Published in final edited form as:

J Struct Biol. 2012 February ; 177(2): 367–381. doi:10.1016/j.jsb.2011.11.032.

Detection of Secondary and Supersecondary Structures of Proteins from Cryo-Electron Microscopy

Chandrajit Bajaj^{a,*}, Samrat Goswami^b, and Qin Zhang^c

^aCenter for Computational Visualization, the Institute for Computational Engineering and Sciences, Department of Computer Science, The University of Texas at Austin, University Station C0200, Austin, TX 78712, USA

^bExa Corporation, 55 Network Drive, Burlington, MA 01803, USA

^cCenter for Computational Visualization, the Institute for Computational Engineering and Sciences, The University of Texas at Austin, University Station C0200, Austin, TX 78712, USA

Abstract

Recent advances in three-dimensional electron microscopy (3D EM) have enabled the quantitative visualization of the structural building blocks of proteins at improved resolutions. We provide algorithms to detect the secondary structures (α -helices and β -sheets) from proteins for which the volumetric maps are reconstructed at 6–10Å resolution. Additionally, we show that when the resolution is coarser than 10Å, some of the super-secondary structures can be detected from 3D EM maps. For both these algorithms, we employ tools from computational geometry and differential topology, specifically the computation of stable/unstable manifolds of certain critical points of the distance function induced by the molecular surface. Our results connect mathematically well-defined constructions with bio-chemically induced structures observed in proteins.

Keywords

protein structures detection; electron microscopy; stable/unstable manifolds; critical points; Delaunay and Voronoi objects

1. Introduction

Three-dimensional electron microscopy (3D EM) reconstruction and, in particular, single particle cryo-EM reconstruction (Frank, 1996, 2002), have advanced rapidly over recent years, such that several macromolecules (e.g., complexes of proteins and ribo-nucleic acids (RNA)) can be routinely resolved at coarse resolution 10–20Å and in some cases at sub-nanometer (intermediate) resolution (6–10Å) (Baker et al., 1999). In the last decade, finer resolution (Zhang et al., 2008; Zhou, 2008; Baker et al., 2010) 3D EM reconstructions (henceforth 3D maps) have been obtained, including GroEL at 4.2Å (Ludtke et al., 2008),

© 2011 Elsevier Inc. All rights reserved.

*Corresponding author. bajaj@cs.utexas.edu (Chandrajit Bajaj), samrat.goswami@gmail.com (Samrat Goswami), zqyork@ices.utexas.edu (Qin Zhang) URL: <http://www.cs.utexas.edu/~bajaj/cvc/index.shtml> (Chandrajit Bajaj).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ϵ -15 virus at 4.5 Å (Jiang et al., 2008), 80S eukaryotic ribosome at 5.5 Å (Armache et al., 2010), and mm-cpn at 4.3 Å (Zhang et al., 2010a), 3.3 Å subviral particle of aquareovirus (Zhang et al., 2010b), cytoplasmic polyhedrosis virus at 3.88 Å (Yu et al., 2008), mammalian chaperonin TRiC/CCT at 4.0 Å (Cong et al., 2010), about 4 Å rotavirus (Chen et al., 2009; Settembre et al., 2011) etc. The finest ultra-structure elucidation of these 3D maps as supported by its resolution, provide not only insights into internal structural fields of individual proteins, nucleic acids, and structural motifs, but, even more importantly, information about how the various structural components interact. In addition, with the increasing capability of determining multiple secondary structures and conformers of a complex (Zhou et al., 2001), there is the promise of studying the dynamics of such interacting systems.

Proteins are polypeptide chains of amino acids which arrange into secondary structures of definitive shape caused by hydrogen bonding of the amide groups. The two most common secondary structures are α -helices and β -sheets. In the α helix, the polypeptide chain is coiled tightly in the fashion of curvilinear chains and thus coarsely visualized as cylinders. The β -sheets consist of several parallel or anti-parallel strands connected laterally by hydrogen bonds and are usually visualized by directional (arrowed) ribbons when resolution permits and the individual strands are discernible. In an intermediate resolution density map, to represent β -sheets by triangulated surface patches for visualization is suitable as they are rarely flat. The association of secondary structures in a particular geometric arrangements is usually called supersecondary structures (or motifs), some of which are known to have a specific biological, or structural roles. Example of motifs include helix hairpins, beta hairpins, helix-loop-helix, helix-turn-helix, beta-alpha-beta unit etc. Protein motifs further build up tertiary folds, like all alpha structures, all beta structures, mixed alpha/beta structures, mixed alpha+beta structures and other folds.

Techniques for the computational determination of secondary structure from near atomic resolution (2–4 Å) macromolecular models reconstructed and refined from X-ray and/or NMR imaging, has been well developed, (Ladd and Palmer, 2003; Cavalli et al., 2007), and encapsulated in popular programs such as DSSP (Kabsch and Sander, 1983), Stride (Frishman and Argos, 1995), and Ribbon (Carson, 1997). Information of various motifs of atomistic resolution protein models, is also widely available from web-based databases including SCOP (Murzin et al., 1995), CATH (Orengo et al., 1997) and DALI/FSSP (Holm and Sander, 1997). Our paper focuses on the computational interpretation and detection of secondary and supersecondary structures of intermediate (6–10 Å) and coarse (10–15 Å) resolution 3D maps reconstructed primarily from single particle cryo electron microscopy (EM) (Frank, 2002).

1.1. Prior Approaches of Detecting Secondary Structure from 3D EM Maps

Jiang et al. (Jiang et al., 2001) devised an algorithm called *Helix Hunter* for detecting α -helices in intermediate resolution 3D maps. They modeled the helices as cylinders and convolved the density function of the cylinder with the original map to detect the location of the helices (namely, the peaks of the cross-correlation). While the algorithm resorts to exhaustive correlation search and is slow, its main disadvantage lies in not being able to automatically segment individual helices from a packing of helices and sheets. A similar approach for β -sheet detection was adopted by Kong et al. (Kong and Ma, 2003; Kong et al., 2004) who modeled a β -sheet as a disk like primitive and searched through the input map to find the possible positions of the disk that matched best with the 3D density. This approach also suffers from the same disadvantages as (Jiang et al., 2001) and moreover fails in detecting the full extent of longer, twisted β -sheets. Yu and Bajaj (Yu and Bajaj, 2008) developed an algorithm for detection of both helices and sheets based on the features present in 3D EM maps. Their algorithm classifies a voxel into planar or cylindrical “features”

based on the relative magnitudes of the eigenvalues of the local structure tensor. Further the cylindrical features are connected to identify the axes of the helices while the planar voxels are patched to form the β -sheets. Their method like the previous volumetric density based approaches (Jiang et al., 2001; Kong and Ma, 2003; Kong et al., 2004), have limited topological and geometric accuracy in correctly locating individual helices and sheets, with their associated densities, from an intertwined collection. This is caused predominately by the corruption of neighboring side chain densities of the individual secondary structure boundaries in a packed collection of helices and sheets, and is problematic even when the secondary structures are not buried deep in the protein core.

1.2. Prior Approaches of Detecting Motifs from 3D EM Maps

The most related work is SPI-EM (Velázquez-Muriel et al., 2005) which applied a probabilistic approach to determine the homologous superfamily defined by CATH for 3D EM maps at a resolution of 8Å till 12Å. Folds or domains detecting methods are also relevant since there is a blurred distinction between super-secondary structure (motifs) and tertiary structure (folds or domains)(cf.

<http://swissmodel.expasy.org/course/text/chapter4.htm>). In (Jiang et al., 2001), the authors developed *Fold Hunter* for fitting a subunit/domain into a 3D map at intermediate/coarse resolution using an exhaustive six-dimensional search scheme. Khayat et al. (Khayat et al., 2010) designed FREDs to identify the crystal structures that best describes the electron density provided by the users, in a fashion similar to SPI-EM. All these superfamily or folds detection schemes share a common model based search technique. First a database of 3D maps of known classification or fold models is constructed. When user provides a 3D map, all the approaches will carry out an exhaustive search for best match with high correlation scores to those maps saved in the database.

In this paper, we present automatic methods for detecting the secondary (see Figure 1) and supersecondary structures of proteins from 3D EM maps of intermediate and coarser resolution in a *de novo* sense. The algorithmic steps for detection of secondary structures from an intermediate resolution 3D EM map are:

- 1) A monomeric protein 3D map Γ is segmented from the overall 3D map of the macromolecular assembly. The description of this step is skipped in this paper as we use the implementation of the algorithms developed by Yu and Bajaj (Yu and Bajaj, 2005) in the publicly available software tool VolRover (Bajaj, 2003).
- 2) A suitable molecular surface Σ is extracted from 3D map Γ of the protein under investigation and the medial axis of this surface is computed (Section 3.1 for definitions and details).
- 3) The *critical points* of distance function h_P to surface Σ are computed (Section 2 for definitions and details).
- 4) The *unstable manifolds* of the index-1 and index-2 critical points of h_P are computed (Section 2 and Appendix A for definitions and details).
- 5) Secondary structures are extracted from these unstable manifolds (Section 3.2 for details).

For the detection of motifs from coarser resolution 3D EM maps, the first three algorithmic steps are the same as the pipeline for secondary structure detection (above). The different steps are:

- 4) The *interior maxima* of the distance function h_P and their associated stable manifolds are computed (Section 2 and Section 4.1 for definitions and details).

- 5) The stable manifolds of the maxima are used to decompose the volume bounded by the molecular surface. Aggregation is performed to produce a set of bio-chemically meaningful supersecondary regions (Section 4.2 for details).

The approach in this paper is novel as the first boundary-based approach for detecting the secondary structure of proteins from 3D EM maps. We exploit the natural connection between the topological features induced by the molecular surface, with its bio-chemical and structural (geometric) features. The key ingredient of the algorithm presented in this paper is the concept of the “distance” function. Variants of the topological structures of this function have been used earlier for shape modeling of three dimensional objects. In computational geometry, the distance function has been used to reconstruct piecewise triangulated surfaces from discrete point samples (Chaine, 2003; Giesen and John, 2003) and to compute the stable medial axis (Chazal and Lieutier, 2004; Siddiqi et al., 1998; Dey and Zhao, 2004). While a surface approximation and its medial axis serve as crude shape descriptors, deeper insights about the feature space of the shape can also be obtained using the critical points of the distance function. This method has been successful in the context of shape segmentation and annotation of points sampled three-dimensional geometry (Dey et al., 2003; Goswami et al., 2006). Another advantage of this method is that it is much faster than those approaches based on exhaustive search idea, which can be easily proved by experimental test.

The method for motif detection is a generalization of the secondary structures detection scheme where the topological features of the distance function with a natural decomposition of the 3D map capturing stable basins of density are connected. We exploit the critical point structure of the distance function and show that a subset of the stable manifolds are connected to the structural motifs of the corresponding protein molecule. The motifs detection scheme is distinct with the others since it is a *de novo* approach where we avoid the general database-searching idea.

Initial results were presented in (Bajaj and Goswami, 2006). The algorithms we present here have been enhanced, and optimized for varying resolutions, between 6–10Å, and 10–15Å. A new and robust curve/surface fitting scheme has further been implemented for secondary structure detection, and parameters have been tweaked for optimal quantitative results. Quantitative evaluation is carried out and critical assessments of the applicability of our method for reconstructed 3D EM maps available from the EMDB are also given.

2. Mathematical Preliminaries of Distance Function

In this section we discuss the mathematical setup in which we design the algorithms to identify the secondary and supersecondary structures. We give an intuitive explanation and a concise mathematical description of the relevant concepts. The interested readers can get a more detailed description of the specific algorithmic contributions of this work in the supplement.

2.1. Voronoi and Delaunay Objects

We use the Euclidean distance $\|\cdot\|$ as metric in this paper. For a finite set of points P in $\mathbb{R}^3$, the Voronoi cell of $p \in P$ is

$$V_p = \{x \in \mathbb{R}^3 : \forall q \in P - \{p\}, \|x - p\| \leq \|x - q\|\}.$$

If the points are in general position, then two Voronoi cells with non-empty intersection meet along a planar, convex Voronoi facet, three Voronoi cells with non-empty intersection meet along a common Voronoi edge and four Voronoi cells with non-empty intersection

meet at a Voronoi vertex. A cell decomposition consisting of the *Voronoi objects* — Voronoi cells, facets, edges and vertices, is the Voronoi diagram $\text{Del}(P)$ of the point set P .

The dual of $\text{Vor}(P)$ is the Delaunay diagram $\text{Del}(P)$ of P which is a simplicial complex when the points are in general position. The tetrahedra are dual to the Voronoi vertices, the triangles are dual to the Voronoi edges, the edges are dual to the Voronoi facets and the vertices (sample points from P) are dual to the Voronoi cells. We also refer to the Delaunay simplices as *Delaunay objects*. In Figure 2, we show examples of Voronoi and Delaunay objects, where Delaunay objects are drawn in green while the Voronoi objects are drawn in magenta.

2.2. Critical Points of Distance Function and Their Stable/Unstable Manifolds

Given a compact surface Σ embedded in $\mathbb{R}^3$ and a point x enclosed by surface Σ , *distance function* $h_\Sigma(x)$ is the distance between point x and surface Σ . If Σ is known via a finite set of points P sampling Σ , $h_\Sigma(x)$ can be approximated by $h_P(x)$ which is the distance from point x to the closest sample point in P . This function encodes some interesting topological and geometric characteristics of Σ . In our case the Σ is the molecular surface extracted from the electron density map.

A point in Euclidean space is *critical* if and only if it is contained in the convex hull (de Berg et al., 2008, chapter 11) of its closest points from P (obviously the closest points need not be unique on lower dimensional Voronoi objects). The convex hull of the closest points is always an face of the Delaunay object. Hence whenever a Delaunay object and its dual Voronoi object intersect, then this intersection is critical. This property points us a way to compute critical points of distance function h_P . Figure 2 further shows the four types of critical points intersected by Delaunay objects and Voronoi objects. In particular, the *maxima* of h_P are the Voronoi vertices which lie inside their dual Delaunay tetrahedra, the *index-2* saddle points of h_P are the intersections of Voronoi edges with their dual Delaunay triangles, the *index-1* saddle points of h_P are the intersections of the Voronoi faces with their dual Delaunay edges and the *minima* of h_P are the sample points (Delaunay vertices) themselves as they lie inside their Voronoi cells. Note, the index of a critical point matches the dimension of the Delaunay simplex that contains the critical point.

At every point $x \in \mathbb{R}^3$ enclosed by surface Σ , a unit vector can be assigned in the direction of the steepest ascent of h_P . The critical points are assigned zero vectors. This vector field in the interior of surface Σ induces a *flow* which tells how a point x moves in $\mathbb{R}^3$ along the steepest ascent of h_P and the corresponding path is called the *orbit* of x . Analogously, one can also think of the *inverted orbit* as the path a point traces when it follows the steepest descent direction. For a critical point c , its *stable manifold* is the union of c and the set of points whose orbits end at c and *unstable manifold* is union of c and the set of points whose inverted orbits end at c . For example, for the maximum critical point in Figure 2 (a), the stable manifold is the Delaunay tetrahedron while the unstable manifold of the maxima is itself. The stable manifold for the index-2 saddle point in Figure 2 (b) is the Delaunay triangle while the unstable manifold of this saddle point is the Voronoi edge. Similarly, the stable manifold for the index-1 saddle point in Figure 2 (c) is the Delaunay edge while the unstable manifold of this saddle point is the Voronoi facet. The stable manifold for the minimum critical point in Figure 2 (d) is itself while the unstable manifold of this critical point is the Voronoi cell.

The unstable manifolds of the saddle points are the subset of the medial axis which can be one-dimensional curve or two-dimensional surface patch. And the stable manifolds of the maxima are subsets of the volume enclosed by the molecular surface. Furthermore, we

exploit the following connections between the stable/unstable manifolds and the molecular surface.

Observation 1: If the molecular surface is locally tubular, then the unstable manifolds (U_2) of the index-2 saddle points give the medial axis of the tube that locally approximates the molecular surface. In other words, the unstable manifolds of the index-2 saddle points are structurally correlated to the α -helices of the molecule.

Observation 2: If the molecular surface is locally flat (or disc like), then the unstable manifolds (U_1) of the index-1 saddle points approximate the planar subset of the skeleton in that region. This indicates that the unstable manifolds of the index-1 saddle points are structurally correlated to the β -sheets of the molecule.

Observation 3: The stable manifolds (S_3) of the maxima are subsets of the volume enclosed by the molecular surface. As will be shown in Section 4, the stable manifolds of the maxima, when suitably combined based on the distance to the molecular surface, help identify the motifs of the molecule.

In Figure 3, we show one example of unstable manifolds U_1 and U_2 of index-1 and index-2 saddle points from a molecular surface generated for a crystal structure of the tyrosine kinase domain of the human insulin receptor.

Therefore our main goal in this work is to compute the unstable manifolds U_1 and U_2 of index-1 and index-2 saddle points of h_P to identify the secondary structures and stable manifolds S_3 of the maxima of h_P to identify the motifs. As we show in this work, the critical points and their stable and unstable manifolds can be efficiently computed based on the Voronoi-Delaunay diagram data structure. The interested reader is refer to Appendix A for the computational and algorithmic details.

3. Secondary Structure Detection

Once the 3D map of a monomeric protein molecule is segmented from that of the macromolecular assembly, the secondary structure detection procedure involves four remaining steps:

1. Extraction of the molecular surface and medial axis (Section 3.1).
2. Computation of the critical points of the distance function to molecular surface (Discussed in subsection 2.2).
3. Computation of unstable manifolds U_1 and U_2 (Described in Appendix A).
4. Detection of α -helices and β -sheets (Section 3.2).

3.1. Molecular Surface Extraction

The input to our molecular surface algorithm is a 3D map Γ of a monomeric protein, most often segmented from a larger macro-molecular assembly, e.g., an asymmetric unit of a virus (Yu and Bajaj, 2008). Starting with Γ , we first extract a molecular surface Σ using an operation known as “contouring” or “isosurface extraction” (Lorensen and Cline, 1987; Bajaj et al., 1999). Given a 3D map, an isocontour is a two-dimensional surface for which all points have same density value or contour level. Different density contour values yield different molecular surfaces, and so the selection of the contour level density is important. We utilize a topology based selection method (van Kreveld et al., 1997; Bajaj et al., 1997) that we address in greater detail, below. Although alternative options are available for producing Delaunay-conforming isosurfaces (Boissonnat and Oudot, 2005; Cheng et al., 2004; Goswami et al., 2007), the marching cubes method is sufficient for generating dense enough point sets for Delaunay triangulation and is thus used. When extracting the

appropriate molecular surface from a 3D density map downloaded from the EMDB (<http://www.ebi.ac.uk/pdbe-srv/emsearch/form>), we consider the recommended isodensity value that is usually deposited along with the map. For example, for the 4.2Å GroEL data we used in Section 5, the recommended isovalue is 0.597, which is a good initial contour value for us to consider and optimize from. For a synthetically constructed map using coarse resolution Gaussian or smooth atomistic kernel blurring of some deposited X-ray structure model, a good initial approximation is to use the contour value (typically 1) which models the atomic density at the region dominated by a single atom. Our selection of the most appropriate molecular contour level density surface is guided by the initial selection above, as well as topological information that we can extract from the *contour tree* (van Kreveld et al., 1997). The contour tree captures the topology of level sets across different contour levels where the edge of the tree for a fixed level yields the number and size (volume) of the connected surface components. We search around the initial selection and choose that contour level that yields a single large connected surface, with minimal number of smaller size components (mostly noise). Finally, if necessary we optimize this topologically based selection and produce a minimal smooth molecular surface using geometric flow evolution (Bajaj et al., 2008, 2009). In Figure 4, we provide the contour tree example for Bluetongue Virus VP7 monomer (see Figure 8 (d)). Figures (a), (b) and (c) show three isosurfaces with contour values 6.08426, 7.81963 and 9.01018, respectively. The contour tree is shown in figure (d), where the horizontal black line segments show the topology of the isosurfaces and three vertical bars with green ends correspond to the three isosurfaces separately. For example, the second vertical line with green end is corresponding to (b) isosurface, with isovalue 7.81963. The other three values belong to the contour spectrum of the density map (Bajaj et al., 1997), which is a signature of the isosurface. The signature consists of a variety of scalar data and contour attributes and three measures are shown here. The area of the isosurface is shown next to contour value (e. g., the area of the isosurface in (b) is $32795.6\text{\AA}^2$). Closed isosurfaces divide the volume into two regions, one we call the minimal volume with density value less than the contour value, and the other region we call the maximal volume, with density value greater than the contour value. For example, the minimal volume separated by the isosurface in figure (b) is $798818\text{\AA}^3$, while the maximal volume is $41918.9\text{\AA}^3$. The number of the intersections of the contour tree with the vertical line with green end is the number of the components of the isosurface and the branch points of the contour tree show that the topology of the isosurface is changing at these corresponding isovalues. From this figure, one can easily see that isosurface component (a) is somewhat inflated, while isosurface (c) is composed of too many connected components. Thus isosurface (b) seems the best qualified, meeting the topological requirements for an initial molecular surface among these three isosurfaces.

Once a suitable molecular surface is extracted, we compute its medial axis using the algorithm in (Dey and Zhao, 2004). Medial axis is usually very noisy and often does not prove to be a very useful and robust descriptor of the shape under investigation. However, since all the critical points of the distance function to molecular surface lie on the medial axis, we can collect the index-1 and index-2 critical points, a subset of the critical points on medial axis, for the computation of the unstable manifolds U_1 and U_2 .

3.2. Detection of α -helices and β -sheets

The unstable manifolds of index-1 and index-2 saddle points restricted to the medial axis of the molecular surface decompose the medial axis into linear (U_2) and planar (U_1) portions. The advantage of such an annotation is that the linear sub-part (U_2) gives the axial description of a tubular part while the planar sub-part (U_1) provides the planar axis of the relatively flat portion of the molecular surface. The next task is to select a subset of U_2 that gives the α -helices of the protein. Typically, an α -helix is a cylindrical subvolume of the

molecule which is of width approximately $2.5\text{\AA}$. Also the subvolume does not deviate much from a straight cylinder for proteins (Branden and Tooze, 1999; Jiang et al., 2001). These two conditions dictate the following computational steps.

The unstable manifold of every index-2 saddle on the medial axis is a polyline with Voronoi vertices at the joints. Every Voronoi vertex has h_P value which can be computed by the circumradius of the dual Delaunay tetrahedron. Locally this gives the radius of the cylinder that best fits the molecular surface in the cylindrical regions. We first populate the set H with the Voronoi vertices whose h_P values fall within $2\text{\AA}$ and $3\text{\AA}$. A $3\text{\AA}$ neighborhood graph is then computed over H that clusters the points. The choice of $3\text{\AA}$ is dictated by the fact that pitch of the helices is $1.5\text{\AA}$ and usually there is more than 2 turns in every helix. The diametral point pair in every cluster is then computed. The maximum deviation of any intermediate point from the straight line joining the diametral points determines how close the fitted cylinder is to the cluster. This way we select those clusters from H and the cylinders fitted to these clusters produce the detected α helices. The process is shown in Figure 5(b).

The selection of β -sheets is similar. U_1 gives the possible candidates for β -sheets. First we notice that there are some tiny components that are created due to sampling artifacts, and they do not correspond to real planar substructures of the molecule. We filter these small clusters out (green patches in Figure 5(c)). After this stage of filtering, we are left with the planar subsets of the medial axis (cyan in Figure 5(c)). At this point we apply the knowledge that β -sheets are of width roughly $1.5\text{\AA}$ (Branden and Tooze, 1999; Kong and Ma, 2003) and we filter out the planar patches which do not satisfy this “thickness” criterion. The width of a facet in U_1 is easy to check as they are the Voronoi facets and therefore have a 1-1 correspondence with their dual Delaunay edges which cross the medial axis. Therefore, we select only those Voronoi facets from U_1 whose dual Delaunay edges are of length between $1\text{\AA}$ and $2\text{\AA}$. The portions of U_1 which are filtered out by this test are shown in magenta in Figure 5(c). The rest of U_1 , which qualify for β -sheets, are shown in cyan. The sheets detected are shown to correspond well with the β -sheets recorded in Protein Data Bank.

4. Supersecondary Structures Detection

We detect, segment, and cluster volumetric densities representing supersecondary structures from the 3D EM maps. The auto-segmentation captures various supersecondary structure 3D densities, mixed in with smaller sized volumes that are individual secondary structures, as well as those that are not entirely discernible as representing a motif. A clustering and classification step allows us to group and filter out the false positives/negatives from more certifiable motifs. We determine also from our experimental tests, that our algorithm reports lower false positives when a coarser resolution 3D EM map of the molecule is available. This is because at coarser resolutions the boundaries of individual secondary structures are blurred sufficiently so that the boundary segmentation of densities better capture a supersecondary collection.

We first describe, by an extension of our secondary structure detection algorithm of the previous section, how a superfamily of the motif densities can be automatically segmented. We again consider the distance function h_P induced by the set of points P sampled on an approximation of the molecular surface Σ extracted from a coarser resolution 3D map.

The steps of motifs detection are as follows:

1. Steps 1) and 2) of the secondary structure algorithm remain as before.
2. The stable manifolds of the interior maximum of h_P are computed.

3. The adjacent maxima are grouped together for appropriate motifs detection.

We have already seen that the maxima of h_P are the intersections of the Delaunay tetrahedra with their dual Voronoi vertices. In other words, these are the circumcenters of the Delaunay tetrahedra whose circumcenters lie inside them (Section 2). The key algorithmic ingredient to identify the motif is the computation of the stable manifolds S_3 of the maxima. Algorithm to compute these stable manifolds is given in (Dey et al., 2003). We briefly describe it here for completeness.

4.1. Stable Manifolds of Maxima

The stable manifolds of the critical points of h_P can be computed from the Voronoi and Delaunay diagram $\text{Del}(P)$ of P . The stable manifolds may not be sub-complex of $\text{Del}(P)$, and so we approximate them with Delaunay simplices.

Given two Delaunay tetrahedra σ and σ' which share a common triangle t , we say $\sigma < \sigma'$, if the circumcenter of σ lies in the half-space defined by t that does not contain the fourth vertex of σ . Figure 6(a) depicts this case. Note that σ' is not always unique (Figure 6(b)). However, a Delaunay tetrahedron σ can have at most two neighbors σ_1 and σ_2 , for which $\sigma < \sigma_1$ and $\sigma < \sigma_2$. Also, it is to be noted that the Delaunay tetrahedron $\sigma_{\max}$ whose dual Voronoi vertex is a maximum, has none of its neighbors σ' for which $\sigma_{\max} < \sigma'$.

The set of maxima is sorted according to h_P . Starting from the biggest maximum, the algorithm collects all σ which fall under the transitive closure of the relation ' $<$ '.

4.2. Clustering of Adjacent Stable Manifolds

The stable manifolds of the maxima are often numerous and therefore they need to be grouped carefully to bring out the underlying features of the molecule. We observe that, when the stable manifolds of two maxima c_1 and c_2 have non-empty intersection and an index-2 saddle s lies at the common intersection, then

1. if c_1 comes from subvolume which is substantially thinner than the subvolume that c_2 belongs to, then

$$h_P(c_2) \gg h_P(c_1) \geq h_P(s),$$

and

2. if both c_1 and c_2 belong to subvolumes which are comparably “fat”, but are connected via a thin channel, then

$$\min\{h_P(c_1), h_P(c_2)\} \gg h_P(s).$$

These two cases are scenarios where the stable manifolds of two maxima (c_1 and c_2) should not be merged for motifs detection. In other words, we merge two neighboring stable manifolds (of c_1 and c_2 with a shared index-2 saddle point s) if and only if their h_P function values are *comparable* – as measured by a scalar parameter ρ –

$$\max\left\{\frac{h_P(c_2)}{h_P(s)}, \frac{h_P(c_1)}{h_P(s)}\right\} \leq \rho.$$

Parameter ρ affects how restrictively we merge the subvolumes to form bigger clusters. The bigger the value of ρ , the less restrictive it is and therefore it allows the coalescence of stable

manifolds with more disparate maxima. At the same time ρ is also crucial in the separation of two or more motifs. We note that similar techniques were also used earlier for free-form object segmentation (Dey et al., 2003) and topological noise removal (Bajaj et al., 2007). In this work, we have seen a value of ρ typically between 1 and 1.5 can separate different motifs effectively. Figures 7(a–d) show an example of the stable manifolds and the motifs they correspond to. While the motifs are not always readily decomposable using our algorithm, we observe that the decomposition algorithm successfully detects the helical regions and separates them from the beta regions.

5. Implementation and Results

For calibrating our structure detection algorithms, we downloaded atomic level descriptions of proteins from the Protein Data Bank (Berman et al., 2000). For each protein, we first “blurred” them into a 3D map (to correspond to a reconstructed 3D EM map) at varying resolutions (5–15Å), using our software T_{EX}MOL (Bajaj et al., 2004), where the relationship between blobbiness b in Gaussian blur function $g(u) = e^{-bu^2}$ and resolution R is $b = (2\pi^2)/(R^2 \ln 2)$. This relationship is consistent with the resolution used in publicly available software EMAN (Ludtke et al., 1999). From these volumetric maps, we extracted the proteins molecular surface using T_{EX}MOL. Next we collected a point set sampling of the molecular surface and computed the medial axis as was done in (Goswami et al., 2006). The next step was to detect the critical points and compute the stable/unstable manifolds of a subset of the critical points. These computations rely on the Voronoi Diagram–Delaunay triangulation of the point set, and were done using the CGAL library (CGAL Consortium, 1996).

The calibration process is essential before we apply them to 3D EM maps of unknown atomic descriptions. We calibrate against a dataset which is similar to the one used in (Jiang et al., 2001). The results are shown in Figure 8 and statistics is given in Table 1. The model data includes bacteriorhodopsin (1C3W, all alpha), Cytochrome C’ (1BBH, all alpha), triose phosphate isomerase (1TIM, α/β), tyrosine kinase domain from the insulin receptor (1IRK, $\alpha + \beta$) and Bluetongue Virus outer shell coat protein (1BVP, a β upper domain and an α lower domain), where the classification is based on the widely accepted Structural Classification Of Proteins (SCOP) (Murzin et al., 1995). The models are blurred to 8Å resolution with blobbiness $b = 0.4450$. In Figure 8, the first column shows the input geometry data and the second column shows the secondary structures detected using our method. The results and the input meshes are shown in the third column with the input meshes transparently displayed. The last column shows our results and the secondary structures from PDB by DSSP (Kabsch and Sander, 1983) rendered by Chimera (Pettersen et al., 2004). We also show the quantification results in Table 1. Column 2 in this table lists the α helix and β sheet numbers documented in PDB files and the helix and sheet numbers assigned by DSSP are listed in column 3. We list the helix and sheet numbers obtained by our methods in column 4 and calculate the numbers of false positives (FP, there should be no helix or sheet but one is wrongly detected) and false negatives (FN, there should be a helix or sheet but one is not detected) comparing to DSSP and list them in column 5. From the figures and the table, we can see most of the helices and sheets can be detected. If the length of helix is very small, it will be usually difficult to detect and this is very common for such as *Helix Hunter* and the method in (Yu and Bajaj, 2008). If we set the length threshold too small, the method will detect more false helices. Thus it is a tradeoff between helices length threshold and number of helices detected.

We next show one result of the secondary structure detection for the segmented capsid proteins of Rice Dwarf Virus at 6.8Å resolution (Zhou et al., 2001). The architecture of the virus suggests that it has a double-layer capsid. The outer capsid layer is composed of

multiple copies of protein P8 and the inner layer is composed of protein P3. Figure 9 shows molecular surface extracted from segmented cryo-EM map passing through the data processing pipeline. The unstable manifolds of the index-1 and index-2 saddle points are computed and subsequently the secondary structures are detected.

Our next results of real data are from the workshop on the validation and modeling of electron cryomicroscopy structures of biological nanomachines held with the Pacific Symposium on Biocomputing (PSB) 2011 on the Big Island of Hawaii during January 3–7, 2011. The first example is a 4.2Å EM map of GroEL (EMDB ID: 5001) (Ludtke et al., 2008) shown in Figure 10. We first segment one subunit from the whole volume. Then from the contoured mesh we detect the secondary structures (see the caption of the figure for details). Another example we show is for an 8Å cryo EM map (EMDB ID: 5140) (Zhang et al., 2010a) of open state *Methanococcus maripaludis* chaperonin (mm-cpn), a group II chaperonin responsible for folding many essential proteins, such as actin and tubulin in eukaryotes. In Figure 11, we show the original volume (a), mesh of segmented subunits (b), and secondary structure detection results (c,d).

Figure 12 shows the performance of our motifs detection algorithm. Top row shows the secondary structure of 1AOR as provided by PDB along with the ensemble of the segments of the protein at 12Å. The tertiary fold of 1AOR is named as *four-layer sandwich*. The two red segments (in c) correspond to the two β -sheets while the top and bottom parts in (b) and (c) show helices. The middle row shows similar results for 1PLQ (at 12Å). The tertiary fold is *box* and the motifs detection can successfully separate the three sheets (red, yellow, green) from the helices (blue). The bottom row shows the decomposition of the *four-layer sandwich* of deoxyribonuclease I (chain A of 2DNJ) at 12Å resolution. The top and bottom helices, the middle sheets are separated (see figures (l) and (m)). In Figure 13, we show a *horseshoe* fold is decomposed into several motifs for procine ribonuclease inhibitor (PDB ID: 2BNH) from a 12 Å cryo-EM map. From figures (e) and (f), one can see that the sheet is separated with different helices.

6. Conclusions

We have presented algorithms for secondary and supersecondary structures detection from 3D EM maps of proteins at varying resolutions. The algorithms work by analyzing the stable and unstable manifolds of a subset of the critical points of the distance function, computed from the molecular surface point set sampling of the protein.

The algorithm presented in this paper relies on a suitable approximation of the molecular surface. We plan to further investigate the choice of such surfaces based on the bond lengths of the atoms present to narrow down the possible range of selection. Also, we believe the critical points of h_p which lie outside the molecular surface carry useful information about the motifs.

During our experiments, we found that detecting the exact secondary and super-secondary structures from cryo-EM maps is quite challenging. The main difficulty is separating backbone and side chain regions which are tightly bounded. From PDB data, one can easily differentiate backbone and side chain atoms but this is not the case for cryo-EM data. Most data used in the paper is virtual cryo-EM data generated from blurring PDB data with prescribed resolution. For real cryo-EM data, even though the resolution is theoretically good enough, it is still quite challenging to detect very good secondary structures and motifs due to noise or some other factors.

Acknowledgments

The authors are supported in part by NIH grants R01 GM074258-021, R01-GM073087 and R01 EB004873. The work performed by Dr. Samrat Goswami was while he was a postdoc at University of Texas at Austin. We thank Dr. Zeyun Yu for valuable discussions on the secondary structure detection of virus proteins and Dr. Alexander Rand for his help with improving the readability of this manuscript. The implementation of all of the above structural interpretation algorithms, as well pictures used here, have been made in our in-house image processing and visualization software tool, called Vol-Rover. The VolRover program is in the public domain and can be freely downloaded from our center's software website (<http://cvcweb.ices.utexas.edu/cvcwp/>).

References

- Armache JP, Jarasch A, Anger AM, Villa E, Becker T, et al. Localization of eukaryote-specific ribosomal proteins in a 5.5-Å cryo-EM map of the 80s eukaryotic ribosome. *PNAS*. 2010; 107(46): 19754–19759. [PubMed: 20974910]
- Bajaj C. VolRover: An interactive volume visualization tool. 2003 http://cvcweb.ices.utexas.edu/cvcwp/?page_id=100.
- Bajaj, C.; Djeu, P.; Siddavanahalli, V.; Thane, A. Interactive visual exploration of large flexible multi-component molecular complexes; *Proc. of the Annual IEEE Visualization Conference*; 2004. p. 243-250.
- Bajaj, C.; Gillette, A.; Goswami, S. Topology based selection and curation of level sets. In: Wiebel, A.; Hege, H.; Polthier, K.; Scheuermann, G., editors. *Topology-in-Visualization*. 2007. p. 45-58.
- Bajaj, C.; Goswami, S. Secondary and tertiary structural fold elucidation from 3D EM maps of macromolecules. *Proceedings of the Fifth Indian Conference on Computer Vision, Graphics & Image Processing*. Vol. 4338 of LNCS; Springer Verlag; 2006. p. 264-275.
- Bajaj, C.; Pascucci, V.; Schikore, D. The contour spectrum. *VIS '97: Proceedings of the 8th conference on Visualization*; IEEE Computer Society Press; 1997. p. 167-173.
- Bajaj, C.; Pascucci, V.; Schikore, D. Accelerated isocontouring of scalar fields. In: Bajaj, C., editor. *Data Visualization Techniques*. Vol. chapter 3. John Wiley and Sons; 1999.
- Bajaj CL, Xu G, Zhang Q. Higher-order level-set method and its application in biomolecule surfaces construction. *Journal of Computer Science and Technology*. 2008; 23(6):1026–1036.
- Bajaj CL, Xu G, Zhang Q. A fast variational method for the construction of resolution adaptive c^2 -smooth molecular surfaces. *Computer Methods in Applied Mechanics and Engineering*. 2009; 198(21–26):1684–1690. special Issue in Honor of Professor J. T. Oden's 70th Birthday. [PubMed: 19802355]
- Baker ML, Zhang J, Ludtke SJ, Chiu W. Cryo-EM of macromolecular assemblies at near-atomic resolution. *Nature Protocols*. 2010; 5(10):1697–1708.
- Baker TS, Olson NH, Fuller SD. Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbio. Mol. Bio. Reviews*. 1999; 63:862–922.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. The Protein Data Bank. *Nucleic Acids Research*. 2000:235–242. [PubMed: 10592235]
- Boissonnat J-D, Oudot S. Provably good sampling and meshing of surfaces. *Graphical Models*. 2005; 67:405–451.
- Branden, C.; Tooze, J. *An Introduction to Protein Structure*. New York: Garland Publishing Inc.; 1999.
- Carson M. Ribbons. *Methods in Enzymology*. 1997; 277:493–505. [PubMed: 18488321]
- Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. USA*. 2007; 104(23):9615–9620. [PubMed: 17535901]
- CGAL Consortium. CGAL: Computational Geometry Algorithms Library. 1996 <http://www.cgal.org>.
- Chaine, R. A geometric convection approach of 3-D reconstruction; *Proc. Eurographics Sympos. Geometry Processing*; 2003. p. 218-229.
- Chazal, F.; Lieutier, A. Stability and homotopy of a subset of the medial axis; *Proc. 9th ACM Sympos. Solid Modeling and Applications*; 2004. p. 243-248.

- Chen JZ, Settembre EC, Aoki ST, Zhang X, Bellamy AR, et al. Molecular interactions in rotavirus assembly and uncoating seen by high-resolution cryo-EM. *Proc. Natl. Acad. Sci. USA*. 2009; 106:10644–10648. [PubMed: 19487668]
- Cheng, S-W.; Dey, T.; Ramos, E.; Ray, T. Sampling and meshing a surface with guaranteed topology and geometry; *Proc. of the 20th Annual Symposium on Computational Geometry*; 2004. p. 280-289.
- Cong Y, Baker ML, Jakana J, Woolford D, Miller EJ, et al. 4.0-Å resolution cryo-EM structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement. *PNAS*. 2010; 107(11): 4967–4972. [PubMed: 20194787]
- de Berg, M.; Cheong, O.; van Kreveld, M.; Overmars, M. *Computational Geometry: Algorithms and Applications*, 3rd Edition. Springer-Verlag; 2008.
- Dey, TK.; Giesen, J.; Goswami, S. Shape segmentation and matching with flow discretization. In: Dehne, F.; Sack, J-R.; Smid, M., editors. *Proc. Workshop Algorithms Data Structures (WADS 03)*. LNCS 2748. Berlin, Germany: 2003. p. 25-36.
- Dey TK, Zhao W. Approximating the medial axis from the Voronoi diagram with convergence guarantee. *Algorithmica*. 2004; 38:179–200.
- Frank, J. *Three-Dimensional Electron Microscope of Macromolecular Assemblies*. San Diego: Academic Press; 1996.
- Frank J. Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* 2002; 31:303–319. [PubMed: 11988472]
- Frishman A, Argos P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics*. 1995; 23:566–579.
- Giesen, J.; John, M. The flow complex: a data structure for geometric modeling; *Proc. 14th ACM-SIAM Sympos. Discrete Algorithms*; 2003. p. 285-294.
- Goswami, S.; Dey, TK.; Bajaj, CL. Identifying at and tubular regions of a shape by unstable manifolds; *Proc. 11th Sympos. Solid and Physical Modeling*; 2006. p. 27-37.
- Goswami S, Gillette A, Bajaj C. Efficient Delaunay mesh generation from sampled scalar function. *Proceedings of the 16th International Meshing Roundtable*. 2007:495–511.
- Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* 1997; 25(1):231–234. [PubMed: 9016542]
- Jiang W, Baker ML, Jakana J, Weigele PR, King J, et al. Backbone structure of the infectious ϵ 15 virus capsid revealed by electron cryomicroscopy. *Nature*. 2008; 451:1130–1135. [PubMed: 18305544]
- Jiang W, Baker ML, Ludtke SJ, Chiu W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* 2001; 308:1033–1044. [PubMed: 11352589]
- Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; 22(12):2577–2637. [PubMed: 6667333]
- Khayat R, Lander GC, Johnson JE. An automated procedure for detecting protein folds from sub-nanometer resolution electron density. *J. Struct. Biol.* 2010; 170:513–521. [PubMed: 20026407]
- Kong Y, Ma J. A structural-informatics approach for mining β -sheets: locating sheets in intermediate-resolution density maps. *J. Mol. Biol.* 2003; 332:399–413. [PubMed: 12948490]
- Kong Y, Zhang X, Baker TS, Ma J. A structural-informatics approach for tracing β -sheets: building pseudo-Ca traces for β -strands in intermediate-resolution density maps. *J. Mol. Biol.* 2004; 339:117–130. [PubMed: 15123425]
- Ladd, MFC.; Palmer, RA. *Structure Determination by X-Ray Crystallography*, 4th Edition. Springer; 2003.
- Lorensen, W.; Cline, H. Marching cubes: A high resolution 3d surface construction algorithm; *SIGGRAPH '87: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*; 1987. p. 163-169.
- Ludtke S, Baldwin P, Chiu W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Bio.* 1999; 128:82–97. [PubMed: 10600563]

- Ludtke SJ, Baker ML, Chen DH, Song JL, Chuang DT, et al. De Novo backbone trace GroEL from single particle electron cryomicroscopy. *Structure*. 2008; 16:441–448. [PubMed: 18334219]
- Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 1995; 247:536–540. [PubMed: 7723011]
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. CATH-a hierarchic classification of protein domain structures. *Structures*. 1997; 5(8):1093–1108.
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. UCSF Chimera - a visualization system for exploratory research and analysis. *J. Comp. Chem.* 2004; 25(13):1605–1612. [PubMed: 15264254]
- Settembre EC, Chen JZ, Dormitzer PR, Grigorieff N, Harrison SC. Atomic model of an infectious rotavirus particle. *EMBO J.* 2011; 30(2):408–416. [PubMed: 21157433]
- Siddiqi K, Shokoufandeh A, Dickinson J, Zucker S. Shock graphs and shape matching. *Computer Vision*. 1998:222–229.
- van Kreveld, M.; van Oostrum, R.; Bajaj, C.; Pascucci, V.; Schikore, D. Contour trees and small seed sets for isosurface traversal; SCG '97: Proceedings of the thirteenth annual symposium on Computational geometry; 1997. p. 212-220.
- Velázquez-Muriel JA, Sorzano COS, Scheres SHW, Carazo JM. SPI-EM: Towards a tool for predicting CATH superfamilies in 3D-EM maps. *J. Mol. Biol.* 2005; 345:759–771. [PubMed: 15588824]
- Yu X, Jin L, Zhou ZH. 3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature*. 2008; 453:415–419. [PubMed: 18449192]
- Yu Z, Bajaj C. Automatic ultra-structure segmentation of reconstructed cryo-EM maps of icosahedral viruses. *IEEE Transactions on Image Processing*. 2005; 14:1324–1337. [PubMed: 16190468]
- Yu Z, Bajaj C. Computational approaches for automatic structural analysis of large bio-molecular complexes. *IEEE/ACM Trans. Comp. Bio. Bioinfo.* 2008; 5:568–582.
- Zhang J, Baker ML, Schröder GF, Douglas NR, Reissmann S, et al. Mechanism of folding chamber closure in a group II chaperonin. *Nature*. 2010a; 463:379–383. [PubMed: 20090755]
- Zhang X, Jin L, Fang Q, Hui WH, Zhou ZH. 3.3 Å cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell*. 2010b; 141:472–482. [PubMed: 20398923]
- Zhang X, Settembre E, Xu C, Dormitzer PR, Bellamy R, et al. Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc. Natl. Acad. Sci. USA*. 2008; 105:1867–1872. [PubMed: 18238898]
- Zhou ZH. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr. Opin. Struct. Biol.* 2008; 18:218–228. [PubMed: 18403197]
- Zhou ZH, Baker ML, Jiang W, Dougherty M, Jakana J, et al. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nature Structural Biology*. 2001; 8(10):868–873.

Appendix

A. Computation of Unstable Manifolds

Given a compact surface Σ embedded in $\mathbb{R}^3$, a distance function h_Σ can be constructed that assigns to each point its distance to Σ . In applications, Σ is often known via a finite set of points P sampling Σ and in such cases h_Σ can be approximated with h_P that assigns the distance to the closest sample point in P .

The unstable manifolds are computed based on an induced *flow* of the function h_P . Crudely the idea is as follows. To compute the unstable manifold of any particular critical point c , we investigate its local neighborhood and see how a *free* (or regular) point x will move so that the function h_P increases most along its trajectory. Note, the critical points themselves do not move because they are assigned zero vectors. Clearly, such a method will not lead to an algorithm as the domain is continuous and therefore this approach is not computationally

feasible. However as we show in this section, a great deal of information about the behavior of a *free* point is encoded in the Voronoi-Delaunay diagram of the set of points P which originally induced the distance function h_p . In fact, simple combinatorial tests on the Voronoi-Delaunay diagram will be defined for computing the flow of any arbitrary regular point, and this leads to a method for computing the unstable manifolds in a discrete setting.

Appendix A.1. Induced Flow via Voronoi-Delaunay Duality

For the rest of this subsection, we assume that x is a regular (free) point which is allowed to move only in the direction of the gradient of the distance function. Depending on which Voronoi element x is in, we give a canonical description of its movement, restricted to that Voronoi element.

Figure A.1: The flow path of a point x when (a) x is inside V_p , and (b) x is inside a Voronoi face at the intersection of V_p and V_q .

Voronoi Cell—For any point x in the Voronoi cell V_p , the function h_p increases most radially along the vector $\vec{px}$. Figure A.1 (a) shows an illustration.

Voronoi Facet—Let F be a Voronoi facet at the intersection of V_p and V_q . The dual Delaunay edge is pq . For a point $x \in F$, the flow is along the vector $\vec{dx}$ where d is the midpoint of pq . Note, by the property of Voronoi-Delaunay duality relations, the point d is the closest point to F . Figure A.1 (b) shows an illustration. When pq intersects F , d is actually an index-1 saddle point.

Voronoi Edge—

Figure A.2: The flow path of a point x (green) on a Voronoi edge. (a) The circumcenter of the dual triangle falls outside the triangle. Two blue points identify a Voronoi edge belonging to a Voronoi facet (skyblue). (b) The circumcenter falls inside but the Voronoi edge segment (two blue points) is above the plane of the triangle. The arrow shows the flow direction starting from the circumcenter to the point x . (c) The Voronoi edge segment

pierces the plane of the triangle through the circumcenter. This results in an index-2 saddle point. Two arrows show different directions depending on the positions of x and y .

Let E be a Voronoi edge at the intersection of V_p, V_q, V_r . The dual Delaunay triangle is denoted as $\Delta_{p,q,r}$. Note by the property of Voronoi-Delaunay duality, the supporting line of E intersects the plane containing p, q, r at the circumcenter of $\Delta_{p,q,r}$. Now consider two cases.

- $\Delta_{p,q,r}$ is obtuse-angled. Let the angle at q be obtuse as seen in Figure A.2 (a). In that case, a point on E shall enter the Voronoi facet dual to the longest side of $\Delta_{p,q,r}$ which, in this case, is pr . The course of the path is then guided by the rule for Voronoi facets as we described earlier. We call such Voronoi facets *acceptors*.
- $\Delta_{p,q,r}$ is acute-angled. Any point x on E flows along the Voronoi edge itself with two subcases.

Case (a) (see Fig. A.2 (b)): the Voronoi edge defined by the two blue points doesn't intersect with the Delaunay triangle $\Delta_{p,q,r}$. The flow direction for point x is from the circumcenter of $\Delta_{p,q,r}$ to point x , as the arrow shows in the figure.

Case (b) (see Fig. A.2 (c)): the Voronoi edge intersects $\Delta_{p,q,r}$ at an index-2 saddle point. The flow direction will be from the saddle point to the point on the edge, as the two arrows shown for two points x and y on the edge.

Therefore for any point $x \in E$, one first determines if the dual $\Delta_{p,q,r}$ is acute or obtuse and then decides if x should move in the direction of the Voronoi edge or if it should enter a Voronoi facet. If it does enter a Voronoi facet (an acceptor), the above rule for Voronoi facet shall dictate its course henceforth.

Figure A.3: The flow path of a point x which is a Voronoi vertex. Voronoi facet is shown in skyblue and v_i, v_j, x are Voronoi vertices. (a) Voronoi vertex x is only blocked by Delaunay facet τ_i . (b) Voronoi vertex x is blocked by τ_i and τ_j and both the bases of perpendiculars from x to the supporting planes of τ_i and τ_j fall outside of Delaunay facets. (c) Voronoi vertex x is blocked by τ_i and τ_j and the base of the perpendicular from x to the supporting plane of τ_i falls inside of τ_i but the base of the perpendicular from x to the supporting plane of τ_j falls outside of τ_j .

Voronoi Vertex—The case of Voronoi vertex is the most involved. Let the point x be one Voronoi vertex which is at the intersection of four Voronoi cells $V_{p_i}, V_{p_j}, V_{p_k}, V_{p_l}$. The dual Delaunay simplex is the tetrahedron σ spanning the four points. Now suppose, x is outside of σ . Note, otherwise x is a maximum and has no *flow*.

Let τ_i be the triangle in σ opposite to p_i . If x and p_i are not in the same half-space of the supporting plane τ_i , we say τ_i *blocks* x (from seeing p_i). By the property of Delaunay triangulation, the number of *blockers* can be at most 2. So let us consider the following two cases; both are illustrated in Figure A.3.

- τ_i blocks x . This case can be derived from previous cases. Let E be the Voronoi edge dual to τ_i . The flow at x is guided by the flow conditions of E (described earlier).
- τ_i and τ_j block x . There are two cases to consider.

Case (a) (see Fig. A.3 (b)): both lines from x which are perpendicular to the supporting planes of τ_i and τ_j fall outside of the triangles. In this situation, the flow at x enters the Voronoi facet dual to the common Delaunay edge shared by τ_i and τ_j .

Case (b) (see Fig. A.3 (c)): one of the perpendicular lines fall inside the triangle. Without loss of generality, we take that facet to be τ_i and in that case, the flow is along the Voronoi edge dual to τ_i .

By the observations above, we know if a regular point x hits a Voronoi vertex v along its flow-path, it either stops there when v is a maximum, or there are simple combinatorial tests that decide the next course of its path.

Appendix A.2. Unstable Manifolds of Index-2 Saddle Points

Once the flow-path of a regular point is fully characterized, we are in a position to build the unstable manifold of an index-2 saddle point. Note, an index-2 saddle point c is at the intersection of a Voronoi edge E and its dual Delaunay triangle τ . Let the two tetrahedra sharing τ be σ_1, σ_2 . The edge E has the endpoints at the dual Voronoi vertices of σ_1 and σ_2 , denoted as v_1, v_2 respectively (two blue circles in Figure A.2 (c)). The unstable manifold $U(c)$ (or U_c) of c , has two intervals - one from c to v_1 and the other from c to v_2 . We look at the structure of one of them, say the one from c to v_1 , and the other one is similar.

At any point x on the subsegment cv_1 , the flow is toward v_1 from c . Once the flow reaches v_1 , the subsequent flow depends on the flow structure of v_1 as described in the previous subsection. Instead of just looking at v_1 , we consider a generic step, where the flow reaches at a Voronoi vertex v and we list the possible situations. If v is a local maximum, the flow stops there and marks the endpoint of U_c . Otherwise depending on the flow structure of v , one can combinatorially decide if the flow line follows along a particular Voronoi edge, incident on v , or if it enters a Voronoi facet. If the flow is along a Voronoi edge, it eventually hits another Voronoi vertex and we apply the same tests. If it enters a Voronoi facet, again by the flow structure, we know the direction of the flow path and it is easy to identify the exit point where it hits the boundary Voronoi edge of the Voronoi facet. The flow structure of the boundary Voronoi edge then decides the next course of the path. Figure A.4 (a) shows an example U_c for an index-2 saddle point c .

Appendix A.3. Unstable Manifolds of Index-1 Saddle Points

The computation of the unstable manifold of an index-1 saddle point is done in a similar fashion as index-2 saddle point. Unstable Manifold of index-1 saddle points are two dimensional. Due to hierarchical structure, they are bounded by the unstable manifold of index-2 saddle points.

Let us consider an index-1 saddle point, c . This point lies at the intersection of a Voronoi facet F and a Delaunay edge. For any point $x \in F \setminus c$, the flow is radially outward in the direction of $\vec{cx}$. Clearly, F is in $U(c)$. As the flow hits the Voronoi edges bounding F , we examine the flow-structure of those edges. Again, by the description of the flow structure given in Subsection Appendix A.1, we know there may be two cases. In the first case, when the flow is along the Voronoi edge, it must be in the boundary of U_c . In the second case, the flow enters one of the neighboring Voronoi facets and we recursively process that face in a similar fashion.

Figure A.4(b) illustrates an intermediate stage of this computation. The index-1 saddle point c is contained in the blue Voronoi facet. The yellow Voronoi facets are already in $U(c)$. The red edges designate the static boundary as they are non-transversal and the green edges designate the active boundary through which the pink facets are included in $U(c)$ in the later stage of the algorithm.

Figure A.4: (a) The unstable manifold $U(c)$ of an index-2 saddle point c , which is drawn with a cyan circle. The portion of $U(c)$ which is a collection of Voronoi edges is drawn in green with intermediate Voronoi vertices drawn in blue. The pink circle is a Voronoi vertex on $U(c)$ where the flow enters a Voronoi facet. The portion of $U(c)$ which lies inside the Voronoi facets is drawn in magenta. The transversal Voronoi edges intersected by this portion of $U(c)$ are dashed. $U(c)$ ends at a local maximum which is drawn in red. (b) An intermediate stage of the computation of unstable manifold of an index-1 saddle point c . The blue facet contains c , yellow facets are currently in $U(c)$ and pink facets are to be included in the subsequent iterations. The red lines denote the boundary of $U(c)$ at this stage. The green edges show the active boundary through which the pink facets will be included in later stages.

Figure 1.

(a) Van der Waals surface rendering of a GroEL protein (PDB ID: 1OEL, chain G). (b) Volume rendering of its blurred 3D map at 8Å resolution. (c) Surface rendering of the protein's molecular surface. (d) The unstable manifolds of the index-1 and index-2 saddle points. (See Appendix A for details of computation of unstable manifolds.) (e) The secondary structure detection results (17 helices and 6 sheets). The magenta cylinders represent α -helices and the other randomly colored surface patches represent β -sheets. (f) The secondary structure detection results with a transparent rendering of the mesh in (c). (g) The secondary structural motifs from the PDB, where the helices are shown as cyan ribbon coils and the sheets are sets of green ribbon strands. (h) Combination of (e) and (g).

Figure 2.

The possible cases of Voronoi-Delaunay duality and the four types of critical points of distance function h_p . (a) Delaunay tetrahedron and Voronoi vertex which is a maximum critical point. (b) Delaunay triangle and Voronoi edge and their intersection is an index-2 saddle point. (c) Delaunay edge and Voronoi facet and their intersection is an index-1 saddle. (d) Delaunay vertex and Voronoi cell. The Delaunay vertex is a minimum critical point. The (non-zero dimensional) Delaunay objects are drawn in green and the Voronoi objects are drawn in magenta.

Figure 3.

A collage of U_1 (yellow) and U_2 (red) of the Insulin Receptor Tyrosin Kinase domain (PDB ID:1IRK). (a) The unstable manifold U_1 of index-1 saddle points with transparent molecular surface. (b) The unstable manifold U_2 of index-2 saddle points with transparent molecular surface from a different view. (c) U_1 and U_2 only. (d) An enlarged view of (c). Note U_2 bounds U_1 .

Figure 4. Contour tree for Bluetongue Virus VP7 monomer density map. (a), (b) and (c) show three isosurfaces with contour values 6.08426, 7.81963 and 9.01018, respectively. (d) shows the contour tree, three vertical bars for different isovalues and three measures of contour spectrum.

Figure 5.

- (a) The molecular surface of the Insulin Receptor Tyrosin Kinase domain (PDB ID:1IRK).
 (b) The selected Voronoi vertices on U_2 and the fitted cylinder. (c) The filtering of U_1 which are small (green) or do not satisfy the width test (magenta). (d) Result of the secondary structures obtained from the PDB and its correspondence with the computed structure (b,c).

Figure 6.

(a) $\sigma_1 < \sigma_2$. (b) The situation when σ has two neighbors σ' and σ'' for both of which $\sigma < \sigma'$ and $\sigma < \sigma''$.

Figure 7.

(a) Helices surround the sheets to form the tertiary structure called α/β -barrel (PDB ID: 1TIM). (b) Molecular surface of 1TIM at 15Å resolution. (c,d) The initial segmentation and further refinement to bring out the β -fold of the barrel from the surrounding helices (yellow, magenta and blue).

Figure 8.

Several tests of the algorithm on model data: (a) Cytochrome C' (PDB ID: 1BBH), (b) Bacteriorhodopsin (PDB ID: 1C3W), (c) Insulin Receptor Tyrosin Kinase domain (PDB ID: 1IRK), (d) Bluetongue Virus VP7 monomer (PDB ID: 1BVP), and (e) Triose Phosphate Isomerase (PDB ID: 1TIM). The columns contain (*.1) the input geometry, (*.2) the detected secondary structures, (*.3) the detected secondary structure with a transparent input mesh and (*.4) the detected secondary structure with the secondary structure assigned from the PDBs. False positive (FP) and false negative (FN) results are indicated in red and blue, respectively.

Figure 9. The different stages of the secondary structure detection process on the outer-capsid layer protein P8 of Rice Dwarf Virus (EMDB ID: 1060): (a) the molecular surface, (b) the unstable manifolds U1 and U2, (c) an enlarged view of (b), (d) the detected helices and sheets, and (e) the ribbon diagram of the secondary structure recorded in Protein Data Bank.

Figure 10. Detection of secondary structures from a 4.2Å GroEL cryo-EM map (EMDB ID: 5001): (a) the original volume, (b) one subunit after segmentation, (c) the medial axis of the subunit, (d) the unstable manifolds U_1 and U_2 , (e) the detected secondary structure with 16 helices and 9 sheets, and (f) the mesh and the secondary structures.

Figure 11. Detection of secondary structures from an 8Å mm-cpn map (EMDB ID: 5140): (a) the original volume, (b) segmented subunits and (c/d) the detected secondary structures.

Figure 12.

Three examples of the performance of the motifs detection algorithm. (a) The four-layer sandwich fold of the complex of Aldehyde Ferredoxin Oxidoreductase protein with Molybdopterin (PDB ID: 1AOR). The red parts in figures (b) (c) and (d) show the two β -sheets. (e) The box fold of the Eukaryotic DNA polymerase processivity factor PCNA (PDB ID: 1PLQ). (f) The biggest sheet in this fold. (g)–(i) The other three sheets (red, yellow and green) with the helices (blue) are detected. (j) Four-layer sandwich of deoxyribonuclease I (PDB ID: 2DNJ). (k) Molecular surface at 12Å. (l) Motifs detection result. Top and bottom parts show the isolated helices. The middle part shows sheets. (m) Another view of the five helices at the top part.

Figure 13. Motifs detection of Procine ribonuclease inhibitor. (a) The molecule (PDB ID: 2BNH). (b) The horseshoe tertiary fold. (c) The extracted molecular surface from a 12Å cryo-EM map. (d) The molecular surface with the tertiary fold. (e/f) The detected motifs.

Table 1

The performance of our helix and sheet detection algorithm. Column 1 displays the PDB ID and the chain(s) we used. The α helix and β sheet numbers documented in the PDB files are listed in column 2. The helix and sheet numbers assigned by DSSP are listed in column 3. From the 8Å resolution blurred map, the helix and sheet numbers by our method are listed in column 4. Comparing our results with the results of DSSP, we list the number of false positive/false negative helices and sheets in column 5.

PDB ID(chain(s))	PDB values (# α , # β)	DSSP values (# H, # S)	Our values (# H, # S)	Comparison to DSSP (+/- # H, +/- # S)
1BBH _A	(4,0)	(7, 0)	(6, 0)	(0/1, 0/0)
1C3W _A	(7,1)	(7,1)	(7,1)	(0/0, 0/0)
1IRK	(9,3)	(9,3)	(10, 2)	(2/1, 0/1)
1BVP _{1,2,3}	(27,1)	(30,1)	(28,1)	(2/4, 0/0)
1TIM _A	(12,1)	(11,1)	(11,1)	(2/2, 0/0)