Spring 2015

Computational Development for Secondary Structure Detection From Three-Dimensional Images of Cryo-Electron Microscopy

Dong Si
Old Dominion University

COMPUTATIONAL DEVELOPMENT FOR SECONDARY STRUCTURE DETECTION FROM THREE-DIMENSIONAL IMAGES OF CRYO-ELECTRON MICROSCOPY

by

Dong Si
B.S. June 2007, Nanjing University, China
M.S. June 2010, Chang'an University, China
M.S. May 2014, Old Dominion University, USA

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
May 2015

Approved by:

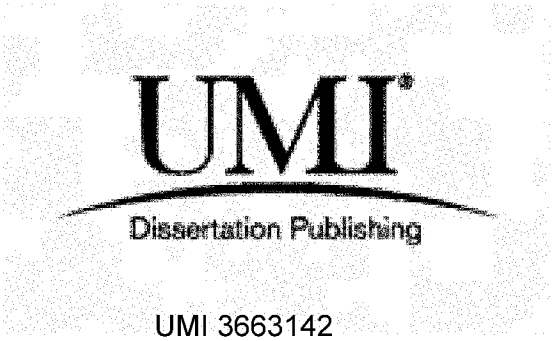
Jing He (Director)

Nikos Chrisochoides (Member)

Shuiwang Ji (Member)

Desh Ranjan (Member)

Lesley Greene (Member)



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

COMPUTATIONAL DEVELOPMENT FOR SECONDARY STRUCTURE DETECTION FROM THREE-DIMENSIONAL IMAGES OF CRYO-ELECTRON MICROSCOPY

Dong Si
Old Dominion University, 2015
Director: Dr. Jing He

Electron cryo-microscopy (cryo-EM) as a cutting edge technology has carved a niche for itself in the study of large-scale protein complex. Although the protein backbone of complexes cannot be derived directly from the medium resolution (5–10 Å) of amino acids from three-dimensional (3D) density images, secondary structure elements (SSEs) such as alpha-helices and beta-sheets can still be detected. The accuracy of SSE detection from the volumetric protein density images is critical for *ab initio* backbone structure derivation in cryo-EM. So far it is challenging to detect the SSEs automatically and accurately from the density images at these resolutions. This dissertation presents four computational methods - *SSEtracer*, *SSElearner*, *StrandTwister* and *StrandRoller* for solving this critical problem.

An effective approach, *SSEtracer*, is presented to automatically identify helices and β-sheets from the cryo-EM three-dimensional maps at medium resolutions. A simple mathematical model is introduced to represent the β-sheet density. The mathematical model can be used for β-strand detection from medium resolution density maps. A machine learning approach, *SSElearner*, has also been developed to automatically identify helices and β-sheets by using the knowledge from existing volumetric maps in the Electron Microscopy Data Bank (EMDB). The approach has been tested using

simulated density maps and experimental cryo-EM maps of EMDB. The results of *SSElearner* suggest that it is effective to use one cryo-EM map for learning in order to detect the SSE in another cryo-EM map of similar quality.

Major secondary structure elements such as α-helices and β-sheets can be computationally detected from cryo-EM density maps with medium resolutions of 5-10Å. However, a critical piece of information for modeling atomic structures is missing, since there are no tools to detect β-strands from cryo-EM maps at medium resolutions. A new method, *StrandTwister*, has been proposed to detect the traces of β-strands through the analysis of twist, an intrinsic nature of β-sheet. *StrandTwister* has been tested using 100 β-sheets simulated at 10Å resolution and 39 β-sheets computationally detected from cryo-EM density maps at 4.4-7.4Å resolutions. *StrandTwister* appears to detect the traces of β-strands on major β-sheets quite accurately, particularly at the central area of a β-sheet.

β-barrel is a structure feature that is formed by multiple β-strands in a barrel shape. There is no existing method to derive the β-strands from the 3D image of β-barrel. A new method, *StrandRoller*, has been proposed to generate small sets of possible β-traces from the density images at medium resolutions of 5-10Å. The results of *StrandRoller* suggest that it is possible to derive a small set of possible β-traces from the β-barrel cryo-EM image at medium resolutions even when it is not possible to visualize the separation of β-strands.

This thesis is dedicated to my parents, my wife and my first baby – Nolan Haochen Si.

ACKNOWLEDGMENTS

I would like to acknowledge the inspirational instruction and guidance of my advisor and committee chair Dr. Jing He. I would like to thank her for the countless hours of mentoring, encouraging, and most of all patience throughout the entire research.

I wish to thank my committee members - Dr. Shuiwang Ji, Dr. Nikos Chrisochoides, Dr. Desh Ranjan and Dr. Lesley Greene, who were more than generous with their expertise and precious time on providing the invaluable counseling and serving in my advisory committee.

I would like to express my deepest gratitude to my colleagues - Kamal H. Al Nasr, Lin Chen, Hao Ji, Rongjian Li, Wei Li, Abhishek Biswas, and all my friends in the US, China, and worldwide. They were always willing to help and give their best suggestions.

TABLE OF CONTENT

Page

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Proteins carry out vital functions within cells and make up more than half the cells dry weight. Its functions vary from acting as enzymes, to cellular signaling, and to molecular transportation. They follow energetically favorable pathways to form a unique and stable three-dimensional (3D) structure known as its native conformation. These folded protein structures are critical in biological functions as they are required to be in specific folded state [1-5]. The sequence of amino acids that constructs the protein ultimately determines its native structure. The structure can be categorized to four levels as follows:

Primary Structure: The primary structure of a protein refers to the linear sequence of amino acids in the polypeptide chain. It is held together by covalent bonds (e.g. peptide bonds) made during the process of protein biosynthesis or translation. The primary structure is determined by the gene corresponding to the protein. Figure 1A depicts a portion of the primary structure.

Secondary structure: The secondary structure of a protein refers to a regular sub-conformational structure formed by consecutive amino acids stabilized by hydrogen bonds (H-bonds). The most common examples of secondary structures are alpha-helices (α-helices), beta-sheets (β-sheets), and turns/loops (see Figure 1B). Helices and sheets are geometrically stabilized by hydrogen bonds between peptide groups. Different regions on

the polypeptide chain may adopt different secondary structures according to the primary sequence of amino acids in the protein.

a. **Helix:** The helix is the most common and most predictable secondary structure based on the amino acid sequence. The orientation of such a conformation produces a helical coiling of the peptide backbone causes the side chain groups to stem out of the helix coil and sit perpendicular to the axis. Not all amino acids are optimal in forming helices due to constraints of their side chains. Amino acids such as alanine, asparatic acid, glutamic acid, isoleucine, leucine, and methionine favor the formation of α-helices, whereas, glycine and proline disrupt helix formation. Figure 1B (red) shows the geometry of a helix.

b. **Beta-sheet:** The second most common secondary structure, β-sheets are composed of two or more different strands of amino acids connected by backbone hydrogen bonds. β-sheets are either parallel or anti-parallel. Parallel sheets that following the peptide chain proceed in the same direction, whereas, anti-parallel sheets that following the chain are aligned in opposite directions. Figure 1B (blue) shows an example of a β-sheet with two anti-parallel strands.

c. **Turns/loops:** Turns and loops play an important role in protein 3D structures by connecting together β-strands, strands to helices, or helices to one other. The amino acid sequences in turn regions may vary. Figure 1B (yellow) shows examples of turns and loops.

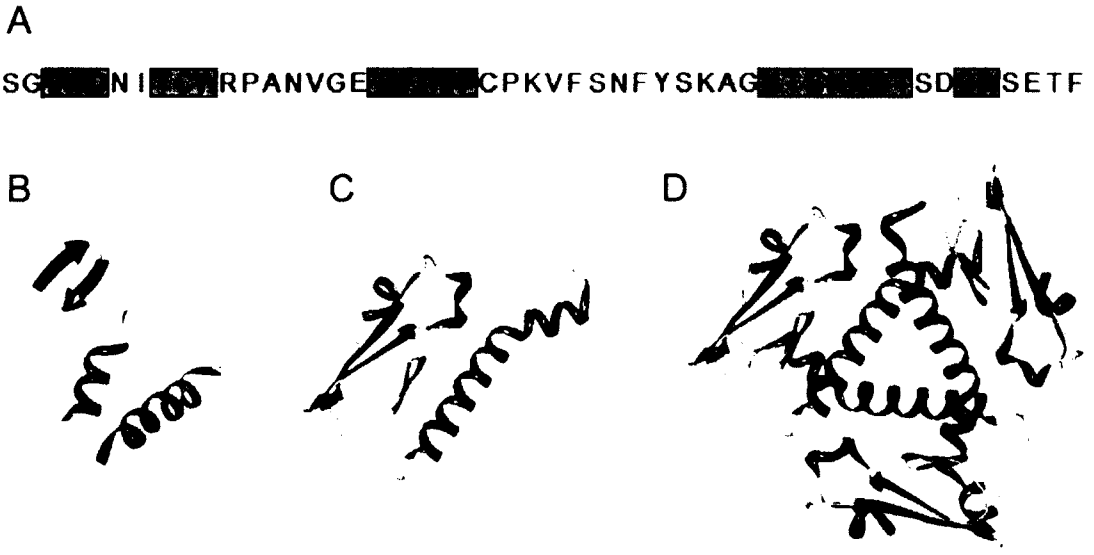


Figure 1. The four levels of protein structures. (A) The primary structure, only ordered sequence of amino acids; (B) secondary structures, β-sheet (blue) is shown as segments of stretches, helices (red) are spiral, and loop/turn (yellow) connects other secondary structures; (C) tertiary structure, complete 3D structure of a single protein molecule; (D) quaternary structure, multiple polypeptides.

Tertiary Structure: The tertiary structure of a protein refers to the formation of a complete 3D structure of a single protein molecule. It defines the spatial relationship of different secondary structures to one another within a polypeptide chain. It also describes the relationship of different domains to one another within a protein. The physics of the intra-protein and the environment governs the interaction between different domains such as hydrogen bonding, hydrophobic interactions, electrostatic interactions, and van der Waals forces. An example of tertiary structure is shown in Figure 1C.

Quaternary Structure: The quaternary structure of a protein refers to multiple polypeptide chains that may form the protein molecule. The quaternary structure is stabilized by disulfide bonds and the same non-covalent interactions as the tertiary structure. Figure 1D shows one example of quaternary structure.

1. Protein Structure Determination and Prediction

A number of experimental techniques are used to determine the structure of proteins. Two such techniques are X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. The more common X-ray crystallography measures the 3D density distribution of electrons in the protein and accounts for the prediction of approximately 90% of proteins found in the Protein Data Bank (PDB) [6, 7]. Unfortunately, in some cases, both techniques can be very expensive and time consuming (sometimes longer than a year). Therefore, developing new computational methods to predict the structure of proteins has been given considerable attention and effort [8].

Protein determination techniques are expensive, time-consuming, and not always successful with every type of protein. Membrane proteins are an example of a type that is hard to be successfully determined by experimental methods [9, 10]. The success of X-ray crystallography is limited to the existence of suitable crystals from the protein, and unfortunately, large proteins cannot easily produce crystals. On the contrary, the sequencing of proteins is fast, simple, and relatively less expensive. As the number of genome projects increase worldwide, the difference between number of sequences and known 3D structures is rapidly increasing. The number of protein sequences available at

the time of writing this dissertation is more than 87 million[1] while the number of structure determined and posted on Protein Data Bank[2] is only 107,436. Furthermore, the sequence of amino acids, together with the physics of the intra-protein and the environment interactions, play an important role in determination of protein structure. Therefore, the prediction of a protein native structure from its amino acids sequence (primary structure) has been given more attention [8]. The need for faster and more cost effective computational methods is critically important. It is one of the most important goals in bioinformatics and theoretical chemistry. The design of drugs and novel enzymes are two important examples of the applications of protein structure prediction in medicine.

Protein structure prediction is still extremely hard to process for some proteins. Two main difficulties are calculation of the good energy function and finding the global minimum of the energy function. The search space of the prediction method for the problem is astronomically large. Cyrus Levinthal stated in "Levinthal"s Paradox" that, due to the large number of degrees of freedom in the primary structure of the protein, the molecule has an astronomical number of possible conformations [11]. For example, if a protein of length 100 residues is sequentially sampled by all the possible conformations (3^{198} different conformations), it would require a time longer than the age of the universe to arrive at its native conformation. The huge search space can be pruned by comparative modeling or *ab initio* modeling. When the target primary structure is assumed to adopt a similar structure of another experimentally determined protein, comparative modeling would narrow the search space and guide the prediction method accordingly. Otherwise,

[1] The information is collected from the website http://www.uniprot.org/uniparc/ as of March 2015
[2] From the website of Protein Data Bank www.pdb.org as of March 2015

ab initio modeling is used to predict the structure from scratch. The accuracy and performance of current prediction methods is assessed by Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment every two years [12-15].

Ab initio modeling is a computational method aimed at predicting and characterizing the structure/function of the protein using the information of primary structure as the only input. Due to the difficulty of the problem and the astronomical size of the search space, most of *ab initio* approaches use knowledge-based and physics-based potentials to guide the protein folding prediction process. The usage of this information is helpful to discover important features regarding secondary structures, distant constraints, and conformational preferences taken from the sequences. The majority of *ab initio* approaches focus on three aspects of this problem. First, suitable protein representation and corresponding protein conformation space in that representation. Second, an accurate energy function that is able to distinguish good conformations from bad ones and is compatible with the representation. Third, an efficient approach that is able to search the conformational search space and minimize the energy term [16]. Numerous sophisticated algorithms such as Monte Carlo, genetic algorithms, and molecular dynamics are used to search the conformational space.

In contrast to *ab initio* modeling, comparative modeling uses previously determined structures as a template. This template modeling seems to be effective because of the limited number of tertiary structures motifs available even though the number of proteins in nature is incredibly large. Many proteins with good sequence similarity have similar

functions and structures; when a query protein shares 30% sequence identity with a protein of known structure, comparative modeling can predict the structure to a fairly good accuracy [17-20]. Most comparative modeling consists of four steps [21]: First, find a good template from previously determined structures in the protein data bank. Second, align query sequence with the template structure. Third, build the structural framework based on alignment by copying aligned regions. Fourth, fill the gaps found on the framework.. The first two steps are performed simultaneously in the threading (or fold recognition) phase [22, 23]. Similarly, the last two steps are also performed simultaneously [16].

Although homology-based comparative modeling is the most successful methods for structure prediction to date [8, 24, 25], identifying the correct template and refining it is still an important condition. The appropriate template in the PDB is a crucial condition for the success of this model otherwise *ab initio* modeling should be used.

2. Cryo-Electron Microscopy

Electron cryo-microscopy (cryo-EM) has become a major experimental technique to study the structures of large protein complexes [26, 27]. It is a structure determination technique complementary to X-ray crystallography and NMR. A number of large molecular complexes, such as ribosome and viruses, have been resolved to near atomic resolutions (2-5Å) [28-31]. Many more have reached medium resolutions (5-10Å) [32, 33]. Resolution in terms of electron density is a measure of the resolvability in the electron density map of a molecule.

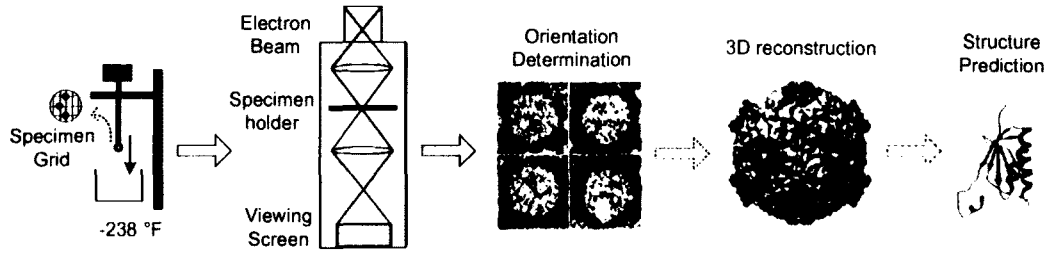


Figure 2. Cryo-EM technique - from freezing the specimen to atomic structure.

Cryo-EM involves a process of freezing the sample in ethane slush to produce specimen"s non-crystalline ice (Figure 2). These frozen specimens studied at extremely low show a structure similar to the native conformation [34]. The advantage of freezing the sample is to view it without any distortions or artifacts such as redistribution of elements or removal of substances and its ability to visualize different functional states [35, 36]. Averaging and processing multiple 2D images (i.e. thousands) lead to relatively good resolution information (between 5 and 15Å) of the 3D object (3D reconstruction). Unfortunately, at such 5-15 Å, atom positions are difficult to interpret directly from the volumetric density map. However, Hong Zhou et al. recently reported an image of a virus structure at a high enough resolution (3.3Å) to see atoms effectively [37]. They used a single-particle cryo-EM to report the structure of a primed, infectious subvirion particle of aquareo virus. The volumetric density map they have generated reveals side-chain densities of all types of amino acids except glycine . It allowed them construct a full-atom model of the viral particle.

Many volumetric density maps of large protein complexes have been generated at low and/or intermediate resolution using cryo-EM technique [37-42]. There are 2858 cryo-

EM experimental density maps that have been deposited to EMDataBank, a Unified Data Resource for 3Dimensional Electron Microscopy[3]. Most of the entries are at intermediate resolution range such as 5-10Å. However, the intermediate resolution density maps are not resolved well enough to determine the atomic information of the protein. Recent works show the ability of volumetric density maps to help in discriminating between models built by *ab initio* and/or comparative modeling and in building final models as well [40, 43-51]. Given an initial structural model obtained by either *ab initio* or comparative modeling, the volumetric density map is used to refine and fit the model structure to generate a high-resolution, all-atom protein model. The Refinement process is done by heuristic methods such as conjugate gradients minimization (CG) and simulated annealing molecular dynamics (MD). A fitting scoring function measures how well the model fits into the volumetric density map to guide structure refinement process and identify mismatch regions between the model and the map [44, 49].

3. Problem of Secondary Structure Elements Extraction from Cryo-EM Density Maps

At medium resolutions, molecular features are not resolved and it is challenging to derive atomic structures from the density maps. In some special situations, particularly for small proteins with mostly α-helices, direct modeling is possible to derive the backbone of a protein [52]. A major approach is to start with a homologous model and to adjust the model through fitting [51, 53-56]. The initial model can be a homologous structure or a model built from a template structure [46, 57, 58]. Fitting methods have evolved from previous rigid fitting to flexible fitting [44, 59-62]. Although fitting a homologous model

[3] From the website of EMDataBank http://www.emdatabank.org/ as of March 2015

has been fairly successful resolving structures from density maps at medium resolutions, it is still challenging to find suitable known structures as templates in many cases.

Ab initio modeling aims to derive atomic structures from electron density maps without a template structure. Although the connection between SSEs, such as α-helices and β-sheets, is ambiguous at medium resolutions, likely connections may still be derived. Given the positions of α-helices and β-strands in a density map, one can match them with secondary structure sequence segments that can be predicted from the amino acid sequence to derive the overall topology of a protein chain [47, 63-67]. Once the topology is determined (Figure 3), backbone and side chains can be constructed and evaluated using energy functions [66, 68, 69].

The location of secondary structures is critical in modeling atomic structure from a density map. Although it is not possible to distinguish the amino acid at medium resolutions, SSEs such as α-helices and β-sheets can be visually and computationally identified using image processing techniques.

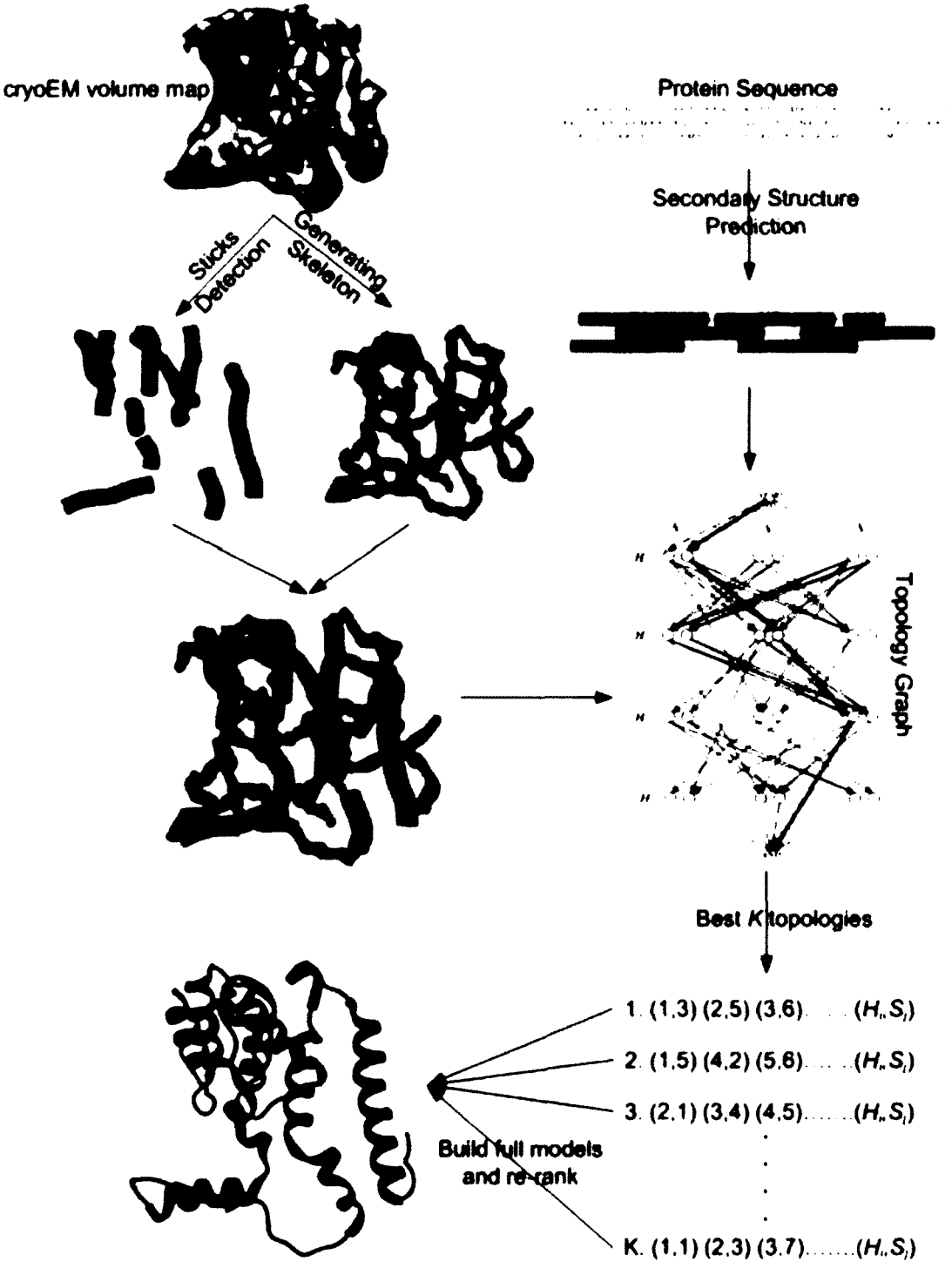


Figure 3. *Ab initio* protein structure prediction from the volumetric density maps.

A. Challenge of Automatic and Efficient SSE Detection

It was first demonstrated using *HelixHunter* that α-helices can be computationally detected from a density map at sub-nano resolution [70]. After that, a number of approaches have been developed to detect the α-helices from the medium resolution electron density maps [63, 71-75]. A few approaches have also been developed to detect the β-sheets [63, 72, 74, 76]. Most of the computational approaches use automatic detection, while a few of them are semi-automatic guided by user interpretation [63].

Although multiple methods have been developed to detect SSE from the density maps, accurate detection either needs user intervention or the careful adjustment of various parameters. It is still challenging to detect the SSE automatically and accurately from cryo-EM density maps at medium resolutions (~5-10Å). A detected β-sheet can be represented by either the voxels of the β-sheet density or by many piece-wise polygons to compose a rough surface. However, none of these is effective in capturing the global surface feature of the β-sheet.

Two computational methods, *SSEtracer* and *SSElearner*, are used for solving this critical problem. An effective approach, *SSEtracer*, is presented to automatically identify helices and β-sheets from the cryo-EM three-dimensional (3D) maps at medium resolutions. A simple mathematical model is introduced to represent the β-sheet density. The mathematical model can be used for β-strand detection from medium resolution density maps. A machine learning approach, *SSElearner*, has also been developed to

automatically identify helices and β-sheets by using the knowledge from existing volumetric maps in the Electron Microscopy Data Bank (EMDB).

B. Challenge of β-strand Detection

A β-sheet contains multiple β-strands. Although β-sheets can be identified from cryo-EM density maps at 5-10Å resolutions, it is almost impossible to detect the β-strands of a β-sheet. The spacing between two neighboring β-strands is between 4.5 and 5Å, and β-strands are only visible when the resolution is higher than 4.7Å [77, 78]. Without knowing the location of β-strands, the representation of a protein is purely dependent on the relative location of helices [73]. *De novo* modeling has been successful in deriving the backbone from the density map of GroEL (4.2 Å resolution) [79] and gp10 (4.5 Å resolution) [80]. However, there has not been an α/β structure that is resolved using *ab initio* modeling from a density map at a medium resolution. One of the challenges is the inability of detecting β-strands from the density maps.

Two computational methods - *StrandTwister* and *StrandRoller* for solving this challenging problem. A new method, *StrandTwister*, has been proposed to detect the traces of β-strands through the analysis of twist, an intrinsic nature of β-sheet. β-barrel is a structure feature that is formed by multiple β-strands in a barrel shape. There is no existing method to derive the β-strands from the 3D image of β-barrel. A new method, *StrandRoller*, has been proposed to generate small sets of possible β-traces from the density images at medium resolutions of 5-10Å.

CHAPTER II

TRACING THE SECONDARY STRUCTURE FROM CRYO-EM DENSITY MAPS

Secondary structure elements (SSEs) refer to the density elements corresponding to α-helices and β-sheets of the protein. At the medium resolutions, an α-helix appears as a cylindrical stick and a β-sheet appears as a thin layer of density that is often twisted. The identification of SSEs from volumetric maps is critical for *ab initio* backbone structure derivation from cryo-EM maps. Many methods have been developed to identify the SSEs at medium resolutions [70] [72, 73, 81-85] [86, 87]. Among which more identify α-helices and less identify β-sheets [72, 76, 83, 86, 87]. Most of the computational approaches use automatic detection, while a few of them are semi-automatic guided by user interpretation [72, 83]. Previous automatic methods usually need multiple user-defined parameters which make them hard to use. In general, α-helices are easier to be detected than β-sheets. In fact, the first method of SSE detection from low resolution density maps detected only helices [70]. β-sheets in medium resolution density maps usually do not adopt a single characteristic shape like the cylindrical shape of α-helices which make them much more difficult to identify.

This chapter is a summary of the *SSEtracer* methodology published in paper [88].

1. Motivation

By now, it is known that β-sheets are not flat as first proposed by Pauling in 1951 [89]. β-sheets in proteins are almost always twisted. This conspicuous feature of β-sheet has been recognized after several β-sheets had been seen in three-dimensional protein structures. The right-handed twist of β-sheet was first described by Cyrus Chothia in 1973 [90]. After that, Salemme suggested in 1981 that the spatial configuration of β-sheets is isotropically stressed surface [91]. Some methods have been proposed to show that the β-sheet atomic structures can be modeled as different types of 3D surface [92-94].

In addition to the need for accurate detection of β-sheets, accurate detection of β-strands from a β-sheet is needed for modeling the atomic structure. There has not been an effective method for β-strands detection from the cryo-EM map at the medium resolutions. Due to the closeness of β-strands, cryo-EM density of β-sheets that at such resolution range almost has no indication of single β-strand with any threshold. In this chapter, a simple and effective method is presented to detect both α-helices and β-sheets from such cryo-EM maps. More importantly, the first method to represent the detected β-sheet using a mathematical model is developed. In order to derive β-strands from the β-sheet density, it is important to have a mathematical model that accurately captures the overall surface pattern of the β-sheet density. The details of how the mathematical model assists the detection of β-strands from β-sheet have been included in a separate chapter – chapter IV. The focus of this chapter is to demonstrate that it is possible to use a simple mathematical model to represent β-sheet density voxels detected from cryo-EM maps.

2. Methodology

Based on the local shape characteristics of α-helices and β-sheets, *SSEtracer* performs a series of local feature analysis to detect the SSEs. The detected helix voxels are used to generate a spline to represent the helix central axis. The detected β-sheet voxels are used to generate a mathematical polynomial model to represent the overall surface of β-sheet (Figure 4).

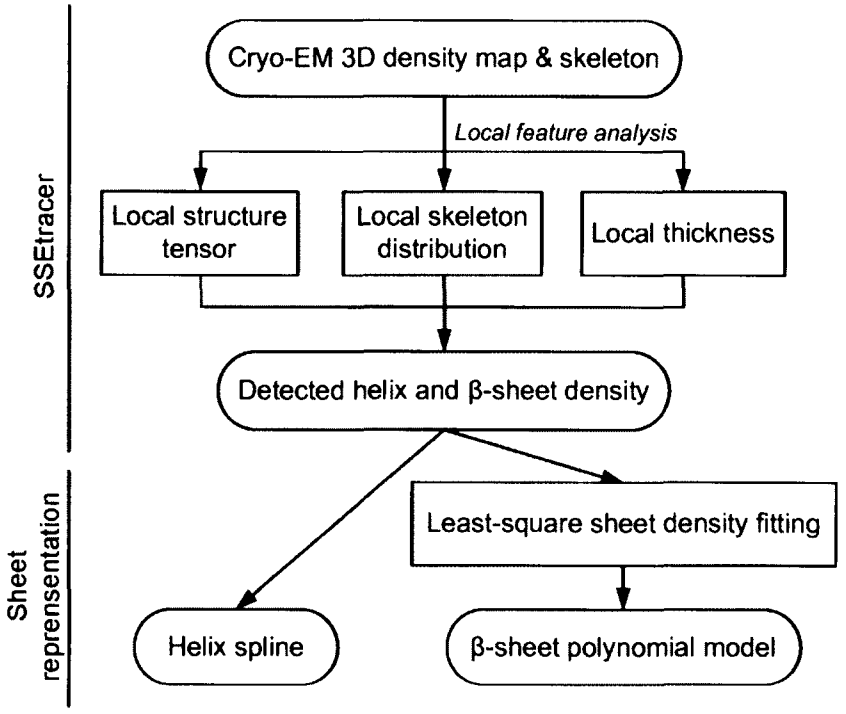


Figure 4. Flowchart of *SSEtracer* and β-sheet representation.

C. Local Feature Analysis

SSEtracer takes an iso-surface threshold as the only user input parameter, and automatically detects the location of α-helices and β-sheets. It is designed to be a tool that is both effective and easy to use. All the calculations in this local feature analysis step are based on the iso-surface threshold for the density map. The iso-surface threshold can be obtained from the EMDB database [32] for experimentally derived cryo-EM maps.

Skeletonization is a powerful method to extract the descriptive structural information from the density maps [83, 87, 95]. The skeleton is a set of grid points, or voxels. It refers to a medial, geometric representation that approximates the overall shape and connects topology on the map. The skeleton can be extracted by using *Gorgon*, which is a GUI and semi-automatic tool for skeletonization [83]. The skeleton used in this chapter was generated by using *Gorgon*, because of the ability of *Gorgon* on building the surface skeleton and generating clearer skeleton with less redundancy.

The skeleton density voxels was firstly grouped into local clusters based on a distance cutoff, which is equals to the spacing of skeleton density map times 1.732. The centers of these voxel clusters were used to speed up the processing, instead of working on each single voxel of the original density map. The sparseness of cluster center points along a skeleton can be used to describe the local geometric shape of the density map. Three local structure features: local structure tensor, local skeleton distribution and local thickness are calculated at each cluster center in the local shape analysis step.

Local structure tensor

Local gradient is often used to characterize the geometrical features in volumetric density maps [70, 81, 82]. The local structure tensor has been applied to describe the local shape [82, 86, 96].

Let $I(x, y, z)$ denote the density at voxel (x, y, z). The local structure tensor is a symmetric positive semi-definite matrix given by:

$$K_\alpha * \begin{bmatrix} I_x^2 & I_x I_y & I_x I_z \\ I_x I_y & I_y^2 & I_y I_z \\ I_x I_z & I_y I_z & I_z^2 \end{bmatrix}$$

where I_x, I_y, and I_z are the derivatives (or gradient) along x, y and z direction respectively. The symbol "*" stands for component wise convolution, and K_α is a Gaussian convolution kernel, with standard deviation α over which the local structure is averaged. The orthogonal eigenvectors of the structure tensor v_1, v_2, v_3 provide the preferred local orientations. The corresponding eigenvalues λ_1, λ_2, λ_3 $(\lambda_1 \geq \lambda_2 \geq \lambda_3)$ provide the average contrast along these directions. The eigenvalues and eigenvectors can be calculated by using Jacobi eigenvalue algorithm [97]. The first eigenvector v_1 represents the direction with the maximum variance of the density, whereas v_3 represents the direction with the minimum variance. The three eigenvalues could therefore be used, based on their relative eigenvectors, to describe the local density nature in three classes: cylinder-like, plane-like or isotropic structure:

- Cylinder-like structures: $\lambda_1 \approx \lambda_2 \gg \lambda_3$
- Plane-like structures: $\lambda_1 \gg \lambda_2 \approx \lambda_3$
- Isotropic structures: $\lambda_1 \approx \lambda_2 \approx \lambda_3$

Based on the above local structure measurements, two ratios of the eigenvalues λ_1/λ_2 and λ_2/λ_3 are calculated at each cluster point.

Local skeleton distribution

The sparseness of cluster points along a skeleton can be used to describe the local geometry shape of the density map. Two shape descriptors are calculated to capture the local geometric shape of the density map: (1) Number of neighbors and (2) The local distribution angle. A local distribution angle is formed by a cluster point and its two neighbors. The cluster points are considered neighbors if their Euclidian distance is less than a threshold. The cluster point is predicted to be located on the surface-like region if its number of neighbors is more than two. And also the smallest angle among all local distribution angles for this cluster point is smaller than 90 degree. Otherwise the cluster point is predicted to be on the curve-like region (Figure 5B and C).

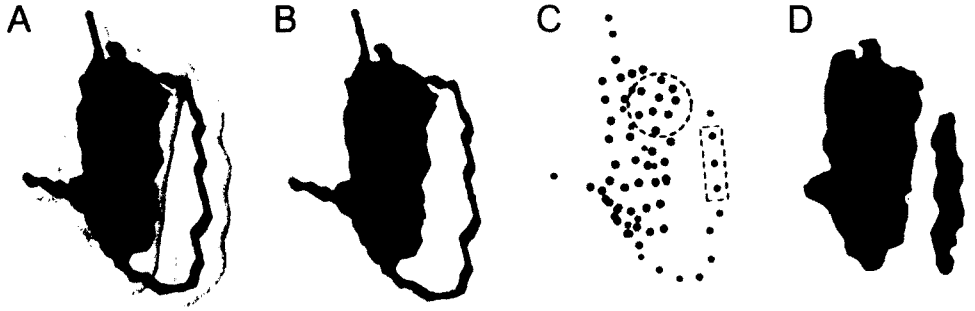


Figure 5. Data representation in *SSEtracer*. (A) The simulated density map of protein 2AW0 at 8Å resolution (gray) and the skeleton (green); (B) the cluster points (dark balls) generated from the density skeleton; (C) the curve-like cluster points (red balls) and surface-like cluster points (blue balls) in the skeleton distribution calculation of local shape analysis step; (D) the detected helix (red) and β-sheet (blue) density area.

Local thickness

The thickness at each cluster point was calculated using volume-based estimation [98, 99]. The method does not depend on the assumption of the structural type, thus it is suitable to assess the thickness distribution of any object. The local thickness $\tau(\underline{p})$ is defined as the diameter of the largest sphere that fits completely inside the density map and contains the cluster point ($\underline{p}$):

$$\tau(\underline{p}) = 2 * \max\left(\left\{r \middle| \underline{p} \in sph(\underline{x}, r) \subseteq \Omega, \underline{x} \in \Omega\right\}\right) \quad (1)$$

Where $sph(\underline{x}, r)$ is the set of voxels inside a sphere with center $\underline{x}$ and radius r, as shown in Figure 6.

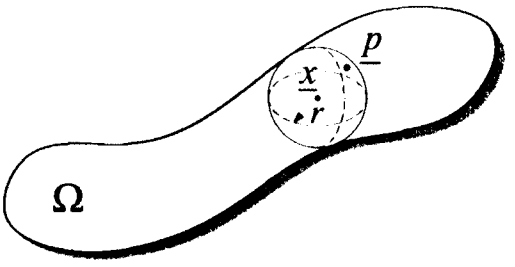


Figure 6. Local thickness of an object **Ω** determined by finding maximal sphere to the object.

One of the most characteristic features of β-sheet density is the relative small thickness. To measure the thickness, *sheetminer* uses a template searching scheme which is computationally expensive [76]. The measurement of local thickness was introduced into *SSEtracer*. The general thickness definition for arbitrary structures allowing *SSEtracer* to calculate the mean structure thickness and the thickness distribution of 3-D objects in a direct way and independently of an assumed structure model [98, 99]. Since the β-sheet density region is usually thinner than the helix density region at medium resolution density maps. The efficient implementation of the local thickness method was used to help on distinguishing between the β-sheet and helix density region.

D. Secondary Structure Voting

The three local features were used to conduct a simple voting scheme to determine if a particular cluster point belongs to a helix or a β-sheet.

Two ratios of the eigenvalues were compared for the local structure tensor feature. The helix vote of a cluster point was increased by 1 if $\lambda_1/\lambda_2 < \lambda_2/\lambda_3$; otherwise the β-sheet

vote was increased by 1. For the local skeleton distribution feature, the helix vote was increased by 1 if the cluster point is detected to be on the curve-like region (Figure 5C, red balls); while the β-sheet vote was increased by 1 if it is detected to be on the surface-like region (Figure 5C, blue balls). For the local thickness feature, the overall average thickness on the density map for curve-like region and surface-like region was first calculated respectively. The helix vote was increased by 1 if the local thickness at certain cluster point is within a range from the average thickness of curve-like region; the β-sheet vote was increased by 1 if the local thickness at certain cluster point is within a range from the average thickness of surface-like region.

The total SSE vote at a particular cluster point was then summed up. The cluster point is finally detected to be on helix area if the vote for helix features is greater or equal to 2, and it is detected to be on β-sheet area if the vote for the β-sheet features is greater or equal to 2. The maximum votes a cluster point can get is 3.

Finally, the original density voxels around the cluster points were retrieved and grouped by using the pre-determined parameter - distance cutoff that mentioned before. Any voxels that within this distance cutoff were brought back. The size of the voxel group was then estimated by the number of voxels and the maximum distance within the group. The small or short voxel groups will be filtered out. For example, helix that detected as shorter than 3Å will be discarded. The generated voxel groups will be kept as detected helix and sheet density area. The detected helix can be simply represented by an

interpolated spline that within the detected helix voxels, the spline is often close to the central axis of the helix.

E. Mathematical Models for β-sheet

Many studies have shown that a variety of saddle shaped surfaces can be used to model β-sheets in atomic structures. β-barrels has been modeled as highly twisted hyperboloid surfaces [92]:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 \qquad (2)$$

Helicoids have been used to fit small β-sheets using the principle of minimal surfaces [93]. Additional models involve catenoid for β-barrels and β-sandwiches [94].

Rather than using different forms for different types of surface, a more general model was proposed for polynomial surface [100]. Although the order-two polynomial surface can already describe some surface pattern for β-sheets, it is sometimes not good enough to capture the flexibility and curvature for highly twisted β-sheets. Higher order polynomials (order-four or even higher) may exaggerate minor fluctuations in the density data. *SSEtracer* uses order-three 3D polynomial surface (Formula 3) to represent the β-sheet surface.

$$z = Ax^3 + By^3 + Cx^2 + Dy^2 + Ex^2y + Fy^2x + Gxy + Hx + Iy + J \qquad (3)$$

Since Formula (3) is a function that maps coordinate x and y to coordinate z, the 3D surface can be best fitted when the β-sheet density area is approximately parallel to X-Y plane and the normal vector of β-sheet density is along the Z direction. Due to the folded shape of β-sheet, the geometry center of β-sheet density may not be on the density itself. Some scattered cluster points were first searched from the density voxels based on a distance cutoff 5Å, and defined the sheet center as the closest voxel to the density geometry center. The three cluster points that are closest to the sheet center were picked to build a center plane for finding the rough normal vector of the β-sheet density (Figure 7A). The β-sheet density was then rotated so that the normal vector of sheet density is aligned with the Z direction (Figure 7A). The β-sheet density was then fitted with the polynomial surface model (Formula 3) using least-square method, as shown in Figure 7B. The (x, y, z) in this formula is the voxel coordinate of the β-sheet density. All the ten coefficients in this formula can be optimized using least-square fitting method. Finally, the β-sheet density was rotated back after the modeling has done.

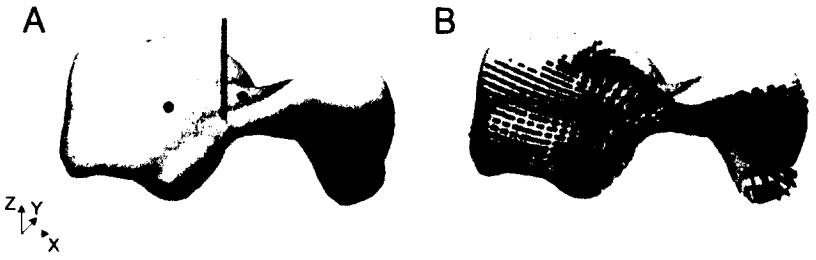


Figure 7. Fitted polynomial surface to the β-sheet density. (A) The center plane that decided by three cluster points (blue balls) with its normal vector (red line); (B) fitted 3D surface model (yellow surface points were generated by the model).

3. Result

The performance of *SSEtracer* was tested on ten simulated density maps and five experimental derived cryo-EM density maps from EMDB. An identified helix is defined as if its length is within one turn difference from the length of helix in the PDB structure which is measured by the central axis of the helix. A identified β-sheet is defined as if the detected β-sheet area visually overlays on the β-sheet of the PDB structure [72]. Although not included in this chapter, alternative tests could also be conducted if the detected SSE location is compared using the Cα atom [86].

SSEtracer was tested using simulated density maps of the representative structures from the top 10 most commonly occurring folds [101], which were generated to 8Å resolution using the program *pdb2mrc* of EMAN [102] with a sampling size of 1 Å/pixel. The 10 proteins were used for testing *SSEhunter* at the same resolution [72]. Our method successfully identified 73 of the 74 helices that have more than four amino acids (Table 1). Most of the missed helices have 3 amino acids in length, presumably of the 3_{10} helices. *SSEtracer* detected 14 of the 17 β-sheets. All 3 missed β-sheets have only two strands.

Compared to the semi-automatic method *SSEhunter* (Table 1), our fully automatic *SSEtracer* appears to be slightly better on detecting the short helices (<5 amino acids and 5-8 amino acids) and 2-stranded β-sheets. Since *SSEhunter* is a semi-automatic method, it requires user intervention and careful adjustment of various parameters. The comparison of the performance of *SSEtracer* with the performance of *SSEhunter* is based on the latest

public result from *SSEhunter* paper [72]. *SSEtracer* is a fully automatic method which does not require user intervention. In this dataset, it outperforms on short helix (<5 amino acids and 5-8 amino acids), and 2-stranded β-sheets (Table 1).

Table 1. The number of secondary structure identified by *SSEtracer* on 10 most commonly occurring folds, compared with *SSEhunter* [72].

PDB ID	Our *SSEtracer*					*SSEhunter**		
	Hlx <5aa	Hlx 5–8aa	Hlx >8aa	Sht= 2strd	Sht> 2strd	Hlx <5aa	Hlx 5–8aa	Sht= 2strd
1AJW	0/0	1/1	0/0	0/0	2/2	0/0	1/1	0/0
1AJZ	0/1	3/3	7/7	0/1	1/1	0/1	3/3	0/1
1AL7	1/3	4/4	10/10	0/2	1/1	1/3	4/4	0/2
1CV1	1/1	1/2	8/8	0/0	1/1	1/1	0/2	0/0
1DAI	2/2	2/2	5/5	2/2	1/1	2/2	2/2	0/2
1ENY	0/0	1/1	9/9	0/0	1/1	0/0	1/1	0/0
1WAB	2/3	0/0	6/6	0/0	1/1	1/3	0/0	0/0
2AW0	0/0	0/0	2/2	0/0	1/1	0/0	0/0	0/0
2ITG	0/0	1/1	5/5	0/0	1/1	0/0	1/1	0/0
3LCK	1/4	2/2	6/6	1/1	1/1	1/4	0/2	1/1
Totals	7/14	15/16	58/58	3/6	11/11	6/14	12/16	1/6

*As a comparison, the columns for *SSEhunter* can be found in the supplementary table 1 of *SSEhunter* paper [72].

In addition to the simulated maps, the performance of *SSEtracer* was also tested using five experimental derived cryo-EM density maps that were downloaded from the EMDB database. The "recommended contour level" from EMDB was used as the iso-surface threshold for the input parameter. The test of five cryo-EM maps suggests that the helices longer than eight amino acids and the β-sheets with more than two strands can be detected well. *SSEtracer* detected 19 of 20 such helices and all 11 such β-sheets (Table 2).

Table 2. The number of secondary structure identified by *SSEtracer* on the experimental derived cryo-EM maps.

EMD_PDB ID, Resolution	Hlx <5aa	Hlx 5–8aa	Hlx >8aa	Sht= 2strd	Sht> 2strd
1237_2GSY_A, 7.2Å	1/3	3/5	3/3	1/1	5/5
1733_3C91_H, 6.8Å	0/0	0/0	5/5	0/1	2/2
1740_3C92_A, 6.8Å	0/1	0/0	5/6	0/0	2/2
1780_3IZ6_K, 5.5Å	0/0	0/1	2/2	0/0	1/1
5030_3FIN_R, 6.4Å	0/0	0/0	4/4	0/0	1/1
Totals	1/4	3/6	19/20	1/2	11/11

For helices with no more than eight amino acids, *SSEtracer* was only able to detect 4 of 10 such helices (Table 2). Our test using five experimentally derived cryo-EM maps shows the challenges in small helices. A variety of possible errors could be introduced from the experimentally derived density maps. As shown in Figure 8, the missing density for the short helix makes the detection of that helix very difficult.

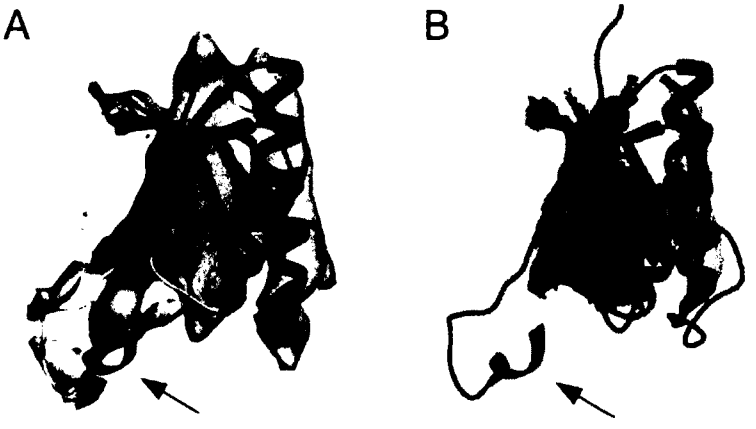


Figure 8. Detected SSEs from experimental derived cryo-EM map by *SSEtracer*. (A) EMDB entry EMD-1780 at resolution 6.4Å, corresponding true structure (chain K of protein 3IZ6) shown as colored ribbon; (B) identified and modeled helices and β-sheets.

The experimentally derived cryo-EM density maps often contain noises and bumps at the edge area of β-sheet, due to the closeness to parts of other structures such as loops or turns (Figure 8B). The polynomial model of β-sheet represents the overall twisted surface pattern (Figure 9, right column). The surface points shown in Figure 8 and 9 are generated by this polynomial model. As an example, the ten coefficients that calculated for the polynomial surface model (Formula 3) of β-sheet shown in Figure 9A are listed: $A = 0.0013, B = 0.0022, C = 0.0017, D = -0.0004, E = 0.0731, F = -0.0084, G = 0.0699, H = 3.2652, I = -1.8834, J = 1.9414$. In this model, each parameter (A to J) can be associated with a feature of the 3D surface. For example, the combinations of parameter A to G produce the complex of surfaces. The parameter H and I simply tilt the surface in x and y respectively, and J sets the base level. As shown in Figure 8 and 9, the polynomial surface model visually fits in the detected β-sheet cryo-EM density area well and represents the 3D surface feature of the β-sheets. It is a simplified representation over the density voxel representation and other piecewise polygon representation [72]. Furthermore, the mathematical model can be used to represent the twist of β-sheet, which is an important feature of the β-sheet structure.

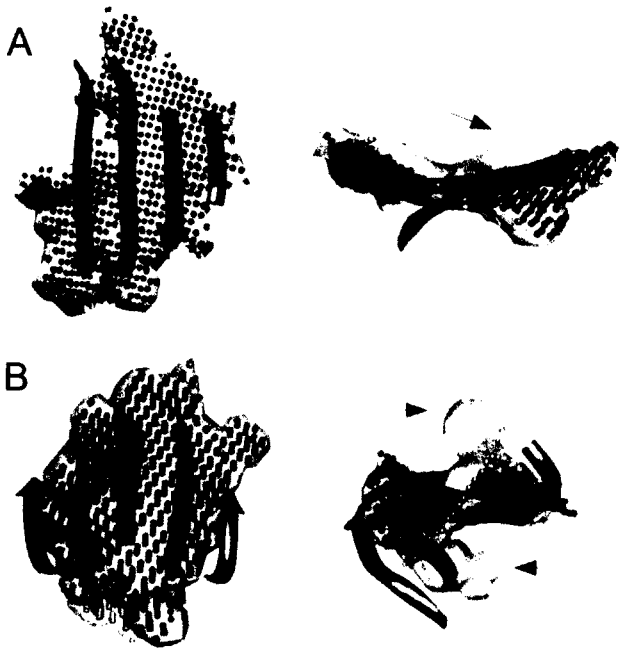


Figure 9. Polynomial model that fits in the β-sheet density. (A) Generated points (yellow) by the polynomial model to show the 3D surface, superimposed on the true structure (cyan ribbon, sheet A of protein 3C92) and the detected sheet density (gray, EMDB entry 1740); (B) sheet W of protein 3IZ6 and the detected sheet density from EMDB entry 1780.

To further quantify the performance of the polynomial surface fitting method, the fitting error by measuring the vertical offsets from the modeled surface to the density voxels was calculated. The root-mean-square-error (RMSE) was used to represent the overall error of polynomial fitting, which is similar to the previous measurements that were used for β-sheet atomic structure fitting [92, 94]. The error being minimized in the previous method is the sum of the squared distance between the center of mass of each peptide bond (reference point) in a β-strand and the intersection of the catenoid surface with a line normal to the z axis and passing through the reference point. The error being

minimized in our polynomial surface fitting is the distance between the fitted surface and the density voxels.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \overline{z}_i)^2} \quad (4)$$

Where N is the total number of density voxels, z_i is the Z coordinate of i-th density voxel and $\overline{z}_i$ is the Z coordinate of its corresponding fitted surface point.

Table 3. Polynomial fitting error for the β-sheets in cryo-EM density maps.

EMD_Sheet ID	# Strands	RMSE(Å)
1237_2GSY_A	4	2.19
1237_2GSY_B	5	2.30
1237_2GSY_C	6	2.37
1237_2GSY_E	4	1.80
1237_2GSY_G	5	2.16
1733_3C91_O	5	1.66
1733_3C91_Q	5	1.72
1740_3C92_A	5	1.29
1740_3C92_B	5	1.62
1780_3IZ6_W	5	2.27
5030_3FIN_AE	3	1.31
Average		**1.88**

Table 3 shows the 3D surface fitting result for eleven β-sheets that were identified by *SSEtracer*. The eleven β-sheet density maps are experimentally derived cryo-EM density maps with resolution between 5.5Å and 7.2Å. Note that the fitting procedure in our method is based on the 3D cryo-EM density voxels instead of the true atoms of PDB structures [92, 94]. The fitting error is related to the threshold of density maps. The "recommended contour level" from EMDB was used as the iso-surface threshold. Most

of the errors are from the bumps on the edge of β-sheets. It is known that the van der Waals radius of common atoms is between 1Å and 2Å. Considering the electron density of protein backbone and side-chain at medium resolutions, the average fitting error 1.88Å in Table 3 is fairly small. It shows the accuracy of our β-sheet 3D surface modeling method on cryo-EM density maps.

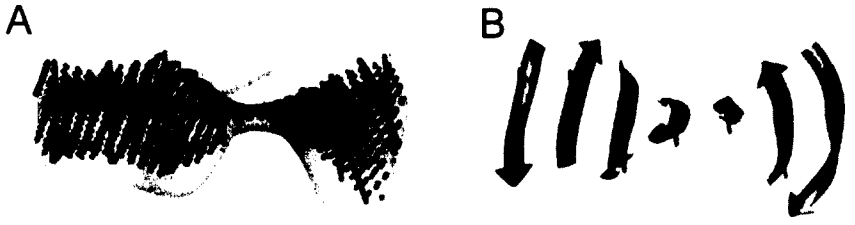


Figure 10. Detected β-strands using the mathematical model of β-sheet. (A) 3D polynomial β-sheet surface model (yellow) and the possible β-strand samples (blue and red curve) that on the modeled surface; (B) best detected β-strand position (red curve) that superimposed on the true structure (cyan ribbon).

One of the significant contributions of the β-sheet mathematical model is for identifying β-strands. This is due to the simplicity of the model yet capturing the overall curvature of the β-sheet. Figure 10B shows an example of the detected β-traces (red curve) based on the points generated from the mathematical model. The detected β-trace aligns well with the true β-strands (blue ribbon). The details of our β-strand detection method are included in a separated chapter.

As expected, accurate detection of β-strands depends on accurate identification of a β-sheet. The boundary of the identified β-sheet may affect β-strand detection. In most cases, the inaccurate boundary can result in a longer/shorter detected β-strand, or

extra/missing β-strand. Intuitively, the wrongly detected β-sheet density (edge areas in Figure 9) may affect the curvature/size of the β-sheet and the β-strand modeling.

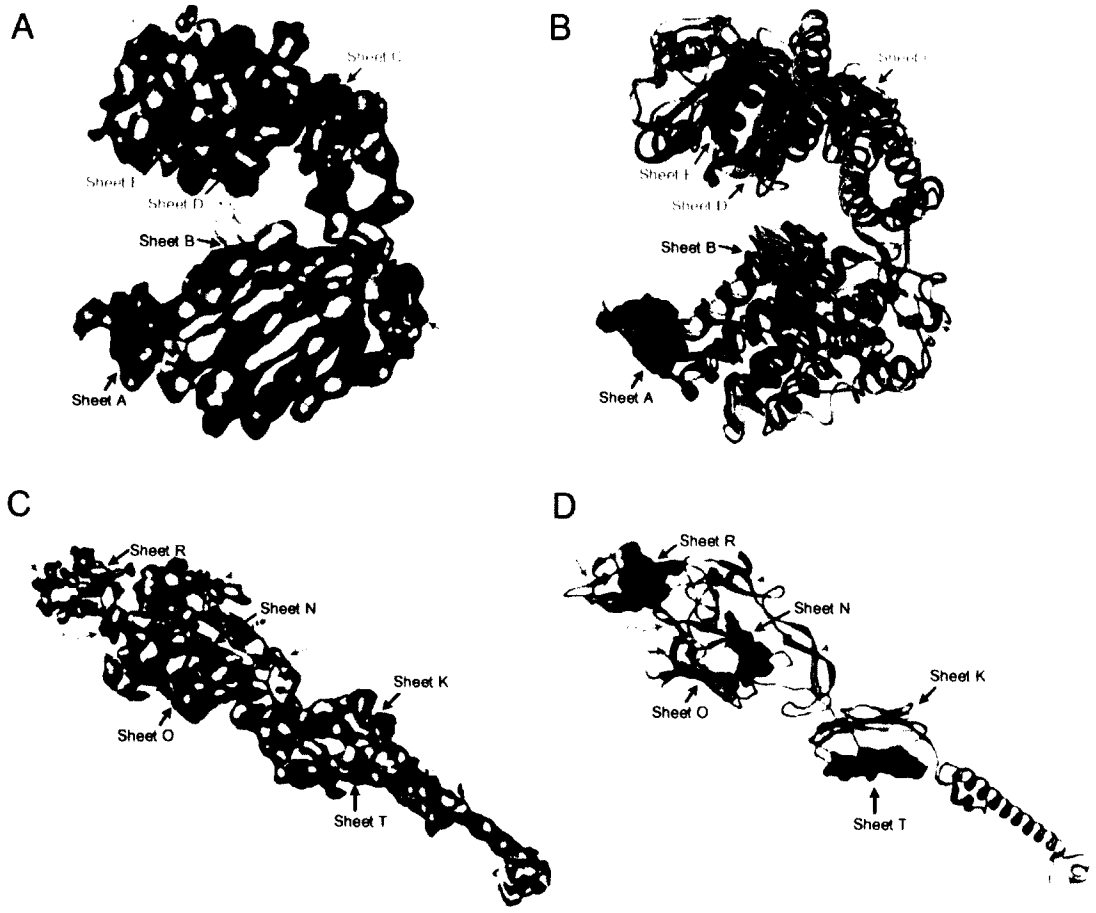


Figure 11. Challenge of β-sheet detection from the cryo-EM density map at medium resolutions. (A) The monomer density (gray) of GroEL extracted from density map EMD_5001; (B) Five β-sheet density regions (colored density) identified using *SSEtracer* are superimposed on chain A of PDB_3CAU (purple Cα trace) and chain A of PDB_1SS8 (cyan ribbon); (C) E2 monomer density (gray) in Encephalitis Virus (EMD_5276) at 4.4Å resolution; (D) β-sheet density regions (colored density) identified using *SSEtracer* are superimposed on chain B of PDB_3J0C (cyan ribbon).

In addition to short helix (<5 amino acids and 5-8 amino acids) that shown in Table 1 – 2 and Figure 8, small β-sheets are also quite challenging and hard to detect. Current version of *SSEtracer* detects the secondary structures based on the density features of SSEs at medium resolutions. The detection could fail if there is a missing density, wrong density or inaccurate density (as example shown in Figure 8 and Figure 11). *SSEtracer* detected five of the seven sheets from the density monomer of EMD_5001 (Figure 11A). Two 2-stranded β-sheets (F and G) were missed due to the fact that a 2-stranded sheet can be confused with a helix (Figure 11B, pointed by orange arrows). The structure of Venezuelan equine encephalitis virus (VEEV) was resolved from the 4.4Å resolution cryo-EM density map (EMD_5276). The monomer of E2 which aligned with chain B of 3J0C was isolated from the density map. *SSEtracer* detected five larger β-sheets (N, K, O, T and R) (Figure 11D). Three 2-stranded β-sheets and two 3-stranded β-sheets were missed. Sheet Q (3-stranded) is mostly a 2-stranded twist and appears as a helix in the density. Sheet S (3-stranded) is located at the outer surface of E2 (Figure 11C) where the density is weak and has no obvious sheet property.

Current version of *SSEtracer* takes density skeleton as an input. The quality of skeleton generated from *Gorgon* also depends on the quality of density maps. The quality of skeleton would affect the SSE detection result. In order to build a clear and accurate skeleton from *Gorgon* with both surface for the β-sheet region and linear curves for the helix/loop regions, careful adjustment of the parameters (such as threshold, step count, minimum curve/surface length, curve/surface radius, skeleton radius, and etc.) would be needed.

CHAPTER III

A MACHINE LEARNING APPROACH FOR THE DETECTION OF SECONDARY STRUCTURE FROM CRYO-EM MAPS

The current secondary structure detection methods are mostly based on image-processing techniques, these methods search for cylinder-like regions for helices and plane-like regions for β-sheets [63, 70-74, 76]. Although such methods can recognize most of the helices and β-sheets, they face difficulties in recognizing the border-line cases. These methods do not have the capability of using existing data to assist with the detection. As more and more protein backbones are derived for the cryo-EM maps, learning from the existing data is more and more important. It has been suggested recently that machine learning improves the helix detection in RENNSH [75]. RENNSH method uses the nested k Nearest Neighbors (kNN) classifiers in machine learning only for the detection of α-helices. It uses the training data and the test data from different proteins of the same cryo-EM map. However, when the true PDB structures of same cryo-EM map are not available for training, data from different maps should be used for machine learning. In this chapter, it will be demonstrated that the training process and the test process can use different cryo-EM maps in EMDB. Our *SSElearner* detects both helices and β-sheets through the supervised learning from the cryo-EM density map that is estimated to have a similar nature to that in the target cryo-EM map.

This chapter is a summary of the *SSElearner* methodology published in paper [86].

1. Motivation

Although a number of methods have been developed to identify the SSEs, it is still challenging to identify them automatically and accurately. In general, the long α-helices and large β-sheets can be detected more accurately. Small helices appear to be similar to turns in the density maps at the medium resolution and they are hard to distinguish. A β-sheet with two strands can be confused with a helix. Ideally, the detection methods should be tested using a large number of experimentally derived cryo-EM density maps for which the backbone structures are known. However, due to the lack of such paired data, the current detection methods were predominantly tested using simulated volumetric density maps. Without a test of a large number of cryo-EM maps, the effectiveness of the current methods is still not clear when the experimentally derived cryo-EM maps are presented.

2. Methodology

There are three major components in our method (Figure 12). The first component develops the features using image processing concepts. The second component performs the multi-task classification using Support Vector Machine (SVM). The post-processing step performs additional filtering and clustering based on the relationships among the classified voxels.

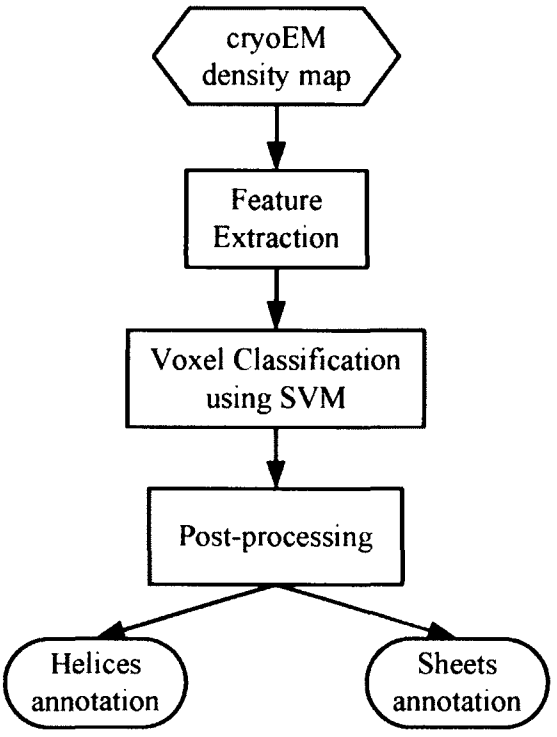


Figure 12. The flowchart of *SSElearner*.

A. Preprocessing and Data Preparation

The performance of *SSElearner* has been tested on ten simulated density maps and thirteen experimental cryo-EM density maps from EMDB. The selected EMDB density maps are between 3.8Å and 9Å resolution. Two types of evaluation were performed. One measures the number of identified secondary structures [70-72] and the other measures the number of Cα atoms [70, 76] that falls in the neighborhood of the secondary structures. A helix is identified if its length is within one turn difference from the length of the helix in the PDB structure. A β-sheet is identified if the identified β-sheet voxels visually overlay on the β-sheet of the PDB structure. In order to present a more

quantitative estimation about the size of the identified helices and β-sheets, the number of Cα atoms that are close to the identified helix voxels and β-sheet voxels was estimated. In particular, a Cα is considered as an identified helix Cα, if it is within 2.5Å distance from an identified helix voxel. A Cα is considered as an identified sheet Cα, if it is within 3Å distance from an identified sheet voxel. The definition of the secondary structures was based on the PDB file that is the authors'' annotation of the protein structure. Note that the authors'' annotation in the PDB file may be slightly different from the annotation using DSSP [103]. Although the definition of a helix and a β-sheet is clear in almost all the PDB files in our tests, it is necessary to visually decide the number and length in rare cases. For example, there is an overlap in the annotated helices with amino acid index 92-107 and 106-111 of 1CV1 (PDB ID). Three strands with amino acid index 37-48, 362-375, 96-110 of 2GSY were annotated in two β-sheets.

SSElearner has been tested using ten simulated density maps that were generated to 8Å resolution using the program pdb2mrc of EMAN [104] with a sampling size of 1Å/pixel. The ten proteins were used for testing *SSEhunter* at the same resolution [72]. The training dataset contains four other proteins (PDB ID: 1C3W, 1IRK, 1TIM and 2BTV) previously used for testing *SSEhunter* [72].

Although it is essential to test the SSE detection methods using experimentally derived cryo-EM maps, it has been challenging to collect a large number of such data. Fourteen cryo-EM maps have been collected from the EMDB with resolutions between 3.8Å and 9Å, out of which thirteen were used to test *SSElearner* (Table 4). Four of the thirteen

cryo-EM maps were selected from the cryo-EM Modeling Challenge 2010 (http://ncmi.bcm.edu/challenge). They are EMD-5030 (3FIN_chain R), EMD-5030 (3FIN_chain F), EMD-5140 (3IYF_chain A) and EMD-5001 (3CAU_chain A).

B. Geometric Processing and Machine Learning

Geometric Feature Extraction

The feature extraction step characterizes each voxel based on its local geometrical features. Local gradient is often used to characterize the geometrical features in volumetric density maps [70, 71, 74]. Local structure tensor is applied to describe the local shape [74, 105].

Let $I(x, y, z)$ denote the density at voxel (x, y, z). The local structure tensor is a symmetric positive semi-definite matrix given by:

$$K_\alpha * \begin{bmatrix} I_x^2 & I_x I_y & I_x I_z \\ I_x I_y & I_y^2 & I_y I_z \\ I_x I_z & I_y I_z & I_z^2 \end{bmatrix}$$

where I_x, I_y, and I_z are the derivatives (or gradient) along x, y and z direction respectively. The symbol "*" stands for component wise convolution, and K_α is a Gaussian convolution kernel, with standard deviation α over which the local structure is averaged. The orthogonal eigenvectors of the structure tensor v_1, v_2, v_3 provide the preferred local orientations. The corresponding eigenvalues λ_1, λ_2, λ_3 $(\lambda_1 \geq \lambda_2 \geq \lambda_3)$

provide the average contrast along these directions. The first eigenvector v_1 represents the direction with the maximum variance of the density, whereas v_3 represents the direction with the minimum variance. The three eigenvalues could therefore be used, based on their relative eigenvectors, to describe the local density nature in three classes: cylinder-like, plane-like or isotropic structure:

- Cylinder-like structures: $\lambda_1 \approx \lambda_2 \gg \lambda_3$
- Plane-like structures: $\lambda_1 \gg \lambda_2 \approx \lambda_3$
- Isotropic structures: $\lambda_1 \approx \lambda_2 \approx \lambda_3$

Instead of using the three eigenvalues of the structure tensor to distinguish different local structures, Yu and Bajaj proposed a practical parameter – thickness [74]. The thickness that applied here is defined by the width of the region above a pre-chosen threshold along the eigenvector. Let t_1, t_2, t_3 be the thicknesses along direction v_1, v_2, v_3. The typical thicknesses for different local structures have the following criteria:

- Cylinder-like structures: $t_1 \approx t_2 \ll t_3$
- Plane-like structures: $t_1 \ll t_2 \approx t_3$
- Isotropic structures: $t_1 \approx t_2 \approx t_3$

Based on the above local structure measurements, five features for each voxel in the density map were derived: two ratios of the eigenvalues λ_1/λ_2 and λ_2/λ_3, two ratios of the thickness $t_1/\ t_2$ and $t_2/\ t_3$, and the normalized density value of this voxel.

Multi-Class Classification of the Voxels Using Support Vector Machine

First introduced by Boser, Guyon, and Vapnik in 1992 [106]. SVM is one of the most commonly used supervised learning methods. It employs a maximum margin criterion and is a powerful tool for classification and regression tasks. The SVM was applied to classify the voxels from the test density map into three different classes: helix, sheet and background voxels (Figure 13). Given a training set of instance-label pairs (x_i, y_i), $i = 1, \ldots, I$ where $x_i \in R^n$ is an n-dimensional feature vector and y_i is the corresponding class label of that instance. SVM finds the parameters of a decision function $D(x) = w^T\phi(x) + b$ during a learning phase, where $\phi(x_i)$ maps x_i into a higher dimensional space [106]. The idea is to find a linear separating hyper plane with maximal margin between the classes in this higher dimensional space [106, 107]. All the parameters found during this learning phase can be stored in a model for future prediction on the test data.

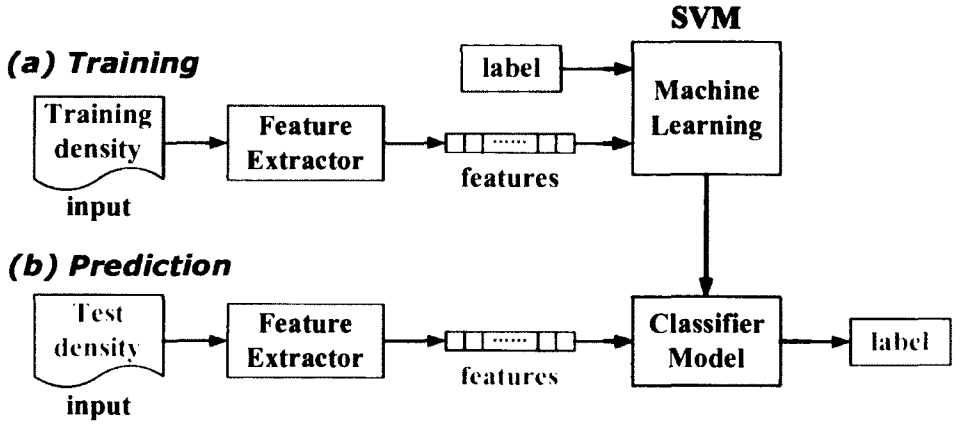


Figure 13. The training and prediction using the SVM

In the secondary structure identification problem, each voxel in the training density maps is associated with five features and one class label. The class label of each training voxel is determined based on its estimated proximity to the secondary structures. The cut-off values are empirical, by taking the consideration of the typical thickness of a helix (~5Å in diameter), and the distance between two adjacent β-stands (~4.5Å). In particular, the three classes were defined as the following.

- +1 for a helix voxel, if it is within 3Å from the axis of a helix;
- -1 for a sheet voxel, if it is within 4.5Å from the Cα atoms of a β-sheet;
- 0 for a background voxel, if it is not a helix voxel nor a sheet voxel.

SVM is inherently two-class classifiers. Multiple two-class problems can be converted to a multi-class problem using the concept of voting. LIBSVM [108] was employed in this method to solve the three-class prediction problem.[4] LIBSVM uses the "one-against-one" approach for multi-class classification [109]. If k is the total number of classes, this approach trains $k * (k-1)/2$ classifiers for all the possible combinations of the class pairs. A voting strategy was applied in which each 2-class classification is considered as a vote [108]. Each voxel from the test density map was then classified according to the class with the highest number of "votes".

[4] The version that downloaded from LIBSVM homepage http://www.csie.ntu.edu.tw/~cjlin/libsvm/ was in November 2011.

Post-processing

The SVM classification determines the class label for each density voxel in the target map. The post-processing takes the class labels as input and determines the exact position for the helices and β-sheets (Figure 14). A helix was represented as a set of voxels that are often near the central axis of the helix. A β-sheet was represented as a set of critical voxels on the sheet. The post processing includes two steps: filtering and clustering.

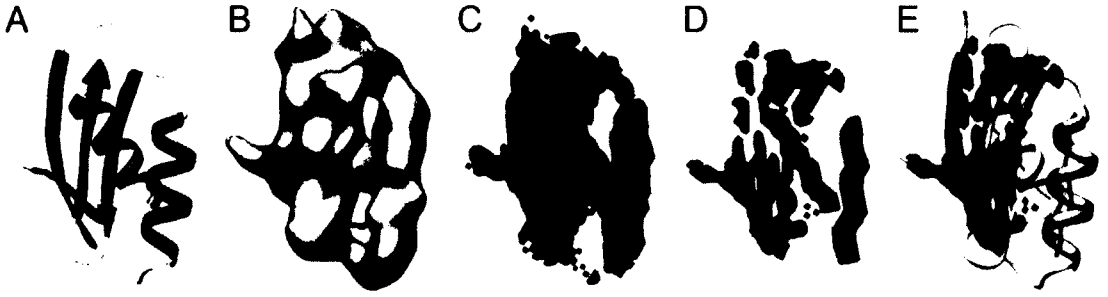


Figure 14. Post-processing. (A) the structure of 2AW0 (PDB ID) with helices (red ribbon) and b-sheet (blue ribbon); (B) the simulated density map at 8Å resolution; (C) the helix (red) and sheet (blue) voxels labeled by SVM; (D) the helix (red) and sheet (blue) voxels after post-processing; (E) the detected secondary structures superimposed on the PDB structure.

The filtering step aims at identifying the voxels with high density in a small neighborhood. It is observed that such voxels are often more reliable representatives for the SSEs. A filter has been applied using the local-peak-counter (LPC) proposed in SheetTracer [110]. For each voxel, the average density was calculated within a sphere of 3Å in radius. Those voxels in the sphere with density value greater than the average have their LPC incremented. All the voxels were sorted according to their LPC numbers after

counting. A threshold parameter was used to select the top ranked voxels. For example, the LPC filtering step selected the top 50% of voxels as the candidate representatives for the helices in the simulated density map. The top 75% of the voxels were selected as the candidate representatives for the β-sheets (Figure 14D).

The candidate voxels were further clustered to select the more reliable clusters for the annotation of secondary structures. The clusters were created based on the adjacency of voxels and then the size of each cluster was measured. A cluster size parameter was used to discard the small clusters that are often related to the turns. As an example, the size of 3Å and 8Å has been used to discard the small clusters for the helix and the sheet respectively in the simulated density map. The two threshold parameters can be adjusted by the user depending on the quality of the density maps.

Finally, a central axial line of helix voxel cluster was generated to represent the helix. This was done by travelling along the locally highest density voxels between the two ends of the helix voxel cluster. Since the shape of β-sheets is different for different sheets, the sheet voxels after post processing are used to represent the sheets (Figure 14E).

3. Result

The performance of *SSElearner* has been tested on ten simulated density maps and thirteen experimental cryo-EM density maps from EMDB. The selected EMDB density maps are between 3.8Å and 9Å resolution. Two types of evaluation were performed. One measures the number of identified secondary structures [70-72] and the other measures the number of Cα atoms [70, 76] that falls in the neighborhood of the secondary

structures. A helix is identified if its length is within one turn difference from the length of the helix in the PDB structure. A β-sheet is identified if the identified β-sheet voxels visually overlay on the β-sheet of the PDB structure. In order to present a more quantitative estimation about the size of the identified helices and β-sheets, the number of Cα atoms that are close to the identified helix voxels and β-sheet voxels is estimated. In particular, a Cα is considered as an identified helix Cα, if it is within 2.5Å distance from an identified helix voxel. A Cα is considered as an identified sheet Cα, if it is within 3Å distance from an identified sheet voxel. The definition of the secondary structures was based on the PDB file that is the authors'' annotation of the protein structure. Note that the authors'' annotation in the PDB file may be slightly different from the annotation using DSSP [103]. Although the definition of a helix and a β-sheet is clear in almost all the PDB files in our tests, it is necessary to visually decide the number and length in rare cases. For example, there is an overlap in the annotated helices with amino acid index 92-107 and 106-111 of 1CV1 (PDB ID). Three strands with amino acid index 37-48, 362-375, 96-110 of 2GSY were annotated in two β-sheets.

SSElearner has been tested using ten simulated density maps that were generated to 8Å resolution using the program *pdb2mrc* of EMAN [104] with a sampling size of 1Å/pixel. The ten proteins were used for testing *SSEhunter* at the same resolution [72]. The training dataset contains four other proteins (PDB ID: 1C3W, 1IRK, 1TIM and 2BTV) previously used for testing *SSEhunter* [72].

Our method successfully identified all the 74 helices that have more than four amino acids (Table 4). Since 3Å is used as the minimum helix length in the post-process step, only 4 out of the 14 extremely short helices were identified. Most of the missed helices have 3 amino acids in length, presumably of the 3_{10} helices. Our method detected all the 17 β-sheets, 6 of which have only two strands. Compared to *SSEhunter*'s result (Table 4), our *SSElearner* appears to be able to detect more 2-stranded β-sheets and is at least comparable in helix identification. Note that the same criteria is used to measure the number of the detected helices and β-sheet as indicated in the *SSEhunter* paper [72].

Table 4. The comparison of the number of detected secondary structures from the simulated maps.

PDB ID	*SSElearner*					*SSEhunter**		
	Helix < 5aa	**Helix 5 – 8aa**	**Helix > 8aa**	**Sheet = 2 strands**	**Sheet > 2 strands**	**Helix < 5aa**	**Helix 5 – 8aa**	**Sheet = 2 strands**
1AJW	0/0	1/1	0/0	0/0	2/2	0/0	1/1	0/0
1AJZ	0/1	3/3	7/7	1/1	1/1	0/1	3/3	0/1
1AL7	0/3	4/4	10/10	2/2	1/1	1/3	4/4	0/2
1CV1	1/1	2/2	8/8	0/0	1/1	1/1	0/2	0/0
1DAI	1/2	2/2	5/5	2/2	1/1	2/2	2/2	0/2
1ENY	0/0	1/1	9/9	0/0	1/1	0/0	1/1	0/0
1WAB	1/3	0/0	6/6	0/0	1/1	1/3	0/0	0/0
2AW0	0/0	0/0	2/2	0/0	1/1	0/0	0/0	0/0
2ITG	0/0	1/1	5/5	0/0	1/1	0/0	1/1	0/0
3LCK	1/4	2/2	6/6	1/1	1/1	1/4	0/2	1/1
Totals	4/14	16/16	58/58	6/6	11/11	6/14	12/16	1/6

* As a comparison, the columns for *SSEhunter* can be found in the supplementary table 1 of the *SSEhunter* paper [72].

In order to quantify the size of the detected secondary structures, particularly for β-sheets, the specificity and sensitivity have been calculated based on the detected helix and sheet Cα atoms similar to the estimation used in *SheetMiner* paper [76]. Table 5 shows the number of identified Cα atoms for the dataset in Table 4. A Cα is considered as an

identified helix Cα, if it is within 2.5Å distance from an identified helix voxel. A Cα is considered as an identified sheet Cα, if it is within 3Å distance from an identified sheet voxel. The sensitivity and the specificity of helix identification is 95.8% and 94.9% respectively. The sensitivity and specificity for β-sheet identification is 96.4% and 86.7% respectively.

Table 5. The accuracy of identified Cα atoms from the simulated maps.

PDB ID	to[a]	H[b]	tp H[c]	m H[d]	fp H[e]	S[f]	tp S[g]	m S[h]	fp S[i]	Sp. H[j]	Se. H[k]	Sp. S[l]	Se. S[m]
1AJW	145	5	5	0	1	63	50	13	10	99.3%	100.0%	87.8%	79.4%
1AJZ	282	124	120	4	6	37	37	0	31	96.2%	96.8%	87.4%	100.0%
1AL7	350	159	145	14	7	46	46	0	26	96.3%	91.2%	91.5%	100.0%
1CV1	162	123	114	9	3	14	14	0	11	92.3%	92.7%	92.6%	100.0%
1DAI	219	84	81	3	7	47	47	0	32	94.8%	96.4%	81.4%	100.0%
1ENY	268	126	121	5	4	66	56	10	25	97.2%	96.0%	87.6%	84.9%
1WAB	212	96	90	6	6	24	24	0	21	94.8%	93.8%	88.8%	100.0%
2AW0	72	22	22	0	3	25	25	0	11	94.0%	100.0%	76.6%	100.0%
2ITG	160	66	66	0	10	21	21	0	18	89.4%	100.0%	87.1%	100.0%
3LCK	270	107	98	9	9	30	30	0	33	94.5%	91.6%	86.3%	100.0%
Average										94.9%	95.8%	86.7%	96.4%

a: The total number of Cα atoms in the protein;
b: The total number of Cα atoms in the helices;
c: The number of true positive Cα atoms of helices;
d: The number of missed Cα atoms that are on helices but not detected;
e: The number of false positive Cα atoms for helices;
f: The total number of Cα atoms in the β-sheets;
g: The number of true positive Cα atoms for β-sheets;
h: The number of missed Cα atoms that are on β-sheets but were not detected;
i: The number of false positive Cα atoms for β-sheets;
j: The specificity of helix detection, calculated by the formula: $1 - (e/(a - b))$;
k: The sensitivity of helix detection, calculated by the formula: c/b;
l: The specificity of sheet detection, calculated by the formula: $1 - (i/(a - f))$;
m: The sensitivity of sheet detection, calculated by the formula: g/f.

Although it is essential to test the SSE detection methods using experimentally derived cryo-EM maps, it has been challenging to collect a large number of such data. Fourteen cryo-EM maps have been collected from the EMDB with resolutions between 3.8Å and 9Å, out of which thirteen were used to test *SSElearner* (Table 6). Four of the thirteen cryo-EM maps were selected from the cryo-EM Modeling Challenge 2010

(http://ncmi.bcm.edu/challenge). They are EMD-5030 (3FIN_chain R), EMD-5030 (3FIN_chain F), EMD-5140 (3IYF_chain A) and EMD-5001 (3CAU_chain A). Since SVM is a supervised machine learning method, the best identification accuracy should be expected when the training data and the target/test data have similar models. The selection of training data for a target density map is an important step in this classification. The quality of the cryo-EM density maps can be different for those at the similar resolutions. A training density map has been carefully selected for each target cryo-EM map that is to be tested (Table 6). A number of factors were taken into consideration in the selection of training data. These factors include the resolution, the estimated range of helix density, the estimated range of sheet density and the estimated noise level in the density map.

Table 6. The target cryo-EM density maps (EMDB ID, PDB ID and resolution) and their corresponding training data.

Test data	Training data
5030 (3FIN_R), 6.4Å	5168 (3MFP_A), 6.6Å
5030 (3FIN_F), 6.4Å	5168 (3MFP_A), 6.6Å
5140 (3IYF_A), 8Å	5030 (3FIN_F), 6.4Å
1733 (3C91_H), 6.8Å	1780 (3IZ6_K), 5.5Å
5168 (3MFP_A), 6.6Å	5030 (3FIN_R), 6.4Å
1237 (2GSY_A), 7.2Å	1780 (3IZ6_K), 5.5Å
5100 (3IXV_A), 6.8Å	5030 (3FIN_R), 6.4Å
5199 (3N09_C), 3.8Å	1780 (3IZ6_K), 5.5Å
1780 (3IZ6_K), 5.5Å	1740 (3C92_A), 6.8Å
5223 (3IZ0_A), 8.6Å	1340 (2P4N_A), 9Å
1340 (2P4N_A), 9Å	5223 (3IZ0_A), 8.6Å
1780 (3IZ6_T), 5.5Å	5030 (3FIN_F), 6.4Å
5001 (3CAU_A), 4.2Å	1740 (3C92_A), 6.8Å

An example of the detected secondary structures is shown in Figure 15. In this case the cryo-EM density map EMD-5030 was aligned with 3FIN_chain R. *SSElearner* detected

all the four helices and the β-sheet (row 1, Table 7). The specificity and the sensitivity for the detected helix Cα atoms are 93.1% and 96.6%, and those for the sheet detection are 89.3% and 100% respectively (row 1, Table 8).

Figure 15. Secondary structures detected using *SSElearner*. (A) Part of EMDB entry EMD-5030 at resolution 6.4 Å with fitted secondary structure of protein 3FIN_chain R; (B) identified helix and sheet locations.

The test of thirteen cryo-EM maps suggests that the helices longer than eight amino acids and sheets with more than two strands can be mostly detected. *SSElearner* detected 89 of 107 such helices and all 26 such β-sheets (Table 7). Note that *SSElearner* detected 100% such helices and sheets in the simulated data (Table 4). The contrast shows the challenges to the SSE detection method for the experimentally derived density maps. This is only visible when a large number of the experimental cryo-EM density maps are used for testing. Our test also suggests that *SSElearner* detects the β-sheets fairly well in the cryo-EM maps. It detected all the 26 β-sheets that have more than two strands, and 9 of 16 β-sheets with two strands (Table 7). For helices with no more than eight amino acids, *SSElearner* was only able to detect 30 of 61 such helices (Table 7).

Table 7. The identified secondary structures from the experimental cryo-EM density maps.

EMDB (PDB) ID	Helix < 5aa	Helix 5 – 8aa	Helix > 8aa	Sheet = 2 strands	Sheet > 2 strands
5030 (3FIN_R)	0/0	0/0	4/4	0/0	1/1
5030 (3FIN_F)	1/1	0/0	6/6	2/2	1/1
5140 (3IYF_A)	0/0	4/8	9/16	1/3	4/4
1733 (3C91_H)	0/0	0/0	5/5	1/1	2/2
5168 (3MFP_A)	0/0	6/9	8/10	0/1	2/2
1237 (2GSY_A)	0/3	2/5	3/3	1/1	4/4
5100 (3IXV_A)	0/2	5/11	12/13	0/1	2/2
5199 (3N09_C)	0/2	3/5	6/6	1/3	2/2
1780 (3IZ6_K)	0/0	0/1	2/2	0/0	1/1
5223 (3IZ0_A)	3/3	1/2	8/11	0/0	2/2
1340 (2P4N_A)	0/1	2/4	9/11	0/0	2/2
1780 (3IZ6_T)	0/0	0/0	4/4	1/1	0/0
5001 (3CAU_A)	0/0	3/4	13/16	2/3	3/3
Totals	4/12	26/49	89/107	9/16	26/26

The performance of our SSE detection method has been analyzed using the number of identified Cα atoms to reveal the size accuracy of the SSE (Table 8). The overall specificity and sensitivity are 91.8% and 74.5% respectively in helix detection. The main reason for the reduced sensitivity between the simulated maps verses the EMDB maps is in the short helix detection. For example, 8 out of 11 helices in EMD-1237 and 13 out of 26 helices in EMD-5100 are no more than eight amino acids. Another reason is the reduced quality of the experimental cryo-EM maps compared to that of the simulated density maps. The experimental density maps often have incomplete density data, particularly for the short helices. The overall specificity and sensitivity in sheet identification are 85.2% and 86.5% respectively. Our test using thirteen experimentally-derived cryo-EM maps shows the challenges in the SSE detection from the real cryo-EM maps. It is not possible to detect them as accurately as in the simulated maps at this point.

Table 8. The identified Cα atoms from the experimental cryo-EM maps.

EMDB (PDB) ID	to[a]	H[b]	tp H[c]	m H[d]	fp H[e]	S[f]	tp S[g]	m S[h]	fp S[i]	Sp. H[j]	Se. H[k]	Sp. S[l]	Se. S[m]
5030 (3FIN R)	117	59	57	2	4	14	14	0	11	93.1%	96.6%	89.3%	100.0%
5030 (3FIN F)	208	80	71	9	4	37	37	0	15	96.9%	88.8%	91.2%	100.0%
5140 (3IYF A)	491	263	136	127	42	74	47	27	69	81.6%	51.7%	83.5%	63.5%
1733 (3C91 H)	203	86	69	17	0	62	54	8	25	100.0%	80.2%	82.3%	87.1%
5168 (3MFP A)	374	186	138	48	17	60	41	19	56	91.0%	74.2%	82.2%	68.3%
1237 (2GSY A)	428	69	46	23	5	187	180	7	71	98.6%	66.7%	70.5%	96.3%
5100 (3IXV A)	626	277	195	82	35	99	86	13	51	90.0%	70.4%	90.3%	86.9%
5199 (3N09 C)	397	144	118	26	24	128	114	14	41	90.5%	81.9%	84.8%	89.1%
1780 (3IZ6 K)	119	37	25	12	1	29	29	0	11	98.8%	67.6%	87.8%	100.0%
5223 (3IZ0 A)	412	186	116	70	34	55	42	13	24	85.0%	62.4%	93.3%	76.4%
1340 (2P4N A)	412	202	124	78	21	55	39	16	75	90.0%	61.4%	79.0%	70.9%
1780 (3IZ6 T)	82	54	45	9	0	7	7	0	8	100.0%	83.3%	89.3%	100.0%
5001 (3CAU A)	526	255	214	41	61	82	71	11	72	77.5%	83.9%	83.8%	86.6%
Average										91.8%	74.5%	85.2%	86.5%

a: The total number of Cα atoms in the protein;
b: The total number of Cα atoms in the helices;
c: The number of true positive Cα atoms of helices;
d: The number of missed Cα atoms that are on helices but not detected;
e: The number of false positive Cα atoms for helices;
f: The total number of Cα atoms in the β-sheets;
g: The number of true positive Cα atoms for β-sheets;
h: The number of missed Cα atoms that are on β-sheets but were not detected;
i: The number of false positive Cα atoms for β-sheets;
j: The specificity of helix detection, calculated by the formula: $1 - (e/(a - b))$;
k: The sensitivity of helix detection, calculated by the formula: c/b;
l: The specificity of sheet detection, calculated by the formula: $1 - (i/(a - f))$;
m: The sensitivity of sheet detection, calculated by the formula: g/f.

CHAPTER IV

MODELING BETA-STRAND FROM CRYO-EM DENSITY MAPS

Ab initio modeling aims to derive near atomic structures from electron density maps without a template structure. Although the connection between secondary structure elements (SSEs), such as α-helices and β-sheets, is ambiguous at medium resolutions, the likely connections may be derived. Given the positions of α-helices and β-strands in a density map, one can match them with secondary structure sequence segments that can be predicted from the amino acid sequence to derive the overall topology of a protein chain [47, 63-66]. Once the topology is determined, backbone and side chains can be constructed and evaluated using energy functions [66, 68, 69]. *Pathwalking* derives Cα trace directly from pseudo atoms extracted from the density map [111]. The main drawback of this method appears to be at β-sheet regions where the β-strands are not resolved.

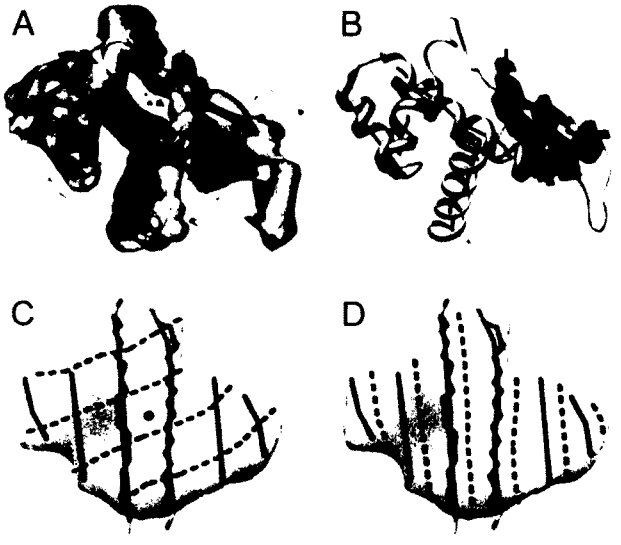


Figure 16. The problem of β-strand detection from medium-resolution density maps. (A) The density corresponding to Chain R of 3FIN (PDB ID) was extracted from cryo-EM density map EMD_5030 (6.4Å resolution) and was superimposed with its corresponding PDB structure (ribbon). (B) The

Figure 16. (Continued)
detected helices (red lines) and β-sheet region (surface view of blue voxels) using *SSElearner* [86]. Two possible sets of β-traces (black solid lines and red dashed lines) may differ in orientation (C) and/or position (shift) (D). The backbone structure is superimposed on the observed β-traces for the middle two β-strands in (C) and (D).

This chapter is a summary of the *StrandTwister* methodology published in paper [112].

1. Motivation

The location of secondary structures is critical in modeling atomic structure from density map and as an overall shape descriptors in identifying similar structures [33]. Although it is not possible to distinguish the amino acid at medium resolutions, secondary structure such as α-helices (red lines in Figure 16B) and β-sheets (blue density voxels in Figure 16B) can be identified [70, 72, 74, 76, 86, 113, 114]. A β-sheet contains multiple β-strands. Although β-sheets can be identified from cryo-EM density maps at 5-10Å resolutions, it is almost impossible to detect the β-strands of a β-sheet. The spacing between two neighboring β-strands is between 4.5 and 5Å, and β-strands are only visible when the resolution is higher than 4.7Å [77, 78]. Without knowing the location of β-strands, the representation of protein is purely dependent on the relative location of helices [73]. *Ab initio* modeling has been successful deriving the backbone from the density map of GroEL (4.2 Å resolution) (Ludtke et al. 2008) and gp10 (4.5 Å resolution) [80]. However, there has not been an α/β structure that is resolved using *ab initio* modeling from a density map at a medium resolution. One of the challenges is the inability of detecting β-strands from the density maps. In addition to the secondary structural elements, skeleton that represents possible connections can be identified [95, 115], although ambiguous points exist in the skeleton.

A helix detected from the medium resolution data is often represented as a line, referred here as an α-trace that corresponds to the central axis of a helix (red lines in Figure 16B). A β-trace (black line in Figure 16C and D) is defined as the central line along a β-strand. In particular, an observed β-trace is the line interpolating all geometrical centers of three consecutive Cα atoms on a β-strand plus two Cα atoms at the end of the β-strand. At medium resolutions, the Cα trace of a β-strand is not resolved in the density map. The problem of detecting β-strands from the density of a β-sheet is to find the orientation (Figure 16C) and location (Figure 16D) of β-traces.

Figure 17. Density of a β-sheet at different resolutions. The density was simulated using atomic structure of 1A12 (PDB ID) and EMAN [102] at 6Å in (A), 8Å in (B), and 9Å in (C). A β-sheet detected from an experimentally derived cryo-EM density map (EMD_5030) at 6.4Å using *SSEtracer* [116] and visualized in Chimera [117] in (D). The surface representation of density is shown at a lower (left) and a higher (right) threshold in (A)-(D).

As more experimentally determined cryo-EM maps accumulate in the Electron Microscopy Data Bank (EMDB) [32, 118], it is clear that such cryo-EM maps are more challenging than the density maps simulated at the same resolution. A simulated density map at 6Å resolution often reveals the separation

of β-strands at a proper range of density thresholds (Figure 17A). Even at 8Å resolution, a simulated density map partially reveals the separation of β-strands (Figure 17B) and it has much better quality than an experimentally-obtained cryo-EM map at a similar resolution. However, a simulated density map at 9 or 10Å resolution is often challenging for visual detection (Figure 17C), and so is a cryo-EM map at 6.4Å resolution (Figure 17D). The question this chapter addresses is if it is possible to derive the β-traces from cryo-EM maps at the medium resolutions when no separation of β-strands is detectable.

Currently there are no tools to derive the β-traces from cryo-EM density maps at medium resolutions. *Sheettracer* is the first attempt to derive β-traces [110]. It uses de-convolution to enhance the separation of β-strands while filtering and clustering follow. However, this method was predominantly tested using simulated density maps with resolutions of 6Å and 8Å. For simulated density maps at such resolutions, the separation of β-strands may be visible or partially visible (Figure 17A and B); however, this is not true for experimentally-obtained cryo-EM maps (Figure 17D). *Sheettracer* may be suitable for density maps in which the separation of β-strands is partially visible. *Gorgon* uses a semi-automated method allowing a user to determine the position of β-strands, which is challenging to apply in the cryo-EM maps at medium resolutions [119]. *Pathwalking* derives backbone from cryo-EM maps at near-atomic resolutions such as 3-5Å. It performs well in α-helical domains but fails at β-sheet regions at the medium resolutions [111].

The detection of β-strands from medium-resolution cryo-EM maps remains an open problem since the first attempt in 2004 [110]. In addressing this challenge, a new method - *StrandTwister* is proposed which does not rely on the separation of β-strands and therefore is applicable to much lower resolutions. *StrandTwister* has been tested using 100 β-sheets simulated to 10Å resolution and 39 β-sheets that were

computationally detected from cryo-EM maps at 4.4-7.4Å resolutions. It has been observed that *StrandTwister* can detect the traces of β-strands fairly well for many β-sheets with three or more β-strands. The detection of β-traces is limited by the identification of β-sheet. This means 2-stranded β-sheets and those β-sheets at low quality regions of a density map remain challenging. The results and challenges in β-strands detection will be discussed using three cases: gp10 of bacteriophage epsilon15 map (7.3Å resolution), GroEL density (4.2Å resolution), and E2 of Venezuelan equine encephalitis virus map (4.4Å resolution).

2. Methodology

There are two major steps in our *StrandTwister*. The first is to generate a polynomial surface to fit β-sheet voxels [116]. The second step identifies right-handed β-twist from the polynomial surface model.

A. Polynomial Fitting of β-sheet Density

A β-sheet generally appears as a thin layer of density at medium resolutions. However, bumps and missing density in β-sheet region are often observed in the experimentally derived cryo-EM maps, presumably due to errors in the experimental data. In order to capture the overall shape of β-sheet density and make strand-detection less sensitive to errors in density data, a polynomial surface (5) was first determined by fitting the density voxels (see description in Figure 18). Here (x, y, z) is the coordinate of a voxel point.

$$z = Ax^3 + By^3 + Cx^2 + Dy^2 + Ex^2y + Fy^2x + Gxy + Hx + Iy + J \quad (5)$$

In order to fit the polynomial surface (Formula 5) into β-sheet density, a normal vector of the β-sheet density was first determined using the points near the center of β-sheet. Translation and rotation were performed such that the center becomes the origin and the normal vector aligns with the z-axis (Figure 18A). Least-square fitting was then performed to determine the parameters in the polynomial surface. The fitting error was reported in [116]. The resulting polynomial surface (Figure 18B) was then used to calculate the two main features of β-sheet – handedness and twist angles.

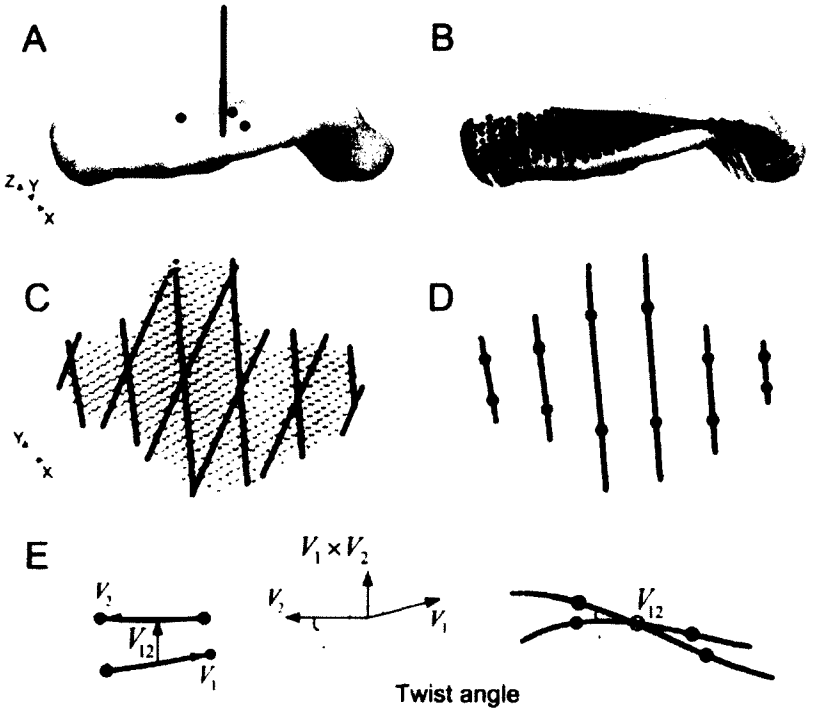


Figure 18. Sampling of the β-traces and calculation of the twist angles. (A) The density of a β-sheet (gray) with the center and the normal vector (red); (B) The surface points (yellow) derived from polynomial fitting; (C) Two sampled orientations; (D) Vectors of β-traces (green to red); (E) Calculation of handedness and twist angle for two neighboring strands.

The handedness and twist angles were calculated for each set of β-traces. Each set of β-traces was generated based on the observation that β-strands are roughly parallel to each other with two

neighboring strands forming a small twist angle. A set of parallel lines (Figure 18C) with 4.5Å spacing was first created on a plane perpendicular to the normal vector of the β-sheet density. Eighteen sets of parallel lines were created on the plane to sample the orientation space by every 10°. Each set of the parallel lines was then shifted 1.5Å left / right to sample the translation freedom. Each set of parallel lines was used as an initial reference to generate non-parallel lines on the surface model. To do this, the parallel lines were projected back to the polynomial surface model. Note that the resulting β-traces are not parallel anymore due to the twisted curvature of β-sheet. Since the central area of a β-sheet is often more reliably detected than the edge, each resulting curved β-trace was divided into four segments with equivalent length and the central two were used (Figure 18D) to represent the vector for a β-trace (V_1 pointing from the green to the red dot in Figure 18E).

Let V_1 and V_2 represent two neighboring β-traces, and let V_{12} represent the vector pointing from line segment V_1 to V_2 and having the shortest distance between them. A right-handed twist requires the following.

- $(V_1 \times V_2) \cdot V_{12} > 0$, if V_1 and V_2 are on the anti-parallel β-strands
- $(V_1 \times V_2) \cdot V_{12} < 0$, if V_1 and V_2 are on the parallel β-strands

In principle, $V_1 \times V_2$ is on the line of V_{12}, either having the same direction as V_{12} in which $((V_1 \times V_2) \cdot V_{12}) > 0$, or having the opposite direction as V_{12} in which $((V_1 \times V_2) \cdot V_{12}) < 0$. Suppose V_1 and V_2 are on antiparallel β-strands as in the first case, a right-handed twist will have $V_1 \times V_2$ in the same direction as V_{12} (Figure 18E). A left-handed twist will result in $((V_1 \times V_2) \cdot V_{12}) < 0$. Since a twist angle is often

small, the acute angle formed by V_1 and V_2 was used as the inter-strand twist angle (Figure 18E). Similar principle applies in the second case in which V_1 and V_2 are on the parallel β-strands.

B. Right-handed β-twist and β-strand Detection

Right-handed twist of a β-sheet was first described by Cyrus Chothia in 1973 [90]. Salemme et al. suggested that the spatial configuration of a β-sheets is isotropically stressed surface [91]. To understand the right-handed twist, one may thread the right-hand fingers along the β-strands (Figure 19D). The natural curvature of our right hand would lift up the index finger and lower down the pinky (Figure 19D).

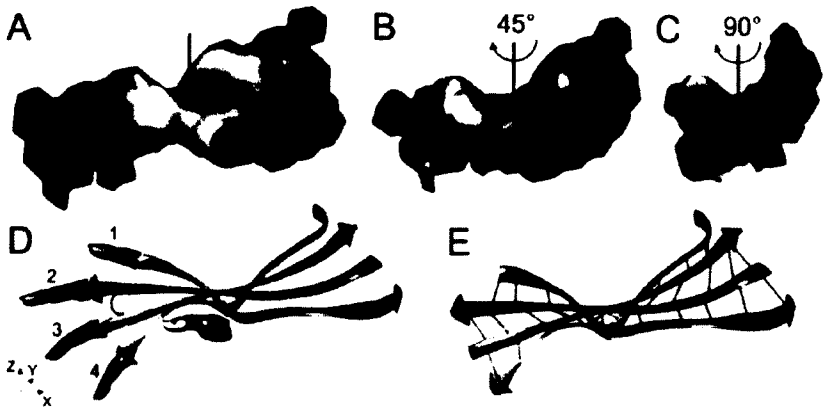


Figure 19. The right-handed twist of a β-sheet. The β-sheet density was detected using *SSEtracer* from cryo-EM map (EMD_1237), and is shown in 0° (A), 45°(B), and 90° (C) view around the normal vector (red line); (D) The right-handed twist of β-sheet 2GSY_sheet G. The strands are labeled such that the index finger (1) is at forefront and the pinky (4) is the farthest from the page. The correspondent PDB structure of (A) is shown with hydrogen bonds (thin black lines) in (E).

The key idea in our method is the observation that the density of a β-sheet reveals the right-handed twist very well (Figure 19A, B, C), particularly at the medium resolutions. In addition, it has been discovered

that wrongly positioned β-strands in the β-sheet density can produce left-handed twist. For example, if one attempts to thread fingers into the paper and fit the β-sheet density in Figure 19A, it has to be the left hand, not the right hand. This suggests that the correct orientation of the β-strands can be identified by measuring the handedness and twist angles [120].

The surface point (yellow in Figure 18B and C) generated by the mathematical model in formula (3) is a simple representation of β-sheet density, and was used to measure handedness and twist angles. The handedness and twist angles were calculated for each set of β-traces. Each set of β-traces was generated based on the observation that β-strands are roughly parallel to each other with two neighboring strands forming a small twist angle (see description in Figure 18).

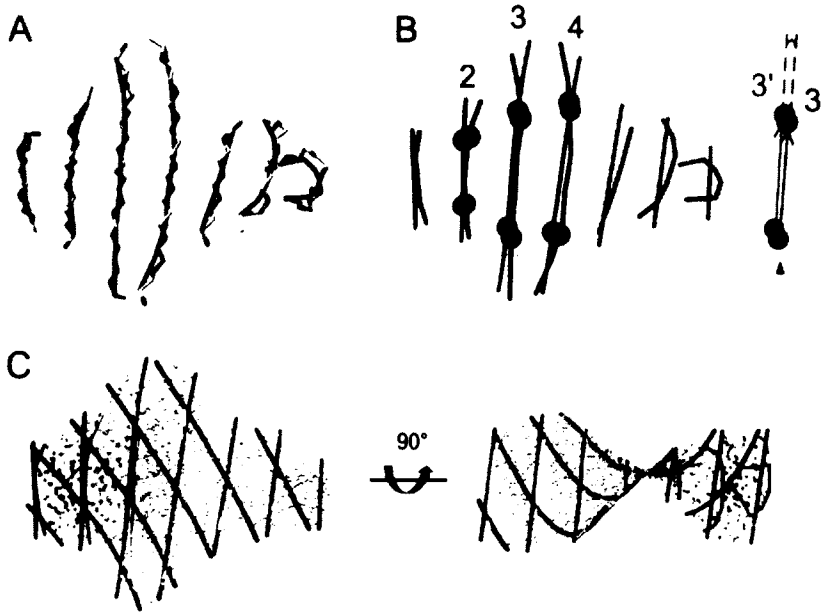


Figure 20. The set of β-traces with the maximum twist. (A) The observed β-traces (black) are superimposed on the backbone of sheet A of 1FX2 (PDB ID). (B) The set of β-traces with the largest AMT (red). The central 2/4 of the traces is indicated by two spheres. The smallest angle between the detected β-trace and the corresponding observed β-trace among all pairs of detected/observed β-traces is shown in dashed box. (C) The points on the fitted surface (dots), the observed β-traces (white lines) and

two potential sets of β-traces (red, green) are shown in two views. Among all the sets of sampled traces, the set of red traces has the largest AMT.

Since there is a handedness and a twist angle for each pair of neighboring β-traces, the overall handedness and twist angle of the entire set of β-traces need to be determined. Those sets having right-handed twist for all neighboring-strands were first selected. Such sets were further evaluated by the Average-Main-Twist (AMT) angle to select the best one (Figure 20). AMT is the average of three consecutive inter-strand twist angles: the largest inter-strand twist angle and that to its left and to its right respectively. The assumption is that the correct set of β-traces is expected to have near maximum overall twist (see Table 9). The top ten (first ten) sets of β-traces with the largest AMT were detected as potential sets.

Table 9. Maximum twist angle and main orientation difference (MOD) for the set of β-traces with the maximum twist.

No.	PDB ID[a]	#Det./#Obs.[b]	Max tw.[c]	MOD[d]
1	1A12_B	2 / 3	29.6	12.7
2	1A12_D	2 / 3	31.5	8.5
3	1A4I_E	3 / 3	14.1	7.8
4	1A4I_F	3 / 4	14.4	20.0
5	1AOP_1	5 / 5	26.0	13.0
6	1AOP_2	5 / 5	22.3	8.9
7	1B5E_D	5 / 5	17.5	6.3
8	1RV9_B	5 / 6	22.8	8.1
9	1T8H_B	6 / 6	14.1	19.8
10	1VLY_A	5 / 6	22.9	15.1
11	1YT3_A	5 / 6	26.1	9.0
12	2HKE_A	5 / 6	16.7	4.5
13	1CHD_SH1	5 / 7	23.7	24.7
14	1D5T_D	5 / 7	24.4	12.2
15	1FX2_A	7 / 7	27.3	4.2
16	1DTD_A	8 / 8	22.9	8.9
17	1HDO_AA	7 / 8	21.0	7.0
18	1JL0_A	7 / 8	11.6	5.1
19	1UD9_B	7 / 9	16.3	23.9

a. The PDB and sheet ID;
b. The number of the detected β-traces / the observed β-traces in the set with the maximum twist;

c. The largest twist measured as the AMT angle among all sets of right-handed β-traces sampled with difference orientations;
d. MOD between the set of β-traces with the maximum twist and the observed set.

3. Result

In order to test the feasibility of β-trace detection using handedness and twist angles, the relationship between β-strand orientations and the maximum twist angle was first investigated using atomic structures of β-sheets. *StrandTwister* was then evaluated using 100 density maps simulated to 10Å resolution. It was eventually tested using 39 β-sheets detected from experimentally derived cryo-EM maps at 4.4-7.4Å resolutions. To evaluate the accuracy of β-trace detection, the 2-way distance between the set of detected β-traces and the set of observed β-traces was calculated (see definition of β-trace in Introduction and Figure 16C).

A. The Main Orientation of β-strands and the Maximum Twist Angle

In order to identify the correct β-strand orientation among many orientation samples, the relationship between the observed orientation and the maximum twist angle was investigated. Nineteen β-sheets were randomly selected from the structures in PDB with less than 40% sequence similarity. It appears that an observed β-trace roughly represents the central axis of the β-strand in terms of the backbone atoms (Figure 20A and Figure 16C). A surface model was derived by fitting the polynomial of formula (1) to all points on the observed β-traces of the sheet. All sets of β-traces with right-handed twists were selected and the AMT angle was calculated for each set. In general, the AMT angle represents the average of three twist angles nearby the largest twist angle, and it is more stable than either the largest twist angle or the average of all twist angles of a sheet (data not shown). The set of β-traces with the largest AMT was identified; this was referred as the set with the maximum twist (Table 9). It appears

that the maximum twist angle (the largest AMT) among all sets of β-traces is between 11° and 32° for the nineteen test cases (Column 3 Table 9). Interestingly, most of the largest twist angles are between 15° and 30°, a popular range of twist angles measured previously using the atoms of the β-strands [121]. Note that the maximum twist angle in Table 9 was not directly calculated using positions of atoms, but rather through the sampling of different orientations and selecting for the largest AMT. This suggests that the largest AMT roughly corresponds to the twist angle of β-strands, and the set of β-traces with the largest AMT roughly reflects the orientation of the true β-strands.

Figure 20B shows an example of the detected set of seven β-traces with the maximum twist (red lines). The AMT of this set is 27.3° (row 15 of Table 9), and it is the largest among all the sets of β-traces sampled using different orientations and translations. It appears that such a set (red) aligns well with the observed set (black), particularly for the central region of the β-sheet. Some portions near the edge of β-sheet are not well detected. In this case, strand 3 appears to be best detected, followed by strand 2 and 4. In order to see if the maximum twist can be used as an indicator to detect the β-traces, the Main Orientation Difference (MOD) was calculated for the set with the largest twist. MOD is defined as the average of three angles, the smallest angle between the detected β-trace and the corresponding observed β-trace among all pairs of detected/observed β-traces (3'' and 3 in Figure 20B), and two similar angles for its neighboring two pair of β-traces, the angle between 2'' and 2 and the angle between 4'' and 4 in this case. The MOD for the set of β-traces that has the maximum twist is only 4.2° for the β-sheet of 1FX2 (Figure 20B and Table 9), suggesting a very good estimation of the β-traces for strand 2, 3 and 4. To determine whether or not the small MOD happened by chance, sets of β-traces of difference orientations were randomly generated. As expected, MOD varies from 0° to 90° (data not shown), with an average of about 45° which is much larger than the MOD of the set with maximum twist angle (Table

9 column 4). Sixteen of the nineteen cases have their MOD less than 20°. This result suggests that the β-trace with the maximum twist is a close estimation of the β-strands for the central region in a β-sheet. Note that a β-trace is often curved, and the curvature appears to be more at the edge of the β-sheet. However, our results suggest that it is possible to detect the major β-trace orientation by the maximum twist for the central area of the sheet. Due to the quasi-parallel nature of β-strands, the other strands were derived using the major orientation as a guide. The results are shown in later sections. Our results support a hypothesis stating the actual β-strand orientation roughly follows the orientation that creates the maximum twist at central area of β-sheet, where longer stretches of hydrogen bonds restrict the flexibility of the conformation. It is possible that the maximum-twist conformation may represent a stable conformation for the β-sheet to fold into a compact protein structure.

B. Two-way Distance

In order to calculate the 2-way distance, 1-1 correspondence between the β-traces in the detected set and those in the observed set was first determined based on the overall smallest 2-way distance. This ensures that the same number of detected β-traces ($S_1, S_2, \ldots, S_T$) are compared to same number of observed traces ($S'_1, S'_2, \ldots, S'_T$) in which S_k is compared with S'_k for k = 1, ..., T. The number of miss-detected (or wrongly detected) β-strands can be inferred from the difference between the total number of the observed and that of the detected β-traces. The 2-way distance of a β-strand k, D_k was calculated for each pair of lines S_k and S'_k. The overall 2-way distance D reflects the quality of detected β-traces that are corresponding to their observed ones.

$$D_k = \left(\sum_{i=1}^{N} D_i^{SS'}/N + \sum_{j=1}^{M} D_j^{S'S}/M\right)/2 \quad (6)$$

$$D = \left(\sum_{s=1}^{T} D_k\right)/T \quad (7)$$

In formula (4), N and M are the numbers of points on the detected (S_k) and the observed β-traces (S'_k) respectively. i and j are the indices of a point along line S_k and S'_k respectively. $D_i^{SS'}$ is the projected distance from point i of S_k to S'_k. The projection of point i is required to be within line S'_k. In the case it is outside, the distance between i and an end of S'_k was used as an approximate distance. In order to estimate how much of a β-strand was detected, the percentage of the detected Cα atoms of an observed β-strand was calculated. An amino acid of a β-strand is considered detected if the projection distance from its Cα atom to the corresponding detected β-trace is less than 2.5Å, which is about half β-strand spacing.

C. Performance on the Simulated Maps

The purpose of this test is to investigate if the β-strands can be traced from the density maps simulated at 10Å resolution, at which the separation of β-strands is not visible. The dataset was collected randomly from PDB with the following requirements: (a) The β-sheets are regular sheets rather than barrels; (b) The number of strands is between 3 and 10. The atomic structures of β-sheets were used to generate β-sheet density maps at 10Å resolution using EMAN, a popular software to produce simulated density [102] with step size of 1Å/pixel. A polynomial surface (1) was generated to fit the β-sheet density.

Figure 21 shows the best of top ten sets of detected β-traces (red lines) for three cases with 3, 6 and 9 β-strands respectively. In the case of sheet B of PDB structure 1T8H, one of the top ten detected sets appears to align with the β-strands very well (Figure 21B). In this case all six β-traces were detected with a small 2-way distance of 0.70Å (Table 10 row 3 of 6-stranded). It is observed that, the top ten detected sets always include a set with close orientations to the actual β-strand orientations. The

detection is slightly better for smaller β-sheets such as those with 3-7 strands, if the number of detected β-traces and the 2-way distance are considered. However, some large β-sheets were still well detected, such as the 9-stranded sheet IUD9_B (Figure 21C). The 2-way distance is only 1.07Å in this case (Table 10), and most of the β-traces are accurately detected. The error appears to be at the edge of the β-sheets.

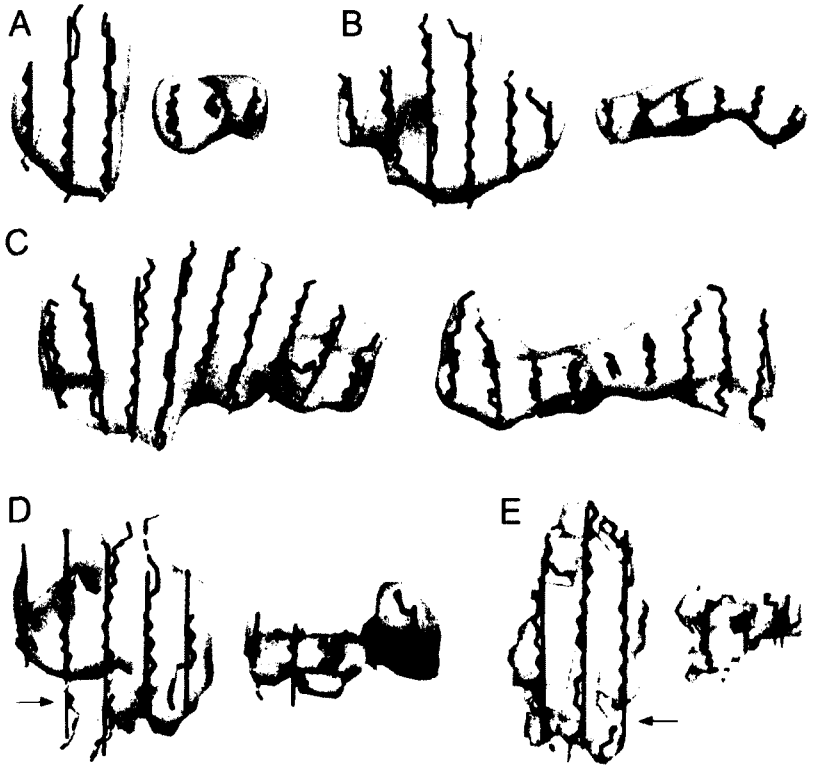


Figure 21. β-strand detection from simulated density maps at 10Å and experimental cryo-EM maps. The best of the top ten sets of detected β-traces (red lines) are superimposed with the backbone of the β-strands and the density maps (gray) for β-sheets 1A4I_B in (A), 1T8H_B in (B), 1UD9_B in (C), EMD_2165_4B4T_1C in (D) and EMD_1780_3IZ5_Z in (E). The top view (left) and the side view (right) are shown in each case. The density (gray) of β-sheet in (D) and (E) was detected using *SSEtracer*.

The test of 100 simulated β-sheet density maps shows that one of the top ten ranked sets of β-traces aligns very well with the observed set of β-traces, with an overall 2-way distance of 1.24Å (Table 10 last

two rows) for the detected β-traces. To analyze the sensitivity of the detection, the number of amino acids that were missed in the detection was measured (see definition early in Results). For example, 1ATZ_A has five of the six β-strands detected (Table 10 row 1 of 6-stranded), and four amino acids were missed. For the five β-traces detected, the 2-way distance is 1.18Å. Among the 100 test cases, *StrandTwister* appears to be able to detect 81.48% of the β-strands in one of the top ten ranked sets of β-traces (Table 10).

Table 10. β-trace detection from simulated density maps at 10Å resolution.

PDB ID[a]	Fit	#Det.[c]	2-w	#Det./#Obs.	PDB ID[a]	Fit	#Det.[c]	2-w	#Det./#Obs.
3-stranded					6-stranded				
1A12_B	1.29	3	0.95	16 / 17	1ATZ_A	1.51	5	1.18	36 / 40
1A12_D	1.25	3	0.93	15 / 17	1RV9_B	1.28	6	0.95	27 / 29
1A4I_B	1.30	3	0.64	18 / 18	1T8H_B	1.13	6	0.70	28 / 28
1A8D_B	1.18	3	0.75	12 / 12	1VLY_A	1.47	6	0.99	25 / 29
1ATG_B	1.39	3	1.07	10 / 12	1YT3_A	1.20	6	0.90	29 / 31
1AZO_A	1.52	3	0.82	16 / 18	2HKE_A	1.32	6	0.80	31 / 32
1AZO_B	1.52	3	0.84	13 / 13	2P51_A	1.36	6	0.98	28 / 30
1B3A_A	1.45	3	0.82	14 / 15	2QTR_A	1.56	5	1.85	22 / 31
1B5E_A	1.22	3	1.01	13 / 13	2VBF_BA	1.51	6	0.96	29 / 31
1BM8_A	1.81	3	1.11	16 / 18	2VOA_AB	1.38	6	1.00	29 / 32
1BTE_B	1.43	3	1.41	22 / 24	2ZSG_A	1.47	6	0.82	28 / 30
1BUP_C	1.54	3	1.45	16 / 19	3BL9_D	2.33	5	1.84	21 / 29
1E0M_1	1.61	3	1.08	15 / 18	7-stranded				
4-stranded					1CHD_SH1	1.47	7	1.42	34 / 39
1A12_A	1.37	4	0.97	24 / 25	1D5T_D	1.78	6	1.48	47 / 56
1A12_C	1.28	4	1.07	18 / 20	1ELU_B	1.57	6	1.84	27 / 36
1A12_J	1.33	4	0.93	18 / 20	1FX2_A	1.87	7	0.95	40 / 45
1A4I_C	1.20	4	0.76	18 / 18	1FYE_A	1.36	7	1.00	28 / 31
1A8D_E	1.52	4	1.03	25 / 27	1G8K_B	1.43	7	1.06	28 / 29
1AOP_1	1.56	4	1.08	18 / 21	2A6Z_A	1.49	7	0.94	42 / 47
1AQZ_D	1.59	3	1.29	17 / 23	2APJ_A	1.56	6	1.58	22 / 37
1BUP_D	1.57	4	0.93	28 / 32	2DKJ_C	1.58	7	1.29	29 / 32
1CID_C1	1.25	4	0.74	27 / 28	3BA1_B	1.41	7	0.92	25 / 26
1CC8_A	1.50	4	0.98	25 / 27	8-stranded				
1CCW_B	1.43	4	1.36	17 / 22	1DTD_A	1.68	8	1.15	40 / 47
1DD9_A	1.40	3	1.74	15 / 21	1H2W_AJ	2.36	8	1.12	44 / 51
1DS1_C	1.29	3	0.95	13 / 17	1HDO_AA	1.73	7	0.92	34 / 37
1Q38_1	1.43	4	0.94	23 / 27	1JL0_A	1.56	8	0.85	48 / 52
1S04_1	1.56	4	1.37	14 / 16	1JOV_D	1.64	7	1.32	52 / 60
2P8Y_1	1.37	4	0.98	19 / 20	1JUH_A	1.73	8	1.04	40 / 44
2VZ1_AE	1.46	4	0.78	17 / 18	1JW9_A	1.63	8	1.76	33 / 45

	5-stranded				1KMV_B	1.96	7	1.65	34 / 49
1AKY_A	1.44	5	1.22	22 / 25	1LAM_B	1.45	8	1.04	50 / 56
1AOP_1	1.54	5	1.03	26 / 29	1M15_A	1.70	7	1.76	37 / 46
1AOP_2	1.75	5	1.78	21 / 38	1M4L_A	1.80	8	0.99	41 / 48

Table 10. (Continued)

PDB ID[a]	Fit	#Det.[c]	2-w	#Det./#Obs.	PDB ID[a]	Fit	#Det.[c]	2-w	#Det./#Obs.
1AOP_3	1.63	5	1.17	26 / 31	1NKG_C	1.64	7	1.11	45 / 54
1B5E_D	1.40	5	1.88	18 / 29	1ZLI_1	1.82	8	1.16	41 / 47
1BUP_A	1.51	5	1.12	26 / 27	3RL6_1	1.73	7	1.89	35 / 51
1C1D_A	1.57	4	1.19	29 / 37	8DFR_S1A	1.88	8	1.46	42 / 56
1C7K_A	1.66	5	1.59	24 / 30			9-stranded		
1CXQ_A	1.53	5	0.95	24 / 27	1QNA_C	1.62	7	1.90	39 / 56
1DGW_C	1.46	4	1.00	27 / 32	1UD9_B	2.02	9	1.07	49 / 58
1DMH_C	1.46	5	0.77	26 / 27	1UWC_BA	1.68	8	1.35	41 / 53
1DTD_B	1.53	4	1.10	23 / 28	2ABS_A	1.47	10	1.29	39 / 49
1E2K_A	2.59	4	2.47	13 / 28	2EAB_A	2.04	8	1.18	59 / 75
1MOL_S1	2.44	5	1.91	34 / 50	2VVG_AB	1.92	7	1.67	43 / 67
	5-stranded						9-stranded		
1ZH2_A	1.34	5	1.19	20 / 24	3DB7_A	2.21	9	1.90	37 / 51
2JKX_AD	1.23	4	0.74	19 / 23	3EN0_A	1.46	8	1.66	32 / 39
2VZ1_AD	1.23	4	0.66	19 / 23	3FCX_A	2.19	9	1.71	48 / 66
	10-stranded				3H9M_C	2.10	7	2.38	30 / 53
1IG0_B	1.81	9	2.13	33 / 55	3HID_A	1.94	6	1.66	25 / 40
1PE9_A	1.34	9	1.68	40 / 52	3HO6_A	2.03	9	1.89	38 / 53
1V7W_A	1.77	9	1.26	59 / 70	**Average**	**1.59**		**1.24**	**2834 / 3478 = 81.48%**
2B0T_A	2.17	8	2.47	22 / 44	**Std_dev**	**0.29**		**0.42**	

a. The PDB and sheet ID;
b. The RMSE (root-mean-square-error, in Å) for the polynomial surface fitting.
c. The number of detected β-traces;
d. The 2-way distance (in Å) for the best of the top ten sets of detected β-traces.
e. The number of detected / the total number of amino acids in the β-sheet.

D. β-strand Detection from Cryo-EM Maps

The performance of *StrandTwister* on the error-free simulated β-sheet density shows the potential of our β-strands detection using the principle of β-sheet twist. This section examines the performance of β-strand detection using 39 β-sheets, a large dataset from experimentally derived cryo-EM maps. Seven cryo-EM maps from EMDB with resolutions between 4.4 and 7.4Å were

collected. All seven density maps were aligned with their corresponding PDB structures at download except EMD_1237 which was aligned manually with the help of "Fit in Map" function of UCSF Chimera [117]. The density region belonging to a chain of a protein was isolated from the molecular complex, such as a virus, using the PDB structure as an envelope. Once the density of the entire chain of a protein was obtained, *SSEtracer* [116] was used to identify β-sheets from it. Such identified β-sheet density voxels (shown as a blue surface view in Figure 16B) were then forwarded to *StrandTwister*. The results in this section represent the performance of both *SSEtracer* and *StrandTwister*, since the error at either step will affect the results in Table 11.

Figure 21D shows the density detected by *SSEtracer* from a cryo-EM map at 7.4Å resolution. At this resolution, the separation of β-strands is not visible, and some regions may have weak/missing density (arrows) due to the error in data or in the identification of β-sheet. However, the mathematical surface fitting appears to compensate the missing density to some extent, since the surface is based on the overall density distribution of the detected β-sheet. *StrandTwister* was able to detect all five strands in 2165_4B4T_1C (Figure 21D), and they align fairly well with the observed β-traces. In this case, the 2-way distance for the five strands is only 1.60 Å, and it detected 24 of 31 amino acids on the β-sheet (Table 11).

The number of strands was correctly detected in 28 of the 39 cases (shown in Table 11 for the best of the top ten detections), in spite of the challenge from missing/extra density in the experimentally derived data. This is notable, since *StrandTwister* does not require the knowledge of the number of strands in β-sheet during detection. Two different sampling sets of β-traces may

differ in the number of strands that are determined by the width of sheet perpendicular to the strand orientation. However, right-handed twist correctly distinguished between a 3-stranded β-sheet (with longer strands) verses a 6-stranded β-sheet (with shorter strands) in the case of 1780_3IZ5_Z (Figure 21E), since sampling along the orientation perpendicular to the true orientation may result in six shorter β-strands with left-handed twist. This suggests that the number of β-strands and the position of the β-traces are intrinsic characters of β-sheet density and they are reflected in the β-twist.

Table 11. Accuracy of β-strand detection for the experimentally derived cryo-EM maps.

EMDB_PDB	Res.[b]	Fit	#Det./#Obs.	2-w	#Det./#Obs.
1237_2GSY_A	7.2	2.19	4 / 4	2.34	13 / 22
1237_2GSY_B		2.30	5 / 5	1.90	26 / 28
1237_2GSY_C		2.37	5 / 6	1.90	36 / 41
1237_2GSY_E		1.80	5 / 4	3.03	23 / 47
1237_2GSY_G		2.16	5 / 4	1.85	36 / 48
1740_3C92_A	6.8	1.29	5 / 5	1.71	24 / 27
1740_3C92_B		1.62	5 / 5	1.62	29 / 31
1740_3C92_O		1.46	4 / 5	1.04	24 / 28
1740_3C92_Q		1.47	5 / 5	1.59	26 / 27
1780_3IZ5_AC	5.5	2.15	4 / 4	1.58	21 / 25
1780_3IZ5_AH		1.33	4 / 4	1.60	14 / 15
1780_3IZ5_AI		1.60	3 / 3	1.56	18 / 22
1780_3IZ5_AS		1.44	4 / 3	1.72	12 / 19
1780_3IZ5_AT		1.35	4 / 4	1.53	19 / 20
1780_3IZ5_AY		1.61	4 / 5	1.96	12 / 16
1780_3IZ5_F		1.45	5 / 5	1.41	27 / 30
1780_3IZ5_H		1.28	3 / 3	1.32	15 / 17
1780_3IZ5_I		2.14	3 / 4	1.37	17 / 22
1780_3IZ5_J		1.35	3 / 3	1.46	12 / 14
1780_3IZ5_K		1.31	4 / 4	1.63	16 / 19
1780_3IZ5_L		1.81	4 / 5	2.17	19 / 21
1780_3IZ5_R		1.75	4 / 4	1.73	23 / 32
1780_3IZ5_W		2.00	4 / 4	1.93	19 / 23
1780_3IZ5_Z		1.65	3 / 3	1.00	27 / 28
1780_3IZ6_AF		1.51	3 / 3	0.89	14 / 14
1780_3IZ6_D		1.64	3 / 3	1.58	12 / 13
1780_3IZ6_F		2.01	4 / 4	1.95	15 / 20
1780_3IZ6_I		1.51	4 / 5	1.57	15 / 20
1829_2WWL_CA	5.6	1.46	3 / 4	1.72	15 / 21
1829_2WWL_CB		1.59	4 / 4	1.91	22 / 24
1829_2WWQ_SA		2.18	3 / 3	1.80	21 / 26
1829_2WWQ_TA		1.11	3 / 3	1.03	13 / 13
2165_4B4T_1A	7.4	1.39	4 / 5	2.18	19 / 31
2165_4B4T_1C		1.47	5 / 5	1.60	24 / 31
2165_4B4T_AA		1.28	5 / 5	1.38	23 / 26
5036_3FIH_P	6.7	1.02	3 / 3	1.94	13 / 17
5276_3J0C_K	4.4	1.16	4 / 4	1.47	20 / 22
5276_3J0C_O		1.52	3 / 3	1.60	12 / 15
5276_3J0C_T		1.57	3 / 3	1.13	17 / 17
Average		**1.62**		**1.66**	**763/932**
Standard deviation		**0.35**		**0.40**	

a. EMDB_PDB_sheet ID;
b. Resolution of the density map;
c. The RMSE (in Å) for the polynomial surface fitting.
d. The number of β-traces in the best of the top ten detected sets / the number
e. The 2-way distance (in Å) between the observed β-traces and the detected
f. The number of detected / the total number of amino acids in the β-sheet.

The results in Table 11 suggest that it is possible to detect traces of β-strands from many β-sheets identified from cryo-EM maps at medium resolutions. The overall 2-way distance for the 39 test cases is 1.66Å for the detected β-traces (Table 11), suggesting fairly good specificity. *StrandTwister* was able to detect 81.87% of the β-strand amino acids overall. Table 11 shows that once a β-sheet region is identified roughly correct, traces of β-strands can be detected fairly well. All of the sheets in Table 11 have three or more β-strands, since the identification of 2-stranded sheet is not reliable. In the case of EMD_1237_2GSY (7.2Å resolution), there is one unique chain (chain A) that contains six β-sheets. *SSEtracer* identified five β-sheets (Table 11), but missed a 2-stranded β-sheet (sheet D) that contains four amino acids. Two extra short strands were wrongly detected by *StrandTwister* due to inaccurate boundary of identified β-sheets. Note that the annotation of sheet F and sheet G in PDB file corresponds to similar region, and hence they are counted as one sheet. EMD_1740_3C92 (6.8Å resolution) has two unique chains (chain A and H) that contain five sheets. Four of the five sheets were detected well, although a 2-stranded sheet (sheet P, with seven amino acids) was missed. Three case studies (Figure 22, 23, 24, 25, 26 and 27) will analyze the details for each β-sheet in the entire chain of gp10, GroEL and E2 protein.

Detection of Four β-strands from gp10 Protein of Epsilon15

The backbone Cα trace of proteins in bacteriophage epsilon 15 was derived from the cryo-EM density map (EMD_5678) at 4.5Å resolution [80]. Its staple protein gp10 appears to contain two β-sheets (dashed outlines in Figure 23D). The larger β-sheet was predicted to contain four

strands [80]. In order to see if *StrandTwister* is able to detect the β-strands from density maps at the medium resolution, *StrandTwister* was applied on gp10 density which was extracted from epsilon 15 map (EMD_1557) [122] at 7.3Å resolution. To extract the density of gp10, the chains of PDB structure 3J40 were first manually fitted into the 7.3Å map and refined with "Fit in Map" option (Figure 22) in Chimera. The density region of gp10 (gray in Figure 23A) was manually extracted using the guidance of the fitted gp10 structure (see Figure 22). The envelope of gp10 protein can be distinguished at 7.3Å resolution (box in Figure 22B).

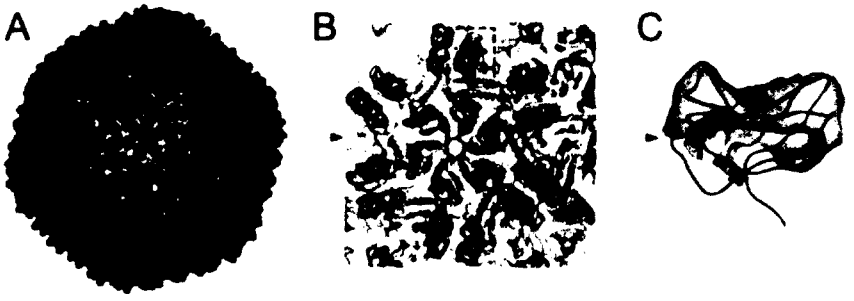


Figure 22. Staple protein gp10 density that was isolated from the cryo-EM density map of epsilon-15 bacteriophage 7.3Å resolution (EMD_1557). (A) The density map of epsilon-15 (EMD_1557) is shown in blue, green and yellow from outer surface to inner surface. (B) Zoom-in view of the highlighted region in (A) shows that gp10 is located at the outer surface of the virus. The isolation of gp10 density region was guided by the PDB structure (3J40, purple chain), which was fitted into the density map using UCSF Chimera [117]. (C) The isolated gp10 density that aligned with the PDB structure (Cα trace in purple).

At 4.5Å resolution, the two sheets are shown as separate sheets (Figure 23D), and the separation of β-strands is visible. However, this is not true for the 7.3Å resolution map (Figure 23A). The lower sheet region has weak/missing density and *SSEtracer* detected only the upper β-sheet (Figure 23A and B). The detected β-sheet voxels appear to show the twist nature of that β-sheet

(Figure 23B). *StrandTwister* successfully detected four β-traces from the detected β-sheet of gp10 (red lines in Figure 23C). Our analysis of β-twist in the 7.3Å resolution map further supports the finding of 4-stranded sheet in gp10. Note that only the Cα trace of the backbone is available in the PDB file and therefore the β-sheet is not defined. However, the Cα chain appears to resemble a 4-stranded β-sheet [80]. This case study demonstrated that it is possible to detect β-traces from a medium resolution map where the separation of the β-strands is not available. The fact that only one of the two β-sheets was detected by *SSEtracer* may suggest that the other one of them is smaller.

The best of top ten detected sets of β-traces is evaluated using the observed β-traces in corresponding PDB structures (column 1). For example, the best set detected four strands in sheet D of 1OEL (row 16). There are four strands annotated in sheet D of 1OEL. The 2-way distance for the detected four strands is 2.12Å.

"NA" refers to one of the two situations: (1) the calculation of 2-way distance is not applicable due to the missing annotation of β-sheets (beginning and ending position of strands) in the PDB file. (2) The sheet is not identified from density.

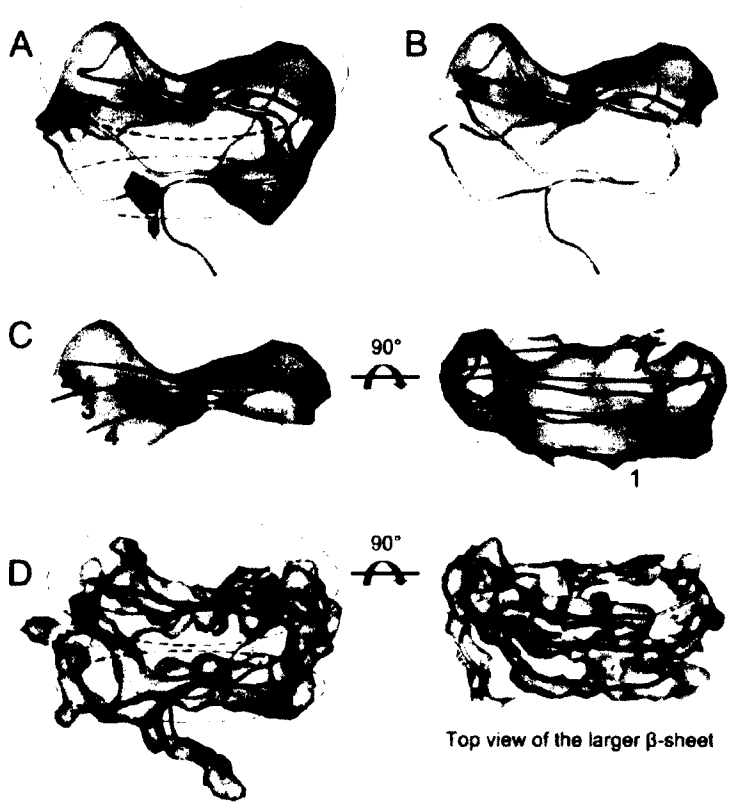


Figure 23. β-strand detection from the 7.3Å resolution map of epsilon15. (A) The density region (gray surface) of gp10 protein was extracted from bacteriophage epsilon15 density map EMD_1557 at 7.3Å resolution. The Cα chain of gp10 (chain I of PDB_3J40) is superimposed with the density; (B) The β-sheet density region detected using *SSEtracer*; (C) The β-strands detected (red lines) using *StrandTwister* are superimposed on the density of β-sheet (left: side view) and the Cα trace of β-strands (right: top view). (D) The superposition of gp10 density at 4.5Å resolution (EMD_5678) and PDB structure (chain I of 3J40) is shown as a side view (left) and also as a top view (right) at the larger sheet region. See also Figure 22 and Table 12.

The annotation of β-sheets in PDB structures can be sometimes complicated due to the flexibility in forming β-structures. If a β-strand is annotated as a single "U" shape without having a turn (AA91-103 in the original annotation of 3J0C sheet N). The annotation was forced to be considered as two separate parallel strands (AA91-97, AA98-103) in order to calculate the 2-way distance for two corresponding detected β-traces. If the strands are annotated twice in two β-

sheets, the strands was considered only once in one β-sheet (AA33-37 and AA44-55 are originally annotated in both 3J0C sheet M and sheet N in the PDB file, now they are counted only in 3J0C sheet N). In this way the strand number in column 3 of Table 12 will not be double counted.

Detection of β-strands from GroEL Density Map EMD_5001

The quality of a density map may vary from region to region, thus it is possible that not all β-strands are well detected in a map at 4-5Å resolution. The cryo-EM density map of GroEL (EMD_5001) was obtained at 4.2Å resolution, from which the Cα trace (PDB_3CAU) was derived using *ab initio* modeling [79]. There are three other GroEL structures (1SS8, 1OEL and 1XCK) that have been solved by X-ray crystallography [123-125]. Although these four structures are slightly different, they all appear to have seven β-sheets at approximately the same locations. The Cα trace of 3CAU was aligned with the three crystal structures using "Matchmaker" in Chimera. The main difference among the four structures appears to be at the upper domain, which contains sheet C, D and E (Figure 24A and B).

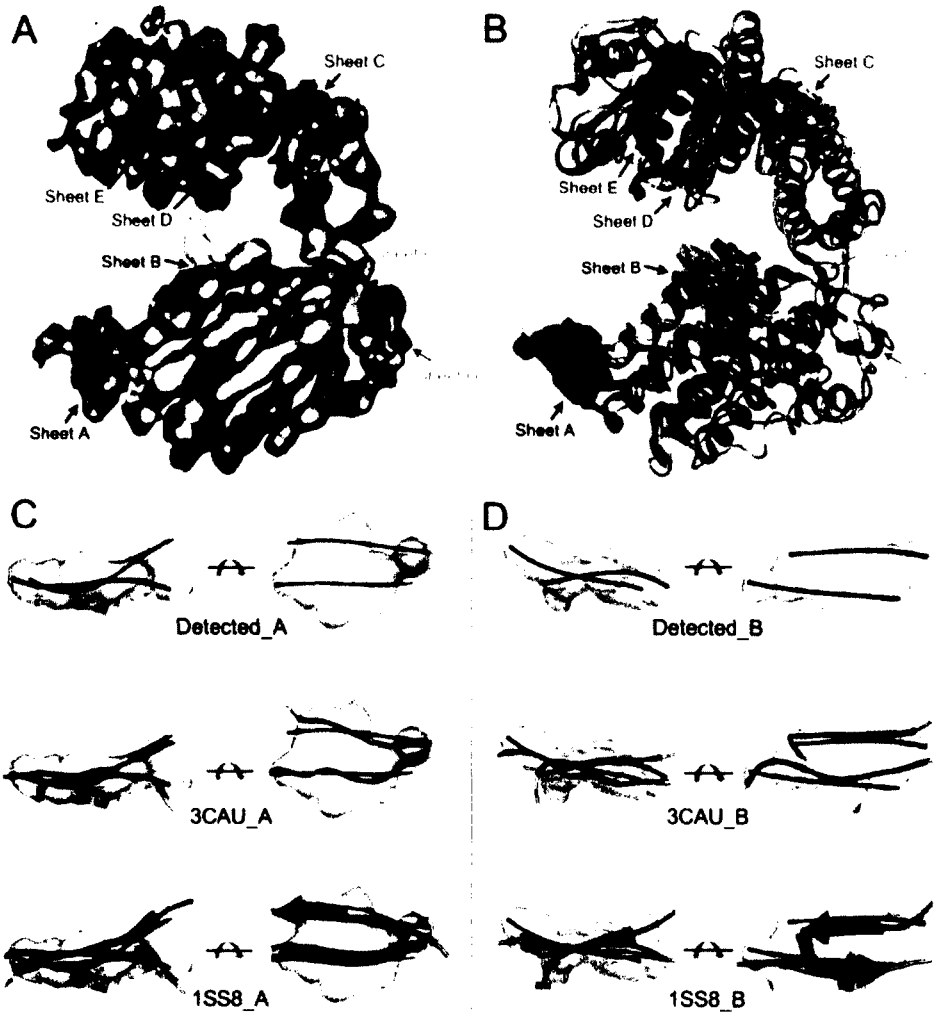


Figure 24. β-strand detection from the density map of GroEL at 4.2Å resolution. (A) The monomer density (gray) of GroEL extracted from density map EMD_5001; (B) Five β-sheet density regions (colored density) identified using *SSEtracer* are superimposed on chain A of PDB_3CAU (purple Cα trace) and chain A of PDB_1SS8 (cyan ribbon). The side view (left) and the top view (right) of the β-traces detected (red lines) using *StrandTwister* are superimposed with PDB structures of 3CAU (purple Cα trace) and 1SS8 (cyan ribbon) for sheet A in (C) and sheet B in (D). See also Figure 25 and Table 12.

SSEtracer detected five of the seven sheets from the density monomer of EMD_5001 (Figure 24B). Two 2-stranded β-sheets (F and G) were missed due to the fact that a 2-stranded sheet can be confused with a helix (Figure 24A, pointed by orange arrows). Since the Cα trace of 3CAU

does not have an annotation of secondary structures, the annotation was used in the other three X-ray structures to estimate the number of strands in the β-sheets. However, the beginning and ending position of β-strands vary among the three structures. *StrandTwister* detected fourteen of nineteen traces of β-strands from the density monomer of GroEL (Table 12 and Figure 24 and S4). The detected β-traces for β-sheet A and B appear to agree well with those in the four structures (Figure 24C and D). In fact, the 2-way distance is only 1.28Å and 1.13Å respectively for 1SS8_A and 1SS8_B (Table 12). In terms of sheet C, D, and E, the annotation of β-sheets is different among the X-ray structures (see description in Figure 25).

The Cα model (3CAU) was derived directly from the cryo-EM density map of GroEL at 4.2Å [79]. Three other GroEL structures (1SS8, 1OEL, and 1XCK) were solved using X-ray crystallography [123-125]. 3CAU (second row) appears to be different from the other three X-ray structures (3^{rd}-5^{th} row) in the region of sheet C, D, and E. The annotation of β-sheets and β-strands are also slightly different (arrows in B and C) among the three X-ray structures. The detected β-traces appear to align better with 1SS8, 1OEL and 1XCK in sheet C and sheet D than with 3CAU. The orientation and the position of the detected β-traces appear to align well with the X-ray structures, particularly in sheet C and D. The length of a detected β-trace may not be accurate when the outline of the β-sheet is not accurately identified (arrows in Figure 25). *StrandTwister* detected three of four β-traces in sheet E (Figure 25C) because the β-sheet was identified smaller using *SSEtracer* with respect to the X-ray structures.

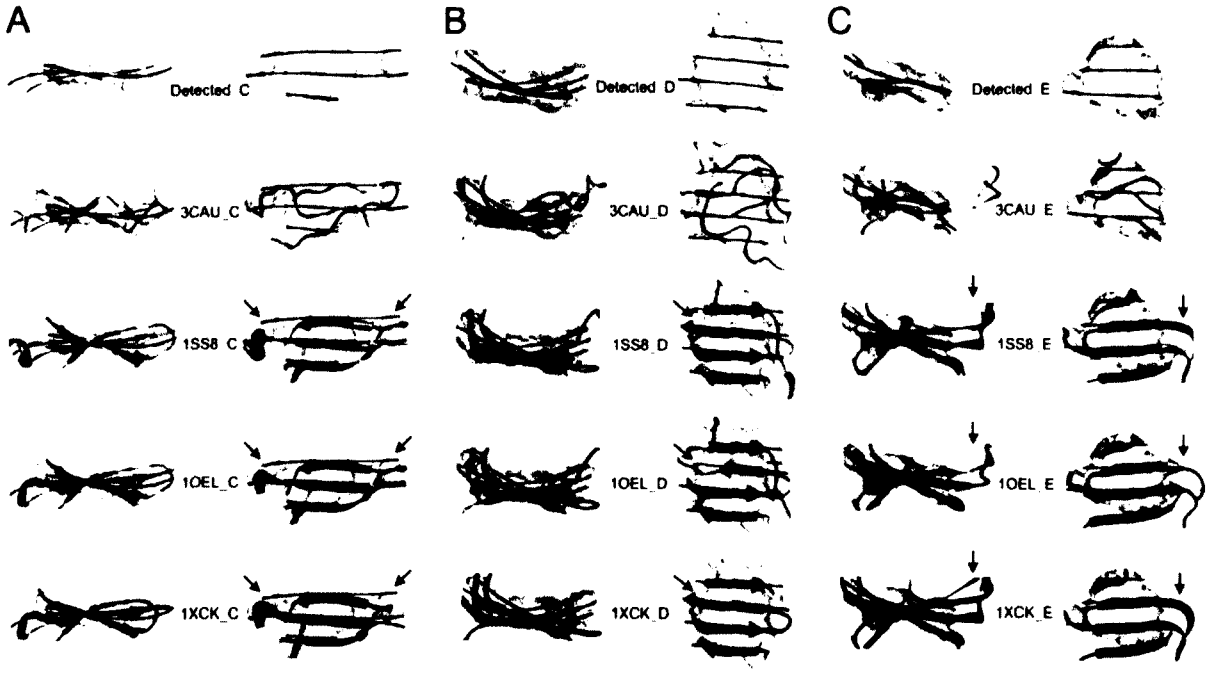


Figure 25. β-strand detection on β-sheet C, D and E of GroEL cryo-EM density map at 4.2Å resolution (EMD_5001). The side view (left) and the top view (right) of the detected β-traces (red lines) using *StrandTwister* are superimposed with the β-sheet density identified using *SSEtracer* and the observed β-strands of sheet C in (A), sheet D in (B) and sheet E in (C) for PDB structure 3CAU (2nd row), 1SS8 (3rd row), 1OEL (4th row) and 1XCK (5th row). The arrows in (A) indicate the extra density identified as β-sheet. Arrows in (B) and (C) indicate the annotation difference among the three PDB structures.

Figure 25 shows the set of β-traces best align with strands in 1OEL. Note that the detected β-traces in Figure 25 agrees well with that of the X-ray structures (1SS8, 1OEL, 1XCK), although the beginning and ending positions may differ. This suggests that β-twist captures the property of orientation and position of the strands but not precise enough about the start and end of a β-strand. The accurate outline of the β-sheet region is currently required. When the start and end of a strand is not accurately detected, 2-way distance will be affected. For example, the 2-way distance for 1SS8_C is 2.46Å (Table 12), which is larger than that of 1.13Å for 1SS8_B. This is

mostly due to the fact that the detected traces are much longer than the observed ones (arrows in Figure 25A). Note that the detected β-traces do not align with the Cα trace of 3CAU at sheet D (row 2 in Figure 25B). In fact, none of the top ten detected sets align well with 3CAU at sheet D.

Table 12. Accuracy of the detected β-traces in gp10, GroEL and E2 with respect to differently annotated β-sheets.

No.	EMDB_PDB_Sheet ID	Figure ID	#Det./#Obs. Strd	2-w Dist. (Å)
1	1557_3J40_Gp10_upper	Figure 23	4 / 4	NA
2	1557_3J40_Gp10_lower		0 / 4	NA
3	5001_3CAU_A	Figure 24	2 / NA	NA
4	5001_3CAU_B		2 / NA	NA
5	5001_1SS8_A		2 / 2	1.28
6	5001_1SS8_B		2 / 2	1.13
7	5001_3CAU_F		0 / NA	NA
8	5001_3CAU_G		0 / NA	NA
9	5001_3CAU_C	Figure 25	3 / NA	NA
10	5001_3CAU_D		4 / NA	NA
11	5001_3CAU_E		3 / NA	NA
12	5001_1SS8_C		3 / 3	2.46
13	5001_1SS8_D *		4 / 4 *	2.17
14	5001_1SS8_E *		3 / 4 *	2.73
15	5001_1OEL_C		3 / 3	2.51
16	5001_1OEL_D		4 / 4	2.12
17	5001_1OEL_E		3 / 4	1.99
18	5001_1XCK_C		3 / 3	2.76
19	5001_1XCK_D *		4 / 4 *	2.16
20	5001_1XCK_E *		3 / 4 *	2.65
21	5276_3J0C_N *	Figure 27	3 / 4 *	2.21
22	5276_3J0C_K		4 / 4	1.47
23	5276_3J0C_O		3 / 3	1.60
24	5276_3J0C_T		3 / 3	1.13
25	5276_3J0C_L		0 / 2	NA
26	5276_3J0C_M *		0 / 2 *	NA
27	5276_3J0C_P		0 / 2	NA
28	5276_3J0C_Q		0 / 3	NA
29	5276_3J0C_S		0 / 3	NA
30	5276_3J0C_R	Figure 26	4 / 4	2.14

*Note:
1SS8_D and 1XCK_D are annotated as 4-stranded sheet: AA192-195, AA330-335, AA320-325, AA213-216;
1SS8_E and 1XCK_E are annotated as 4-stranded sheet: AA318- 319, AA219-227, AA247-254, AA273-277;
3J0C_N is annotated as 4-stranded sheet: AA33-37, AA44-55, AA91-97, AA98-103;
3J0C_M is annotated as 2-stranded sheet: AA61-70 and AA73-78.

Detection of β-strands from E2 of Venezuelan Equine Encephalitis Virus Density Map EMD 5276

The structure of Venezuelan equine encephalitis virus (VEEV) was resolved from the 4.4Å resolution cryo-EM density map. E2 contains a transmembrane helix and thirty β-strands that are on ten β-sheets [126] (see Table 12 for details of annotation). The density monomer of E2 was isolated which aligned with chain B of 3J0C. *SSEtracer* detected five larger β-sheets (N, K, O, T and R) (Figure 27). Three 2-stranded β-sheets and two 3-stranded β-sheets were missed. Sheet Q (3-stranded) is mostly a 2-stranded twist and appears as a helix in the density. Sheet S (3-stranded) is located at the outer surface of E2 (Figure 27A, B and 26A) where the density is weak and has no obvious sheet property. Gaussian filter (Figure 26) was applied to the outer-most domain to enhance the weak density and was able to detect sheet R (Figure 27G). *StrandTwister* detected 17 of 30 β-traces in E2, suggesting that the majority of the β-strands can be detected for β-sheets with three or more β-strands. The detected β-traces align well with the corresponding observed β-strands on sheet K, O and T, with 1.47Å, 1.60Å and 1.13Å 2-way distance respectively (Table 12).

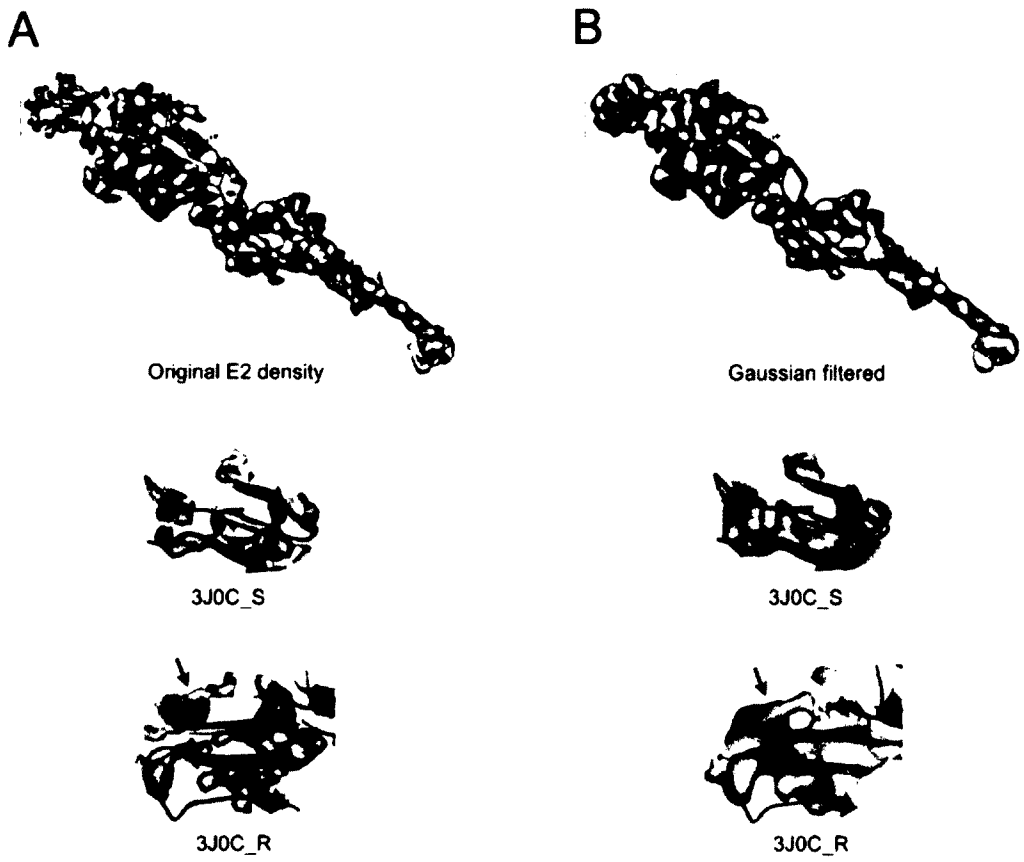


Figure 26. Density quality variation in E2 extracted from cryo-EM density map of Venezuelan Equine Encephalitis Virus at 4.4Å resolution (EMD_5276). (A) The density of E2 in EMD_5276 (top) shows that its outer-most domain (box) containing sheet R and S has weaker density than other regions of E2. The density region at sheet R and S is superimposed with the corresponding structures. (B) The density of E2 after Gaussian filtering and is represented similarly as in (A). Note that there is an extra density area for which no corresponding PDB structure can be found even from the neighboring chain (arrow).

The variation of density quality in E2 was observed in [126] and it is also shown in Figure 26A. In order to identify sheet S and R, Gaussian filter (at Width=1.07Å) was applied to enhance connectivity in the density. *SSEtracer* was able to detect sheet R after Gaussian filter was applied. Even after smoothing, sheet S does not show the characters of a typical β-sheet (Figure 26B middle).

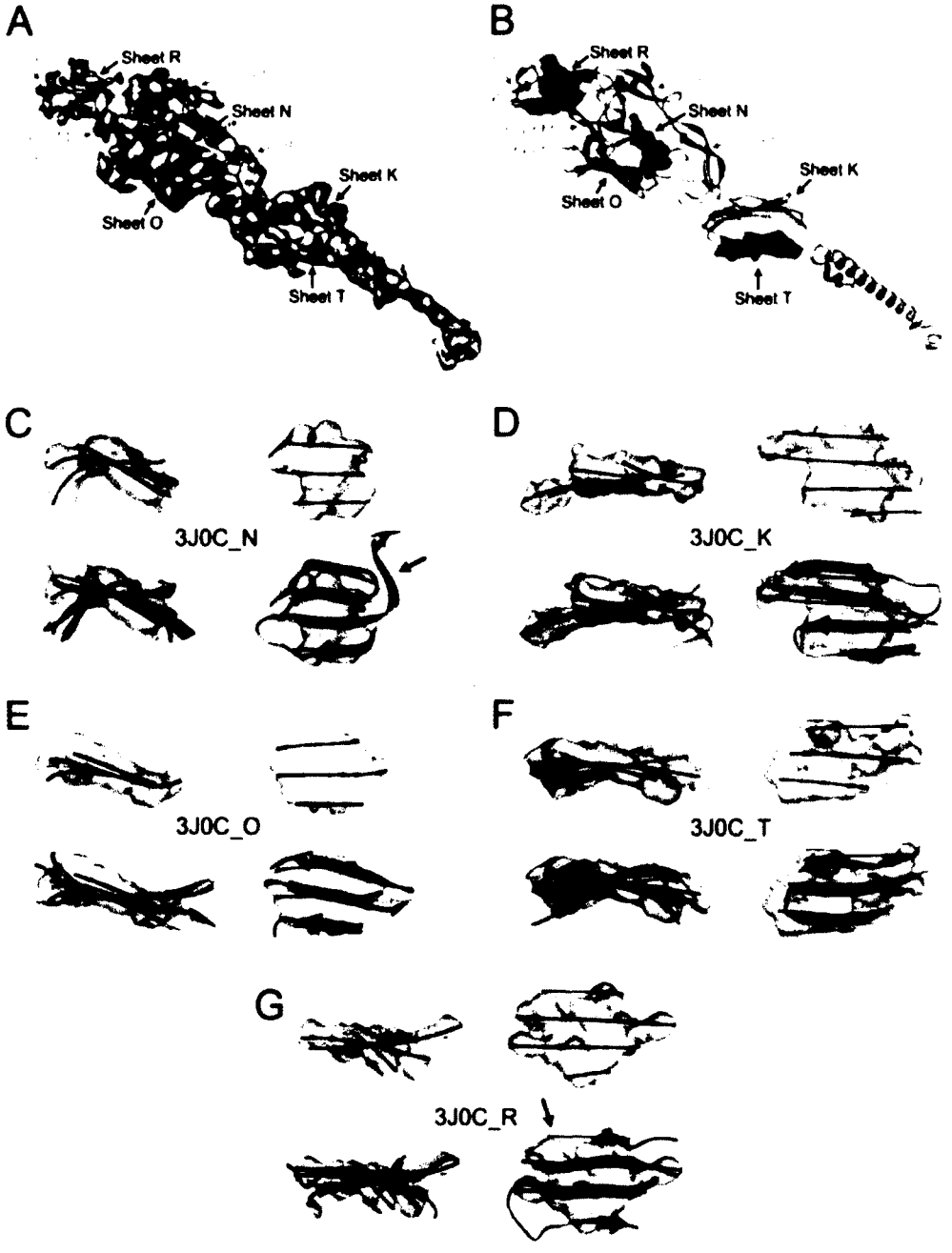


Figure 27. β-strand detection from the density map of E2 in Encephalitis Virus (EMD_5276). (A) E2 monomer density (gray) at 4.4Å resolution; (B) β-sheet density regions (colored density) identified using *SSEtracer* are superimposed on chain B of PDB_3J0C (cyan ribbon); The side view (left) and the top view (right) of detected β-traces (red lines) using *StrandTwister* are superimposed with the observed β-strands of sheet N in (C), K in (D), O in (E), T in (F) and R in (G). The density shown in (G) was obtained after applying Gaussian filter to enhance density connectivity at the outer domain. See also Figure 26 and Table 12.

E. Cα Models Derived from β-traces

A rapid method has been developed to construct the backbone from the density of a β-sheet for situations when the resolution of the density map is high enough to resolve the density pattern of a β-strand (i.e. ~3.8Å resolution) [127]. At the medium resolutions however, multiple possible Cα models may be derived from a β-trace. *StrandTwister* produces a possible Cα model from a set of β-traces. A test using 39 sets of β-traces detected from cryo-EM density maps shows that the models have good overall accuracy of 2.56Å RMSD for 84% of the Cα atoms in the true β-sheets. This error is reasonable since the input density maps have resolutions around 4.4-7.4Å.

Construction of the Cα model from β-traces

A method has been investigated to construct the Cα model by enforcing both β-traces and the general rules observed from true structures of β-sheets, so that the model is along the β-traces and appears as β-strands. To investigate the effectiveness of this method, the accuracy of the Cα models built for 39 sets of β-traces detected from cryo-EM maps was evaluated. Each set of β-traces is the best detected β-traces with the accuracy reported in Table 11.

Table 13. Accuracy of the Cα model built for the β-sheet density.

No.	PDB_ID	#Det./#Obs.[a]	RMSD-M[b]	#Match/#Total[c]	#Match/#Built[d]
1	1237_2GSY_A	4/4	2.76	17/22	17/20
2	1237_2GSY_B	5/5	2.43	28/28	28/39
3	1237_2GSY_C	5/6	2.53	34/41	34/46
4	1237_2GSY_E	5/4	4.04	40/47	40/56
5	1237_2GSY_G	5/4	3.35	38/48	38/40
6	1740_3C92_A	5/5	2.87	27/27	27/39
7	1740_3C92_B	5/5	2.72	30/31	30/40
8	1740_3C92_O	4/5	1.81	24/28	24/31
9	1740_3C92_Q	5/5	2.75	27/27	27/38
10	1780_3IZ5_AC	4/4	1.86	20/25	20/23
11	1780_3IZ5_AH	4/4	2.35	14/15	14/19
12	1780_3IZ5_AI	3/3	2.19	18/22	18/18
13	1780_3IZ5_AS	3/3	2.82	15/19	15/16
14	1780_3IZ5_AT	4/4	2.74	18/20	18/24
15	1780_3IZ5_AY	4/5	1.98	12/16	12/19
16	1780_3IZ5_F	5/5	2.16	26/30	26/34
17	1780_3IZ5_H	3/3	1.94	16/17	16/18
18	1780_3IZ5_I	3/4	2.04	18/22	18/21
19	1780_3IZ5_J	3/3	2.59	12/14	12/13
20	1780_3IZ5_K	3/4	2.21	13/19	13/15
21	1780_3IZ5_L	4/5	2.40	19/21	19/31
22	1780_3IZ5_R	4/4	2.57	22/32	22/22
23	1780_3IZ5_W	3/4	2.79	13/23	13/22
24	1780_3IZ5_Z	3/3	2.13	22/28	22/22
25	1780_3IZ6_AF	3/3	2.04	14/14	14/17
26	1780_3IZ6_D	3/3	2.21	11/13	11/16
27	1780_3IZ6_F	4/4	2.70	17/20	17/23
28	1780_3IZ6_I	4/5	2.34	16/20	16/22
29	1829_2WWL_CA	4/4	2.77	16/21	16/20
30	1829_2WWL_CB	4/4	2.96	23/24	23/29
31	1829_2WWQ_SA	4/3	3.23	20/26	20/28
32	1829_2WWQ_TA	3/3	2.38	12/13	12/14
33	2165_4B4T_1A	4/5	3.45	21/31	21/22
34	2165_4B4T_1C	5/5	2.85	28/31	28/30
35	2165_4B4T_AA	5/5	2.56	21/26	21/21
36	5036_3FIH_P	2/3	2.73	10/17	10/10
37	5276_3J0C_K	4/4	2.29	21/22	21/25
38	5276_3J0C_O	3/3	3.17	13/15	13/15
39	5276_3J0C_T	3/3	2.27	17/17	17/24
	Average		**2.56**	**783/932 = 84.0%**	**783/982 = 79.7%**

a. The number of detected β-traces / the number of β-strands in the true structure;
b. The RMSD distance (in Å) of matched Cα atoms between the model and the β-strands;
c. The number of Cα atoms in the model that are matched to their corresponding Cα atoms in the β-sheet / the total number of Cα atoms in the true β-sheet structure;
d. The number of Cα atoms in the model that are matched to their corresponding Cα atoms in the β-sheet / the total number of Cα atoms built in the model.

The accuracy of the Cα model was evaluated mostly from two aspects: the number of the matched Cα atoms out of the total number of atoms in the observed β-sheet (column 5 of Table 13), and the RMSD for the matched Cα atoms. For example, there are 28 Cα atoms in the β-sheet of 1740_3C92_O (row 8 of Table 13). The Cα model has 1.81Å RMSD for 24 matched Cα atoms. It missed 4 atoms which are not included in the calculation of the RMSD. The model contains 31 Cα atoms suggesting that one or more β-strands in the model are slightly longer than that in the true structure (Figure 28B). Overall, the models has fairly good accuracy of 2.56Å RMSD for 84% of the β-sheet structure. The models have similar size as the true β-sheets overall, since the number of matched atoms covers about 80% of the atoms built in the models.

β-traces detected from the density of a β-sheet were used to derive the Cα model for a β-sheet. Cα atoms were generated starting from the middle of the center β-trace, with later generations moving towards the two ends of the β-trace. The rise of Cα atoms along a β-trace was used as 3Å [128], an approximated rise in a β-strand. Cα atoms were built around each β-trace according to the following rules:

(1) In order to approximate the alignment between two sets of Cα atoms from two neighboring strands, the initial Cα atom on each strand should be approximately aligned with that of a neighboring strand.

(2) The distance between two adjacent Cα atoms is between 3.75Å and 3.8Å.

(3) The angle formed by three consecutive Cα atoms is between 100° and 150° [129].

(4) An ending Cα of a β-trace is assigned if it is at least 3Å away from the last built Cα atom. Note that this distance might be slightly shorter from the typical distance between two Cα atoms.

A translation of 1.5Å up/down from the initial Cα was performed and three models were generated for each β-sheet. The one with the best RMSD-M is summarized in Table 13, although there is not much difference among the three models in terms of RMSD.

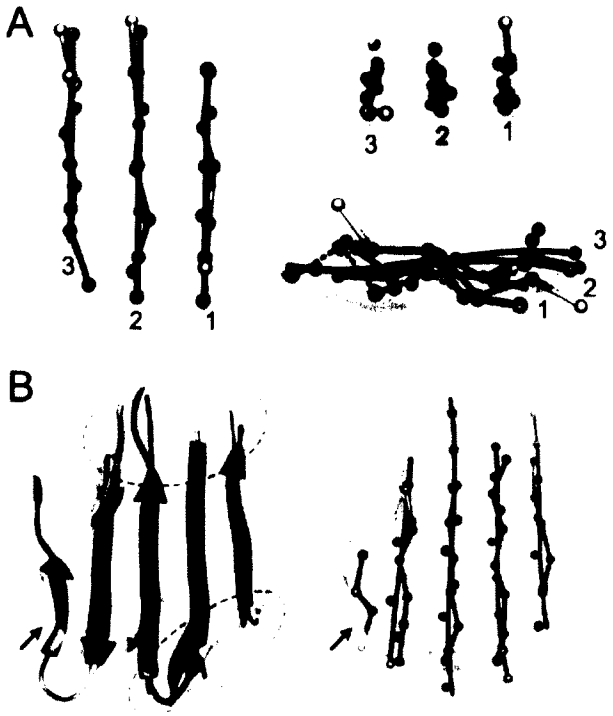


Figure 28. Cα model derived from the detected β-traces. (A) The Cα atoms (pink, green and blue) derived from the β-traces (red lines) detected from the β-sheet density (gray) are superimposed on the true β-strands (golden) of sheet 1780_3IZ5_H. The Cα atoms built for different strands are shown in different colors. For easy viewing, the Cα atoms built for β-trace 1 (blue) were connected with dashed lines. (B) The miss-detected strand (arrow) and the over-detected region (dashed lines) are indicated for sheet 1740_3C92_O. The Cα atoms derived from the β-sheet density are shown on the right panel using similar color labeling scheme as in (A).

Examination of the Accuracy for the Cα Model Built for the Density of a β-sheet

RMSD of Cα atoms is a common parameter to estimate the accuracy of a conformation. Typically, the two models have the same number of points; the two sets of points also have one-to-one correspondence. Since the model constructed from the β-sheet density often has different number of Cα atoms from that of the observed β-sheet, a subset of Cα atoms was first searched so that they can be one-to-one corresponded to a subset of Cα atoms on the true β-sheet. For example, if the model misses a β-strand in the detection (arrow in Figure 28B), the Cα atoms in the missed strand are not included in the set "matched Cα atoms". Once the detected β-traces had a one to one correspondence with the β-strands in the true structure, the Cα atoms on each strand were examined for one-to-one correspondence. Given a model β-strand C_i and its corresponding true β-strand β_i, the smaller number of Cα atoms was adopted as the matched number of Cα atoms, M_i, for the particular strand. To determine the matched Cα atoms between C_i and β_i, the best subset containing M_i atoms were searched based on the overall distance. RMSD was calculated for the matched Cα atoms (Table 13).

CHAPTER V

BUILDING THE BETA-BARREL STRUCTURE FROM 3D CRYO-EM DENSITY IMAGES

Electron cryo-microscopy (Cryo-EM) has become a major experimental technique to study the structures of large protein complexes, such as ribosomes and viruses [26, 130]. It is a structure determination technique complementary to the X-ray Crystallography and Nuclear Magnetic Resonance (NMR). At the medium resolutions such as 5-10Å, detailed molecular features are not resolved. However, secondary structure features such as α-helices and β-sheets (Figure 29) can be computational identified. The α-helix appears as a stick (red in Figure 29A) and can be identified using image processing methods [70, 72, 86, 113]. A β-sheet appears as a thin layer of density and can be detected computationally (blue in Figure 29B) [72, 76, 86, 131]. Some β-sheets curve into β-barrels. β-barrel structures are commonly found in porins and other proteins that span cell membranes [132]. A β-barrel is composed of multiple β-strands (ribbon of Figure 29C and Figure 30) that twist and coil to form a closed structure in which the first strand is hydrogen bonded to the last.

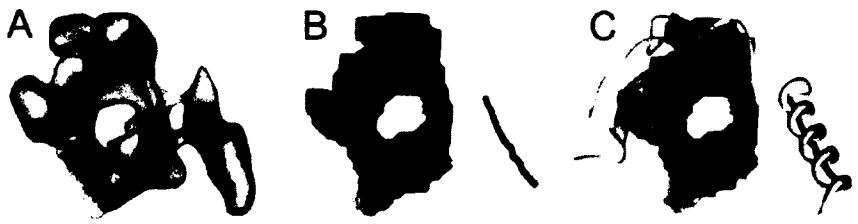


Figure 29. Three-dimensional protein density image, the secondary structures (SSEs), and a β-barrel. (A) Protein denisty image simulated using EMAN [102] with protein 3GP6 from the Protein Data Bank; (B) the computationally detected helix (red line) and β-barrel region

Figure 29. (Continued)
(blue voxels) using *SSEtracer* [131]; (C) the atomic structure of the protein (shown in ribbon) overlaped with the detected SSEs.

This chapter is a summary of the *StrandRoller* methodology published in paper [133].

1. Motivation

Although β-sheets can be detected from cryo-EM density images at 5-10Å, it is almost impossible to detect the β-strands, the components of a β-sheet. The spacing between two neighboring β-strands is between 4.5 and 5Å, and therefore they are not visible when the resolution is at 5-10 Å [77, 78]. Image processing techniques can be used to model the β-strands when the separation of β-strands is visible, if the resolution of the image is higher than 5Å [127]. However, such method failed to detect β-strands at the medium-lower resolutions when there is no separation at all. The detection of β-strands from medium-lower resolution images has been a hard problem since it was first proposed in 2004 [110]. There has been no solution to this problem. In this chapter, an alternative approach was proposed to incorporate a modeling method for addressing this problem. Although the exact β-strands are impossible to detect directly from such images, *StrandRoller* shows that it is possible to generate a small sets of possible traces for the β-strands. A novel method *StrandRoller* was proposed, to generate the traces of β-strands based on the intrinsic nature of β-barrel.

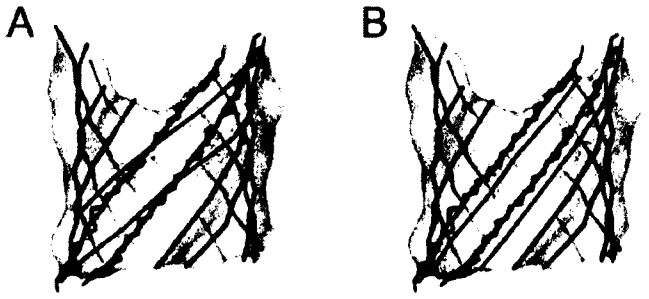


Figure 30. β-strands of a β-barrel image. A set of β-traces is shown in black lines at the front and gray lines at the back in (A) and (B). Two sets of possible β-traces (represented by two black lines and two red lines) may have different orientations shown in (A) or locations shown in (B). The atomic structure of the two β-strands is superimposed on the two representative β-traces in (A) and (B).

A helix detected from the medium resolution data is often represented as a line (red line in Fig 1B and C), referred as an α-trace that corresponds to the central axis of a helix. The β-trace (black line in Figure 30) is defined as the central line along a β-strand. In particular, the observed β-trace is the line interpolating all geometrical centers of three consecutive Cα atoms on a β-strand plus the two Cα atoms at the end of the β-strand. An observed β-trace represents the line along the atomic structure of β-strand. Given the image of β-barrel image voxels, the problem of β-strands detection is to find the orientation (Figure 30A) and location (Figure 30B) of the β-traces from the three-dimensional density image.

2. Methodology

A. β-barrel Surface Modeling from Cryo-EM Image

β-barrels have characteristic shapes and have been modeled mathematically in previous studies. The atomic structure of a β-barrel has been modeled as hyperboloid surfaces [92, 134, 135] and catenoid surfaces [136]. All these methods concentrated on the fitting of a particular

mathematical model to the β-barrel structures by using linear or non-linear fitting procedure. Although these models can approximate the major area of a β-barrel, the cryo-EM images of β-barrels often deviate from the rigid mathematical models in certain area. An adaptive method was presented to generate a surface that can fit in the three-dimensional image of a β-barrel. The idea is to use a rigid model for area that fit well and then optimize the model on where it does not fit.

For the region of density that related to a β-barrel, least-square procedure was first performed to find the central axis of the barrel by fitting an elliptical cylinder to it (Figure 31A).

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{8}$$

The purpose of fitting a cylinder here is to find the Z axis of the β-barrel. Then the surface model of the β-barrel was built from bottom to top. The density voxels on each cross-section of Z axis (Figure 31B, gray) look like a round belt, and the fitted elliptical cylinder on each cross-section is an ideal ellipse. The voxels that located roughly around the ellipse were saved as the surface model (Figure 31B, yellow), if the ellipse is within the density voxels on each cross-section. For the region where the fitted elliptical cylinder is outside the density (arrow in Figure 31B), the closest voxels to the ellipse were searched and saved into the surface model. The surface model was built one layer by one layer until reach the top of the β-barrel density.

Our modeled barrel surface clearly follows to the morphed regions (arrows in Figure 31D) of the density image. The more accurate barrel surface makes the β-strand modeling more accurate.

B. Strand Traveling on the Modeled Barrel Surface

It was first noticed by MacLachlan in 1979 that the specification of the number of strands and their relative stagger completely determines the overall structure of a β-barrel [137]. The main structural characteristics of ideal β-barrel have been discussed based on a cylindrical barrel [137-139]. In short, β-barrel forms a closed cylindrical barrel. In all known β-barrel structures, the β-strands are right-twisted, and in order to satisfy hydrogen bonding, each β-strand is right-tilted with respect to the membrane normal axis. Studies have shown that the tilt angles α of the β-strands can vary within certain bounds, between 30° and 60° that relative to the barrel axis as reflected in the known structures of membrane proteins [139-141]. Note the tilt angles α of β-strands can even vary by $\pm 15°$ around the same β-barrel [141]. However, the inter-strand distance d remains the same due to the hydrogen bonding pattern of β-strands. In general, the inter-strand distance d is roughly between 4.5 and 5Å. These two important statistical parameters (tilt angle α and inter-strand distance$d$) formed the fundamental basis of our β-barrel modeling from cryo-EM density images. Since the real β-barrel is not always an ideal cylinder with fixed radius, the β-strands was built based on the surface model and combined with the right-handed tilt feature of β-barrel proteins.

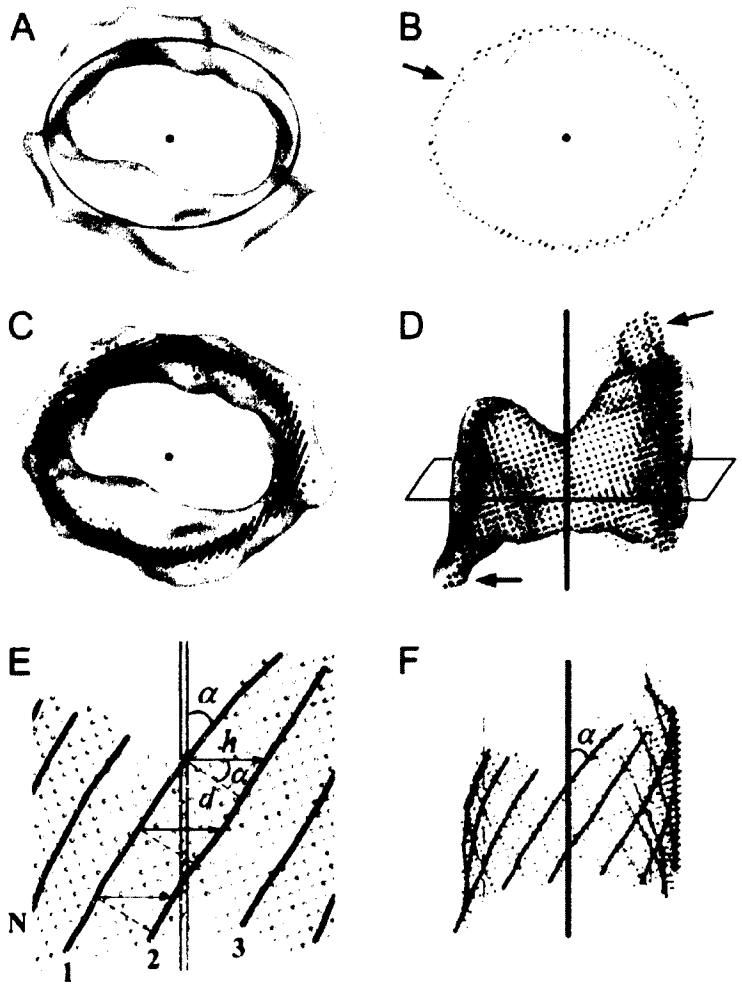


Figure 31. Modeling the surface and building the β-strands from 3D β-barrel image. (A) The barrel axis (red) was searched by fitting an elliptical cylinder (line) to the density image (gray), shown as the top view; (B) one cross-section of the barrel, arrow shows the shrinking area of the surface model according to the morphed density; (C) top view of the modeled barrel surface (yellow); (D) side view of the barrel surface that modeled from the density (yellow), the barrel axis (red) and one cross-section of the β-barrel; (E) recursive generation of β-traces based on the tilt angle α and the side-way distance d of β-strands; (F) the entire set of β-traces for a certain tilt angle, the zoomed in view of the front portion is shown in (E).

Firstly, an initial β-trace was generated on the modeled surface by tilting the barrel axis to a certain initial angle α and then projected the tilted axis onto the surface (blue in Figure 31E). The second β-trace was then generated by traveling a horizontal distance h on the surface (Figure 31E). Here the horizontal distance was estimated as:

$$h = d/\cos\alpha, d = 4.8\,\text{Å} \quad (9)$$

Iteratively, the entire set of β-traces was built on the barrel surface one by one around the barrel surface (Figure 31F), until the last β-trace (Figure 31E). The tilt angles were sampled for every 5° among a small range of 35° to 55°, and three translations were also sampled for each tilt angle sampling. The horizontal distance h was equally divided into three segments and each segment equals a translation distance. Totally there are fifteen sample sets of β-traces.

3. Result

Our method *StrandRoller* was tested on eleven density images of β-barrel that simulated to 10Å resolution and one experimental derived cryo-EM density image from EMDB (http://www.emdatabank.org/) at 6.7Å resolution. To evaluate the accuracy of our β-trace detection, the 2-way distance was calculated between the set of detected β-traces and the set of observed β-traces. The observed β-trace is the line interpolating all geometrical centers of three consecutive Cα atoms on a β-strand plus the two Cα atoms at the end of the β-strand, as shown in Figure 30. The concept of 2-way distance was previously used to measure the error between two sets of points [40]. In order to calculate the 2-way distance, the 1-1 correspondence between the β-traces in the detected set and those in the observed set was first determined based on the overall smallest distance. This ensures that the same number of detected β-traces $(S_1, S_2, \ldots, S_T)$ are compared to the observed traces $(S'_1, S'_2, \ldots, S'_T)$ in which S_k is compared with S'_k for $k = 1, \ldots, T$. The number of miss-detected (or wrongly detected) β-strands can be inferred from the difference between the total number of the observed and that of the detected β-strands. D_k, the 2-

way distance of a β-strand k, was calculated for each pair of lines S_k and S'_k. The overall 2-way distance D reflects the quality of the detected β-traces that are corresponded to their observed ones.

$$D_k = \left(\sum_{i=1}^{N} D_i^{SS'}/N + \sum_{j=1}^{M} D_j^{S'S}/M\right)/2 \quad (10)$$

$$D = \left(\sum_{s=1}^{T} D_k\right)/T \quad (11)$$

In formula (3), N and M are the numbers of points on the detected and the observed β-traces S_k and S'_k respectively. i and j are the indexes of the point along the lines S_k and S'_k respectively. $D_i^{SS'}$ is the projected distance from point i of line S_k to line S'_k. The projection of point i was required to be within line S'_k. In case it is outside, the distance between point i and the end of line S'_k was used to approximate the error.

The purpose of this test is to investigate if the β-strands of β-barrel can be modeled by our method from the medium resolution density images simulated at 10Å resolution, at which the separation of β-strands is completely not visible. The dataset was mainly collected from the β-barrel transmembrane super family of Orientations of Proteins in Membranes (OPM) database [142] with less than 40% sequence similarity. The atomic structures of β-barrels were used to generate the β-barrel density images at 10Å resolution using EMAN [102], a popular software to produce simulated density, with a sampling of 1Å/pixel.

Figure 32 shows the best of the fifteen modeled β-traces (red) for three cases with 12, 16 and 5 β-strands respectively. In the case of sheet A of PDB structure 4FQE, the detected set of β-traces

appears to align with the β-strands very well (Figure 32A). In this case all the twelve strands were detected with a small 2-way distance of 1.25Å (Table I). It is observed that, the fifteen sampled sets of β-traces always include a set with close orientation and location to the actual β-strands" orientation and location. The test shows the ability of our *StrandRoller* for modeling various sizes of β-barrels with range from 5 strands to 16 strands. Although the detection is slightly better for the smaller β-barrels, some large β-barrels were still well detected, such as the 16-stranded β-barrel 2J1N_AA18 (Figure 32B). The 2-way distance is only 1.67Å in this case (Table I), and all of the strands are accurately detected. The error appears to be at the edge of the β-barrels (arrows in Figure 32B), where the β-strands tend to be more flexible.

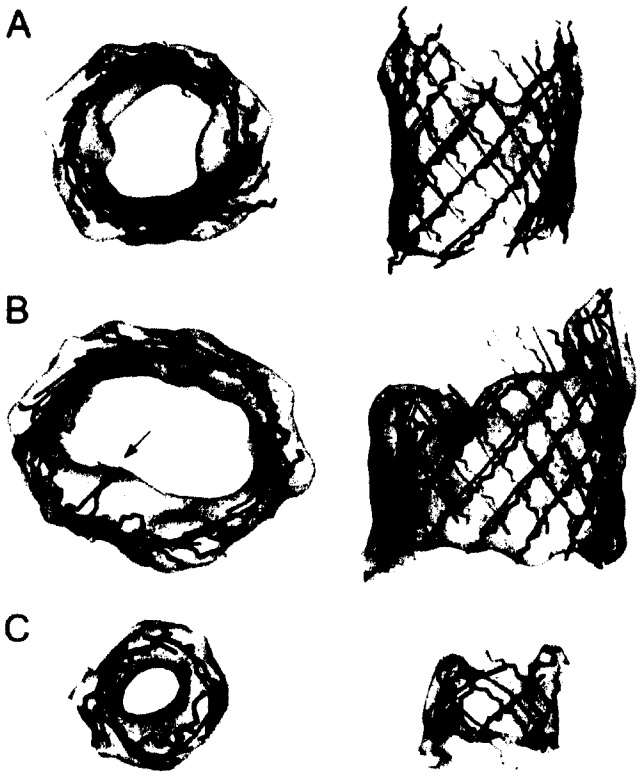


Figure 32. β-strands modeling from the simulated density image at 10Å and one experimental derived image of β-barrel. The best of the fifteen sets of modeled β-traces (red) are superimposed with the back-bone of the β-strands (blue) and the density images (gray) for β-barrels 4FQE_A in

Figure 32. (Continuerd)
(A) and 2J1N_AA18 in (B). The top view (left) and the side view (right) are shown in each case. A experimental derived cryo-EM image of β-barrel (EMD_5036, sheet H of protein 3FIK) is shown in (C).

Figure 32C shows the density region of a β-barrel (3FIK_H) from the experimentally derived cryo-EM image at 6.7Å resolution (EMD_5036). At this resolution, single β-strands are not visible. *StrandRoller* was able to detect all five strands, and they align fairly well with the observed β-traces. In this case, the 2-way distance for the five β-strands is only 1.6Å, and it detected 27 of 29 amino acids on the β-barrel.

Table 14. Accuracy of β-barrel modeling from simulated density images at 10Å resolution.

PDB ID[a]	#Det./#Obs. Strd[b]	2-w Dist.[c]	#Det./#Obs. AA[d]
1G7K_A13	11 / 11	1.53	103 / 124
1QJP_A	8 / 8	1.81	74 / 107
1RRX_A12	11 / 11	1.71	95 / 118
1TX2_B	7 / 8	1.30	30 / 34
2ERV_A10	8 / 8	1.11	79 / 94
2J1N_AA18	16 / 16	1.67	148 / 181
2QDZ_C17	15 / 16	1.50	165 / 198
2QOM_C	12 / 12	1.71	153 / 190
2WJR_AA15	11 / 12	1.59	103 / 130
3FID_A14	12 / 12	1.36	127 / 155
4FQE_A	12 / 12	1.25	122 / 134
Average		**1.50**	**1199/1465 = 81.84%**
Standard deviation		**0.22**	

a. PDB_Sheet ID;
b. The number of β-traces in the best of the fifteen modeled sets / the
c. The 2-way distance (in Å) between the observed β-traces and the
d. The number of detected / total number of amino acids in the β-barrel.

The test of 11 simulated β-barrel density images shows that one of the fifteen sets of β-traces aligns very well with the observed true set of β-traces, with an overall 2-way distance of 1.5Å for

the detected β-traces (Table 14). To analyze the sensitivity of the detection, the number of amino acids that were missed in the detection was measured. An amino acid was considered detected if its Cα atom is within 2.5Å from the detected β-trace that corresponds to the strand where the Cα resides. For example, 1TX2_B has seven of the eight β-strands detected (Table 14 row 5). It missed four amino acids. For the seven detected strands, the 2-way distance is 1.3Å. Among the 11 test cases, *StrandRoller* appears to be able to detect 81.84% of the β-strands fairly accurately in one of the fifteen sampled sets of β-traces (Table 14).

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

Two fully automatic methods, *SSEtracer* and *SSElearner*, have been developed for the detection of helices and β-sheets from cryo-EM density maps at the medium resolution range of 5-10Å. The *SSEtracer* was tested using ten simulated maps and five experimental cryo-EM maps from EMDB. The *SSElearner* has been tested using ten simulated density maps as well as thirteen cryo-EM maps from the EMDB. Our results show that although the detection can be fairly accurate in the simulated density maps, the accuracy decreases significantly for the short helices and small β-sheets from the experimentally derived density maps. The overall detection accuracy of *SSElearner* demonstrated that it is feasible to select a specific density map from the current EMDB as training data to detect the SSE of a target cryo-EM map.

The supervised machine learning approach, *SSElearner*, requires the selection of a training density map for each target testing density map. It requires certain knowledge about the nature of the cryo-EM density map. *SSElearner* have demonstrated that it is possible to find a training density map that shares similar density nature with that of the target map in the current EMDataBank. In addition to the existing features, more sophisticated and advanced features could be further discovered and added to the feature vector of *SSElearner* to improve the accuracy. Also, it is notable that LIBSVM training time is longer for some density maps. According to the website of LIBSVM, slow convergence may happen for some difficult cases.[5] Currently, LIBSVM uses grid search for cross validation to select the best parameters. Future

[5] According to LBSVM FAQ as of updated on 25 Feb 2015: http://www.csie.ntu.edu.tw/~cjlin/libsvm/faq.html

work may include tuning and adjusting those parameters to reduce the number of iterations. Parallelizing the grid parameter search by using multi-cores may also accelerate the training process.

Compared to *SSElearner*, *SSEtracer* is faster and easier to use. Without selecting a training density map for a target map and taking the time to train a particular model for a target map, it can detect both helices and β-sheets in a simple voting way. It has been noticed that the local thickness feature is more sensitive to the local density variance, especially in the experimental cryo-EM density maps which contain noises and errors. Even small density errors (missing density and extra density) would immediately affect the calculation and analysis of the local thickness. Also, it has been observed that the local thickness of helices and β-sheets could be very similar under a certain threshold. The local thickness of some helix regions could be even smaller than the thickness of β-sheets in some experimental cryo-EM data. This would also affect the accuracy of the voting procedure in *SSEtracer*, since the local thickness feature would not help much for distinguishing between helices and β-sheets. Let $\Omega \subset R^3$be the structure. The local thickness at point $p \in \Omega$ is the diameter of the largest sphere that contains p and is completely inside the structure. The local thickness can also characterize a 3D binary image of a complex structure such as bone, cell, or paper fiber [98]. Such images are available from, micro-computed tomography [143].

A novel method, *StrandTwister*, has also been developed for the detection of β-strands from cryo-EM density maps. As expected, accurate detection of β-strands depends on accurate identification of a β-sheet. The boundary of the identified β-sheet may affect β-strand detection.

In most cases, inaccurate boundary can result in a longer/shorter detected β-strand, and such an error is reflected on the 2-way distance value. The missing density at the upper region of the β-sheet in Figure 21D resulted in missing three amino acids in the detection. The extra density (box in Figure 33) affected the curvature/size of the β-sheet and the detected set was slightly off with a 2-way distance of 2.34Å (first row of Table 11), compared to the average 2-way distance of 1.66Å in Table 11. Our implementation ignored all the short strands less than 6Å in length. This may also be responsible for some of the missing strands.

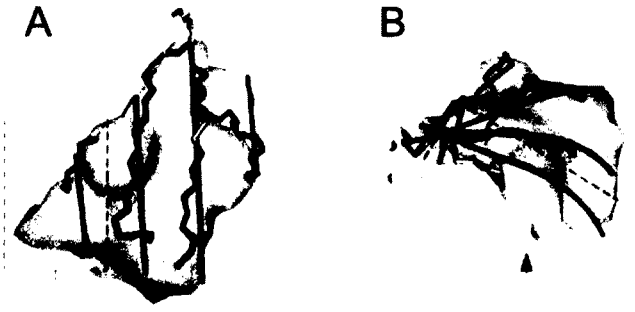


Figure 33. The accuracy of β-sheet density identification affects the accuracy of β-strands detection. The identification of β-sheet density affects the detection of β-strands in EMDB_1237_2GSY_A. The color scheme is the same as that in Figure 21 in the main manuscript. (A) Top view; (B) The side view with roughly 90° rotation. The extra density detected in β-sheet is highlighted in a box.

StrandTwister was tested using 39 β-sheets and the results were analyzed in details for three case studies. The conclusion from the tests appears 2-fold. (1) Many β-traces can be detected from density maps at medium resolutions. Our proposed idea to use β-twist in detection appears to be effective once a β-sheet region is identified approximately. One of the top ten sets of β-traces contains a set with close estimation to the observed β-traces, particularly at the central region of a β-sheet. Fine adjustment is needed to improve the detection near the edge of a β-sheet where

special properties have been observed [144]. (2) The detection error mainly comes from two situations, the inaccurate boundary of β-sheets and the miss-identified β-sheets. Both are long-standing challenges for β-sheet identification. The limitation of β-strand detection is mostly from the identification of a β-sheet.

In spite of the errors in identifying short helices and 2-stranded β-sheets, major/larger helices and sheets can be detected from the medium-resolution maps. Our results in this thesis add to the statement that major helices and those β-strands on larger sheets can be traced. Our previous results and those from other studies have shown that the topology of major secondary structures may not rely on the detection of all secondary structures. In many cases, the topology of major helices is correctly predicted without the detection of short helices [47, 66, 145]. Deriving atomic structures from density maps at medium resolutions will inevitably involve sophisticated modeling of uncertainties. The methodology in *ab initio* modeling from the medium resolution density maps has been improved recently to work with a large number of helices [146], to work with complicated skeletons [115], and to build the atomic chains in modeling [66, 111]. However, the work has been mostly tested using the true position of β-strands for density maps at the medium resolutions. *StrandTwister* detects the traces of β-strands for major β-sheets. With the β-traces, *ab initio* modeling is expected to move a significant step ahead. The current topology determination method will be extended to both α-traces and β-traces. It has been shown that additional constraints can be added to represent popular β-strand pairing during topology determination [146]. An effective method has been illustrated to build a Cα backbone using β-traces. However, density errors in the cryo-EM maps at medium resolutions determine that multiple sets of possible β-strands have to be generated. The correct set has to be identified when

it is modeled together with other parts of the chain. Another potential impact of β-traces lies in the representation of secondary structures in density map. The relative location of α-traces has been used as signatures to search for atomic structures with a similar fold [73]. Now it is possible to include β-traces as well.

Deriving atomic structures from the medium resolution cryo-EM density maps is a challenging problem. Although a number of methods exist to detect α-helices, the detection of β-strands from medium resolution cryo-EM maps has been an open problem since the first attempt in 2004 [110]. A novel method has been proposed to detect both the number of β-strands and the β-traces directly from medium resolution density data using the intrinsic twist of a β-sheet. To our knowledge, this is the second attempt to address the problem of β-strand detection in ten years, and *StrandTwister* gave an optimistic answer to this problem using a completely different approach.

A novel approach, *StrandRoller*, has also been proposed by using image processing and geometric modeling to generate a small set of possible positions of β-strands from the medium resolution β-barrel density images. Our preliminary results show that it is possible to derive such small sets. Each possible set of β-strands can be further evaluated for the best choice when more atomic details are added in modeling. This method does not require the resolution of the density to be high enough (<5Å) to resolve the separation of β-strands in β-barrel images. It applies to the images with lower resolutions. In the test containing eleven β-barrel images, *StrandRoller* detected about 81.84% of the amino acids in the β-strands with an overall 1.5Å 2-way distance between the detected β-traces and the observed ones, if the best of the fifteen detections is

considered. The results suggest that β-strands can be generated from the medium resolution cryo-EM images of β-barrel proteins. To our knowledge, this is the first method to address the problem of β-strands detection from medium resolution β-barrel images.

Although *StrandRoller* can build β-traces for an entire β-barrel based on the cryo-EM density, it is still challenging for detecting a complete chunk of β-barrel density from the medium resolution cryo-EM density maps accurately. *StrandRoller* requires the density to be continuous all around the barrel without having disconnection or missing side of the barrel, since the current method uses strand-walking for generating β-strands one by one around the barrel. Future work would include further development of the *StrandRoller* to overcome this limitation. Piece-wise regional modeling would be an option for dealing with incompletely detected β-barrel density. Instead of modeling the entire β-barrel as one whole surface, local piece-wise surface modeling and β-strand generation could be implemented to break down the large and complex β-barrel into multiple regions. Although local pieces could describe the regional curvatures and features more precisely, how to effectively, smoothly and seamlessly merge multiple single pieces into a global twisted surface and build the β-strands on that surface would be another challenging problem.

In addition to single-layer β-sheet and β-barrel, there are also many other β-structures like β-sandwich, propeller, trefoil, prism, solenoid, and etc.[6] In order to solve all types of β-structures from the medium resolution cryo-EM density maps, especially those large and complex β-structures. Different mathematical or geometric models should be utilized to capture the features of different β-structures. Instead of using one polynomial equation to model the entire β-sheet and β-barrel, piece-wise surface fitting and modeling could be one of the applicable approaches

[6] According to the CATH classification of proteins: http://www.cathdb.info/

for describing various local details and curvatures on very large and complex β-structures. Although the center area of β-sheet and β-barrel can be modeled by *StrandTwister* and *StrandRoller*, more effort could be put on the modeling of challenging areas such as the edge area of β-sheet and β-barrel. In addition to the 2-way distance calculation for the accuracy of β-strand detection, some standard measurement such as Hausdorff Distance could also be implemented to calculate the distance between detected β-trace and true β-trace.

As an emerging technology, cryo-EM has shown to be powerful in solving the 3D structure of large macromolecular assemblies and cellular complexes. Since some of the biological molecules are sensitive to high energy electron radiations, imaging must be conducted using low dose conditions to keep the sampling species in an *in vivo* status. Non-particle images - i.e. ice, dust, contaminations, or noises - in a dataset can lead to severe distortions in the result, including erroneous electron densities. Therefore, the images obtained by cryo-EM technique are extremely noisy compared to other imaging techniques [147]. The success of 3D reconstruction crucially depends on the number and the quality of 2D particle images. Before putting the 3D volume data into EMDataBank, the modeling and computational errors from the 3D reconstruction step also introduce extra levels of noise. Therefore, noise reduction and image enhancement are desirable for 3D reconstruction, segmentation, and/or structural analysis, such as skeletonization and SSE detection. A large number of image filters have been developed to decrease the noise, such as low pass filter, wavelet transforms, median filters, and so on. What makes denoising so challenging is that a successful approach must also preserve characteristic singular features of local details such as the flexible edge areas of β-sheets. Future work would

include compromise between denoising and feature preservation to improve the accuracy of protein structure detection from 3D cryo-EM density data.

Secondary structure detection from the cryo-EM density map at the medium resolution is still a challenging problem despite the multiple proposed methods. Small β-sheet like 2-stranded sheet and short helix (< 5 amino acids) are still quite hard to detect [72, 88]. Edges of β-sheets and complex β-structures are still open problems in the structure prediction from medium resolution cryo-EM density maps. In order to speed-up the research in this direction, coordinated effort is needed to promote the public sharing of the developed software and the development of the benchmark data that is available to the public. However, the comparison of the software is still challenging. Some of the methods are automatic and others are semi-automatic [63]. The continuing maintenance of the software has also been inadequate in this area.

REFERENCES

[1] T. L. Blundell, S. Bedarkar, E. Rinderknecht *et al.*, "Insulin-like growth factor: a model for tertiary structure accounting for immunoreactivity and receptor binding," *Proc Natl Acad Sci U S A*, vol. 75, no. 1, pp. 180-4, Jan, 1978.

[2] I. T. Weber, "Evaluation of homology modeling of HIV protease," *Proteins*, vol. 7, no. 2, pp. 172-84, 1990.

[3] A. K. Rapp , and C. J. Casewit, *Molecular mechanics across chemistry*, Sausalito, Calif.: University Science Books, 1997.

[4] G. J. Siegel, *Basic neurochemistry : molecular, cellular, and medical aspects*, 7th ed., Amsterdam ; Boston: Elsevier, 2006.

[5] R. K. Murray, D. K. Granner, and V. W. Rodwell, *Harper's illustrated biochemistry*, 27th ed., New York ; London: Lange Medical Books/McGraw-Hill, 2006.

[6] H. Berman, J. Westbrook, Z. Feng *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235-242, 2000.

[7] J. L. Sussman, D. Lin, J. Jiang *et al.*, "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules," *Acta Crystallogr D Biol Crystallogr*, vol. 54, no. Pt 6 Pt 1, pp. 1078-84, Nov 1, 1998.

[8] J. Moult, K. Fidelis, A. Kryshtafovych *et al.*, "Critical assessment of methods of protein structure prediction-Round VII," *Proteins*, vol. 69 Suppl 8, pp. 3-9, 2007.

[9] Z. Xiang, "Advances in homology protein structure modeling," *Curr Protein Pept Sci*, vol. 7, no. 3, pp. 217-27, Jun, 2006.

[10] M. S. Johnson, N. Srinivasan, R. Sowdhamini *et al.*, "Knowledge-based protein modeling," *Crit Rev Biochem Mol Biol*, vol. 29, no. 1, pp. 1-68, 1994.

[11] Levintha.C, "Are There Pathways for Protein Folding," *Journal De Chimie Physique Et De Physico-Chimie Biologique*, vol. 65, no. 1, pp. 44-&, 1968.

[12] D. Fischer, C. Barret, K. Bryson *et al.*, "CAFASP-1: critical assessment of fully automated structure prediction methods," *Proteins 1999*, vol. Suppl 3, pp. 209-17, 1999.

[13] T. Defay, and F. E. Cohen, "Evaluation of current techniques for ab initio protein structure prediction," *Proteins*, vol. 23, no. 3, pp. 431-45, Nov, 1995.

[14] D. T. Jones, "Progress in protein structure prediction," *Curr Opin Struct Biol*, vol. 7, no. 3, pp. 377-87, Jun, 1997.

[15] S. A. Benner, M. A. Cohen, and D. Gerloff, "Correct structure prediction?," *Nature*, vol. 359, no. 6398, pp. 781, Oct 29, 1992.

[16] Y. Zhang, "Progress and challenges in protein structure prediction," *Curr Opin Struct Biol*, vol. 18, no. 3, pp. 342-8, Jun, 2008.

[17] D. Baker, and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, no. 5540, pp. 93-6, Oct 5, 2001.

[18] K. Ginalski, "Comparative modeling for protein structure prediction," *Current Opinion in Structural Biology*, vol. 16, no. 2, pp. 172-177, 2006.

[19] B. John, A. Sali, and O. Journals, "Comparative protein structure modeling by iterative alignment, model building and model assessment," *Nucleic Acids Research*, vol. 31, no. 14, pp. 3982-3992, 2003.

[20] M. Tress, I. Ezkurdia, O. Grana *et al.*, "Assessment of Predictions Submitted for the CASP6 Comparative Modeling Category," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 7, pp. 27-45, 2005.

[21] R. Sanchez, and A. Sali, "Comparative protein structure modeling as an optimization problem," *Journal of Molecular Structure-Theochem*, vol. 398, pp. 489-496, Jun 30, 1997.

[22] J. U. Bowie, R. Luthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure," *Science*, vol. 253, pp. 164-170, 1991.

[23] D. T. Jones, W. R. Taylor, and J. M. Thornton, "A new approach to protein fold recognition," *Nature*, vol. 358, no. 6381, pp. 86-9, Jul 2, 1992.

[24] R. Bonneau, and D. Baker, "Ab initio protein structure prediction: progress and prospects," *Annu Rev Biophys Biomol Struct*, vol. 30, pp. 173-89, 2001.

[25] J. Moult, K. Fidelis, A. Kryshtafovych *et al.*, "Critical assessment of methods of protein structure prediction - Round VIII," *Proteins*, vol. 77 Suppl 9, pp. 1-4, 2009.

[26] W. Chiu, M. L. Baker, W. Jiang *et al.*, "Electron cryomicroscopy of biological machines at subnanometer resolution," *Structure*, vol. 13, no. 3, pp. 363-72, Mar, 2005.

[27] C. F. Hryc, D. H. Chen, and W. Chiu, "Near-atomic-resolution cryo-EM for molecular virology," *Curr Opin Virol*, vol. 1, no. 2, pp. 110-7, Aug, 2011.

[28] W. Jiang, M. L. Baker, J. Jakana *et al.*, "Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy," *Nature*, vol. 451, no. 7182, pp. 1130-4, Feb 28, 2008.

[29] X. Yu, L. Jin, and Z. H. Zhou, "3.88 A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy," *Nature*, vol. 453, no. 7193, pp. 415-9, May 15, 2008.

[30] A. M. Anger, J. P. Armache, O. Berninghausen *et al.*, "Structures of the human and Drosophila 80S ribosome," *Nature*, vol. 497, no. 7447, pp. 80-5, May 2, 2013.

[31] X. K. Zhang, P. Ge, X. K. Yu *et al.*, "Cryo-EM structure of the mature dengue virus at 3.5-angstrom resolution," *Nature Structural & Molecular Biology*, vol. 20, no. 1, pp. 105-U133, Jan, 2013.

[32] C. L. Lawson, M. L. Baker, C. Best *et al.*, "EMDataBank.org: unified data resource for CryoEM," *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D456-64, Jan, 2011.

[33] J. Esquivel-Rodriguez, and D. Kihara, "Computational methods for constructing protein structure models from 3D electron microscopy maps," *J Struct Biol*, vol. 184, no. 1, pp. 93-102, Oct, 2013.

[34] L. Wang, and F. J. Sigworth, "Cryo-EM and single particles," *Physiology (Bethesda)*, vol. 21, pp. 13-8, Feb, 2006.

[35] H. R. Saibil, "Conformational changes studied by cryo-electron microscopy," *Nat Struct Biol*, vol. 7, no. 9, pp. 711-4, Sep, 2000.

[36] K. Mitra, and J. Frank, "Ribosome dynamics: Insights from atomic structure modeling into cryo-electron microscopy maps," *Annual Review of Biophysics and Biomolecular Structure*, vol. 35, pp. 299-317, 2006.

[37] X. Zhang, L. Jin, Q. Fang *et al.*, "3.3 angstrom Cryo-EM Structure of a Nonenveloped Virus Reveals a Priming Mechanism for Cell Entry," *Cell*, vol. 141, no. 3, pp. 472-482, Apr 30, 2010.

[38] C. D. Ludtke SJ, Song JL, Chuang DT, Chiu W., "Seeing GroEL at 6 A resolution by single particle electron cryomicroscopy.," *Structure*, vol. 12, no. 7, pp. 1129-36, Jul, 2004.

[39] J. F. Conway, N. Cheng, A. Zlotnick *et al.*, "Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy," *Nature*, vol. 386, no. 6620, pp. 91-4, 1997.

[40] M. L. Baker, W. Jiang, W. J. Wedemeyer *et al.*, "Ab initio modeling of the herpesvirus VP26 core domain assessed by CryoEM density," *PLoS Comput Biol*, vol. 2, no. 10, pp. e146, Oct 27, 2006.

[41] A. G. Martin, F. Depoix, M. Stohr *et al.*, "Limulus polyphemus hemocyanin: 10 angstrom cryo-EM structure, sequence analysis, molecular modelling and rigid-body fitting reveal the interfaces between the eight hexamers," *Journal of Molecular Biology*, vol. 366, no. 4, pp. 1332-1350, Mar 2, 2007.

[42] E. Villa, J. Sengupta, L. G. Trabuco *et al.*, "Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis," *Proc Natl Acad Sci U S A*, vol. 106, no. 4, pp. 1063-8, Jan 27, 2009.

[43] Y. Lu, C. E. M. Strauss, and J. He, "Incorporation of Constraints from Low Resolution Density Map in Ab Initio Structure Prediction Using Rosetta," *Proceeding of 2007 IEEE international Conference on Bioinformatics and Biomedicine Workshops*, pp. p67-73, 2007.

[44] M. Topf, K. Lasker, B. Webb *et al.*, "Protein structure fitting and refinement guided by cryo-EM density," *Structure*, vol. 16, no. 2, pp. 295-307, Feb, 2008.

[45] M. Topf, M. L. Baker, B. John *et al.*, "Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy," *J Struct Biol*, vol. 149, no. 2, pp. 191-203, Feb, 2005.

[46] M. Topf, M. L. Baker, M. A. Marti-Renom *et al.*, "Refinement of protein structures by iterative comparative modeling and CryoEM density fitting," *J Mol Biol*, vol. 357, no. 5, pp. 1655-68, Apr 14, 2006.

[47] Y. Lu, J. He, and C. E. Strauss, "Deriving topology and sequence alignment for the helix skeleton in low-resolution protein density maps," *J Bioinform Comput Biol*, vol. 6, no. 1, pp. 183-201, Feb, 2008.

[48] F. DiMaio, M. D. Tyka, M. L. Baker *et al.*, "Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta," *Journal of Molecular Biology*, vol. 392, no. 1, pp. 181-190, Sep 11, 2009.

[49] W. Wriggers, R. A. Milligan, and J. A. McCammon, "Situs: A package for docking crystal structures into low-resolution maps from electron microscopy," *J Struct Biol*, vol. 125, no. 2-3, pp. 185-95, Apr-May, 1999.

[50] G. F. Schröder, A. T. Brunger, and M. Levitt, "Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution," *Structure (London, England : 1993)*, vol. 15, no. 12, pp. 1630-1641, 2007.

[51] J. Zhang, M. L. Baker, G. F. Schroder *et al.*, "Mechanism of folding chamber closure in a group II chaperonin," *Nature*, vol. 463, no. 7279, pp. 379-83, Jan 21, 2010.

[52] R. A. Crowther, N. A. Kiselev, B. Bottcher *et al.*, "Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy," *Cell*, vol. 77, no. 6, pp. 943-50, Jun 17, 1994.

[53] Y. Hashem, A. des Georges, J. Fu *et al.*, "High-resolution cryo-electron microscopy structure of the Trypanosoma brucei ribosome," *Nature*, vol. 494, no. 7437, pp. 385-9, Feb 21, 2013.

[54] G. Zhao, J. R. Perilla, E. L. Yufenyuy *et al.*, "Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics," *Nature*, vol. 497, no. 7451, pp. 643-6, May 30, 2013.

[55] K. Lasker, F. Forster, S. Bohn *et al.*, "Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach," *Proc Natl Acad Sci U S A*, vol. 109, no. 5, pp. 1380-7, Jan 31, 2012.

[56] F. Beck, P. Unverdorben, S. Bohn *et al.*, "Near-atomic resolution structural model of the yeast 26S proteasome," *Proc Natl Acad Sci U S A*, vol. 109, no. 37, pp. 14870-5, Sep 11, 2012.

[57] A. Sali, and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *J Mol Biol*, vol. 234, no. 3, pp. 779-815, Dec 5, 1993.

[58] K. Arnold, L. Bordoli, J. Kopp *et al.*, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling," *Bioinformatics*, vol. 22, no. 2, pp. 195-201, Jan 15, 2006.

[59] M. G. Rossmann, "Fitting atomic models into electron-microscopy maps," *Acta Crystallographica Section D-Biological Crystallography*, vol. 56, pp. 1341-1349, Oct, 2000.

[60] L. G. Trabuco, E. Villa, K. Mitra *et al.*, "Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics," *Structure*, vol. 16, no. 5, pp. 673-683, May, 2008.

[61] G. F. Schroder, A. T. Brunger, and M. Levitt, "Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution," *Structure*, vol. 15, no. 12, pp. 1630-1641, Dec, 2007.

[62] K. Y. Chan, L. G. Trabuco, E. Schreiner *et al.*, "Cryo-Electron Microscopy Modeling by the Molecular Dynamics Flexible Fitting Method," *Biopolymers*, vol. 97, no. 9, pp. 678-686, Sep, 2012.

[63] M. L. Baker, S. S. Abeysinghe, S. Schuh *et al.*, "Modeling protein structure at near atomic resolutions with Gorgon," *Journal of Structural Biology*, vol. 174, no. 2, pp. 360-373, 2011.

[64] K. Al Nasr, D. Ranjan, M. Zubair *et al.*, "Ranking Valid Topologies of the Secondary Structure Elements Using a Constraint Graph," *Journal of Bioinformatics and Computational Biology*, vol. 09, no. 03, pp. 415-430, Jun, 2011.

[65] A. Biswas, D. Si, K. Al Nasr *et al.*, "Improved efficiency in cryo-EM secondary structure topology determination from inaccurate data," *J Bioinform Comput Biol*, vol. 10, no. 3, pp. 1242006, Jun, 2012.

[66] S. Lindert, N. Alexander, N. Wotzel *et al.*, "EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps," *Structure*, vol. 20, no. 3, pp. 464-78, Mar 7, 2012.

[67] J. He, and D. Si, "Towards De Novo Folding of Protein Structures from Cryo-EM 3D Images at Medium Resolutions."

[68] W. Sun, and J. He, "Native secondary structure topology has near minimum contact energy among all possible geometrically constrained topologies," *Proteins*, vol. 77, no. 1, pp. 159-73, Oct, 2009.

[69] K. Al Nasr, W. Sun, and J. He, "Structure prediction for the helical skeletons detected from the low resolution protein density map," *BMC Bioinformatics*, vol. 11 Suppl 1, pp. S44, 2010.

[70] W. Jiang, M. L. Baker, S. J. Ludtke *et al.*, "Bridging the information gap: computational tools for intermediate resolution structure interpretation," *J Mol Biol*, vol. 308, no. 5, pp. 1033-44, May, 2001.

[71] A. Del Palu, J. He, E. Pontelli *et al.*, "Identification of Alpha-Helices from Low Resolution Protein Density Maps," *Proceeding of Computational Systems Bioinformatics Conference(CSB)*, pp. 89-98, 2006.

[72] M. L. Baker, T. Ju, and W. Chiu, "Identification of secondary structure elements in intermediate-resolution density maps," *Structure*, vol. 15, no. 1, pp. 7-19, Jan, 2007.

[73] K. Lasker, O. Dror, M. Shatsky *et al.*, "EMatch: discovery of high resolution structural homologues of protein domains in intermediate resolution cryo-EM maps," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 4, no. 1, pp. 28-39, Jan-Mar, 2007.

[74] Y. Zeyun, and C. Bajaj, "Computational Approaches for Automatic Structural Analysis of Large Biomolecular Complexes," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 5, no. 4, pp. 568-582, 2008.

[75] L. Ma, M. Reisert, and H. Burkhardt, "RENNSH: A Novel alpha-Helices Identification Approach for Intermediate Resolution Electron Density Maps," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. PP, no. 99, pp. 1-1, 2011.

[76] Y. Kong, and J. Ma, "A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps," *J Mol Biol*, vol. 332, no. 2, pp. 399-413, Sep 12, 2003.

[77] Z. H. Zhou, "Towards atomic resolution structural determination by single-particle cryo-electron microscopy," *Current Opinion in Structural Biology*, vol. 18, no. 2, pp. 218-228, Apr, 2008.

[78] M. L. Baker, M. R. Baker, C. F. Hryc *et al.*, "Gorgon and pathwalking: macromolecular modeling tools for subnanometer resolution density maps," *Biopolymers*, vol. 97, no. 9, pp. 655-68, Sep, 2012.

[79] S. J. Ludtke, M. L. Baker, D. H. Chen *et al.*, "De novo backbone trace of GroEL from single particle electron cryomicroscopy," *Structure*, vol. 16, no. 3, pp. 441-8, Mar, 2008.

[80] M. L. Baker, C. F. Hryc, Q. Zhang *et al.*, "Validated near-atomic resolution structure of bacteriophage epsilon15 derived from cryo-EM and modeling," *Proc Natl Acad Sci U S A*, vol. 110, no. 30, pp. 12301-6, Jul 23, 2013.

[81] A. Del Palu, J. He, E. Pontelli *et al.*, "Identification of Alpha-Helices from Low Resolution Protein Density Maps," *Comput Syst Bioinformatics Conf*, pp. 89-98, 2006.

[82] Z. Yu, and C. Bajaj, "Computational approaches for automatic structural analysis of large biomolecular complexes," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 5, no. 4, pp. 568-582, Oct-Dec, 2008.

[83] M. L. Baker, S. S. Abeysinghe, S. Schuh *et al.*, "Modeling protein structure at near atomic resolutions with Gorgon," *J Struct Biol*, vol. 174, no. 2, pp. 360-373, May, 2011.

[84] L. Ma, M. Reisert, and H. Burkhardt, "RENNSH: A Novel alpha-Helices Identification Approach for Intermediate Resolution Electron Density Maps," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 9, no. 1, pp. 228-239, Jan-Feb, 2011.

[85] M. Rusu, and W. Wriggers, "Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions," *J Struct Biol*, vol. 177, no. 2, pp. 410-9, Feb, 2012.

[86] D. Si, S. Ji, K. A. Nasr *et al.*, "A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps," *Biopolymers*, vol. 97, no. 9, pp. 698-708, Sep, 2012.

[87] C. Bajaj, S. Goswami, and Q. Zhang, "Detection of secondary and supersecondary structures of proteins from cryo-electron microscopy," *J Struct Biol*, vol. 177, no. 2, pp. 367-81, Feb, 2012.

[88] D. Si, and J. He, "Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps." pp. 764-770.

[89] L. Pauling, and R. B. Corey, "The pleated sheet, a new layer configuration of polypeptide chains," *Proc Natl Acad Sci U S A*, vol. 37, no. 5, pp. 251-6, May, 1951.

[90] C. Chothia, "Conformation of twisted beta-pleated sheets in proteins," *J Mol Biol*, vol. 75, no. 2, pp. 295-302, Apr 5, 1973.

[91] F. R. Salemme, "Conformational and geometrical properties of beta-sheets in proteins. III. Isotropically stressed configurations," *J Mol Biol*, vol. 146, no. 1, pp. 143-56, Feb 15, 1981.

[92] J. Novotny, R. E. Bruccoleri, and J. Newell, "Twisted hyperboloid (Strophoid) as a model of beta-barrels in proteins," *J Mol Biol*, vol. 177, no. 3, pp. 567-73, Aug 15, 1984.

[93] D. Znamenskiy, K. Le Tuan, A. Poupon *et al.*, "Beta-sheet modeling by helical surfaces," *Protein Eng*, vol. 13, no. 6, pp. 407-12, Jun, 2000.

[94] E. Koh, and T. Kim, "Minimal surface as a model of beta-sheets," *Proteins-Structure Function and Bioinformatics*, vol. 61, no. 3, pp. 559-569, Nov 15, 2005.

[95] T. Ju, M. L. Baker, and W. Chiu, "Computing a family of skeletons of volumetric models for shape description," *Comput Aided Des*, vol. 39, no. 5, pp. 352-360, May, 2007.

[96] J. J. Fernandez, and S. Li, "An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms," *J Struct Biol*, vol. 144, no. 1-2, pp. 152-61, Oct-Nov, 2003.

[97] Rutishau.H, "Jacobi Method for Real Symmetric Matrices," *Numerische Mathematik*, vol. 9, no. 1, pp. 1-&, 1966.

[98] T. Hildebrand, and P. Ruegsegger, "A new method for the model-independent assessment of thickness in three-dimensional images," *Journal of Microscopy-Oxford*, vol. 185, pp. 67-75, Jan, 1997.

[99] T. Saito, and J. I. Toriwaki, "New Algorithms for Euclidean Distance Transformation of an N-Dimensional Digitized Picture with Applications," *Pattern Recognition*, vol. 27, no. 11, pp. 1551-1565, Nov, 1994.

[100] W. R. Taylor, and A. Aszodi, *Protein geometry, classification, topology and symmetry : a computational analysis of structure*, Bristol: Institute of Physics Pub., 2005.

[101] M. Gerstein, "A structural census of genomes: Comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure," *Journal of Molecular Biology*, vol. 274, no. 4, pp. 562-576, Dec 12, 1997.

[102] S. J. Ludtke, P. R. Baldwin, and W. Chiu, "EMAN: semiautomated software for high-resolution single-particle reconstructions," *J Struct Biol*, vol. 128, no. 1, pp. 82-97, Dec 1, 1999.

[103] W. Kabsch, and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577-637, Dec, 1983.

[104] S. J. Ludtke, P. R. Baldwin, and W. Chiu, "EMAN: Semi-automated software for high resolution single particle reconstructions," *J Struct Biol*, vol. 128, no. 1, pp. 82-97, 1999.

[105] J.-J. Fernández, and S. Li, "An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms," *Journal of Structural Biology*, vol. 144, no. 1-2, pp. 152-161, 2003.

[106] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, United States, 1992, pp. 144-152.

[107] C. Cortes, and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273-297, 1995.

[108] C.-C. Chang, and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1-27, 2011.

[109] S. Knerr, L. e. Personnaz, and G. e. Dreyfus, "Single-Layer Learning Revisited: A Stepwise Procedure for Building and Training a Neural Network," *Neurocomputing: Algorithms, Architectures and Applications*, F. and J. H\'erault, eds., pp. 41--50: Springer-Verlag, 1990.

[110] Y. Kong, X. Zhang, T. S. Baker *et al.*, "A Structural-informatics approach for tracing beta-sheets: building pseudo-C(alpha) traces for beta-strands in intermediate-resolution density maps," *J Mol Biol*, vol. 339, no. 1, pp. 117-30, May 21, 2004.

[111] M. R. Baker, I. Rees, S. J. Ludtke *et al.*, "Constructing and validating initial Calpha models from subnanometer resolution density maps with pathwalking," *Structure*, vol. 20, no. 3, pp. 450-63, Mar 7, 2012.

[112] D. Si, and J. He, "Tracing Beta Strands Using StrandTwister from Cryo-EM Density Maps at Medium Resolutions," *Structure*, vol. 22, no. 11, pp. 1665-1676, 2014.

[113] A. Del Palu, J. He, E. Pontelli *et al.*, "Identification of Alpha-Helices from Low Resolution Protein Density Maps," in Proceeding of Computational Systems Bioinformatics Conference(CSB), 2006, pp. 89-98.

[114] M. Rusu, and W. Wriggers, "Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions," *J Struct Biol*, vol. 177, no. 2, pp. 410-419, Feb, 2012.

[115] K. Al Nasr, C. Liu, M. Rwebangira *et al.*, "Intensity-Based Skeletonization of CryoEM Gray-Scale Images Using a True Segmentation-Free Algorithm," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 10, no. 5, pp. 1289-98, Sep-Oct, 2013.

[116] D. Si, and J. He, "Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps," *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pp. 764-770, 2013.

[117] E. F. Pettersen, T. D. Goddard, C. C. Huang *et al.*, "UCSF Chimera--a visualization system for exploratory research and analysis," *J Comput Chem*, vol. 25, no. 13, pp. 1605-12, Oct, 2004.

[118] R. Henderson, A. Sali, M. L. Baker *et al.*, "Outcome of the first electron microscopy validation task force meeting," *Structure*, vol. 20, no. 2, pp. 205-14, Feb 8, 2012.

[119] M. L. Baker, J. J. Zhang, S. J. Ludtke *et al.*, "Cryo-EM of macromolecular assemblies at near-atomic resolution," *Nat Protoc*, vol. 5, no. 10, pp. 1697-1708, 2010.

[120] D. Si, and J. He, "Orientations of beta-strand Traces and Near Maximum Twist."

[121] B. K. Ho, and P. M. G. Curmi, "Twist and shear in beta-sheets and beta-ribbons," *Journal of Molecular Biology*, vol. 317, no. 2, pp. 291-308, Mar 22, 2002.

[122] J. Zhang, N. Nakamura, Y. Shimizu *et al.*, "JADAS: a customizable automated data acquisition system and its application to ice-embedded single particles," *J Struct Biol,* vol. 165, no. 1, pp. 1-9, Jan, 2009.

[123] C. Chaudhry, A. L. Horwich, A. T. Brunger *et al.*, "Exploring the structural dynamics of the E-coli chaperonin GroEL using translation-libration-screw crystallographic refinement of intermediate states," *J Mol Biol,* vol. 342, no. 1, pp. 229-245, Sep 3, 2004.

[124] K. Braig, P. D. Adams, and A. T. Brunger, "Conformational Variability in the Refined Structure of the Chaperonin Groel at 2.8 Angstrom Resolution," *Nature Structural Biology,* vol. 2, no. 12, pp. 1083-1094, Dec, 1995.

[125] C. Bartolucci, D. Lamba, S. Grazulis *et al.*, "Crystal structure of wild-type chaperonin GroEL," *J Mol Biol,* vol. 354, no. 4, pp. 940-951, Dec 9, 2005.

[126] R. Zhang, C. F. Hryc, Y. Cong *et al.*, "4.4 angstrom cryo-EM structure of an enveloped alphavirus Venezuelan equine encephalitis virus," *Embo Journal,* vol. 30, no. 18, pp. 3854-3863, Sep 14, 2011.

[127] T. C. Terwilliger, "Rapid model building of beta-sheets in electron-density maps," *Acta Crystallogr D Biol Crystallogr,* vol. 66, no. Pt 3, pp. 276-84, Mar, 2010.

[128] G. A. Petsko, and D. Ringe, *Protein structure and function*, Oxford: Oxford University Press, 2009.

[129] T. Hamelryck, J. T. Kent, and A. Krogh, "Sampling realistic protein conformations using local structural bias," *Plos Computational Biology,* vol. 2, no. 9, pp. 1121-1133, Sep, 2006.

[130] Z. H. Zhou, "Atomic resolution cryo electron microscopy of macromolecular complexes," *Adv Protein Chem Struct Biol,* vol. 82, pp. 1-35, 2011.

[131] S. Dong, and H. Jing, "Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps," *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics %@ 978-1-4503-2434-2*, ACM, 2013, pp. 764-770.

[132] V. Cherezov, W. Liu, J. P. Derrick *et al.*, "In meso crystal structure and docking simulations suggest an alternative proteoglycan binding site in the OpcA outer membrane adhesin," *Proteins,* vol. 71, no. 1, pp. 24-34, Apr, 2008.

[133] D. Si, and J. He, "Combining Image Processing and Modeling to Generate Traces of Beta-strands from Cryo-EM Density Images of Beta-barrel."

[134] E. Tolonen, B. Bueno, S. Kulshreshta *et al.*, "Allosteric transition and binding of small molecule effectors causes curvature change in central beta-sheets of selected enzymes," *J Mol Model,* vol. 17, no. 4, pp. 899-911, Apr, 2011.

[135] I. Lasters, S. J. Wodak, P. Alard *et al.*, "Structural principles of parallel beta-barrels in proteins," *Proc Natl Acad Sci U S A,* vol. 85, no. 10, pp. 3338-42, May, 1988.

[136] E. Koh, and T. Kim, "Minimal surface as a model of beta-sheets," *Proteins,* vol. 61, no. 3, pp. 559-69, Nov 15, 2005.

[137] A. D. McLachlan, "Gene duplications in the structural evolution of chymotrypsin," *J Mol Biol,* vol. 128, no. 1, pp. 49-79, Feb 15, 1979.

[138] A. G. Murzin, A. M. Lesk, and C. Chothia, "Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis," *J Mol Biol,* vol. 236, no. 5, pp. 1369-81, Mar 11, 1994.

[139] A. G. Murzin, A. M. Lesk, and C. Chothia, "Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures," *J Mol Biol,* vol. 236, no. 5, pp. 1382-400, Mar 11, 1994.

[140] T. Pali, and D. Marsh, "Tilt, twist, and coiling in beta-barrel membrane proteins: relation to infrared dichroism," *Biophys J,* vol. 80, no. 6, pp. 2789-97, Jun, 2001.

[141] G. E. Schulz, "The structure of bacterial outer membrane proteins," *Biochimica Et Biophysica Acta-Biomembranes,* vol. 1565, no. 2, pp. 308-317, Oct 11, 2002.

[142] M. A. Lomize, A. L. Lomize, I. D. Pogozheva *et al.*, "OPM: orientations of proteins in membranes database," *Bioinformatics,* vol. 22, no. 5, pp. 623-5, Mar 1, 2006.

[143] O. A. Peters, A. Laib, P. Ruegsegger *et al.*, "Three-dimensional analysis of root canal geometry by high-resolution computed tomography," *J Dent Res,* vol. 79, no. 6, pp. 1405-9, Jun, 2000.

[144] J. S. Richardson, and D. C. Richardson, "Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation," *Proc Natl Acad Sci U S A,* vol. 99, no. 5, pp. 2754-2759, Mar 5, 2002.

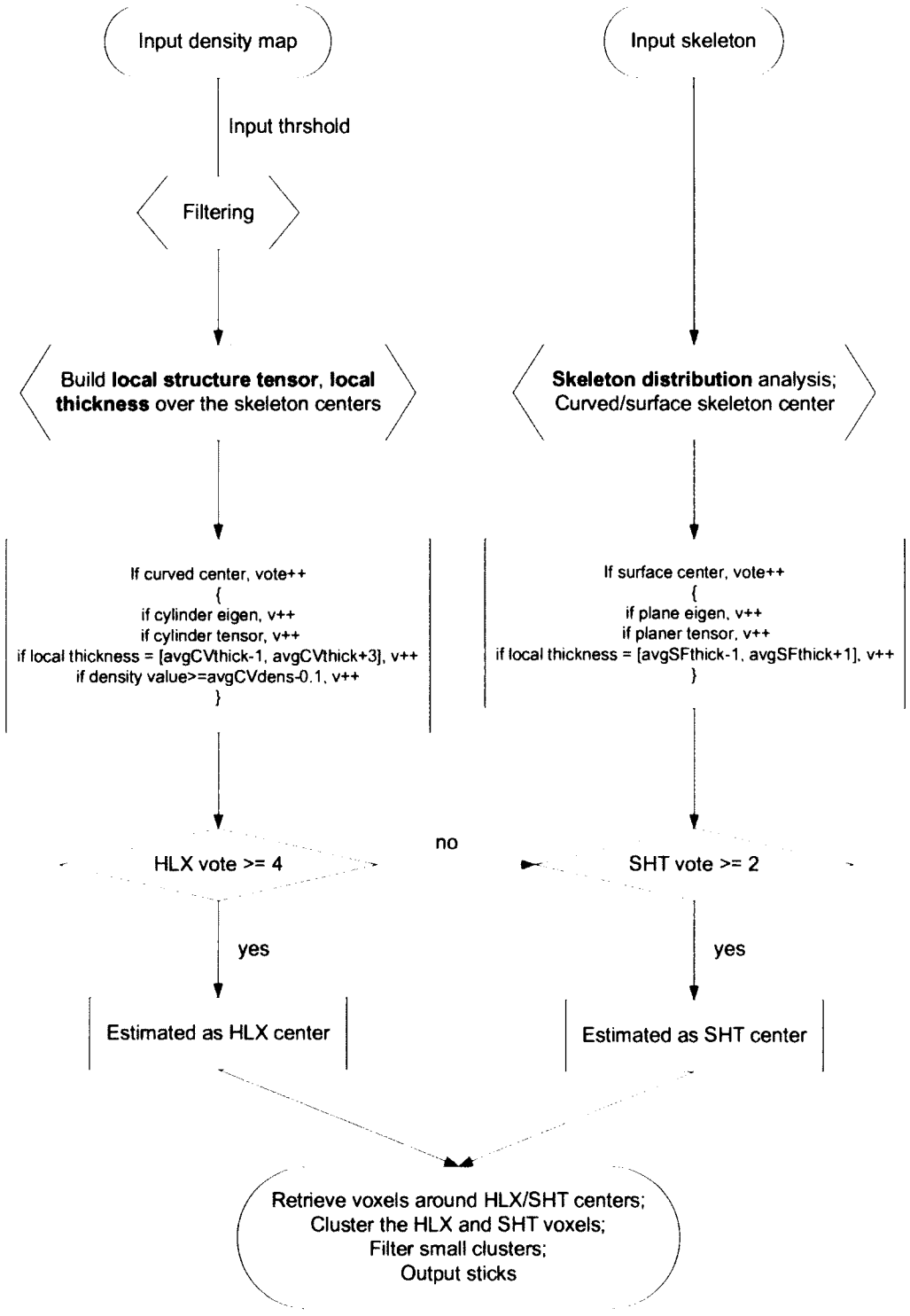
[145] K. Al Nasr, L. Chen, D. Si *et al.*, "Building the initial chain of the proteins through de novo modeling of the cryo-electron microscopy volume data at the medium resolutions," in Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, Orlando, Florida, October 7-10, 2012, pp. 490-497.

[146] K. Al Nasr, D. Ranjan, M. Zubair *et al.*, "Solving the Secondary Structure Matching Problem in Cryo-EM De Novo Modeling Using a Constrained K-Shortest Path Graph Algorithm," *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* vol. 11, no. 2, pp. 419-430, Mar-Apr, 2014.

[147] Q. Zhang, and C. L. Bajaj, "Cryo-Electron Microscopy Data Denoising Based on the Generalized Digitized Total Variation Method," *Far East J Appl Math,* vol. 45, no. 2, pp. 83-161, Aug, 2010.

APPENDIX A

DETAILED FLOW CHART OF *SSETRACER*



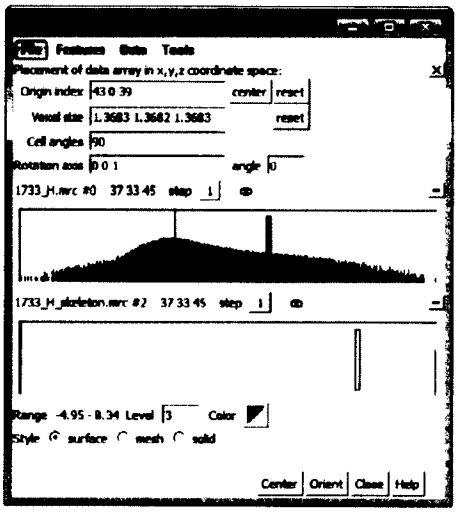
APPENDIX B

MANUAL OF *SSETRACER*

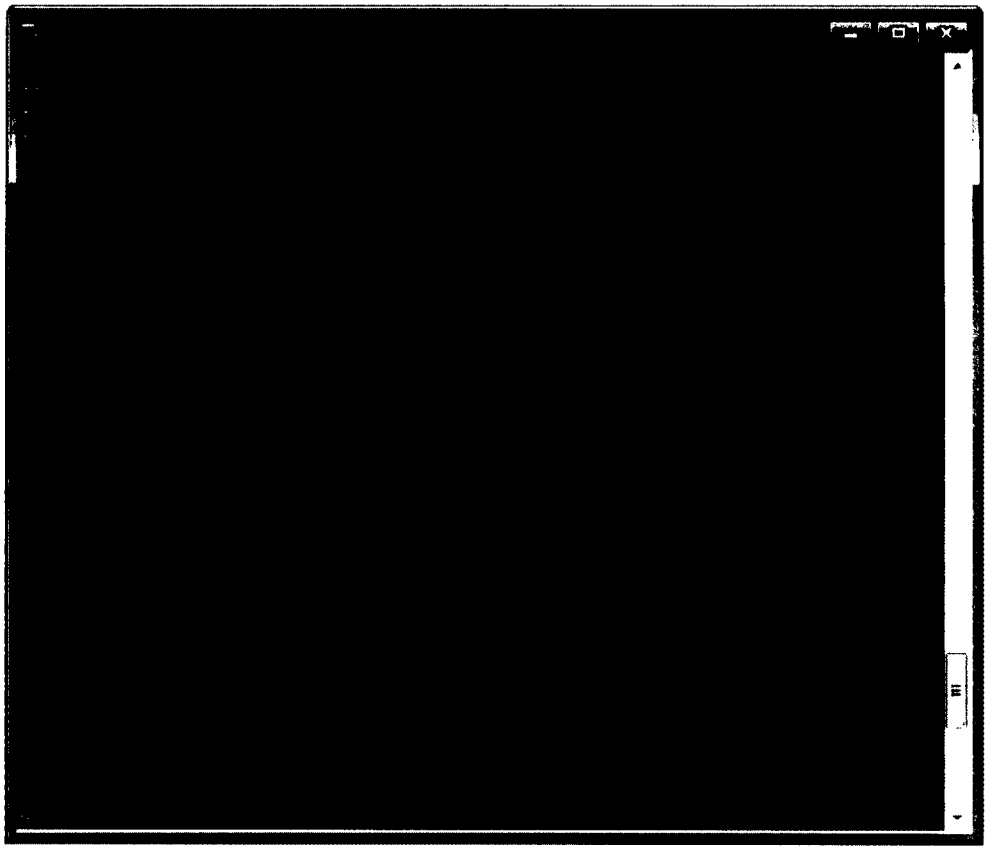
Steps

==========

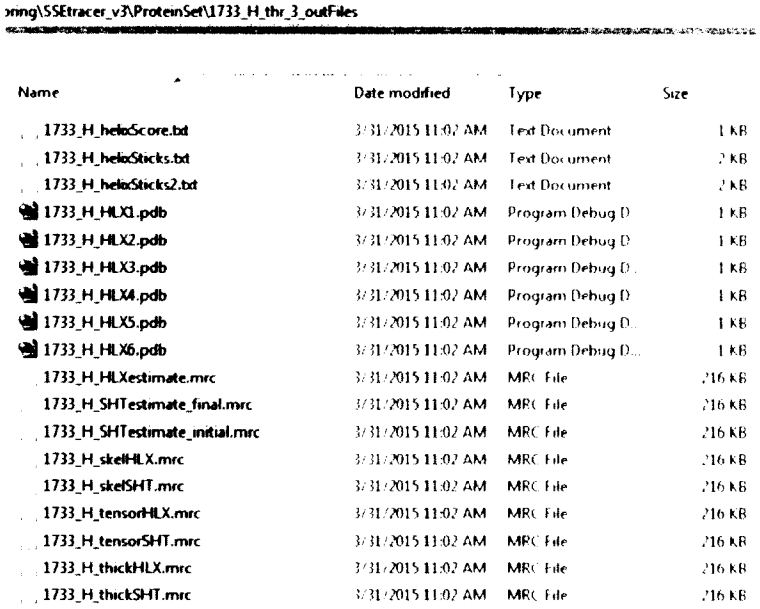
1. To run the latest version of *SSEtracer* in command line. You need to have three inputs: the input density mrc file, the input skeleton mrc file and the protein ID. Make sure the „Origin index" and „Voxel size" of these two inputs density maps have been aligned before you run the program (USCF Chimera → Volume → Volume Viewer → Features → Coordinates). Since the coordinates of skeleton generated from Gorgon is not aligned with the density map by default, you need to manually align it with the coordinates of protein density map. An ideal skeleton would show the sheets as surfaces and helix/loop parts as curves without having extra wrong connections between them. Then simply select a threshold as input the parameter. An ideal threshold would show β-sheet as thin layer of density and helix feature as a cylinder.



2. Run with "tracer_v3_command <relative input path> <pdbID> <thrshold>". Press „Enter" and the program will automatically detect the location of α-helices and β-sheets.



3. The output will be generated in the same directory where you put your input files, but in a separated folder named as "pdbID_threshold_ outFiles".

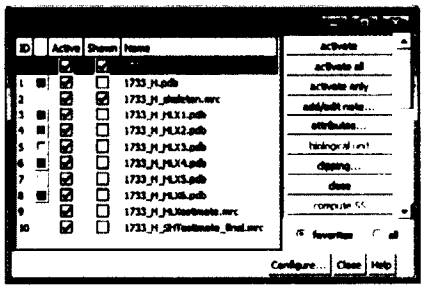


Interpret the result

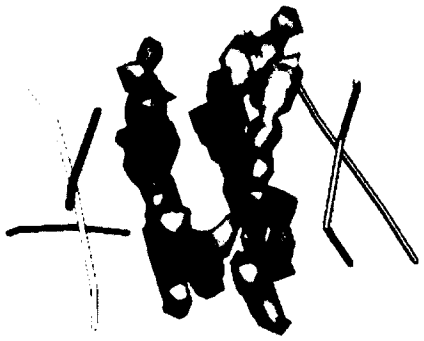
==========

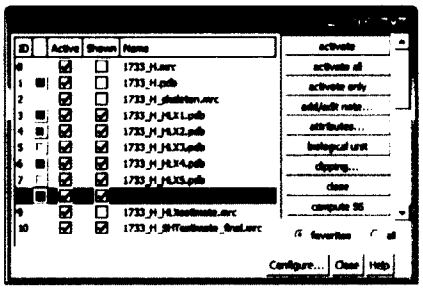
The detected helices are often called "HLX#_.pdb", the detected β-sheet density is called "SHTestimate_final.mrc". Note if there are multiple β-sheets in the protein, *SSEtracer* may detect multiple sheet density regions.

Example of input density map and skeleton:

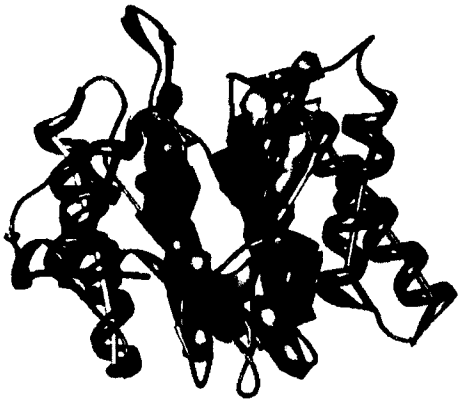


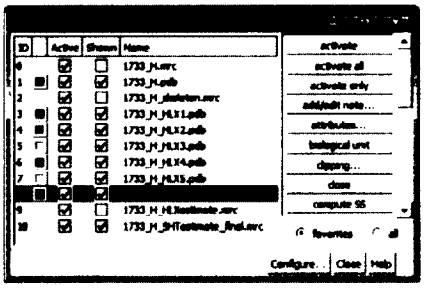
Example of detected SSE locations:





Detected SSE locations overlapped with the true structure:





APPENDIX C

MANUAL OF *SSELEARNER*

Steps

==========

1. Run 'model_generator' first if you want to generate a model from a density map or several maps. The 'Models' folder has already included some pre-generated models, check to see if there is an existing model that suitable for your target density map. A suitable model should have the similar density quality, similar threshold levels for helix/sheet/background, and should show the similar features at these levels with the target density map.

   Make sure you put the PDB file and MRC file into 'ProteinSet' folder.

   Make sure you put all the protein IDs (XXXX_X) into a 'list.txt' file if you want to train multiple maps at a time. The first line of 'list.txt' should contain the name of the model right after '//'.

   All the parameters that required for running the program are located in 'thresholds & parameters.txt'.

   Suggest choosing the cross-validation when you want to train a good model.

2. Run 'predictor' to predict the SSEs from the target density map. It will ask the name of training model that you want to use to predict this target map.

3. Run 'post-processing' to process the rough result after SVM. All the parameters used are located in 'thresholds & parameters.txt'.

   All the output files will be generated to 'Output/XXXX_X_outFiles'.

Example

==========

Use 1780_K as training data to predict 1733_H, do the steps as following:

% ./model_generator

Do you want to train multiple maps or just one single map as a model (m/s): s
Please specify a protein (pdb file without extension): 1780_K
Please enter a threshold for filtering the whole map (0 - 1): 0.1
Please enter a threshold for building thicknesses (>threshold1): 0.545

...

...

```
Do you want to use corss-validation (may take a long time) to compute the best c and g value?
(y/n) : y
      ...
                ...

% ./predictor

Please specify the protein that you want to test : 1733_H
Please enter a threshold for filtering the whole map (0 - 1): 0.1
Please enter a threshold for building thicknesses (>threshold1): 0.425
      ...
                ...
Please specify the model that you want to use to predict 1733_H : 1780_K
      ...
                ...
--- All Done !!

% ./post-processing

Please enter the ID of test density map (without extension): 1733_H
Please enter the minimum length of a helix (default 5): 7
Please enter the minimum length of a sheet (default 8): 10
Please enter the local peak filter divider for helix (int, default 3): 35
Please enter the local peak filter divider for sheet (int, default 4): 2
Please enter the small sheet filter divider (int, default 10): 10
      ...
                ...
Do you want to rebuild the Sheet? (y/n) : y
      ...
                ...
**********************************
  Specificity of Helix= 100.00%
  Sensitivity of Helix= 80.23%

  Specificity of Sheet= 82.27%
  Sensitivity of Sheet= 87.10%
**********************************
Done...
```

Note: This statistics will only be calculated if you put the .pdb file of true structure in the folder.

Minimum length of a helix/sheet

==========

The minimum length of helix/ β-sheet is calculated by measuring the distance between two farthest voxels in the detected helix or sheet regions. This parameter let the user control the minimum size of helix/ β-sheet they want to output.

Local peak filter divider:

==========

Local peak filter is a filter for selecting backbone voxels (highly dense areas). For each voxel, the average density of all voxels contained within a sphere of 3 Å in radius is calculated and those voxels in the sphere with a density value greater than the average have their local-peak-count number increased by 1.The peak counting operation loops over all voxels and assigns each voxel a local peak-count number. Upon completion of this process, all voxels are sorted according to their local-peak-count (lpc) numbers. The voxels that have lpc less than (highest_lpc/divider) are categorized as backbone voxels and discarded.

Code of lpc:

```
if (lpc[i][j][k] < maxCount/divider)  // filter voxels have lower local-peak-count
   mrc.cube[i][j][k] = 0;
```

Small sheet filter divider

==========

The initial detected sheet voxels in one density map are clustered into multiple clusters, the large sheet areas will be selected and the small clusters will be filtered out. The clusters that have number of voxels less than (maxSHTclusterSize/divider) will be discarded.

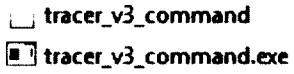
APPENDIX D

MANUAL OF *STRANDTWISTER*

1. Input file

StrandTwister takes a MRC file that contains one chunk of density of single sheet as input. The chunk of density can be the detected density output from *SSEtracer* or *SSElearner*. Note that if the output of *SSEtracer/SSElearner* contains sheet density from multiple β-sheets, then they need to be separated. One way to do so is to cut in Chimera if visual separation is possible. Place the β-sheet MRC file and the downloaded executable file in the same folder to run it. Currently there are two versions:

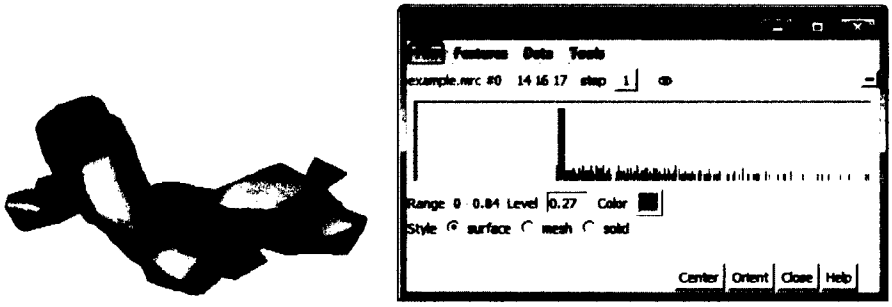
(1) Binary file that compiled under Linux (64bit)
(2) .exe file that compiled under Windows 7 (64bit)



2. Parameter

StrandTwister is a fully automatic tool. The only parameter is a threshold of the density map, which is the user estimation for the size/thickness of the β-sheet density. If your input MRC contains the density that has already been filtered by a given threshold (for example the detected sheet density from *SSEtracer* is filtered by a given input threshold as the Volume View shows below), you can enter 0 as the threshold. Otherwise please input a density threshold for your input sheet density.

The input density:



And the command line:

```
sirius:~/work/13 fall/SheetTwister/strandtwister v2> ./strandtwister_v2_command example 0


Working on example
Filtering the map ...
```

3. Run the program and read the detected result

StrandTwister will output the Top ten detected beta-traces (if there are more than ten right-handed sets been detected) as PDB files in the same folder after few seconds. "trans #" means the sampled possible translations, from 0 to 2 there are three sampled translations (every 1.5Å a translation sampling). "orient #" means the sampled possible orientations, from 0 to 170 degree there are eighteen sampled orientations (every 10 degree an orientation sampling).

```
Top 0 twist 15.4688:    trans 1 - orient 170
Top 1 twist 15.1739:    trans 2 - orient 170
Top 2 twist 13.8909:    trans 1 - orient 0
Top 3 twist 13.8543:    trans 1 - orient 160
Top 4 twist 13.5302:    trans 0 - orient 0
Top 5 twist 13.1992:    trans 2 - orient 0
Top 6 twist 13.0184:    trans 0 - orient 170
Top 7 twist 12.6977:    trans 1 - orient 10
Top 8 twist 12.5161:    trans 2 - orient 160
Top 9 twist 11.7673:    trans 2 - orient 20
```

The statistic results will be saved in a text file named "***_bestResult.txt".

pdbID	fitting Err	best-set	best #strds	#strds	2-way	best #AA	#AA
example	1.47229	trans_0_orient_0	5	5	1.6019	24	31

The best-set with minimum 2-way distance will be recorded in this .txt file with the information of least-square fitting error, detected number of strands in this best set, number of strands in the true structure, the number of amino acids detected in this best set, and the total number of amino acids in the true structure. Note that this statistics will only be calculated if you put the .pdb file of true structure in the folder.

You can also load and view the results in UCSF Chimera, the best detection among the Top ten output in this example is the Top 4 twist set (below left, trans_0_orient_0). You can also load the true structure to check it (below right).

APPENDIX E

MANUAL OF *STRANDROLLER*

1. Input file

StrandRoller takes a MRC file that contains one complete chunk of β-barrel density as input. The chunk of density can be the detected density output from *SSEtracer* or *SSElearner*. Note that if the output of *SSEtracer/SSElearner* contains sheet density from multiple β-sheets, then they need to be separated. One way to do so is to cut in Chimera if visual separation is possible. Place the β-barrel MRC file and the downloaded executable file in the same folder to run it. Currently there are two versions:

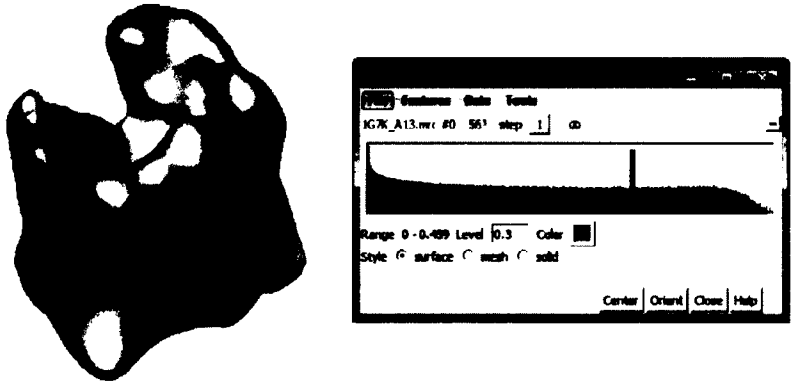
(1) Binary file that compiled under Linux (64bit)
(2) .exe file that compiled under Windows 7 (64bit)

StrandRoller_v1_command
StrandRoller_v1_command.exe

2. Parameter

StrandRoller is a fully automatic tool. The only parameter is a threshold of the density map, which is the user estimation for the size/thickness of the β-barrel density. If your input MRC contains the density that has already been filtered by a given threshold (for example the detected sheet density from *SSEtracer* is filtered by a given input threshold), you can enter 0 as the threshold. Otherwise, please input a density threshold for your input barrel density.

The input density:



And the command line:

```
-> ./StrandRoller_v1 1G7K_A13 0.3
```

3. Run the program and read the detected result

StrandRoller will output the sampled beta-traces with all possible tilt angles as PDB files in the same folder after few seconds.

The statistic results will be saved in a text file named as "***_bestResult.txt".

bestResult.txt

pdbID	best-set	detect #strd	total #strd	2-way	detect #AA	total #AA
1G7K A13	trans 0 orient 40	11	11	1.8015	85	124

The best-set with minimum 2-way distance will be recorded in this .txt file with the information of detected number of strands in this best set, number of strands in the true structure, 2-way distance, the number of amino acids detected in this best set, and the total number of amino acids in the true structure. Note that this statistics will only be calculated if you put the .pdb file of true structure in the folder.

You can also load and view the results in UCSF Chimera, the best detection among the Top ten output in this example is the trans_0_orient_40 set (below left,). You can also load the true structure to check it (below right).

VITA

Dong Si
Department of Computer Science
Old Dominion University
Norfolk, VA 23529

Dong Si received his B.S. in Electronic Information Science and Technology with honors from Nanjing University, China. In fall 2009, Dong joined the Computer Science department of Old Dominion University as a graduate student and then transferred to Ph.D. program in fall 2010.

During his Ph.D. study, a series of fully automatic SSE detection methods has been developed for solving the challenging problem of secondary structure element detection from cryo-EM density images. Dong"s recent research has been focused on a critical and challenging problem of detecting beta-strands from cryo-EM density images at medium resolutions. The results of his work have been published on multiple scientific journals and peer-reviewed conference proceedings. Over the years, Dong"s research has included visual analytics, feature detection and pattern recognition, machine learning, 2D/3D data processing, geometric modeling, and bioinformatics.