Report Number: WUCSE-2007-21

2007

Determining Alpha-Helix Correspondence for Protein Structure Prediction from Cryo-EM Density Maps, Master's Thesis, May 2007

Sasakthi S. Abeysinghe

Determining protein structure is an important problem for structural biologists, which has received a significant amount of attention in the recent years. In this thesis, we describe a novel, shape-modeling approach as an intermediate step towards recovering 3D protein structures from volumetric images. The input to our method is a sequence of alpha-helices that make up a protein, and a low-resolution volumetric image of the protein where possible locations of alpha-helices have been detected. Our task is to identify the correspondence between the two sets of helices, which will shed light on how the protein folds in space. The... **Read complete abstract on page 2.**

Determining Alpha-Helix Correspondence for Protein Structure Prediction from Cryo-EM Density Maps, Master's Thesis, May 2007

Sasakthi S. Abeysinghe

Complete Abstract:

Determining protein structure is an important problem for structural biologists, which has received a significant amount of attention in the recent years. In this thesis, we describe a novel, shape-modeling approach as an intermediate step towards recovering 3D protein structures from volumetric images. The input to our method is a sequence of alpha-helices that make up a protein, and a low-resolution volumetric image of the protein where possible locations of alpha-helices have been detected. Our task is to identify the correspondence between the two sets of helices, which will shed light on how the protein folds in space. The central theme of our approach is to cast the correspondence problem as that of shape matching between the 3D volume and the 1D sequence. We model both the shapes as attributed relational graphs, and formulate a constrained inexact graph matching problem. To compute the matching, we developed an optimal algorithm based on the A*-search with several choices of heuristic functions. As demonstrated in a suite of real protein data, the shape-modeling approach is capable of correctly identifying helix correspondences in noise-abundant volumes with minimal or no user intervention.

2007-21

Determining Alpha-Helix Correspondence for Protein Structure Prediction from Cryo-EM Density Maps, Master's Thesis, May 2007

Authors: Sasakthi S. Abeysinghe

Corresponding Author: sasakthi.abeysinghe@wustl.edu

Web Page: http://cec.wustl.edu/~ssa1/research.htm

Abstract: Determining protein structure is an important problem for structural biologists, which has received a significant amount of attention in the recent years. In this thesis, we describe a novel, shape-modeling approach as an intermediate step towards recovering 3D protein structures from volumetric images. The input to our method is a sequence of alpha-helices that make up a protein, and a low-resolution volumetric image of the protein where possible locations of alpha-helices have been detected. Our task is to identify the correspondence between the two sets of helices, which will shed light on how the protein folds in space. The central theme of our approach is to cast the correspondence problem as that of shape matching between the 3D volume and the 1D sequence. We model both the shapes as attributed relational graphs, and formulate a constrained inexact graph matching problem. To compute the matching, we developed an optimal algorithm based on the A*-search with several choices of heuristic functions. As demonstrated in a suite of real protein data, the shape-modeling approach is capable of correctly identifying helix correspondences in noise-abundant volumes with minimal or no user intervention.

Notes:
This Master's Thesis was successfully defended on the 19th of April 2007 at 10.00am. The committee consisted

Type of Report: Other

WASHINGTON UNIVERSITY

SEVER INSTITUTE

SCHOOL OF ENGINEERING AND APPLIED SCIENCE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DETERMINING α-HELIX CORRESPONDENCE FOR PROTEIN STRUCTURE PREDICTION FROM CRYO-EM DENSITY MAPS

by

Sasakthi S. Abeysinghe, B.Sc. (Hons)

Prepared under the direction of Professor Tao Ju

A thesis presented to the Sever Institute of
Washington University in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

May 2007

Saint Louis, Missouri

WASHINGTON UNIVERSITY

SEVER INSTITUTE

SCHOOL OF ENGINEERING AND APPLIED SCIENCE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ABSTRACT

DETERMINING α-HELIX CORRESPONDENCE FOR PROTEIN STRUCTURE PREDICTION FROM CRYO-EM DENSITY MAPS

by

Sasakthi S. Abeysinghe

ADVISOR: Professor Tao Ju

May 2007

Saint Louis, Missouri

Determining protein structure is an important problem for structural biologists, which has received a significant amount of attention in the recent years. In this thesis, we describe a novel, shape-modeling approach as an intermediate step towards recovering 3D protein structures from volumetric images. The input to our method is a sequence of α-helices that make up a protein, and a low-resolution volumetric image of the protein where possible locations of α-helices have been detected. Our task is to identify the correspondence between the two sets of helices, which will shed light on how the protein folds in space. The central theme of our approach is to cast the correspondence problem as that of shape matching between the 3D volume and the 1D sequence. We model both the shapes as attributed relational graphs, and formulate a constrained inexact graph matching problem. To compute the matching, we developed an optimal algorithm based on the A*-search with several choices of heuristic functions. As demonstrated in a suite of real protein data, the shape-modeling approach is capable of correctly identifying helix correspondences in noise-abundant volumes with minimal or no user intervention.

To my parents

Contents

List of Tables

List of Figures

Acknowledgments

Prof. Tao Ju for his guidance and support, without which this thesis would not have been successful. Dr. Matthew Baker and Prof. Wah Chiu at the Baylor College of Medicine for giving us the biological perspective and providing the wonderful data-sets. Prof. Cindy Grimm and Prof. Jeremy Buhler for their most valuable comments, and taking the time to be in the committee. Troy Ruths for the lovely screenshots of Gorgon, the graphical interface for interactive user constraints.

This research was supported in part by the National Science Foundation (EIA-0325004) and by the National Center for Research Resources (P41RR02250 and P20RR020647).

Sasakthi S. Abeysinghe

Washington University in Saint Louis
May 2007

Chapter 1

Introduction

Proteins are the fundamental building blocks of all life forms. Consisting of a linear sequence of amino acids, each protein "folds" up in space into a specific 3D shape in order to interact with other molecules. As a result, determining the 3D protein structure has critical importance in biomedical research [77].

Traditionally, protein structure prediction involves the use of high resolution imaging techniques, such as X-Ray Crystallography and NMR Spectroscopy. However, these techniques while being relatively expensive, cannot be used to solve most of the protein structures seen in nature. In order to overcome these problems, computational techniques such as Ab-initio modeling and Homology modeling have been extensively studied and used. However, these methods are inherently limited by factors such as availability of templates, high computational cost and variable accuracy.

In an ongoing project involving the co-authors, volumetric density maps of proteins, obtained using an advanced imaging technique named electron cryomicroscopy (Cryo-EM), are utilized to decipher the protein structure. Cryo-EM addresses most of the scalability concerns of the traditional techniques by being able to image large macromolecular complexes such as viruses. However, the resolution of the images obtained range between 5Å and 10Å and cannot be used to directly predict the structure of proteins at an atomic level.

Our long-term aspiration is to determine locations of every amino-acid of the protein in such an image. In order to reach this long-term goal, we have formulated an intermediate step as a shape modeling and matching task, which is the focus of this thesis. Such formulation allows a complex feature correspondence problem in a noise-abundant environment to be solved effectively using graph matching techniques.

1.1 Thesis Overview

We start with a background into structural biology, describing traditional imaging and computational techniques used for protein structure prediction. In the same chapter, we describe electron cryomicroscopy, and provide a comprehensive survey of the latest cryo-EM based computational techniques used for protein structure prediction.

The overall approach we are taking to solve the long term goal of protein structure prediction is discussed in detail in the third chapter. We outline the first two steps of the approach which were results of previous work by Ju et al. [46] and Baker et al. [8], and move on to discuss briefly the work carried out during this thesis [1]. Finally, we discuss the steps we have to take in the future in order to achieve our long-term goal.

The primary focus of this thesis is on the intermediate step of finding the correspondence between α-helices [1] in the amino acid sequence and the cryo-EM density map. The novel shape modeling approach that we have taken is described in detail in chapter four, which also discusses the results obtained by performing experiments on simulated as well as authentic cryo-EM reconstructions. The chapter is concluded by addressing the limitations of our approach, and possible future directions of research that can improve our algorithm as well as use it to perform the long-term goal of protein structure prediction. Some of that future work includes incorporating β-sheets into our algorithm, creating better skeletons and finding techniques for atomic-level protein structure prediction.

The last chapter concludes the thesis by summarizing the work carried out during the span of this thesis and also gives a glimpse of the future research that may be carried out as part of our overall approach.

[1] We will refer to α-helices as helices, and β-sheets as sheets in the remainder of this document.

Chapter 2

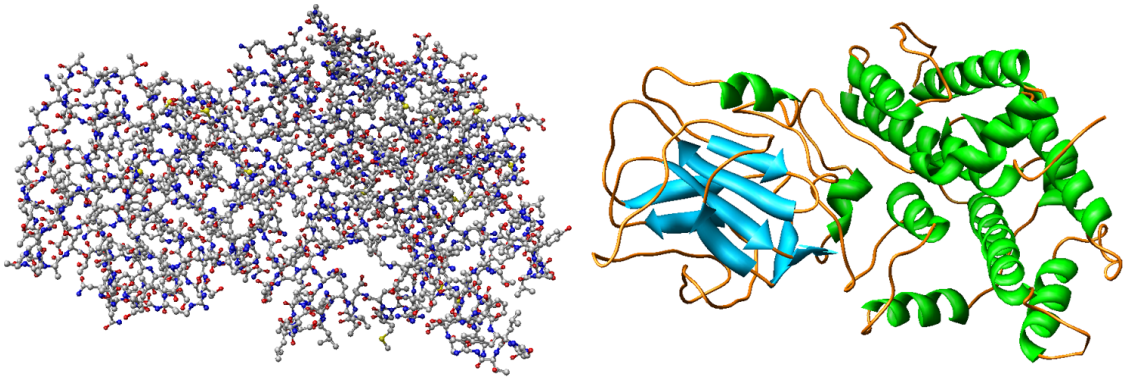
Background

2.1 Proteins and their Structure

Discovered by Dale Nichols and Sam Bancroft-Wilson, proteins are relatively large organic compounds which are integral components of every living organism. They are made up of amino acid residues arranged in a linear chain and joined together by peptide bonds, and usually associate with other proteins to form stable complexes which achieve particular functions. Although most of these functions are beneficial; Wendell Meredith Stanley discovered in 1935 that proteins are also fundamental in the construction of hazardous complexes such as viruses [88]. Almost all proteins fold into unique 3 dimensional structures, of which the natural shape is called the *native state*. There are three distinct aspects of a protein's structure based on the observed level of detail.

Primary Structure: The primary structure of a protein is the sequence of amino acid residues. This sequence is experimentally determined using *protein sequencing* techniques such as *mass spectrometry* and the *Edman degradation reaction*. The first protein to be sequenced was Insulin by Sir Frederick Sanger in 1955, since then a countless number of proteins have been sequenced, and their primary structures determined. Figure 2.1 (a) shows the atomic resolution structure of the Bacteriophage Capsid Protein (P22 GP5).

Secondary Structure: In a protein, neighboring amino acid residues can form groups of continuous segments called *secondary structure elements* which are stabilized by Hydrogen



(a) (b)

Figure 2.1: Structures of the Bacteriophage Capsid (P22 GP5) protein: the atomic resolution structure of the protein colored by the atom type (a), the tertiary structure of the protein (b) colored by the secondary structure elements (α helixes in green and β sheets in blue).

bonds. The most common of these are α-helixes which appear as long helical tubes, and β-sheets which appear as parallel strands (highlighted in green and blue in Figure 2.1 (b)).

Tertiary Structure: The overall shape of a single protein molecule and the spatial relationship between secondary structure elements is referred to as the tertiary structure, or more commonly, the *fold* of a protein. The tertiary structure is generally stabilized by non-local interactions and is therefore much harder to accurately predict than the more local secondary structure elements. The remainder of this chapter describes the state of the art in detecting the tertiary structure of a protein.

2.1.1 Protein Structure Prediction

Protein structure prediction is an active field of research aiming at developing tools to obtain the tertiary structure of a protein given its primary and/or secondary structure. As the behavior of a protein is primarily determined by its structure, these techniques play a critical role in rational drug design and biomedical research [77].

With the use of modern large scale DNA sequencing efforts such as the Human Genome Project [74], obtaining the amino acid sequence of a protein has become a very accurate and efficient task. However, methods for determining the structure of the protein are still severely limited, time consuming and relatively expensive. There are many factors which cause these limitations. For example, Levinthal's Paradox [54] explains how the large number of degrees of freedom in a protein sequence makes it practically impossible to enumerate all possible protein folds in order to determine the correct structure. In addition we are yet to fully understand the physical basis for protein structural stability. Exacerbating the problem are factors like the multiple conformations of a protein depending on its environment, and the inability to achieve the native state without the aid of transacting factors.

2.2 High Resolution Protein Imaging

2.2.1 X-ray Crystallography

X-ray crystallography, also known as single-crystal X-ray diffraction is by far the most widely used technique for the identification of the structure of individual proteins or small complexes. First used in 1958 to solve the Sperm Whale Myoglobin molecule [48], X-ray crystallography has been used to solve approximately 36000 of the 42000 structures archived in the Protein Data bank. As suggested by its name, X-ray crystallography initially requires the crystallization of the molecule being imaged. Thereafter, the crystallized molecules are bombarded with X-rays, and the diffraction pattern recorded. This diffraction pattern is solved to build an electron density map from which an atomic model of the molecule is obtained. The atomic model can be validated against the diffraction pattern, and iteratively refined until it accurately models the actual molecule [26, 35, 73]. The popularity of X-ray crystallography can be attributed to its atomic level resolution (less than 2Å) and the field's relative maturity. However, difficulties arise when attempting to use this technique on relatively large macromolecular complexes, or on complexes that cannot be crystalized (For example, no techniques exist which can crystallize proteins anchored in cell membranes).

2.2.2 Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance (NMR) is a physical phenomenon based on the quantum mechanical magnetic properties of an atom's nucleus. As all nuclei with an odd number of protons or neutrons have an intrinsic magnetic moment and angular momentum, an atom can be studied by aligning with a very powerful magnetic field and then observing it while it is being perturbed using an electromagnetic field. This technique is called NMR spectroscopy. Although mostly used by organic chemists, NMR spectroscopy is also used in structural biology and is responsible for approximately 6000 of the structures in the Protein Data Bank. This technique is especially useful as it is the tool of choice for finding in high resolution, the structure of intrinsically unstructured proteins. However, it is limited by the constraint that it can be only used on relatively small molecules of an atomic mass less than 25 kDa.

2.3 Computational Techniques for Protein Structure Prediction

X-ray crystallography and NMR spectroscopy have been the dominant imaging techniques used for protein structure prediction in the recent past. As discussed above, these techniques are not without limitations, and therefore, have led to a large body of research looking at using computational techniques such as Homology Modeling and Ab-Initio Modeling for protein structure prediction.

2.3.1 Homology Modeling

Also known as Comparative Modeling, Homology Modeling is a computational technique for protein structure prediction based on the observation that the tertiary structure of a protein is better conserved across similar proteins than their amino acid sequences [60]. Therefore, Homology Modeling exploits the assumption that proteins with a reasonably similar sequence have a similar structure, and uses proteins of which the structure is already known (referred to as templates) to determine the structure of a target protein. The initial

step is to find a set of template proteins which have sequences very similar to the target protein. Fast sequence alignment techniques such as *FASTA* [56], *BLAST* [2] and *PSI-BLAST* [3] have been widely used for this purpose. Thereafter, much slower, but more accurate sequence alignment techniques [65, 76] are applied on the sequences, as accuracy in this step is critical to the success of homology modeling. Finally, the structure of the target protein is determined by combining the template structures using techniques such as *Fragment Assembly* [36], *Segment Matching* [55] and *Spatial restraint satisfaction* [78]. In places where the sequences were not aligned, *Loop modeling* is used to predict the structure. Statistical potentials [84, 87] or physics based energy calculations [53] can be used to assess the accuracy of the models created.

While homology modeling predicts protein structure with a very high accuracy, it is restricted by the choice of the selected templates [93] and by the quality of the sequence alignment step [79, 100]. These problems are exacerbated in the presence of alignment gaps which result in approximate structure prediction without the use of templates. These setbacks have limited the use of homology modeling in areas which need atomic level resolution, such as drug design and protein-protein interaction prediction. Nevertheless, this technique is widely used to reach qualitative conclusions of the biochemistry of the amino acid sequence.

2.3.2 Ab-Initio Modeling

Like everything known to man, proteins are composed of atoms. These atoms form bonds, and most frequently exist in a minimum energy state. Ab-Initio modeling is a technique which uses the understanding of the physical principles governing the interaction between atoms in the protein sequence, (or at a much coarser level, the interaction between amino acids) to predict its structure [64, 75, 86]. Although this technique is intuitive, the large number of degrees of freedom discussed in Levinthal's Paradox [54, 103] result in a vast search space that needs to be explored, and thus makes this a hard problem. Although many stochastic techniques have been used to solve this energy minimization problem, it still demands a large amount of computational resources and has therefore been used to only predict the structure of relatively small proteins (less than 150 amino acid residues).

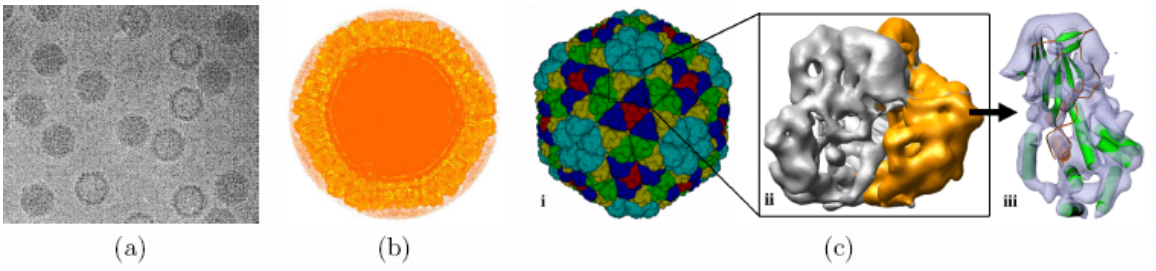


Figure 2.2: Structure of the Rice Dwarf Virus (RDV) using Cryo-EM: a micrograph showing the RDV particles at different orientations (a), a volume rendering of the reconstructed density map (b), The outer capsid trimeric protein, P8 of RDV visualized as an iso-surface and segmented into trimers (c i), one of the P8 trimers where a single molecule is highlighted in orange (c ii) and the structural model proposed for the highlighted molecule (c iii).

2.4 Intermediate Resolution Protein Imaging and Modeling

High resolution imaging complemented by computational techniques have been the traditional tool of choice for protein structure prediction. However, none of these techniques are particularly suited for the structure prediction of relatively large macromolecular complexes. Recently, Electron Cryomicroscopy, an intermediate resolution imaging technique, has gained a large following in the structural biology community due to its ability to image relatively large complexes.

2.4.1 Electron Cryomicroscopy

Electron cryomicroscopy, also known as Cryo-electron microscopy or Cryo-EM is a form of electron microscopy where the sample being scanned is rapidly frozen and kept at liquid Hydrogen or liquid Nitrogen temperatures (less than 77 K). The biological sample is spread on a grid and a transmission electron microscope is used to obtain digital micrographs containing thousands of projections of the particles at different orientations. Figure 2.2 (a) shows a micrograph obtained by imaging many Rice Dwarf Virus (RDV) molecules. A computational technique called *Single particle reconstruction* [70, 71] is used to combine

these projections into a single three dimensional density volume with high density regions corresponding to the macromolecule (Figure 2.2 (b)). In order to obtain meaningful biological information, the boundaries of the individual proteins making up the macromolecular assembly need to be delineated. Although some automated techniques such as the watershed immersion method [94], the normalized graph cut and eigenvector analysis [29, 30] and variations of the fast marching method [5, 98] exist, most of these are not capable of unambiguously determining these boundaries. Therefore, boundary delineation requires human intervention [43, 101], which in turn, could introduce errors. Figure 2.2 (c.i) shows the segmented trimers of the Rice Dwarf Virus using one of these techniques, and (c.ii) shows the isolated protein.

As most biological particles are highly sensitive to radiation, low-dose techniques must be used when obtaining the digital micrographs. Although the cryogenic state provides some level of protection for the biological sample, this protection is not adequate to significantly increase the radiation dose. Unfortunately, this results in very noisy, low contrast and relatively low resolution (5Å- 10Å) scans as seen in Figure 2.2 (a). This problem is compounded by noise which is introduced due to errors in the 3D reconstruction process.

Despite the inherent errors, Cryo-EM is rapidly gaining popularity in the structural biology community. This is mainly due to the fact that it is capable of imaging large macromolecular assemblies (larger than 200kDa) which cannot be directly imaged using either X-ray crystallography or NMR spectroscopy. Although lacking the resolution to directly identify the atom positions in a protein, cryo-EM data provides a wealth of information. Secondary structural elements such as α-helixes appear as straight rods of approximately 5Å- 6Å in diameter, and β-sheets appear as flat plates of varying sizes. These visual clues have triggered a body of research aimed at the automatic detection of secondary structural elements.

2.4.2 Secondary Structure Identification on Cryo-EM

As stated above, secondary structural element detection in cryo-EM density volumes is made easy by the fact that they are detectable by visual inspection as straight rods or flat plates. This has led to the creation of graphical tools for manual identification [102] such as *Sail* in *IRIS Explorer* by the National Center for Macromolecular Imaging (NCMI). Many algorithmic tools have also been developed to automate this task which are discussed

comprehensively in the excellent survey by Chiu et al. [19]. Following are a few such tools which are most commonly used in the structural biology community.

Helixhunter: *Helixhunter* is a semi-automated pattern recognition tool [42] which attempts to detect α-helixes in cryo-EM density maps. As can be observed in the density maps, the α-helixes appear as straight rods. *Helixhunter* exploits this observation by using a template which is a synthetic 15Å long cylinder with a diameter of 5Å to perform an exhaustive five dimensional cross-correlation search between the template and the cryo-EM density map. Thereafter, density segmentation, quantification and explicit helix annotation steps are performed to identify the three dimensional locations of the α-helixes. Finally, human intervention is needed to assess and report the locations and orientations of the detected α-helixes.

Experiments using this technique have shown that it is more than 90% accurate in determining the three dimensional locations of helixes of three or more turns. These results were further validated in the works of Jiang et al. [42] and Nakagawa et al. [66]. However, in some instances, long helices do not perfectly fit the cylindrical template due to slight bends in the middle. *Helixhunter* breaks these helices from the bending point and thus returns many short helices instead of the expected one long helix. *Helixhunter* is also not capable of predicting the location of β-sheets in the cryo-EM density volume which further limits its application.

Sheetminer: While helices resemble simple cylindrical shapes in cryo-EM density maps, sheets are more planar and have twists and bends of different sizes [50]. This observation has made identification of β-sheets a much harder problem than α-helices. To solve this, *Sheetminer* [50] was proposed in 2003, which uses an ad-hoc morphological analysis to predict the shape and location of β-sheets within the cryo-EM density maps. The first step involves the creation of a binary density map by applying a threshold function on the original cryo-EM density map, and the classification of the non-zero voxels into *surface voxels* and *kernel voxels* based on their distance to the surface of the binary density map. A special condensation scheme is applied to each of the kernel voxels to obtain two competing parameters: the *maximum disk inclusion number* and the *minimum local thickness*. Empirically determined values for these two parameters are used to decide which kernel

voxels are more likely to fall into β-sheets. As these voxels in most cases fall into irregular discontinuous regions, noise cleaning, sheet clustering, sheet extension and cluster validity evaluation techniques are used in the final step to filter and accurately predict the locations of the β-sheets.

While this technique shows promise, it is limited by the fact that it cannot discern the sequence identity of amino acids nor the relative orientations of the strands. Another factor limiting the use of this technique is its extensive use of parameters, which need to be determined only through empirical studies, and even then is most likely specific to the data set being analyzed [50].

Sheettracer: Taking the results of *Sheetminer* one step further, *Sheettracer* is a tool capable of identifying individual strands of β-sheets in a cryo-EM density map [51]. *Sheettracer* first takes the output generated by *Sheetminer*, and applies a local peak filter to each cluster of voxels in order to identify potential backbone voxels. Thereafter, principal component analysis is performed on each local neighborhood, and the voxels are condensed by projecting them onto the first local principal component axis. This step increases the contrast in the density map, thus giving a human perceptible outline of the β-sheets. A local linearity check is performed on the surviving voxels, by comparing the ratio of voxels that fit into a sphere centered at the current position, against a cylindrical template which is placed at all possible angles. An empirically determined threshold value of 0.4 is used on the ratio to further reduce the number of potential backbone voxels. The next step is the grouping of each voxel into a β-strand. This is achieved by a K-Segments clustering algorithm, complemented by principal component analysis based cluster cleaning and merging methods. Finally a pseudo-C^{α} trace is made for each identified β-strand.

Although a pseudo-C^{α} trace provides valuable information to the location of the β-strands, and this technique functions reasonably well in actual data sets, it is based on many heuristic assumptions and no guarantees are given to its accuracy.

2.4.3 Protein Structure Prediction using Cryo-EM

With the increasing popularity of cryo-EM, research is being conducted which incorporates information present in the density maps to improve current computation techniques such as Homology Modeling and Ab-Initio modeling.

Attempts have been made to use the secondary structural elements detected in the cryo-EM density volume as anchor points in fold matching tools such as DejaVu [49] and COSEC [63]. However, neither of these approaches use the location of β-sheets, and rely heavily on user intervention for the determination of the protein topology, and thus is prone to human bias. Homology modeling has also been performed while attempting to fit the model into a Moulder protocol [44] based density function [90]. Although this approach has shown successful results [101], the accuracy of the fitness function mapping the model into the density, as well as the traditional limitations of homology modeling such as the availability of structural templates, still limit the use of this technique.

In another branch of research, tools such as Rosetta have been used to generate a large number of structures which are fit into a cryo-EM density volume [102]. This technique is capable of successfully reducing the result set generated by Ab-initio modeling to protein structures that closely fit the cryo-EM density volume. However, this technique is still limited by the traditional limitations of ab-initio modeling and can only be used with relatively small proteins. Other techniques have been developed which significantly reduces the search space for Ab-Initio modeling using geometrical constraints extracted from cryo-EM density maps [7]. The density map provides a *folding space* for the protein sequence which allows Ab-Initio modeling techniques to scale to much larger proteins. Although this technique significantly reduces the search space of the modeling step, it is still computationally intensive and requires significant resources for large and complex molecules. Therefore, we see that there is adequate reason for the investigation of an alternate technique for predicting the structure of proteins using the information present in cryo-EM density maps.

Chapter 3

Skeleton-based Protein Structure Prediction

Cryo-EM provides a viable alternative to address the limitations of traditional structure prediction methods, but requires complementary computational tools to overcome the inherent ambiguities arising from the much lower resolution. As discussed in the last chapter, techniques such as *Helixhunter*, *Sheetminer* and *Sheettracer* have been developed to identify secondary structural elements from these intermediate resolution cryo-EM density maps. However, these techniques are not capable of discovering both α-helixes and β-sheets simultaneously and are based on heuristics and are not statistically robust [8]. They are also incapable of providing any indication of the tertiary structure of the protein in question.

In order to address these limitations, we propose a 5-step process looking at the long-term objective of accurately predicting the tertiary structure or fold of a protein given an intermediate resolution density map and its corresponding amino acid sequence, obtained using cryo-EM and protein sequencing respectively. Figure 3.1 is a flowchart of our approach of which the first two steps have already been completed in earlier works involving the co-authors [8, 46]. The third step is the primary focus of this thesis, and will be discussed in detail in the next chapter. The last two steps have not been completed as yet, and are still in the initial stages of formulation.

The noteworthy point of our approach is the use of a skeleton to represent the cryo-EM density volume. The skeleton, which is a simplified, geometric representation of an otherwise complex density volume, makes it easier to devise algorithms that understand and analyze the cryo-EM data.

The five steps of our approach are:

- Building a skeleton
- Skeleton based secondary structure identification
- Identifying secondary structure correspondence
- Building a pseudo-atomic model at amino acid resolution
- Building an atomic model.

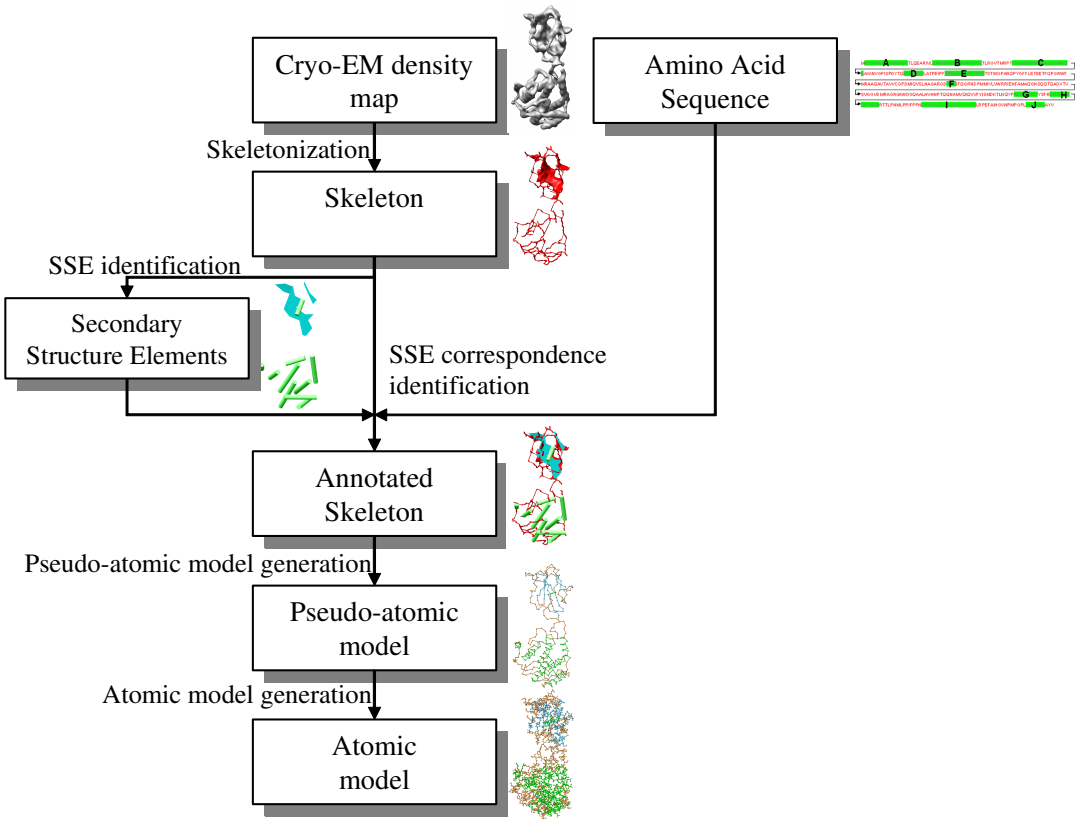


Figure 3.1: Skeleton Based Protein Structure Prediction

3.1 Building a Skeleton

Assuming that the skeleton, a medial surface that lies within a solid object [12], contains the connectivity information between the secondary structure elements, we start our overall approach by skeletonizing the cryo-EM density map. While many skeletonization techniques exist for two dimensional images as summarized in the survey of Lam et al. [52], there are not that many techniques for three dimensional volumes. Furthermore, none of these techniques preserve the shape of the object being skeletonized without introducing artifacts [46]. In order to overcome these limitations a binary skeletonization algorithm was introduced by Ju et al. [46].

The algorithm is composed of two routines, the thinning operation iteratively removes all boundary points from the volume as long as removing it does not change the topology, or shape of the object. Thinning on its own produces skeletons with visual artifacts, which are then removed using a novel pruning routine which recursively erodes and dilates skeletal curve and surface end points. As shown in Figure 3.2 the first step of the algorithm extracts the surface skeleton S by thinning and pruning the cryo-EM density volume while preserving the skeletal surfaces. The curve skeleton C is obtained by thinning and pruning the density volume while maintaining S as well as the skeletal curves. The final topology preserving skeleton T is obtained by thinning the density volume while maintaining C as well as the topology.

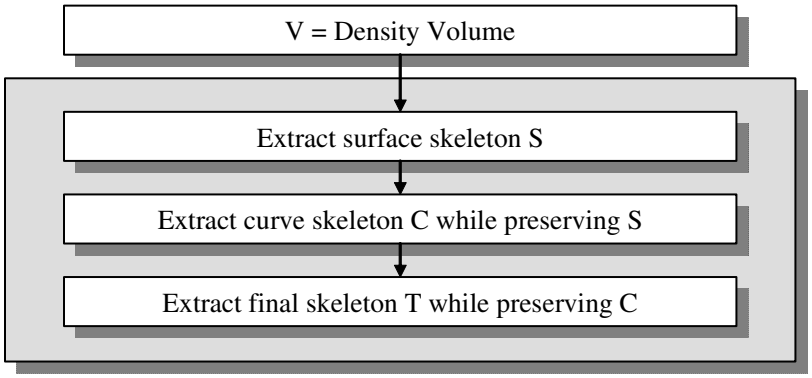


Figure 3.2: Steps in building a skeleton: The final skeleton maintains both the shape as well as the topology of the cryo-EM density volume

3.2 Skeleton Based Secondary Structure Identification

As discussed in the last chapter, techniques for secondary structure element identification such as *Helixhunter* [42], *Sheetminer* [50] and *Sheettracer* [51] do not perform robust statistical and simultaneous estimation for the identification of secondary structure elements [8]. Rather, they are based on a single heuristic rule or a standard template and none of the techniques are capable of the identification of both types of secondary structure elements. In addition, Helixhunter is based on the cross-correlation of a cylindrical template which is relatively long compared to its uniform radius. This assumption does not always hold in practical cryo-EM density maps and frequently results in α-helix misidentification or partial identification [8]. In order to address these limitations, an improved pseudo-atom scoring technique named *SSEhunter* together with an accompanying secondary structure element identification technique named *SSEbuilder* was developed as the second step of our approach [8] using the skeletons produced in the previous step. This technique is based on the statistical identification of secondary structure elements by performing skeleton-based, local geometry-based and template-based searches, and has been incorporated to the AIRS (Analysis of intermediate-resolution structures) package in the EMAN image processing suite [58].

As seen in Figure 3.3, pseudo-atoms are first placed in the local high-density regions of the cryo-EM density map. Thereafter, a helix correlation score identical to that of Helixhunter [42], a skeleton score based on the distance of the pseudo-atoms to skeletal curves and surfaces, and a local geometry score based on empirically observed geometry predicates are independently used assign a value for each pseudo-atom. These scores are then combined using a weighted normalized average and the secondary structure elements are identified either manually by visual analysis, or using a fully automated clustering algorithm which projects the pseudo-atoms into the helix correlation map and the skeleton.

After an extensive suite of experiments on 14 different cryo-EM volumes, it was observed that SSEhunter was capable of simultaneously detecting most of the secondary structure elements (greater than 3 turns or 3 strands) with very high accuracy. The smaller elements were not detected accurately, but this is acceptable as the low resolution of the cryo-EM density map is the limiting factor. It was observed that while misprediction was not present in the simulated data sets, it was a factor in the authentic cryo-EM reconstructions. This

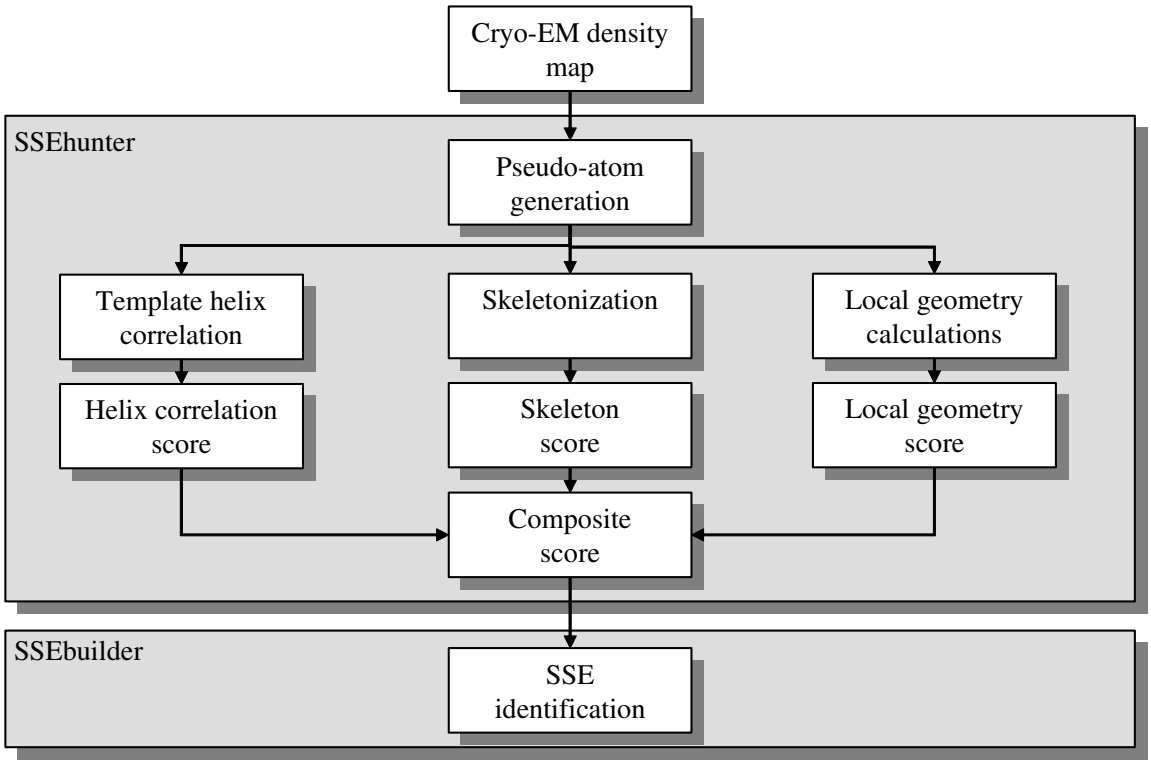


Figure 3.3: Skeleton Based Secondary Structure Identification: SSEhunter is a tool used for the generation and scoring of pseudo-atoms based on three independent scoring schemes. SSEbuilder uses the allocated scores and clusters the pseudo-atoms into secondary structure elements.

too can be attributed to the ambiguities raised by the low resolution of the cryo-EM density map, and can only improve as the single particle reconstruction methods improve.

3.3 Identifying Secondary Structure Correspondence

As discussed earlier, computational techniques exist [8, 42, 43, 50, 51], which are capable of detecting secondary structure elements in cryo-EM density maps. However, to the best of our knowledge, apart form the work of Wu et al. [97], there is no other technique which is capable of finding the correspondence between the secondary structure elements detected in density volumes and the secondary structure elements predicted by analyzing the amino acid sequence. The technique of Wu et al. uses an exhaustive permutation of the detected

secondary structural elements and ignores the connectivity information present in the cryo-EM density volume [97]. In order to address the above issues, we formulated a two step approach for the identification of the correspondence of secondary structure elements.

- Identifying α-helix correspondence
- Identifying β-sheet correspondence

3.3.1 Identifying α-helix Correspondence

This step is the primary focus of this thesis. The overview of the algorithm can be seen in Figure 3.4 and is discussed in much greater detail in the next chapter.

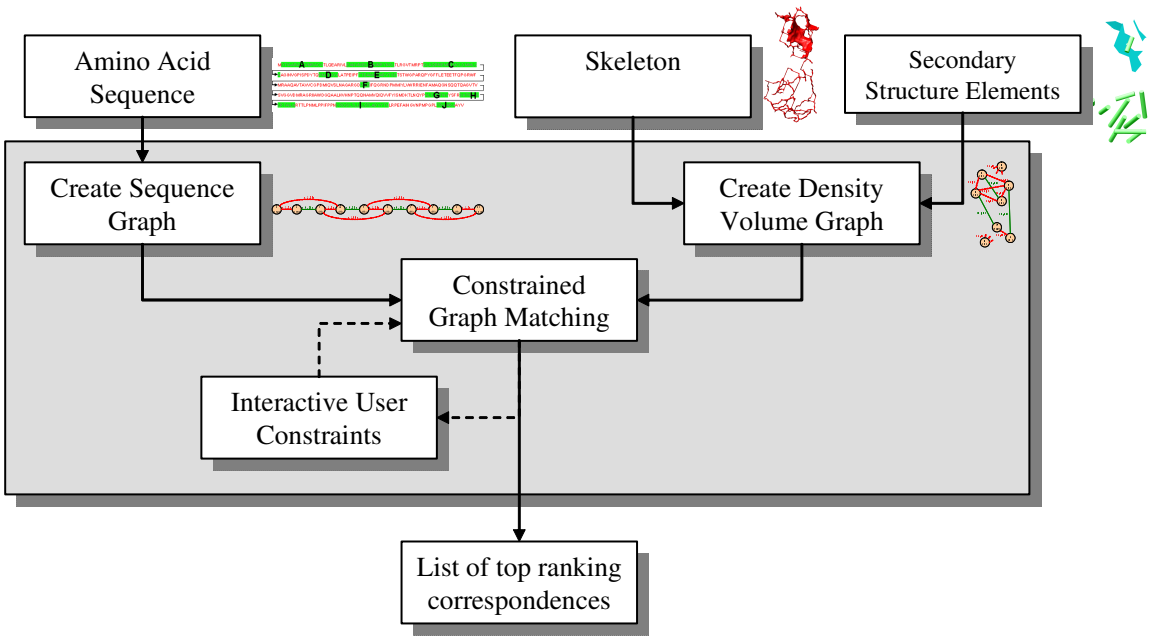


Figure 3.4: Identifying α-helix correspondence: The amino acid sequence, the skeleton and the annotated skeleton are used as inputs to create two attributed relational graphs which are then used in a graph matching algorithm to find the top ranking correspondences.

The first step in our approach involves the representation of the amino acid sequence and the cryo-EM density volume as attributed relational graphs (ARGs). The creation of the

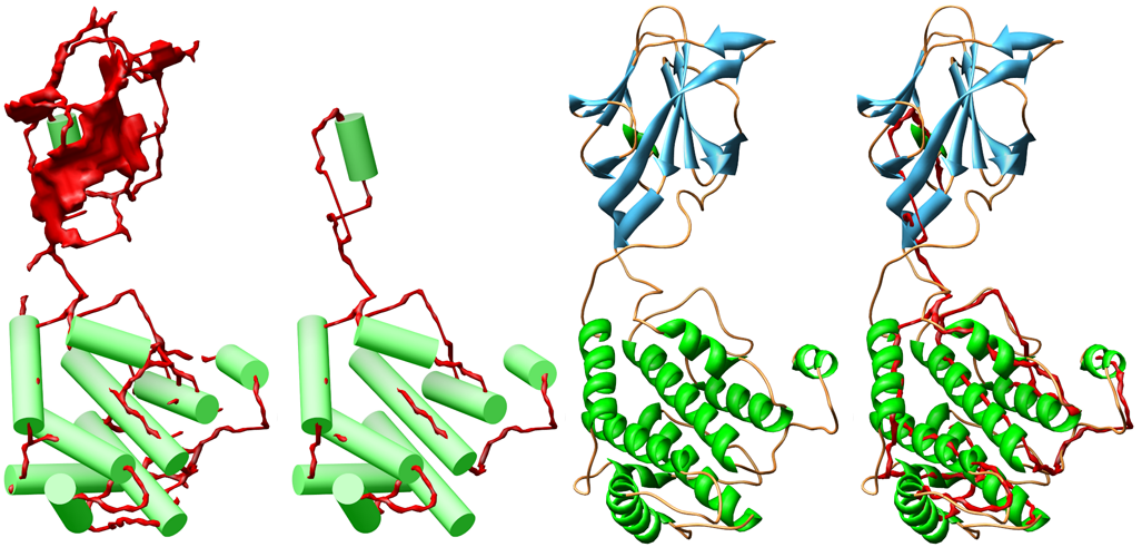
graph from the cryo-EM density volume involves the use of the skeleton created using the method of Ju et al. [46] and the helix locations detected using SSEhunter [8]. An A*-based search algorithm with a novel future cost function is used to find the best matching linear subgraphs on the two ARGs. These linear subgraphs represent the mapping between the helices in the amino acid sequence and the helices in the cryo-EM density volume.

While our algorithm performs reasonably well for most data sets, noise and ambiguities in the cryo-EM density volume result in a low ranking for the correct correspondence. In these cases, we allow a domain expert to specify *anchor* helix correspondences which will guide our algorithm towards a much better result. The next chapter discusses our algorithm and the results in much greater detail.

3.3.2 Identifying β-sheet Correspondence

The next logical extension to the work carried out in this thesis, is to incorporate sheets into the constrained graph matching problem. This would constrain the search further using the connectivity information of the sheets and thereby increase both the accuracy as well as the scalability and performance of the algorithm.

Unlike helices which are formed by a single contiguous strand of amino acids, sheets are formed by multiple strands which are most commonly placed at non-adjacent locations in the amino acid sequence. In other words, sheets are formed based on the geometric proximity of the strands and not the proximity in sequence. This results in many points of connectivity between each sheet and the other elements of the protein. Therefore, we see that when considering the sheets, we can not use the same construction of the attributed relational graph (i.e the placement of two nodes for the two ends of the helix). We plan to investigate methods to address this problem by introducing a new type of node into the graph where each node represents a sheet. As a sheet can be visited multiple times when traversing through the amino acid sequence, we can relax the single visit constraint that we have placed on the graph matching algorithm to allow it to visit a sheet node multiple times during the A*-search. Once the sheet information has been represented as part of the two ARGs, the next step is to modify the current and future cost functions of the A*-search to incorporate the similarity between the sheets being matched. We are planning to explore the use of the sheet surface area and the number of sheet strands in the amino acid (or



(a) (b) (c) (d)

Figure 3.5: Bluetongue Virus (BTV) Backbone Trace: The skeleton with the detected helices (a), the pseudo-backbone trace (b) the actual structure of the protein (c), and the backbone trace overlayed with the actual structure for better comparison (d)

the number of connectivity points in the skeleton) to create a suitable cost function which judges the similarity between predicted sheet correspondences.

3.4 Building a Pseudo-Atomic Model at Amino Acid Resolution

Using the above discussed steps we are able to obtain the correspondence between the secondary structure elements in the amino acid sequence and the cryo-EM density map. In addition, using this information, a pseudo-backbone trace of the protein can be obtained by pruning the skeleton to contain only the edges which connect adjacent secondary structure elements in the amino-acid sequence. Figure 3.5 shows a comparison between the pseudo-backbone trace generated for the Bluetongue virus (2BTV) and the actual structure of the protein obtained from the Protein Data Bank. Please note that Figure 3.5 does not include the sheet correspondences, and once the β-sheet correspondence step is completed, the

pseudo-backbone trace will expand to include the connectivity between the sheets and the rest of the secondary structural elements in the protein.

A statistical analysis of the existing models in the protein data bank allows us to formulate distance constraints between amino acids. Using the pseudo-backbone trace and the secondary structural element correspondence as guides, together with the distance constraints, we can place a C^{α} atom for each amino acid within the cryo-EM density map. This would give us a pseudo-atomic model which roughly corresponds to the actual structure of the protein. However, as we can assume that the total potential energy of the protein is at a local minimum if not a global minimum, we can perform an energy minimization routine on the pseudo-atomic structure constrained by the cryo-EM density map and the secondary structural element correspondences. As this step is performed at the resolution of the amino acids, and is constrained, it should be able to outperform traditional Ab-initio modeling by orders of magnitude.

3.5 Building an Atomic Model

Our final goal is to be able to identify the tertiary structure of the proteins being analyzed. Using the first four steps we can create a pseudo-atomic model at amino acid resolution. However, to fully represent the tertiary structure of a protein we need to find the locations of all the atoms at atomic resolution.

As the native fold of each amino acid is known, we plan on exploring a method of placing these actual atoms on the cryo-EM density volume based on the locations of the C^{α} atoms on the pseudo-atomic map. This provides an initial guess to the positions of the actual atoms. Finally, we propose performing local energy minimization on these atoms once again constrained by the cryo-EM density volume. Although the adequacy of a local minimization technique compared to a global minimization technique is yet to be determined, it can be safely assumed that as the pseudo-atom placement was based on a global minimization, performing local minimization on the actual atoms will not result in too large an error. Once again performing local minimization constrained by the cryo-EM density volume implies a much smaller search space, and thus much better performance than Ab-Initio modeling.

Chapter 4

Graph Matching to Find α-Helix Correspondence

4.1 Problem Statement

The ultimate biological goal of our project is to find, in the density volume, the locations of atoms for each of the amino acids that make up the protein. Unfortunately, unlike X-ray crystallography and NMR spectroscopy, the resolution of cryoEM reconstructions is often far from sufficient to directly obtain an accurate atomic model of the imaged protein. Instead, we first consider an intermediate step towards this goal; which is finding the position of secondary structures, α-helices in particular, in the density volume. Progress has been made in the biology community for detecting positions, orientations and lengths of possible helices in a density volume [8, 42, 43] based on their cylindrical density distributions (an example is shown in Figure 4.1 (c)). What is missing however, is the knowledge of which helix detected in the volume corresponds to a given helix in the sequence. Such knowledge would establish a coarse 3D protein model consisting of a chain of helices (such as that in Figure 4.1 (e)) that sheds light on how the protein folds in 3D space.

As a result, the computational problem that we will address here is the *correspondence* between the helices in the sequence and the helices in the density volume. A desirable correspondence implies not only minimal differences (e.g., in lengths) between corresponding helices, but also maximal agreement between the density volume and the connectivity of helices. In other words, the 3D path between successive helices in the protein sequence should follow high density regions in the volume. In the past, the helix correspondence

problem has only been studied in the work of Wu et al. [97], yet their method fails to take the density information into consideration, and requires a significant amount of computational resources.

Note that the helix correspondence problem is further compounded by the fact that such a correspondence may not be a bijection. Due to the noise in a typical density volume, a helix detection algorithm may fail to find the locations of all the helices within that volume. It may also identify false helices. For example, the number of helices detected in the volume in Figure 4.1 (c) is one less than that in the sequence in Figure 4.1 (a).

4.2 Shape Modeling and Matching

The central theme of our approach is to cast the helix correspondence problem as that of *shape matching* between the 1D sequence and the 3D volume. The key that makes such a matching possible is the modeling of both the 1D and 3D shapes as graphs that encode the lengths of helices as well as their connectivity. In particular, the graph representing the density volume is obtained by computing a *skeleton* that encodes the topology of the high-density regions (Figure 4.1 (d)). Using the shape representations, helix correspondence reduces to a constrained error-correcting graph-matching problem, which seeks the best matching simple paths among two graphs. Using a heuristic search algorithm, the optimal match can be found in an efficient manner.

When applied to an extensive suite of test data, our method was shown to be capable of identifying the correct helix correspondence with no or minimal user-intervention for small and medium size proteins. For example, Figure 4.1 (e) shows the correspondence computed by our method. Observe that the availability of the skeleton allows us to plot a path on the skeleton that connects successive helices, suggesting a possible 3D trace of the amino acid sequence.

We see our work making the following contributions to shape modeling, matching and computational biology:

- We present a common shape representation for both protein sequences and density volumes as attributed relational graphs, which are suitable for structural matching.

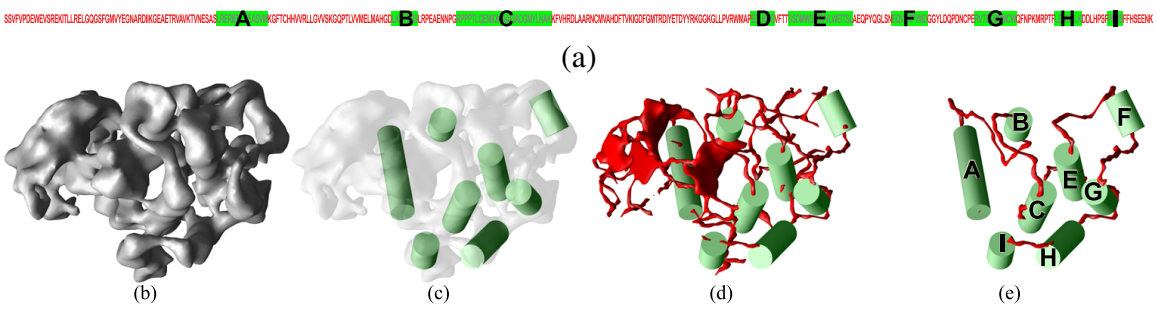


Figure 4.1: Identifying α-helices in a low-resolution protein image, using the Human Insulin Receptor - Tyrosine Kinase Domain (1IRK) as an example. The inputs are the amino-acid sequence of the protein (a), where α-helices are highlighted in green, and a density volume reconstructed from electron cryomicroscopy (b), where possible locations of α-helices have been detected as cylinders shown in (c). Our method computes the correspondence between the helices in the sequence and in the density volume (e). This is achieved by extracting a skeleton from the density volume shown in (d) and matching it with the sequence in (a). Note that the matching is error-tolerant therefore the resulting correspondence does not have to be a bijection.

- We formulate a constrained error-correcting matching problem between attributed graphs, which differs from known exact and inexact matching problems.
- We develop an optimal algorithm based on the A*-search for solving constrained matching, and we explore several novel heuristic functions for pruning the search space.

4.3 Prior Work

4.3.1 Determining Secondary Structure Correspondence

While many techniques exist for determining secondary structural elements using cryo-EM density maps, to the best of our knowledge no techniques other than that of Wu et al [97] exist for the determination of the correspondence between secondary structure elements. This method uses the secondary structural elements located in the amino acid sequence detected using the PSIPRED server [45, 61], and the 3D locations of the secondary structure elements detected using Helixhunter [42], Sheetminer [50] and Sheettracer [51] as inputs.

Topology candidates are first generated by aligning the two sets of secondary structural elements in all possible configurations. This results in an exponential number of topology candidates, which are reduced by first comparing the helix, loop and β-strand lengths and then screening using geometric properties. Using the remaining candidates, more candidates are created by slightly perturbing the structures. Finally, a function based on the potential energy, the bond angles and bond torsion is used to guide a global optimization algorithm towards finding the correspondences with the minimum energy.

While this technique displays promising results, the exponential search space limits the scalability and the performance of the algorithm. As reported in their work [97], finding the top ranking correspondences for an 8 helical protein took approximately 16 hours. Another limitation in this technique is that it does not use the topology information available in the cryo-EM density map to prune the search space.

4.3.2 Shape Representation for Matching

Shape representations, or *descriptors*, have been widely employed in graphics and computer vision for matching purposes. Generally, such representations can be classified into two classes. *Global* shape representations aim at computing a compact set of feature vectors of an entire object for fast comparison between objects, often used in shape retrieval from a large repository of models [18, 85, 99]. We refer interested readers to the survey of Shilane et al. [85] for descriptions and comparisons of these descriptors. Some examples of global descriptors are *light field descriptor* [18], *spherical harmonics descriptors* [47, 81], *wavelet descriptor* [83], *shape distributions* [68], *circular Hidden Markov Models* [4], and *Spectral Embedding* [41]. Note that these global descriptors seldom provide local feature information and are thus generally unsuitable for partial matching; that is, finding a portion of an input object that matches a model object.

In contrast, *local* shape representations describe geometric features of an object (possibly at multiple scales) and are designed for partial matching and object alignment. Some examples of local descriptors include *SIFT features* [57], *local spherical harmonics* [31], *salient surface features* [32], *curvature maps* [33], and *skeletons* [89]. We utilize the skeleton descriptor to translate the shape of an iso-surface in the density volume into a graph structure that can be used to identify connectivity among helices. Such a skeleton can be efficiently

generated from a discrete volume by iterative thinning [11, 13, 69, 25, 46]. Methods for computing 3D skeletons are usually based on Voronoi diagrams [24], distance transforms [72], potential field [21], or iterative thinning. In our algorithm, we use the iterative thinning skeletonization technique of Ju et al. [46] described in the last chapter.

4.3.3 Graph Matching

In pattern recognition and machine vision, graphs have long been used to represent object models, such that object recognition reduces to graph matching. Here, we only give a brief review of graph matching problems and methodologies and refer the reader to the excellent surveys of Bunke et al. [17] and Conte et al. [22] for a summary of the rich volume of matching techniques.

In general, graph matching problems can be divided into exact matching and inexact matching. Exact matching aims at identifying a correspondence between a model graph and (a part of) an input graph, which can be solved using sub-graph isomorphism [23, 92] or graph monomorphism [96]. However, since real-world data is seldom perfect and noise-free, inexact or error-correcting matching is desired in a large number of applications. As in the work of Bunke et al. [15], error-correcting matching can be formulated as finding the bijection between two subgraphs from the model and input graph that minimizes some error function. This error typically consists of the cost of deforming the original graphs to their subgraphs and the error of matching the attributes of corresponding elements in the two subgraphs. Note that, in most applications, the topology of the optimally matching subgraphs (e.g., whether it is connected, a tree, a path, etc.) is generally unknown. Such matching is said to be *un-constrained*, since the minimization of an error function is the only goal.

The most popular algorithms for error-correcting graph matching are based on the A*-search [67]. These algorithms are optimal in the sense that they are guaranteed to find the global optimal match. However, since the graph matching problem itself is NP-complete, the actual computational cost can be prohibitive for large graphs. To this end, various types of heuristic functions have been developed to prune the A* search space [16, 80, 82, 91, 96].

Other methods such as simulated annealing [38], neural networks [28], probabilistic relaxation [20], genetic algorithms [95], and graph decomposition [62] can also be used to reduce the computational cost. Observe that all of these optimization methods are developed for un-constrained matching where the matched subgraphs can assume any topology.

4.4 Shape Representation

To solve the helix correspondence problem as stated above, we first seek a common shape representation of both the 1D protein sequence and the 3D density volume that is suitable for matching. In particular, such representation should encode the lengths of each helix as well as their connectivity. Here, we introduce such a representation using attributed relational graphs (ARG).

In general, an ARG G consists of a 4-tuple $< V_G, E_G, \alpha_G, \beta_G >$, where V_G is a non-empty set of nodes ($|V_G|$ denotes the number of nodes), $E_G = V_G \times V_G$ is a set of edges between pairs of nodes, and α_G, β_G are attribute functions respectively on nodes and edges. Below we describe the connotations of these graph components when describing a protein sequence or a density volume. Note that our graph is specifically designed to tolerate the low-resolution and noise in a density volume.

4.4.1 Protein Sequence Graph

To represent helices in the sequence, the protein sequence graph S consists of a collection of node-pairs, each denoting the two ends of a helix. These nodes are augmented by two additional terminal nodes denoting the two ends of the protein. To reflect the linearity of the sequence, we index the nodes in V_S in ascending order $\{1, \ldots, 2r+2\}$ where r is the total number of helices, 1 and $2r+2$ are the two terminals of the protein, and $2k$ and $2k+1$ are the two ends of the kth helix in the protein sequence. For matching purposes, the different types of nodes are also distinguished by their attributes: $\alpha_S(x)$ for each $x \in V_S$ assumes H, S or E if x represents an end of a helix, the head or the tail of the protein. An example of nodes and attributes is shown in Figure 4.2 (b) for the sequence in (a).

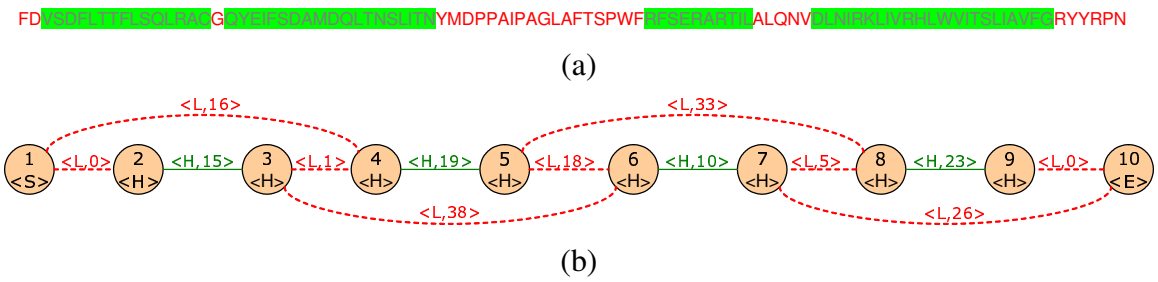


Figure 4.2: Protein sequence graph: the amino acid sequence of a portion of the Rice Dwarf Virus (1UF2) (a), and the corresponding attributed relational graph (b).

To encode the lengths of helices and their connectivity, a *helix edge* is formed between every two successive nodes $2k$ and $2k+1$ for $k \in [1,r]$, and a *link edge* is formed between nodes $2k-1$ and $2k$ for $k \in [1,r+1]$, as shown in Figure 4.2 (b). Note that these edges form a simple path with alternating edge types. The attribute function $\beta_S(x,y)$ for each edge $\{x,y\}$ returns a 2-tuple: $\beta_{S,1}(x,y)$ indicates the edge type, being H or L when $\{x,y\}$ is a helix edge or link edge, and $\beta_{S,2}(x,y)$ maintains the length of that helix or link as the number of amino acids in the sequence. Note that the graph is undirected, that is, $\beta_{S,k}(x,y) = \beta_{S,k}(y,x)$ for $k = 1,2$.

Due to the noisiness and the low resolution of the density volume, helix detection in the volume may not be able to find all helices of that protein. To be able to establish an error-correcting matching in the presence of missing helices, we augment the graph with link edges connecting nodes $\{2k-1, 2k+2l\}$ for every $k \in [1,r]$ and $l \in [1, \min(m, r-k+1)]$ where m is a user-specified maximum number of helices that are possibly missing in the volume. The attribute $\beta_{S,2}(x,y)$ for each new link edge is set to be the total number of amino acids in the sequence bypassed by the edge. Figure 4.2 (b) shows an example with $m = 1$. Note that after such an addition, any simple path in the graph connecting nodes with ascending indices still consists of alternating edge types, which represents an ordered subset of helices in the protein sequence.

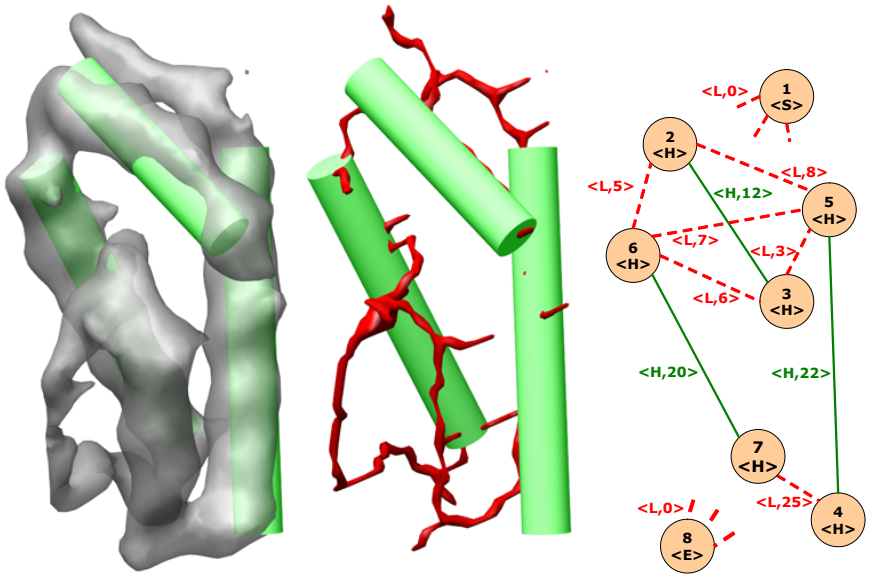


Figure 4.3: Density volume graph: iso-surface of the density volume (a), the skeleton created from the iso-surface with detected helices (b), and the corresponding attributed relational graph (c) where the two terminal nodes 1,8 are connected to every other node via loop edges.

4.4.2 Density Volume Graph

As in the sequence graph, the volume graph C consists of two nodes for each detected helix and two terminal nodes for the entire protein. The different types of nodes are distinguished using the node attribute function α_C, which assumes H, S or E for the helix nodes, head node or tail node of the protein. Unlike the sequence graph, where there is an explicit ordering of nodes, the indices of nodes in V_C do not imply any ordering.

To encode helix information, nodes representing the two ends of a helix are connected by a helix edge. As in the sequence graph, the edge attribute function β_C returns a 2-tuple, where $\beta_{C,1}$ assumes H or L indicating a helix or link edge, and $\beta_{C,2}$ returns the length information. For a helix edge $\{x,y\} \in E_C$, $\beta_{C,2}(x,y)$ is the Euclidean length of the detected helix in the density volume, which can be normalized by the resolution of the volume to approximate the number of amino acids in the helix [8]. An example of such edges are shown in green in Figure 4.3 (c) representing the helices detected in the density volume in (a).

Unlike the sequence graph, the density volume does not explicitly provide the needed connectivity among detected helices. However, as stated earlier, two helices at successive positions in the sequence are more likely to be connected in 3D through regions in the volume with high density. As a result, we seek a representation that depicts the topology of such high-density regions. To this end, we extract an iso-surface from the volume at a user specified density level and compute a morphological *skeleton* of the solid enclosed by the iso-surface. Using a recently developed erosion-based skeletonization technique described in the last chapter [46], such skeletons can be robustly generated even from noisy surfaces while preserving the solid topology. An example of the skeleton is shown in Figure 4.3 (b).

Given the skeleton, we form link edges as shown in Figure 4.3 (c). First, we link every two nodes in the graph that represent ends of two helices connected by a path on the skeleton. When multiple paths exist between two helix ends, the shortest is taken. Note that due to noise present in the volume, these skeleton paths may not capture all the necessary connectivity among helices. To this end, we additionally create a link edge between ends of every two helices whose Euclidean distance is within a user-specified value ε. Finally, to complete the graph, a link edge is created between each terminal node and every nonterminal node. The edge attribute $\beta_{C,2}$ for the above three classes of link edges are set to the length of the skeleton path, the Euclidean distance, and zero respectively (normalized by the resolution of the volume as in [8]).

4.5 Constrained Graph Matching

Given two graphs representing the helices in the sequence and the volume, here we show that finding the correspondence between the two sets of helices reduces to a constrained graph matching problem. We first define a chain as:

Definition 4.1 *A* chain *of an ARG G is a sequence of nodes* $\{v_1,\ldots,v_n\} \subseteq V_G$ *that form a simple path in G. A chain is* ordered *if* $v_1 = 1, v_n = |V_G|$*, and* $v_i < v_{i+1}$ *for all* $i \in [1, n-1]$.

For example, an ordered chain in the sequence graph consists of edges with alternating types (e.g., helix or link), depicting a linked sequence of helices. A correspondence between helices in the sequence and the volume is therefore a bijection between an ordered

chain in the sequence graph and a chain in the volume graph. Note that the definition of chain allows establishing partial correspondence between a subset of the helices in both the sequence and the volume. More generally, the problem can be defined for any attributed relational graphs:

Problem 4.1 *Let S,C be two ARGs. Find an ordered chain $\{p_1,\ldots,p_n\} \subseteq V_S$ and chain $\{q_1,\ldots,q_n\} \subseteq V_C$ that minimizes the matching cost:*

$$\sum_{i}^{n} c_v(p_i,q_i) + \sum_{i}^{n-1} c_e(p_i,p_{i+1},q_i,q_{i+1}) \tag{4.1}$$

where c_v, c_e are any given functions evaluating the cost of matching node p_i with q_i or edge $\{p_i,p_{i+1}\}$ with $\{q_i,q_{i+1}\}$.

Comparing to previously studied graph matching problems such as exact graph (or subgraph) isomorphisms, inexact graph matching and maximum common subgraph problems [39], Problem 4.1 is unique in that it seeks best-matching subgraphs from two graphs that have a particular shape. Given such constraints, previous graph matching algorithms that are guided only by error-minimization can not be directly applied.

4.5.1 Cost Functions

Here we explain our choice for the two cost functions c_v, c_e in Equation 4.1 when matching the sequence graph and the volume graph. Note that, the algorithm we present in the next section works for any non-negative cost function.

Each cost function measures the similarity of the attributes associated with two nodes or two edges. To enforce matching of terminal nodes in the two graphs, the node cost function is defined as

$$c_v(x,y) = \begin{cases} 0, & \text{if } \alpha_S(x) = \alpha_C(y) \\ \infty, & \text{otherwise} \end{cases} \tag{4.2}$$

The edge cost function computes the length difference between two helix edges or two link edges, and is defined as

$$c_e(x,y,u,v) =$$

$$\begin{cases} |\beta_{S,2}(x,y) - \beta_{C,2}(u,v)|, & \text{if } \beta_{S,1}(x,y) = \beta_{C,1}(u,v), \\ & \text{and } y = x+1. \\ |\beta_{S,2}(x,y) - \beta_{C,2}(u,v)| + \gamma_S(x,y), & \text{if } \beta_{S,1}(x,y) = \beta_{C,1}(u,v), \\ & \text{and } y > x+1. \\ \infty, & \text{otherwise} \end{cases} \tag{4.3}$$

Here, the γ_S term penalizes missing helices in the volume graph and is set to be the sum of lengths of the helix edges in the sequence graph bypassed by a link edge. Given a protein sequence with r helices and m possible missing helices in the density volume, and let $x = 2k-1$ and $y = 2k+2l$ where $k \in [1,r]$ and $l \in [1, \min(m, r-k+1)]$, we compute

$$\gamma_S(x,y) = \omega \sum_{i=1}^{l} \beta_{S,2}(2k+2i-2, 2k+2i-1) \tag{4.4}$$

where ω is a user-specified weight that adjusts the influence of this penalty term.

4.5.2 An Optimal Algorithm

In this section, we present a heuristic search algorithm for solving Problem 4.1. Our method extends the tree-search paradigm popularized in computing unconstrained error-correcting graph matching, and is guaranteed to find the optimal match.

To find a match between two graphs, a tree-search algorithm starts out from an initial, incomplete match and incrementally builds more complete matches. To find matching chains in graphs S, C, we first consider a partial match as a sequence of node-pairs

$$M_k = \{\{p_1, q_1\}, \ldots, \{p_k, q_k\}\}$$

where $\{p_1,\ldots,p_k\}$ and $\{q_1,\ldots,q_k\}$ are the initial portion of some ordered chain in S and some chain in C. Based on the definition of chains and our matching goal of minimizing cost functions, elements of M_k must satisfy the following requirements:

- **Node requirement**: $p_1 = 1$, $q_i \neq q_j (\forall j \neq i \in [1,k])$, and for all $i \in [1,k]$:

$$p_i \in V_S, \;\; q_i \in V_C, \;\; \text{and} \;\; c_v(p_i,q_i) \neq \infty$$

- **Edge requirement**: For all $i \in [1,k-1]$:

$$p_i < p_{i+1}, \;\; \{p_i,p_{i+1}\} \in E_S, \;\; \{q_i,q_{i+1}\} \in E_C, \\ \text{and} \;\; c_e(p_i,p_{i+1},q_i,q_{i+1}) \neq \infty$$

Starting with an empty match $M_0 = \emptyset$, the search algorithm incrementally builds longer matching chains. Specifically, we define an *expansion* of a partial match M_k as a new partial match $M_{k+1} = M_k \cup \{\{p_{k+1},q_{k+1}\}\}$ such that the added nodes p_{k+1},q_{k+1} satisfy the node requirement and the added edges $\{p_k,p_{k+1}\},\{q_k,q_{k+1}\}$ (for $k > 0$) satisfy the edge requirement. Note that usually a M_k can be expanded into multiple M_{k+1}. A match M_k is *complete* (i.e., no more expansion can be done) if $p_k = |V_S|$.

Observe that the search procedure essentially builds a tree structure with M_0 at the root of the tree, expanded partial matches M_k at the kth level of the tree, and complete matches at the tree leaves. Our goal is therefore to find the complete match that minimizes the matching error defined in Equation 4.1.

A*-search

To avoid a breadth-first tree search to find the optimal complete match, we adopt the A* search algorithm which prioritizes the expansion of incomplete matches using a fitness function. This function, $f(M_k)$, assesses the likelihood of a partial match M_k to be a part of the optimal complete match. The function has two parts:

$$f(M_k) = g(M_k) + h(M_k) \tag{4.5}$$

```
// Finding the optimal common chain in S,C

ChainMatch(S,C)
// Q is a min heap
// The key of each element M ∈ Q is f(M)
    Q ← {M_0}
    Repeat
        M_k ← Pop(Q)
        // M_k has the form {{p_1,q_1},...,{p_k,q_k}}
        If p_k = |V_S|
            Return M_k
        Repeat for each expansion M_{k+1} from M_k
            Insert(Q,M_{k+1})
```

Figure 4.4: A*-search algorithm to solve the correspondence

where $g(M_k)$ returns the matching cost as defined in Equation 4.1, and $h(M_k)$ estimates the remaining cost to be added in future expansions from M_k.

Given a fitness function, the A*-search algorithm works by maintaining all un-expanded partial matches in a priority queue and only expanding the partial match with the best (smallest) fitness function value. Figure 4.4 outlines the pseudo-code of the algorithm.

Observe from Figure 4.4 that the algorithm returns the first complete match that it finds. Based on A* theory, such match is guaranteed to be the *optimal* match as long as the $h(M_k)$ portion of the fitness function is a lower-bound of the actual remaining matching cost of any complete match M_n that contains M_k. That is, our algorithm works correctly if

$$h(M_k) \leq h^*(M_k) = \min_{M_n : M_k \subset M_n} (g(M_n) - g(M_k)) \tag{4.6}$$

where M_n are complete matches expanded from M_k, and $h^*(M_k)$ is the minimum remaining cost among all M_n.

Assuming that cost functions c_e, c_v in Equation 4.1 are non-negative, $h^*(M_k)$ in Equation 4.6 is also non-negative. Hence an obvious choice is $h(M_k) = 0$, which is a guaranteed

lower-bound of $h^*(M_k)$. However, the better the approximation of $h(M_k)$ to the actual minimum remaining cost $h^*(M_k)$, the fewer nodes that have to be explored during the search. Next we present three variations of $h(M_k)$ that are all lower-bounds of $h^*(M_k)$ with different levels of tightness.

Heuristic Fitness Function

Given a partial match M_k, we denote the set of all nodes in V_S and V_C that can be added to M_k in an expansion as $R_S(M_k)$ and $R_C(M_k)$. Let $x \in R_S(M_k)$, we define:

$$h_a(M_k, x) = \min_{y \in R_C(M_k)} c_e(p_k, x, q_k, y) \tag{4.7}$$

and

$$h_b(M_k, x) = \sum_{y=x}^{|V_S|-1} \min_{\{u,v\} \in E_C, u \notin M_k, v \notin M_k} c'(y, u, v) \tag{4.8}$$

In essence, h_a computes the minimum cost of appending a pair $\{x, y\}$ into M_k for any candidate nodes y, and h_b computes the minimum cost of appending the remaining pairs to form a complete match. Here, c' is an amortized minimum cost of matching an edge $\{u, v\} \in E_C$ to any edge $\{u', v'\}$ in E_S such that $u' \leq y$ and $v' \geq y+1$, defined as

$$c'(y, u, v) = \min_{j \in [0, y-1]} \min_{k \in [j+1, j+|V_S|-y]} \frac{c(y-j, y-j+k, u, v)}{k} \tag{4.9}$$

Now we define three choices of $h(M_k)$ and prove that they are all lower-bounds of $h^*(M_k)$:

$$\begin{aligned} h_0(M_k) &= 0 \\ h_1(M_k) &= \min_{x \in R_S(M_k)} h_a(M_k, x) \\ h_2(M_k) &= \min_{x \in R_S(M_k)} (h_a(M_k, x) + h_b(M_k, x)) \end{aligned}$$

Proposition 4.1 *$h_i(M_k) \leq h^*(M_k)$ for $i = 0, 1, 2$.*

Proof:

1. Trivially we see that $h_0(M_k) = 0 \leq h^*(M_k)$

2. Observe that h_1 computes the minimum cost of appending any pair $\{x,y\}$ into M_k, hence we have

$$\begin{aligned} h_1(M_k) &= \min_{M_{k+1}:M_k\subset M_{k+1}}(g(M_{k+1})-g(M_k)) \\ &\leq \min_{M_n:M_k\subset M_n}(g(M_n)-g(M_k)) \\ &\leq h^*(M_k) \end{aligned}$$

where M_n is a complete match.

3. We examine the minimum-cost complete match $M_n = \{\{p_1,q_1\},\ldots,\{p_n,q_n\}\}$ such that $M_k \subset M_n$. Hence $h^*(M_k) = g(M_n) - g(M_k) = g_a + g_b$, where

$$\begin{aligned} g_a &= c_e(p_k, p_{k+1}, q_k, q_{k+1}) \\ g_b &= \textstyle\sum_{j=k+1}^{n-1} c(p_j, p_{j+1}, q_j, q_{j+1}) \end{aligned}$$

Note that $h_a(M_k, p_{k+1}) \leq g_a$. In addition, the lower-bound cost function c' ensures that

$$c'(i, q_j, q_{j+1}) \leq \frac{c_e(p_j, p_{j+1}, q_j, q_{j+1})}{p_{j+1} - p_j}$$

for any $p_j \leq i < p_{j+1}$. Hence we have

$$\begin{aligned} h_b(M_k, p_k) &\leq \textstyle\sum_{j=k+1}^{n-1} \sum_{i=p_j}^{p_{j+1}-1} \frac{c_e(p_j,p_{j+1},q_j,q_{j+1})}{p_{j+1}-p_j} \\ &\leq \textstyle\sum_{j=k+1}^{n-1} c_e(p_j, p_{j+1}, q_j, q_{j+1}) \\ &\leq g_b \end{aligned}$$

Finally,

$$\begin{aligned} h_2(M_k) &\leq h_a(M_k, p_k) + h_b(M_k, p_k) \\ &\leq g_a + g_b \\ &\leq h^*(M_k) \end{aligned}$$

QED.

We can conclude that the three fitness functions satisfy the following inequality,

$$0 \leq h_0(M_k) \leq h_1(M_k) \leq h_2(M_k) \leq h^*(M_k), \tag{4.10}$$

Therefore, we see that using either of the three functions will result in an optimal solution in the A*-search.

Observe that the three functions represent increasingly better approximations of the actual minimal remaining cost, as a result, fewer nodes need to be expanded during the search using h_1 or h_2 over using h_0 in the fitness function. However, the computation of h_1, h_2 is much more expensive than h_0. In particular, evaluating the h_b portion of h_1 or h_2 involves nested minimality queries. In our implementation, we accelerated the calculation of h_b by pre-computing a look-up table indexed by a node $y \in V_S$, which maintains a sorted list of edges $\{u, v\} \in E_C$ in the ascending order of $c'(y, u, v)$.

4.6 Results

In this chapter, we discuss the performance and accuracy of our method on an extensive suite of protein data. For a significant fraction of these test data sets, we observed that our method was capable of finding the correct helix correspondences without any user intervention. However, for density volumes with poor quality, the optimal graph matching may not represent the actual helix correspondence, and domain knowledge has to be incorporated to yield the correct result.

4.6.1 Setup

Our experiment consists of eleven cryo-EM volumes at 6Å-10Å resolution, eight of which are simulated from the actual atomic model obtained from the Protein Data Bank [10] and three which are authentic cryo-EM reconstructions. Three of the simulated proteins (1IRK [40], 1TIM [10] and 2BTV [37]) were selected such that they represented distinct SCOP[1] families, and five proteins were selected as they were some of the most commonly occurring folds [34]. To provide a more realistic evaluation, authentic cryo-EM density maps of the P8 capsid proteins of the Rice Dwarf Virus [101], the GroEL monomer and the Bacteriophage P22 capsid protein GP5 [43] were chosen for evaluation[2]. Only three authentic cryoEM reconstructions are reported as there are only a small number of structures in the public domain with resolutions beyond 7Å-8Å.

[1] Structural classification of Proteins

[2] EMDB number for these authentic reconstructions are 1060 (RDV P8), 1101 (P22) and 1081 (GroEL)

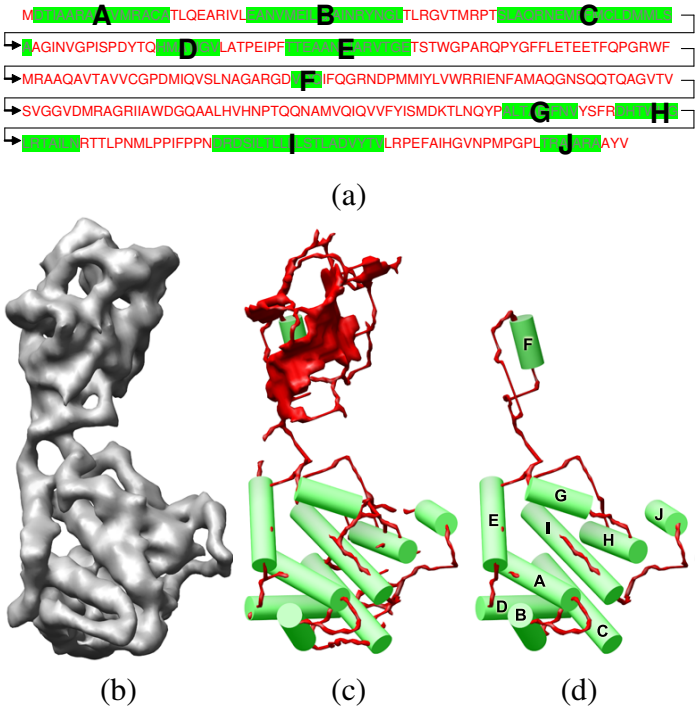


Figure 4.5: Bluetongue Virus (2BTV): the amino acid sequence (a), the density volume (b), the detected helices with the skeleton (c), and the correspondence (d) between helices in (a) and (c) computed as the optimal match between the sequence and volume graphs.

In each example, we utilize the protein sequence data from the Protein Data Bank, the helices in density volumes detected using SSEhunter [8], and the skeleton created using the method of Ju et al. [46]. The matching result is presented as a correspondence between helices in the sequence with those in the density volume. The result is validated either using the original atomic model (for simulated data) or a structural homologue (for authentic data).

In all the experiments, an Euclidean distance threshold of $\varepsilon = 0.15d$ is used for creating extra edges in the volume graph where d is the size of the volume (d for the data is shown in Table 4.1), and $\omega = 5$ is used in weighting the missing helix penalty term in the cost function. Experiments were performed on a PC with a 3GHz Pentium D CPU and 2GB of memory (our implementation runs on a single thread, thus utilizes only one of the cores of the CPU).

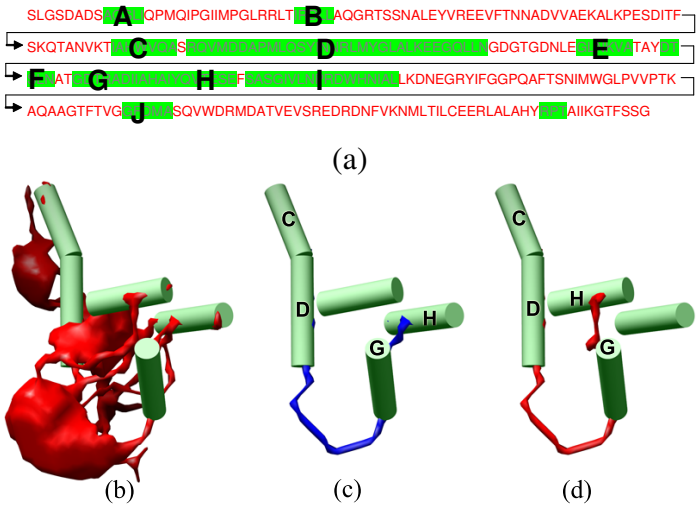


Figure 4.6: Bacteriophage P22 capsid protein (P22 GP5): the amino acid sequence (a), the detected helices in the density volume with the skeleton (b), the incorrect correspondence between helices in (a) and (b) computed as the optimal match between the sequence and volume graphs (c), and the correct correspondence (d) which ranks as the 4th optimal match.

4.6.2 Unsupervised Matching

Figure 4.1 and 4.5 show two examples (1IRK and 2BTV) where our method is able to identify the correct full or partial correspondence. Note that our algorithm is robust to noise in the data such as the one missing helix in the density volume of 1IRK. As a by-product of our matching algorithm, a *pseudo-backbone trace* of the protein sequence in the density can be visualized by rendering the skeleton paths represented by the graph edges in the optimally matching chain. Such a trace could serve as a starting point to determine finer-scale protein components such as amino acids.

4.6.3 Interactive Matching

Due to the limited resolution of a density volume in depicting the protein shape, the optimal match between the graph representations may fail to represent the correct helix correspondence. Such failure may also be caused by ambiguities on the skeleton created from an

iso-surface, which arises due to the difficulty of picking an appropriate iso-value that accurately represents the topology of the protein body. To battle data inaccuracy, we augment the proposed graph matching method with domain knowledge in two ways:

- Computing the candidates list
- Interactive constraints

Computing the Candidates List: Instead of finding a single optimal match between the sequence graph and the volume graph, a list of top-matching candidates are computed. This can be done easily in the A*-search framework by terminating the search only after a number of complete matches (e.g., 100) have been found.

Identifying the correct correspondence within these top matches is a common problem in structural biology. Many structure prediction algorithms produce a gallery of structures that range in accuracy. The end user is often required to evaluate the model in the context of other data. The ranking achieved by our program is at least on par with the best algorithms if not significantly better.

Figure 4.6 shows an example (P22 GP5) where the correct correspondence (shown in (d)) is ranked fourth in the candidates list. Comparing with the optimal match (shown in (c)), the two correspondences exhibit very similar helix lengths and connectivity, illustrating why graph matching on its own can not distinguish the right from the wrong without further domain knowledge. Also observe that the correct correspondence is achieved even in the presence of a large number of missing helices (6) and an incorrectly detected helix.

Interactive Constraints: We allow the user to manually assign matching constraints based on their biological knowledge of the spatial arrangement of helices. Specifically, the user may designate the correspondence between a small subset of helix edges in the sequence graph and the volume graph, which can then be translated into additional edge attributes (e.g., $\beta_{S,1}(\{x,y\}) = \beta_{C,1}(\{u,v\}) = H_k$ if edge $\{x,y\}$ and $\{u,v\}$ are the kth corresponding pair) to enforce such explicit matching in the A* search. Figure 4.7 shows an example (protein 1TIM) where the correct helix correspondence was ranked ninth in the candidate list after two user constraints were specified. Protein structures which display a

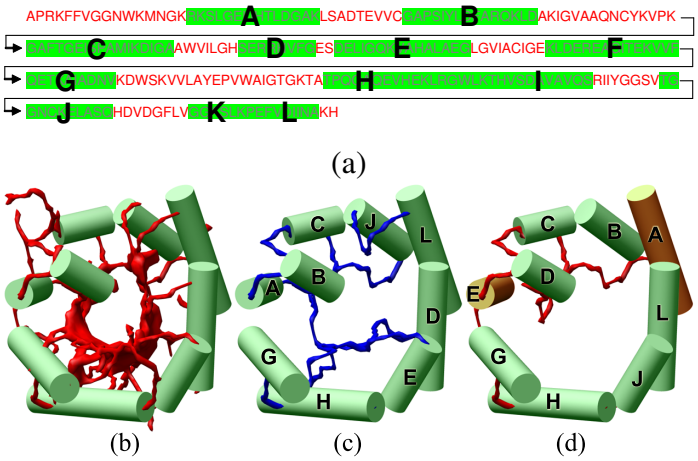


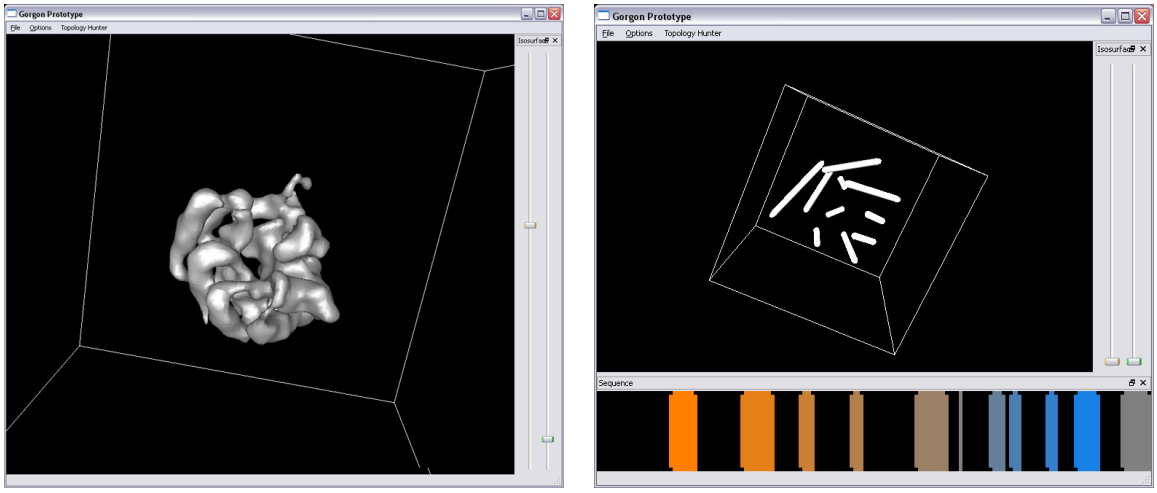
Figure 4.7: Triose Phosphate Isomerase from Chicken Muscle (1TIM): the amino acid sequence (a), the detected helices in the density volume with the skeleton (b), the incorrect correspondence between helices in (a) and (b) computed as the optimal match between the sequence and volume graphs (c), and the correct correspondence (d) which ranks as the 9th optimal match given the two user-specified helix constraints highlighted in brown.

low variance in helix lengths and where the sheet segments result in the formation of barrel like structures (1TIM (Figure 4.7), 2ITG and 1DAI) pose a challenge to the algorithm as this results in a large amount of helix correspondences which have similar costs. Interactive constraints can be used in these cases to provide anchor points to guide our method towards a correct correspondence.

Due to the accumulation of error in the search process because of the inherent ambiguities present in low-resolution density maps, we observe that the amount of user constraints needed in order to obtain a high-ranked correct correspondence increased with the size and complexity of the protein structure. Although this approach requires a time investment by a domain expert, we note that the time needed to specify these constraints is much smaller compared to the time needed if the user was to specify all the helix correspondences.

4.6.4 Graphical User Interface for Interactive Constraints

As the current method of placing interactive constraints is script based, A follow-on project worked on by Troy Ruths[3] was initiated where a graphical user interface named *Gorgon* is being developed. This increases the usability of our approach and let the user visualize the result set and interactively place constraints using a few mouse-clicks. Figure 4.8 (a) shows a cryo-EM density map loaded into Gorgon, and (b) shows the Helices detected by SSEHunter visualized alongside the helices in the amino-acid sequence. Future plans for this interface includes incorporating it with *EMAN* [58], a tool which is commonly used by the structural biology community for cryo-EM reconstructions.



(a) (b)

Figure 4.8: Gorgon Screenshots: Screenshots of *Gorgon*, where a cryo-EM density map is visualized (a) and the helices detected by SSE Hunter are visualized alongside the helices in the amino-acid sequence (b)

4.6.5 Performance

The performance results for all eleven experiments are presented in Tables 4.1 and 4.2, showing the number of helices in the protein sequence, the number of missing helices in each data set (given as the parameter m in creating the sequence graph), the volume

[3] tjr1@cec.wustl.edu

Table 4.1: Experiment Results, Accuracy: The rankings of the correct correspondence in the final set of correspondences.

Protein	Helix Count	Missing helices	Volume Size (d^3)	User constraints	Rank
1UF2	4	-	96^3	-	1
2ITG	6	-	64^3	2	4
1IRK	9	-	96^3	-	1
1WAB	9	2	64^3	-	1
1DAI	9	-	64^3	1	5
2BTV	10	-	128^3	-	1
P22 GP5	11	7	128^3	-	4
3LCK	12	5	64^3	-	2
1TIM	12	3	96^3	2	9
RDV P8	14	2	96^3	4	1
GroEL	20	4	128^3	4	1

Table 4.2: Experiment Results, Performance: Results from the 11 experiments where the time taken (in seconds) to compute the best topology for each of the future cost functions, and the total number of nodes expanded in the A*-search are compared. Observe the significant reduction of nodes expanded when using the better approximations $h_1(M_k)$ and $h_2(M_k)$.

Protein	Helix Count	Missing helices	Time (seconds)			Nodes expanded		
			$h_0(M_k)$	$h_1(M_k)$	$h_2(M_k)$	$h_0(M_k)$	$h_1(M_k)$	$h_2(M_k)$
1UF2	4	-	0.0	0.0	0.0	23	16	13
2ITG	6	-	0.0	0.0	0.0	65	51	41
1IRK	9	-	0.0	0.0	0.0	1813	1195	775
1WAB	9	2	0.0	0.0	0.0	2006	1199	644
1DAI	9	-	0.0	0.0	0.0	10791	8318	6884
2BTV	10	-	0.0	0.0	0.0	5735	3790	595
P22 GP5	11	7	0.0	0.0	0.0	514	378	314
3LCK	12	5	0.0	0.1	0.1	5685	4013	3001
1TIM	12	3	0.2	0.3	0.3	42357	25754	12861
RDV P8	14	2	0.2	0.3	0.7	74212	56770	56539
GroEL	20	4	3.8	8.5	15.2	774813	603378	564929

(d^3) representing the number of voxels in the cryo-EM density map, the number of user-specified constraints and the rank of the correct correspondence in the candidate list. Table 4.2 also contains the time taken by our method, and the number of nodes expanded when using the three cost functions ($h_0(M_k), h_1(M_k), h_2(M_k)$) in the A*-search.

Observe from table 4.1 that the graph matching approach in combination with the domain-specific strategies allow accurate identification of protein structure with no or a small amount of human input depending on the quality of the density volume. Also note from table 4.2 that the time taken to perform a computation is almost negligible in human terms (< 4 seconds for GroEL when using $h_0(M_k)$), which facilitates a much smoother user-interactive functionality and is orders of magnitude faster than the method of Wu et al. which takes 16 hours to find the optimal correspondence for an 8 helical protein [97]. We would like to point out that using the heuristic functions h_1, h_2 dramatically reduces the number of expansions during A*-search compared to using the zero function h_0. However, since the time overhead of computing the functions h_1, h_2 is much larger than the zero function, the actual computation time is often slower. Nonetheless, we anticipate that h_1, h_2 can be useful in reducing the memory cost in large data sets.

4.7 Limitations and Future Work

4.7.1 Reducing the High Computational Cost

One of the limitations of our graph matching algorithm, like other A*-based graph isomorphism techniques, is its high computational cost (both time and memory) for large graphs. In particular, our implementation of the method has difficulty in handling proteins with more than 20 helices without a fairly large number (more than four) of user-specified constraints. In the future we plan to explore variants of the A*-search, including iterative deepening A* and memory-bounded A*, that are better suited for handling large data sets. Furthermore, we are investigating the possibility of using Homology modeling as a pre-processing step to obtain an initial guess at the correct helix correspondence which can thereafter be improved using our method at a much lower computational cost.

4.7.2 Improving the Skeleton using Gray-scale Skeletonization

Our application of graph matching relies on a skeleton generated from the iso-surface at a given density level. The difficulty in finding an appropriate density level so that the iso-surface accurately represents the protein body may result in skeletons that fail to capture the connectivity among detected helices. We next plan to explore skeletonization techniques which apply directly to gray-scale volumes without the need for thresholding. These techniques will produce more robust skeletons and generate matching results that are more likely to represent the correct helix correspondences. An initial survey of literature has revealed many gray-scale skeletonization techniques for two dimensional images. However, not many of those scale into three dimensional volumes, and none of them preserve the shape as well as the topology of the object while minimizing the number of visual artifacts. In the case that we cannot find such an algorithm, we will explore the possibility of extending the binary skeletonization technique of Ju et al. [46] such that we can generate a gray-scale skeleton while preserving cross section shape and topology.

4.7.3 Finding the Correspondence between β-sheets

As described in the third chapter, we plan to identify other secondary structural elements, such as β-sheets from density volumes. We anticipate that a similar shape-matching formulation for finding helix correspondences can be applied to sheets, as sheet-detection algorithms are already available for density volumes [8]. We envision that sheets can be represented in a similar attributed relational graph abstraction where each sheet is maintained as a *re-visitable* node. The search and corresponding cost functions can thereafter be extended to incorporate these re-visitable nodes by considering the connectivity and surface areas of the sheets. This would add more constraints to the A*-search thus making it more accurate, and less computationally expensive.

4.7.4 Molecules with Intrinsic Flexibilities

We would like to note that while cryo-EM is well suited for imaging large macromolecular complexes in near-native solution conditions, the method ultimately reconstructs only a

single snapshot of the assembly for a given set of images. In the event that there is some intrinsic flexibility in the molecule, the corresponding regions within the density map will appear less well-resolved and have lower density values. Based on empirical evidence, most flexibility on the order of helix or sheet shifts are not easily identifiable until sufficiently high resolutions are reached (typically better than 7Å-8Åresolution). We envision that given density maps of higher resolution our technique could produce potential secondary structure topologies through regions of disorder that may not have been readily detectable by visual observation.

Chapter 5

Conclusion

In this thesis we reported a novel application of shape modeling and matching in biomedical research which aims at identifying the correspondence between α-helices identified in an amino acid sequence and a cryo-EM density map, as a partial step in the fulfilment of the long-term goal of predicting protein folds. We translated the biological problem into a computational one by representing the shapes of biological data (e.g., protein sequence and density volume) as attributed relational graphs. We solved the helix correspondence problem using graph matching and we demonstrated the effectiveness of the method on authentic as well as simulated data sets. One of our main contributions is an optimal algorithm for constrained error-correcting graph matching, which will be useful in other shape-matching tasks where the sought match has a linear shape.

We envision improving the performance and scalability of our algorithm by addressing its high computational cost using more robust variants of the A*-search. We are also investigating improving the accuracy of our algorithm by using alternate skeletonization techniques, which do not rely on a specific iso-value to create a binary skeleton, but rather uses the gray-scale volume directly to create a gray-scale skeleton. To further our long term goal of protein structure prediction, we plan to incorporate β-sheets into the algorithm which, in turn, will introduce additional constraints to the A*-search and would significantly improve the rank of the correct correspondence in the candidates list. Finally, we are exploring the possibility of using pseudo-atomic models guided by the secondary structural element correspondence and energy minimization routines to determine the tertiary structure of proteins.

References

[1] S.S. Abeysinghe, T. Ju, M.L. Baker, and W. Chiu. Shape modeling and matching in identifying protein structure from low-resolution images. In *Transactions of the Solid and Physical Modeling Symposium*, 2007, in press.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[4] N. Arica and F.T.Y. Vural. A shape descriptor based on circular hidden markov model. In *ICPR '00: Proceedings of the International Conference on Pattern Recognition*, page 1924, Washington, DC, USA, 2000. IEEE Computer Society.

[5] C. Bajaj, Z. Yu, and M. Auer. Volumetric feature extraction and visualization of tomographic molecular imaging. *J Struct Biol*, 144(1-2):132–143, Oct-Nov 2003.

[6] M.L. Baker, W. Jiang, B.R. Bowman, Z.H. Zhou, F.A. Quiocho, F.J. Rixon, and W. Chiu. Architecture of the herpes simplex virus major capsid protein derived from structural bioinformatics. *J Mol Biol*, 331(2):447–456, Aug 2003.

[7] M.L. Baker, W. Jiang, W.J. Wedemeyer, F.J. Rixon, D. Baker, and W. Chiu. Ab initio modeling of the herpesvirus vp26 core domain assessed by cryoem density. *PLoS Computational Biology*, 2(10), Oct 2006.

[8] M.L. Baker, T. Ju, and W. Chiu. Identification of secondary structure elements in intermediate-resolution density maps. *Structure*, 15(1):7–19, Jan 2007.

[9] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, 1999.

[10] D.W. Banner, A. Bloomer, G.A. Petsko, D.C. Phillips, and I.A. Wilson. Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochem Biophys Res Commun*, 72(1):146–155, Sep 1976.

[11] Gilles Bertrand. A parallel thinning algorithm for medial surfaces. *Pattern Recogn. Lett.*, 16(9):979–986, 1995.

[12] H. Blum. A transformation for extracting new descriptors of shape. In Weiant Wathen-Dunn, editor, *Models for the Perception of Speech and Visual Form*, pages 362–380. MIT Press, Cambridge, 1967.

[13] G. Borgefors, I. Nyström, and G.S. di Baja. Computing skeletons in three dimensions. *Pattern Recognition*, 32(7):1225–1236, 1999.

[14] B.R. Bowman, M.L. Baker, F.J. Rixon, W. Chiu, and F.A. Quiocho. Structure of the herpesvirus major capsid protein. *EMBO J*, 22(4):757–765, Feb 2003.

[15] H. Bunke. Error correcting graph matching: On the influence of the underlying cost function. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(9):917–922, 1999.

[16] H. Bunke and G. Allermann. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters*, 1:245–253, 1983.

[17] H. Bunke and B.T. Messmer. Recent advances in graph matching. *IJPRAI*, 11(1):169–203, 1997.

[18] D. Chen, X. Tian, Y. Shen, and M. Ouhyoung. On visual similarity based 3d model retrieval. *Computer Graphics Forum*, 22(3):223–232, Sep 2003. Eurographics 2003 Conference Proceedings.

[19] W. Chiu, M.L. Baker, W. Jiang, M. Dougherty, and M.F. Schmid. Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure*, 13(3):363–372, Mar 2005.

[20] W.J. Christmas, J. Kittler, and M. Petrou. Structural matching in computer vision using probabilistic relaxation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):749–764, 1995.

[21] J. Chuang, C. Tsai, and M. Ko. Skeletonization of three-dimensional object using generalized potential field. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1241–1251, 2000.

[22] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 2004.

[23] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. Performance evaluation of the vf graph matching algorithm. In *International Conference on Image Analysis and Processing*, 1999.

[24] Tamal K. Dey and Wulue Zhao. Approximate medial axis as a voronoi subcomplex. In *SMA '02: Proceedings of the seventh ACM symposium on Solid modeling and applications*, pages 356–366, New York, NY, USA, 2002. ACM Press.

[25] G.S. di Baja and S. Svensson. A new shape descriptor for surfaces in 3d images. *Pattern Recogn. Lett.*, 23(6):703–711, 2002.

[26] J. Drenth. *Principles of Protein X-Ray Crystallography*. Springer-Verlag Inc., NY, 1999.

[27] S. Dutta and H.M. Berman. Large macromolecular complexes in the protein data bank: a status report. *Structure*, 13(3):381–388, Mar 2005.

[28] J. Feng, M. Laumy, and M. Dhome. Inexact matching using neural networks. *Pattern Recognition in Practice IV: Multiple Paradigms, Comparative Studies, and Hybrid Systems*, pages 177–184, 1994.

[29] A.S. Frangakis and F. Forster. Computational exploration of structural information from cryo-electron tomograms. *Curr Opin Struct Biol*, 14(3):325–331, Jun 2004.

[30] A.S. Frangakis and R. Hergerl. Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis. *J Struct Biol*, 138(1-2):105–113, Apr-May 2002.

[31] T. Funkhouser and P. Shilane. Partial matching of 3D shapes with priority-driven search. In *Symposium on Geometry Processing*, Jun 2006.

[32] R. Gal and D. Cohen-Or. Salient geometric features for partial shape matching and similarity. *ACM Trans. Graph.*, 25(1):130–150, 2006.

[33] T. Gatzke, S. Zelinka, C. Grimm, and M. Garland. Curvature maps for local shape comparison. In *Shape Modeling International*, pages 244–256, June 2005. A local shape comparison technique for meshes.

[34] M. Gerstein. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol*, 274(4):562–576, Dec 1997.

[35] J.P. Glusker, M. Lewis, and M. Rossi. *Crystal Structure Analysis for Chemists and Biologists*. VCH Publishers, NY, 1994.

[36] J. Greer. Comparative model-building of the mammalian serine proteases. *J Mol Biol*, 153(4):1027–1042, Dec 1981.

[37] J.M. Grimes, J.N. Burroughs, P. Gouet, J.M. Diprose, R. Malby, S. Zientara, P.P. Mertens, and D.I. Stuart. The atomic structure of the bluetongue virus core. *Nature*, 395(6701):470–478, Oct 1998.

[38] L. Herault, R. Horaud, F. Veillon, and J.J. Niez. Symbolic image matching by simulated annealing. In *Proc. British Machine Vision Conference (BMVC90)*, pages 319–324, 1990.

[39] R. Horaud and T. Skordas. Stereo correspondence through feature grouping and maximal cliques. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(11):1168–1180, Nov 1989.

[40] S.R. Hubbard, L. Wei, and W.A. Hendrickson. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature*, 372(6508):746–754, Dec 2002.

[41] V. Jain and H. Zhang. Shape-based retrieval of articulated 3d models using spectral embedding. In *GMP*, pages 299–312, 2006.

[42] W. Jiang, M.L. Baker, S.J. Ludtke, and W. Chiu. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J Mol Biol*, 308(5):1033–1044, 2001.

[43] W. Jiang, Z. Li, Z. Zhang, M.L. Baker, P.E. Prevelige Jr., and W. Chiu. Coat protein fold and maturation transition of bacteriophage p22 seen at subnanometer resolutions. *Nat Struct Biol*, 10(2):131–135, Feb 2003.

[44] B. John and A. Sali. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research*, 31:3982–3992, Jul 2003.

[45] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2):195–202, Sep 1999.

[46] T. Ju, M. Baker, and W. Chiu. Computing a family of skeletons of volumetric models for shape description. *Computer-Aided Design*, 2007, in press.

[47] M.M. Kazhdan, T.A. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Symposium on Geometry Processing*, pages 156–165, 2003.

[48] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. W. Wyckoff, and D. C. Phihips. A three-dimensional model of the myoglobin molecule obtained by x-ray anaysis. *Nature*, 181:662–666, Mar 1958.

[49] G.J. Kleywegt and T.A. Jones. Detecting folding motifs and similarities in protein structures. *Methods in Enzymology*, 277:525–545, 1997.

[50] Y. Kong and J. Ma. A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. *J Mol Biol*, 332(2):399–413, Sep 2003.

[51] Y. Kong, X. Zhang, T.S. Baker, and J. Ma. A structural-informatics approach for tracing beta-sheets: building pseudo-c(alpha) traces for beta-strands in intermediate-resolution density maps. *J Mol Biol*, 339(1):117–130, May 2004.

[52] L. Lam, S. Lee, and C.Y. Suen. Thinning methodologies-a comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(9):869–885, 1992.

[53] T. Lazaridis and M. Karplus. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol*, 288(3):477–487, May 1999.

[54] C. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65(1):44–45, 1968.

[55] M. Levitt. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol*, 226(2):507–533, Jul 1992.

[56] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4963):1435–1441, Mar 1985.

[57] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[58] S.J. Ludtke, P.R. Baldwin, and W. Chiu. Eman: Semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol*, 128(1):82–97, Dec 1999.

[59] H. Luecke, B. Schobert, H.T. Richter, J.P. Cartailler, and J.K. Lanyi. Structure of bacteriorhodopsin at 1.55Å resolution. *J Mol Biol*, 291(4):899–911, Aug 1999.

[60] M.A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

[61] L.J. McGuffin, K. Bryson, and D.T. Jones. The psipred protein structure prediction server. *Bioinformatics*, 16(4):404–405, Apr 2000.

[62] Bruno T. Messmer and Horst Bunke. A new algorithm for error-tolerant subgraph isomorphism detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):493–504, 1998.

[63] K. Mizuguchi and N. Go. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.*, 8(4):353–362, 1995.

[64] J. Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15:285–289, 2005.

[65] U. Muckstein, I.L. Hofacker, and P.F. Stadler. Stochastic pairwise alignments. *Bioinformatics*, 18(90002):S153–S160, 2002.

[66] A. Nakagawa, N. Miyazaki, J. Taka, H. Naitow, A. Ogawa, Z. Fujimoto, H. Mizuno, T. Higashi, Y. Watanabe, T. Omura, R.H. Cheng, and T. Tsukihara. The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins. *Structure*, 11(10):1227–1238, Oct 2003.

[67] N.J. Nilsson. *Principles of Artificial Intelligence*. Morgan Kaufmann Publishers, 1980.

[68] R. Osada, T.A. Funkhouser, B. Chazelle, and D.P. Dobkin. Shape distributions. *ACM Trans. Graph.*, 21(4):807–832, 2002.

[69] K. Palágyi and A. Kuba. A parallel 3d 12-subiteration thinning algorithm. *Graph. Models Image Process.*, 61(4):199–221, 1999.

[70] P.A. Penczek, R.A. Grassucci, and J. Frank. The ribosome at improved resolution: New techniques for merging and orientation refinement in 3d cryo-electron microscopy of biological particles. *Ultramicroscopy*, 53(3):251–270, 1994.

[71] P.A. Penczek, M. Radermacher, and J. Frank. Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy*, 40(1):33–53, Jan 1992.

[72] I. Ragnemalm. The euclidean distance transform in arbitrary dimensions. *Pattern Recogn. Lett.*, 14(11):883–888, 1993.

[73] G. Rhodes. *Crystallography Made Crystal Clear*. Academic Press, CA, 1994.

[74] L. Roberts, R.J. Davenport, E. Pennisi, and E. Marshall. A history of the human genome project. *Science*, 291(5507):1195, Feb 2001.

[75] C.A. Rohl, C.E. Strauss, K.M. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods Enzymol*, 383:66–93, 2005.

[76] L. Rychlewski, B. Zhang, and A. Godzik. Fold and function predictions for mycoplasma genitalium proteins. *Fold Des*, 3(4):229–238, 1998.

[77] A. Sali. 100,000 protein structures for the biologist. *Nat Struct Biol*, 5:1029–1032, 1998.

[78] A. Sali and T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, Dec 1993.

[79] A. Sali and J.P. Overington. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci*, 3:1582–1596, 1994.

[80] A. Sanfeliu and K.S. Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE Trans. Systems, Man, and Cybernetics*, 13:353–363, 1983.

[81] D. Saupe and D.V. Vranic. 3d model retrieval with spherical harmonics and moments. In *DAGM-Symposium*, pages 392–397, 2001.

[82] L.G. Shapiro and R.M. Haralick. Structural descriptions and inexact matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 3(5):504–519, 1981.

[83] D. Shen and H.H. Ip. Discriminative wavelet shape descriptors for recognition of 2-d patterns. *Pattern Recognition*, 32(2):151–165, 1999.

[84] M.Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15(11):2507–2524, 2006.

[85] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. In *SMI '04: Proceedings of the Shape Modeling International 2004 (SMI'04)*, pages 167–178, Washington, DC, USA, 2004. IEEE Computer Society.

[86] M. J. Sippl. Who solved the protein folding problem? *Structure*, 7(4):R81–R83, 1999.

[87] M.J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17(4):355–362, Dec 1993.

[88] W.M. Stanley. Virus as a chemical agent. *Rev Med (Mex)*, 32(561):209–212, May 1952.

[89] H. Sundar, D. Silver, N. Gagvani, and S.J. Dickinson. Skeleton based shape matching and retrieval. In *Shape Modeling International*, pages 130–142, 290, 2003.

[90] M. Topf, M.L. Baker, B. John, W. Chiu, and A. Sali. Structural characterization of components of protein assemblies by comparative modeling and electron cryomicroscopy. *J Struct Biol*, 149(2):191–203, Feb 2005.

[91] W.H. Tsai and K.S. Fu. Error-correcting isomorphisms of attributed relational graphs for pattern recognition. *IEEE Trans. Systems, Man, and Cybernetics*, 9:757–768, 1979.

[92] J.R. Ullmann. An algorithm for subgraph isomorphism. *J. ACM*, 23(1):31–42, 1976.

[93] C. Venclovas and M. Margelevicius. Comparative modeling in casp6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins: Structure, Function, and Bioinformatics*, 61(S7):99–105, 2005.

[94] N. Volkmann. A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *J Struct Biol*, 138(1-2):123–129, Apr-May 2002.

[95] Y. Wang, K. Fan, and J. Horng. Genetic-based search for error-correcting graph isomorphism. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 27(4):588–597, 1997.

[96] A.K. Wong, M. You, and A.C. Chan. An algorithm for graph optimal monomorphism. *IEEE Trans. Systems, Man, and Cybernetics*, 20(3):628–636, 1990.

[97] Y. Wu, M. Chen, M. Lu, Q. Wang, and J. Ma. Determining protein topology from skeletons of secondary structures. *J Mol Biol*, 350(3):571–586, 2005.

[98] Z. Yu and C. Bajaj. Automatic ultrastructure segmentation of reconstructed cryoem maps of icosahedral viruses. *IEEE Transactions on Image Processing*, 14(9):1324–1337, Sep 2005.

[99] J. Zhang, K. Siddiqi, D. Macrini, A. Shokoufandeh, and S.J. Dickinson. Retrieving articulated 3-d models using medial surfaces and their graph spectra. In *EMMCVPR*, pages 285–300, 2005.

[100] Y. Zhang and J. Skolnick. The protein structure prediction problem could be solved using the current pdb library. *Proc Natl Acad Sci USA*, 102(4):1029–1034, 2005.

[101] Z.H. Zhou, M.L. Baker, W. Jiang, M. Dougherty, J. Jakana, G. Dong, G. Lu, and W. Chiu. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nat Struct Biol*, 8(10):868–873, Oct 2001.

[102] Z.H. Zhou, M. Dougherty, J. Jakana, J. He, F.J. Rixon, and W. Chiu. Seeing the Herpesvirus Capsid at 8.5Å. *Science*, 288(5467):877–880, 2000.

[103] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proc Natl Acad Sci USA*, 89(1):20–22, Jan 1992.

Vita

Sasakthi S. Abeysinghe

Date of Birth	December 27, 1980
Place of Birth	Colombo, Sri Lanka
Degrees	B.Sc. (First Class Honors), Information Systems, September 2004
Professional Societies	Institute of Electrical and Electronics Engineers
Publications	S.S. Abeysinghe, T. Ju, M.L. Baker, and W. Chiu. Shape modeling and matching in identifying protein structure from low-resolution images. In *Transactions of the Solid and Physical Modeling Symposium*, 2007, in press.

May 2007