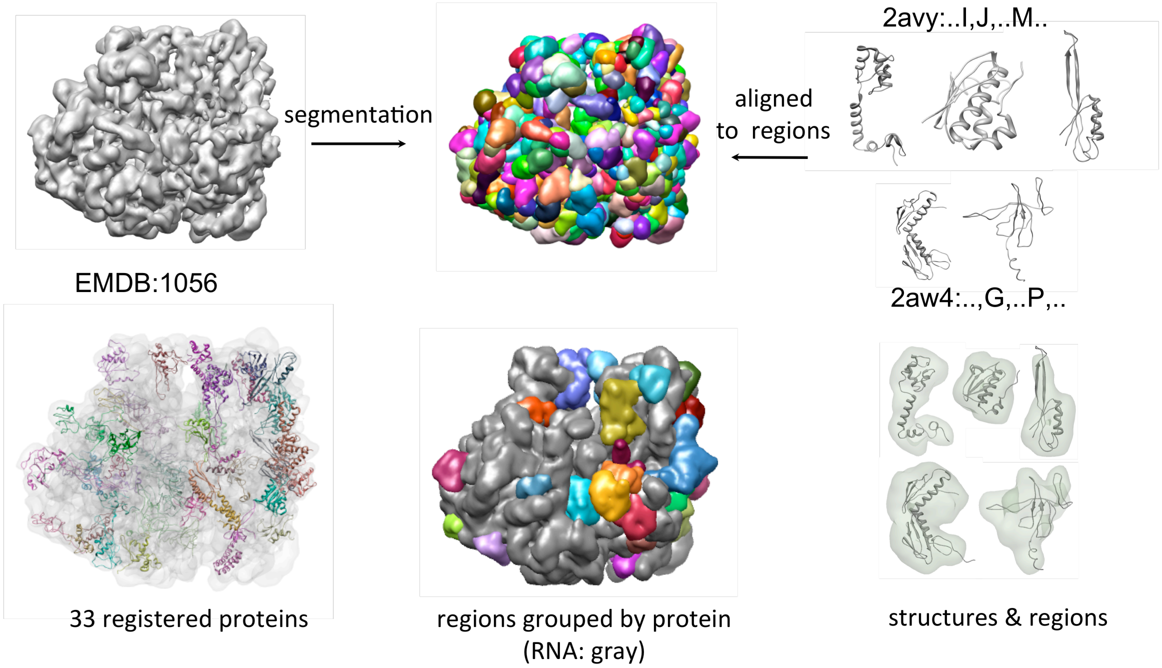

Figure 5.19 Segmentation of the ribosome density map and registration of the 49 proteins from both small and large subunits.

Simulated maps of each of the 49 proteins in both larger and small subunits were registered with regions in the unsmoothed cryo-EM map. Each structure was registered using the automated procedure. About 800 groups were generated for each structure. Of the 49 proteins, 33 were correctly registered (2avy chains B,C,D,E,F,G,I,J,M,O,P,Q,R,T,U, and 2aw4 chains 0,1,2,C,D,E,F,G,K,M,P,Q,R,S,U,V,X,Y,Z), most of them using only the principal-axes transform.

The shape-match scores computed for the regions and the protein-masked regions ranged between 0.436 and 0.784. For the proteins that weren't registered correctly, the potential cause is that the state for the ribosome captured in the cryo-EM map is different than the state captured in the crystal structure, so that some of the proteins may have different conformations, or may not be present at all. In particular, the region in which transcription factor is bound appears substantially different in the cryo-EM density map and crystal structure.

5.7.4 Bacteriophage lambda

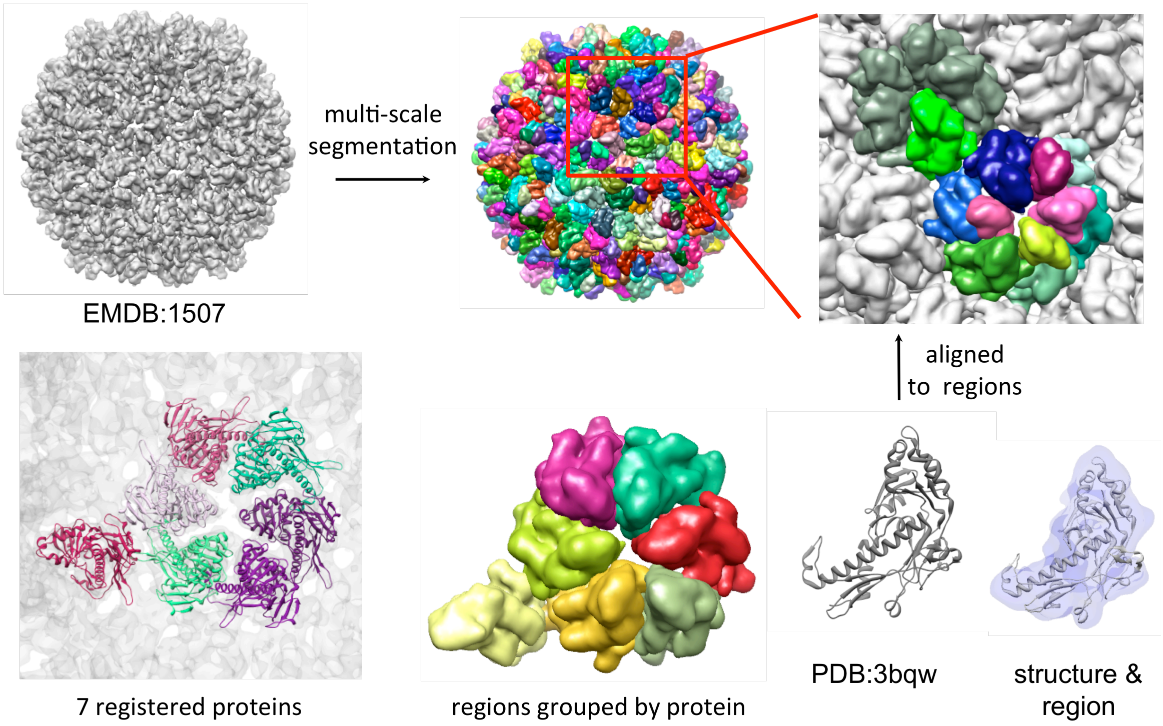

Figure 5.20 Segmentation of the density map of the bacteriophage lambda, and registration of the structure of a single protein with regions from an asymmetric unit.

A total of 10 regions, corresponding to 7 proteins, which make up an asymmetric unit were interactively selected. Amongst these regions, 4 of them corresponded to individual proteins, and the remaining 6, in groups of 2, corresponded to the other 3 proteins.

The structure of a single pro-capsid protein (PDB:3bqw) was registered to these selected regions using the automated procedure. A total of 21 groups of adjacent regions were considered. The principal-axes transform produced correct registrations of the structure to 7 of these groups (some of which were groups containing a single region). The shape-match scores computed between segmented regions and protein-masked regions were quite high, ranging between 0.825 and 0.879.

5.7.5 Rice dwarf virus (RDV)

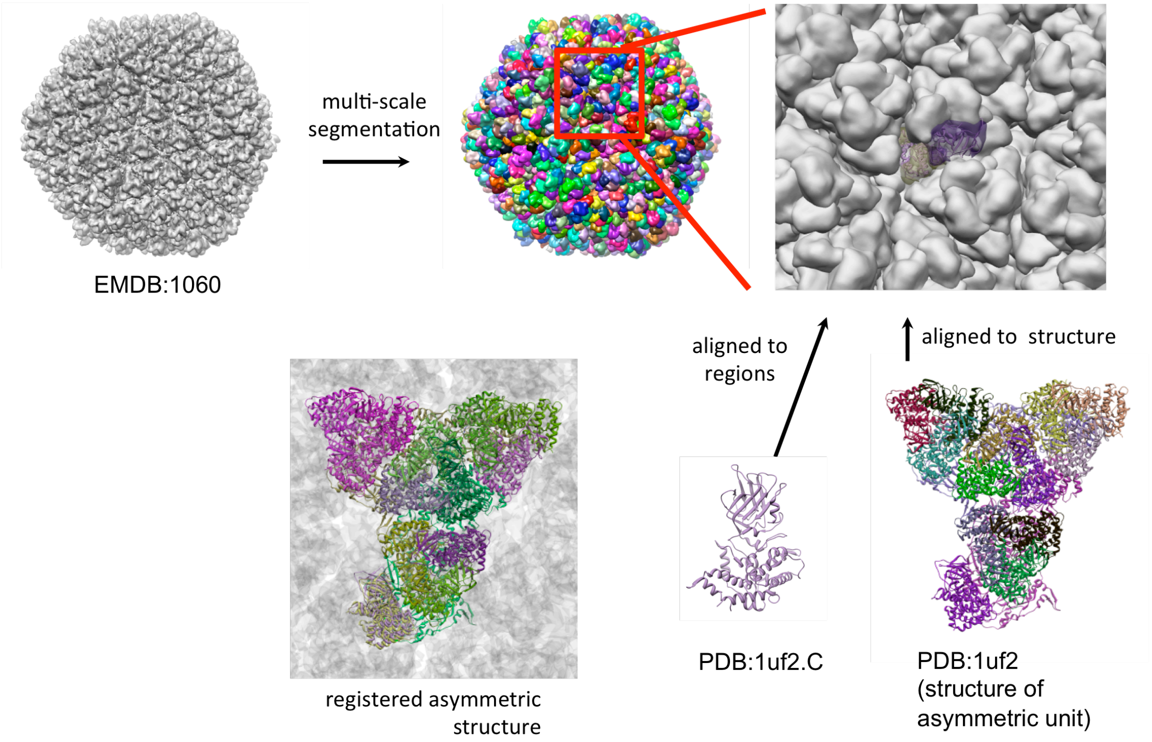

Figure 5.21 Segmentation of the density map of the rice dwarf virus, and registration of the structure of a single asymmetric unit with the map.

The density maps of the rice dwarf virus at 6.8Å resolution [77] (EMDB:1060) contains a symmetric half of the T=15 icosahedral capsid. The segmentation of this map is not shown in the results in chapter 3, since the registration of the entire asymmetric unit, as described below, is required. Segmentation of the entire map alone, without registration, is challenging because the this virus contains both an outer and an inner capsid. However, after extraction of a single asymmetric unit, as described below, the segmentation of this ASU and the registration of individual proteins with it are much easier.

The entire density map was segmented at a threshold of 1.8, resulting in 19,416 regions. The map was smoothed twice with step size of 5.0Å. The most smoothed map produced 1,618 regions, which were sharpened. Individual proteins in the outer capsid could be seen in this segmentation, corresponding with groups of 2 regions each. The crystal structure of the asymmetric unit of this virus (PDB:1uf2) is composed of 13 proteins that form trimers in the outer capsid, and 2 proteins in the inner capsid. A simulated density map of chain C, one of the trimer proteins, was registered correctly using the principal-axes transform with two of the regions, which were selected interactively.

The structure of the entire asymmetric unit was placed into the density map by alignment of the corresponding chain in the structure to the registered chain. The resulting registration was then locally refined. The cryo-EM map was masked with this structure, thus extracting a map the asymmetric unit alone. This was done to simplify further segmentation and registration of structures.

The map of the asymmetric unit alone was then segmented, producing 1155 regions. It was smoothed with 7 steps of size 2.0Å. The most smoothed map produced 65 regions, which were then sharpened. Groups of 2-5 regions corresponded to each protein in this segmentation. Structures of each protein (chains A, B, and C from PDB:1uf2) were aligned with regions using the automated procedure. In total 34, 89, and 94 groups were considered for each structure respectively. All structures were correctly registered using only the principal-axes transform. The shape match scores between segmented regions and protein-masked regions were between 0.561 and 0.704. These scores are quite low, signifying lower segmentation accuracy and/or more substantial differences between crystal and cryo-EM structures.

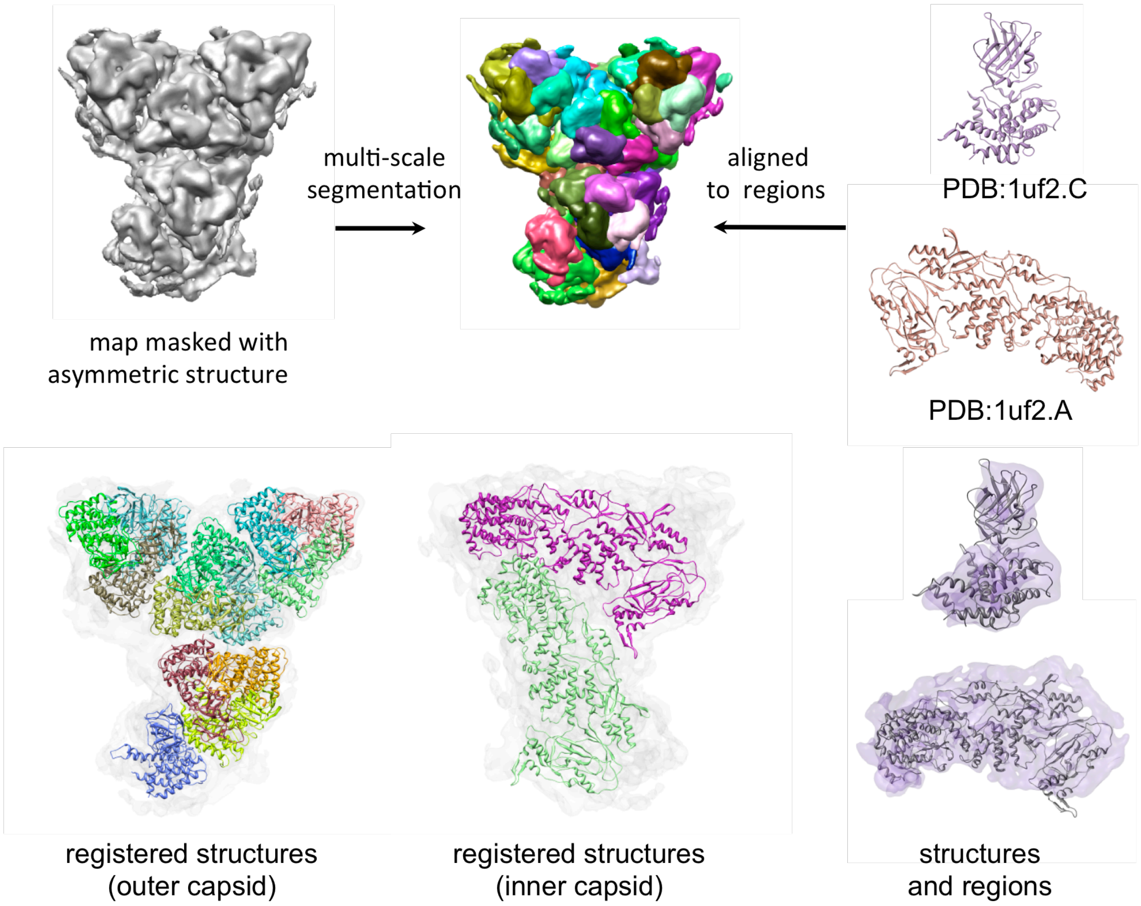

Figure 5.22 Segmentation of the density map of the rice dwarf virus, masked with the structure of a single asymmetric unit, and registration of structures with the resulting regions.

Chapter 6. Conclusions and future work

6.1 Segmentation

A segmentation method that is very easy to use and requires little prior structural knowledge has been presented. This method can produce the segmentation of a density map in several minutes, a process that can otherwise take hours by more interactive approaches. Interactive segmentation is also highly tedious and subjective, requiring a lot of knowledge and skill on the part of the user. The user-interaction required for the method we presented is very minimal. Thus the method is also more objective, requires less skill and knowledge, and gives reproducible results given only three parameters: the initial threshold, the smoothing step size, and the number of smoothing steps.

A metric was used to quantitatively measure segmentation accuracies, by comparison of the segmented regions to protein/subunit-masked regions. Good accuracies were obtained using the multi-scale method. However in some complexes, narrow protrusions were not segmented correctly, and thus lower accuracies were obtained. Maximal accuracies attainable using the watershed method were also computed, showing that by grouping regions obtained using the watershed method, very accurate segmentations are possible. Future studies will attempt to study whether the accuracy of the multi-scale method can be further improved.

A method was also presented allowing the user to subdivide the resulting regions recursively into smaller regions, and to regroup regions so as to locally modify segmented regions. The use of a hierarchy makes the process simple and intuitive. It was shown that the segmentation accuracy could be improved using such edits. It will also be interesting to further study how users respond to these user-editing capabilities, how effective they are, and how easy-to-use other users will find them.

6.2 Registration

Methods were presented which allow structures of individual structures to be accurately, quickly, and reliably registered with a density map, through the alignment of structures to segmented regions. Two alignment methods were used, based on alignment of centers and principal-axes or rotational search. The principal-axes transform is extremely fast since the registration is direct, and it is successful in many of the cases presented here. When it doesn't work, the rotational search is able to find the correct registration, and is also relatively fast since it only searches through 3 degrees of freedom, compared to exhaustive search, which searches through 6 degrees of freedom. These registration methods were shown to be very accurate when used with simulated maps. Their use in experimental density maps was also very successful, producing registrations in which the registered structures closely matched the segmented regions.

For future work, it will be important to allow flexibility in the structure being registered, so that it better captures different conformations of the components in cryo-EM density maps. This is an important task, since it will allow us to discover structures of complexes in a wider variety of states seen in cryo-EM density maps. The use of the methods described in this work will help with this task. Firstly, the initial registration for a structure can be created using the registration methods presented here. Moreover, the target shape of the structure of a single component can be obtained by segmentation of the density map.

6.3 Public use of contributed methods

Aside from focusing extensively on providing accurate and efficient methods for segmentation and registration, we have also aimed to make the methods presented here easy to use and widely accessible to the public, through the *Segger* software [70]. Continued effort in this direction should lead to improved tools allowing us to more quickly and accurately extract important biological information from the wide variety of density maps obtained by the increasingly popular cryo-EM method.

References

[1] W. Massa, *Crystal structure determination*, Springer, 2004.

[2] L.D. Landau and E.M. Lifshitz, *Electrodynamics of Continuous Media*, Oxford: Pergamon, 1960.

[3] C. Brändén and J. Tooze, *Introduction to protein structure*, New York (N.Y.): Garland, 1999.

[4] "http://www.pdb.org."

[5] B. Horn, "Density reconstruction using arbitrary ray-sampling schemes," *Proceedings of the IEEE*, vol. 66, 1978, pp. 551-562.

[6] S.J. Ludtke, P.R. Baldwin, and W. Chiu, "EMAN: semiautomated software for high-resolution single-particle reconstructions," *Journal of structural biology*, vol. 128, Dec. 1999, pp. 82-97.

[7] T.R. Shaikh, H. Gao, W.T. Baxter, F.J. Asturias, N. Boisset, A. Leith, and J. Frank, "SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs," *Nature Protocols*, vol. 3, 2008, pp. 1941-1974.

[8] S. Deans, *The Radon Transform and Some of Its Applications*, Krieger Publishing Company, .

[9] W.E. Lorensen and H.E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *SIGGRAPH Comput. Graph.*, vol. 21, 1987, pp. 163-169.

[10] "http://www.emdatabank.org," 2002.

[11] M. van Heel and M. Schatz, "Fourier shell correlation threshold criteria," *Journal of Structural Biology*, vol. 151, Sep. 2005, pp. 250-62.

[12] W. Jiang, M.L. Baker, J. Jakana, P.R. Weigele, J. King, and W. Chiu, "Backbone structure of the infectious [epsi]15 virus capsid revealed by electron cryomicroscopy," *Nature*, vol. 451, Feb. 2008, pp. 1130-1134.

[13] S.J. Ludtke, M.L. Baker, D. Chen, J. Song, D.T. Chuang, and W. Chiu, "De novo backbone trace of GroEL from single particle electron cryomicroscopy," *Structure (London, England : 1993)*, vol. 16, Mar. 2008, pp. 441-8.

[14] X. Yu, L. Jin, and Z.H. Zhou, "3.88[thinsp]A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy," *Nature*, vol. 453, May. 2008, pp. 415-419.

[15] X. Zhang, E. Settembre, C. Xu, P.R. Dormitzer, R. Bellamy, S.C. Harrison, and N. Grigorieff, "Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction.," *Proc Natl Acad Sci U S A*, Jan. 2008.

[16] W. Wriggers and P. Chacón, "Modeling tricks and fitting techniques for multiresolution structures," *Structure (London, England: 1993)*, vol. 9, Sep. 2001, pp. 779-88.

[17] L.G. Shapiro and G.C. Stockman, *Computer Vision*, Prentice Hall, 2002.

[18] D. Pham, C. Xu, and J. Prince, "A Survey of Current Methods in Medical Image Segmentation," *Annual Review of Biomedical Engineering*, 2000, pp. 338, 315.

[19] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 8, 1.

[20] P. Dollar, Z. Tu, and S. Belongie, "Supervised Learning of Edges and Object Boundaries," *Proceedings of the 2006 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition - Volume 2, IEEE Computer Society, 2006, pp. 1964-1971.

[21] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” *International Journal of Computer Vision*, vol. 1, Jan. 1988, pp. 321-331.

[22] R. Malladi, J.A. Sethian, and B.C. Vemuri, “Shape modeling with front propagation: A level set approach,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, 1995, pp. 158--175.

[23] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, 2000, pp. 888--905.

[24] P.F. Felzenszwalb and D.P. Huttenlocher, “Efficient Graph-Based Image Segmentation,” *Int. J. Comput. Vision*, vol. 59, 2004, pp. 167-181.

[25] L. Grady, “Random Walks for Image Segmentation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, 2006, pp. 1768-1783.

[26] K. Fukunaga and L. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition,” *Information Theory, IEEE Transactions on*, vol. 21, 1975, pp. 32-40.

[27] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, 2002, pp. 603-619.

[28] S. Paris and F. Durand, “A Topological Approach to Hierarchical Segmentation using Mean Shift,” *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1-8.

[29] L. Vincent and P. Soille, “Watersheds in digital spaces: an efficient algorithm based on immersion simulations,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, 1991, pp. 583-598.

[30] Beucher, S. and Lantuejoul, C, “Use of watersheds in contour detection,” Rennes, France: 1979.

[31] A. Witkin, “Scale-space filtering: A new approach to multi-scale description,” 1984, pp. 153, 150.

[32] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, 1990, pp. 629-639.

[33] L.M. Lifshitz and S.M. Pizer, “A Multiresolution Hierarchical Approach to Image Segmentation Based on Intensity Extrema,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, 1990, pp. 529-540.

[34] T. Lindeberg, “Edge detection and ridge detection with automatic scale selection,” *International Journal of Computer Vision*, vol. 30, 1996, pp. 465--470.

[35] X. Ren, “Multi-scale Improves Boundary Detection in Natural Images,” *Proceedings of the 10th European Conference on Computer Vision: Part III*, Marseille, France: Springer-Verlag, 2008, pp. 533-545.

[36] U. Braga-Neto and J. Goutsias, “Object-based image analysis using multiscale connectivity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, Jun. 2005, pp. 892-907.

[37] M.L. Baker, Z. Yu, W. Chiu, and C. Bajaj, “Automated segmentation of molecular subunits in electron cryomicroscopy density maps,” *Journal of Structural Biology*,

vol. 156, Dec. 2006, pp. 432-441.

[38] N. Volkmann, "A novel three-dimensional variant of the watershed transform for segmentation of electron density maps," *Journal of Structural Biology*, vol. 138, 2002, pp. 123-129.

[39] J. Liu, D.W. Taylor, E.B. Krementsova, K.M. Trybus, and K.A. Taylor, "Three-dimensional structure of the myosin V inhibited state by cryoelectron tomography," *Nature*, vol. 442, Jul. 2006, pp. 208-211.

[40] B. Marcotegui, S. Beucher, and C. De Morphologie Mathématique, "Fast implementation of waterfall based on graphs," *Volume 30 of Computational Imaging and Vision*, vol. 30, 2005, pp. 177--186.

[41] T.D. Goddard, C.C. Huang, and T.E. Ferrin, "Visualizing density maps with UCSF Chimera," *Journal of structural biology*, vol. 157, Jan. 2007, pp. 281-7.

[42] J.B. Heymann and D.M. Belnap, "Bsoft: Image processing and molecular modeling for electron microscopy," *Journal of Structural Biology*, vol. 157, Jan. 2007, pp. 3-18.

[43] S. Pruggnaller, M. Mayr, and A.S. Frangakis, "A visualization and segmentation toolbox for electron microscopy," *Journal of Structural Biology*, vol. 164, Oct. 2008, pp. 161-165.

[44] B. J, A.P. Witkin, M. Baudin, and R.O. Duda, "Uniqueness of the Gaussian kernel for scale-space filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, 1986, pp. 26-33.

[45] B. Zitova, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, Oct. 2003, pp. 1000, 977.

[46] D. Lowe, "Object Recognition from Local Scale-Invariant Features," 1999, pp. 1150--1157.

[47] N.M. Alpert, J.F. Bradshaw, D. Kennedy, and J.A. Correia, "The Principal Axes Transformation--A Method for Image Registration," *J Nucl Med*, vol. 31, Oct. 1990, pp. 1717-1722.

[48] B.K.P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *JOURNAL OF THE OPTICAL SOCIETY OF AMERICA A*, vol. 4, 1987, pp. 629--642.

[49] W.K. Pratt, *Digital image processing (2nd ed.)*, John Wiley \&amp; Sons, Inc., 1991.

[50] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, *Multi-modal volume registration by maximization of mutual information*, 1996.

[51] E.D. Castro and C. Morandi, "Registration of translated and rotated images using finite Fourier transforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, 1987, pp. 700-703.

[52] J.A. Kovacs and W. Wriggers, "Fast rotational matching," *Acta Crystallographica. Section D, Biological Crystallography*, vol. 58, Aug. 2002, pp. 1282-6.

[53] T.A. Jones, J.Y. Zou, S.W. Cowan, and M. Kjeldgaard, "Improved methods for building protein models in electron density maps and the location of errors in these models," *Acta Crystallographica. Section A, Foundations of Crystallography*, vol. 47 (Pt 2), Mar. 1991, pp. 110-119.

[54] W. Wriggers and S. Birmanns, "Using situs for flexible and rigid-body fitting of multiresolution single-molecule data," *Journal of Structural Biology*, vol. 133, Mar.

2001, pp. 193-202.

[55] M.G. Rossmann, R. Bernal, and S.V. Pletnev, “Combining Electron Microscopic with X-Ray Crystallographic Structures,” *Journal of Structural Biology*, vol. 136, Dec. 2001, pp. 190-200.

[56] A.M. Roseman, “Docking structures of domains into maps from cryo-electron microscopy using local correlation,” *Acta Crystallographica. Section D, Biological Crystallography*, vol. 56, Oct. 2000, pp. 1332-40.

[57] Wriggers W. [1] [2], Milligan R.A. [2], and McCammon J.A. [1], “Situs: A Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy,” *Journal of Structural Biology*, vol. 125, 1999, pp. 185-195.

[58] J. Navaza, J. Lepault, F.A. Rey, C. Alvarez-Rúa, and J. Borge, “On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation,” *Acta Crystallographica. Section D, Biological Crystallography*, vol. 58, Oct. 2002, pp. 1820-1825.

[59] W. Jiang, M. Baker, S. Ludtke, and W. Chiu, “Bridging the information gap: Computational tools for intermediate resolution structure interpretation,” *Journal of Molecular Biology*, vol. 308, 2001, pp. 1033-1044.

[60] J.A. Kovacs, P. Chacón, Y. Cong, E. Metwally, and W. Wriggers, “Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom,” *Acta Crystallographica. Section D, Biological Crystallography*, vol. 59, Aug. 2003, pp. 1371-1376.

[61] J.I. Garzon, J. Kovacs, R. Abagyan, and P. Chacon, “ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage,” *Bioinformatics*, vol. 23, Feb. 2007, pp. 427-433.

[62] S. Birmanns and W. Wriggers, “Multi-resolution anchor-point registration of biomolecular assemblies and their components,” *Journal of Structural Biology*, vol. 157, Jan. 2007, pp. 271-280.

[63] P. Chacón and W. Wriggers, “Multi-resolution contour-based fitting of macromolecular structures,” *Journal of Molecular Biology*, vol. 317, Mar. 2002, pp. 375-384.

[64] H. Ceulemans and R.B. Russell, “Fast Fitting of Atomic Structures to Low-resolution Electron Density Maps by Surface Overlap Maximization,” *Journal of Molecular Biology*, vol. 338, May. 2004, pp. 783-793.

[65] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali, “Protein structure fitting and refinement guided by cryo-EM density,” *Structure (London, England: 1993)*, vol. 16, Feb. 2008, pp. 295-307.

[66] M. Topf, M.L. Baker, M.A. Marti-Renom, W. Chiu, and A. Sali, “Refinement of Protein Structures by Iterative Comparative Modeling and CryoEM Density Fitting,” *Journal of Molecular Biology*, vol. 357, Apr. 2006, pp. 1655-1668.

[67] H. Chen, “Gradient-based approach for fine registration of panorama images,” *J. Comput. Sci. Technol.*, vol. 19, 2004, pp. 691-697.

[68] T.D. Goddard, C.C. Huang, and T.E. Ferrin, “Software Extensions to UCSF Chimera for Interactive Visualization of Large Molecular Assemblies,” *Structure*, vol. 13, Mar. 2005, pp. 473-482.

[69] W.H. Press, *Numerical recipes*, Cambridge University Press, 2007.

[70] “http://people.csail.mit.edu/gdp/segger,” 2009.

[71] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin, “UCSF Chimera--a visualization system for exploratory research and analysis,” *Journal of computational chemistry*, vol. 25, Oct. 2004, pp. 1605-12.

[72] N.A. Ranson, D.K. Clare, G.W. Farr, D. Houldershaw, A.L. Horwich, and H.R. Saibil, “Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes,” *Nat Struct Mol Biol*, vol. 13, Feb. 2006, pp. 147-152.

[73] J. Zhang, “Personnal Communication.”

[74] M. Valle, A. Zavialov, W. Li, S.M. Stagg, J. Sengupta, R.C. Nielsen, P. Nissen, S.C. Harvey, M. Ehrenberg, and J. Frank, “Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy,” *Nat Struct Mol Biol*, vol. 10, Nov. 2003, pp. 899-906.

[75] P.A. Thuman-Commike, B. Greene, J.A. Malinski, J. King, and W. Chiu, “Role of the Scaffolding Protein in P22 Procapsid Size Determination Suggested by T = 4 and T = 7 Procapsid Structures,” *Biophysical Journal*, vol. 74, Jan. 1998, pp. 559-568.

[76] G.C. Lander, A. Evilevitch, M. Jeembaeva, C.S. Potter, B. Carragher, and J.E. Johnson, “Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM,” *Structure (London, England: 1993)*, vol. 16, Sep. 2008, pp. 1399-1406.

[77] Z.H. Zhou, M.L. Baker, W. Jiang, M. Dougherty, J. Jakana, G. Dong, G. Lu, and W. Chiu, “Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus,” *Nat Struct Mol Biol*, vol. 8, Oct. 2001, pp. 868-873.