A Topological Characterization of Protein Structure

Bala Krishnamoorthy[1], Scott Provan[2], and Alexander Tropsha[3]

[1] Department of Mathematics
Washington State University
kbala@wsu.edu
[2] Department of Statistics and Operations Research
University of North Carolina
scott_provan@unc.edu
[3] School of Pharmacy
University of North Carolina
alex_tropsha@unc.edu

Summary. We develop an objective characterization of protein structure based entirely on the geometry of its parts. The three-dimensional alpha complex filtration of the protein represented as a union of balls (one per residue) captures all the relevant information about the geometry and topology of the molecule. The neighborhood of a strand of contiguous alpha carbon atoms along the back-bone chain is defined as a "tube" which is a sub-complex of the original complex that has been sub-divided. We then define a retraction for the tube to another complex that is guaranteed to be a 2-manifold with boundary. We capture the topology of the retracted tube by computing the most persistent connected components and holes in the entire filtration. A "motif" for a 3D structure is characterized by the number of persistent 0- and 1-cycles, and the relative persistences of these cycles in the filtration of the "tube" complex. These motifs represent non-random, recurrent, tertiary interactions between parts of the protein back-bone chain that characterize the overall structure of the protein. A basis set of 1300 motifs are identified by analyzing the alpha complex filtrations of several proteins. Any test protein is represented by the number of times each motif from the basis set occurs in it. Preliminary results from the discrimination of protein families using this representation are provided.

Key words: Protein structure, simplicial complexes, homology groups, topological persistence.

Structural Similarity Between Proteins

Understanding the similarities and differences between protein structures is central to the study of connections between the sequence, structure, and the

function of the proteins, and also for detecting possible evolutionary relationships. With the number of proteins with known structures currently exceeding 25,000 [21], and rapidly increasing by the day, the need for reliable and automated methods for structural comparison has never been greater. Various techniques for structural comparison have emerged, ranging from those which try to match the geometric coordinates of the back-bone [23], to those which use vector approximations to secondary structure elements [16, 14]. Then there are domain-based methods, which try to classify proteins based on the units of structure (or domains) that they contain. Even though there is no exact definition available, a structural domain is usually considered as a compact and semi-independent unit of a protein, which consists of a small number of contiguous segments of the peptide chain, and forms a structurally "separate" region in the whole three-dimensional structure of the protein. Widely used structural databases such as SCOP [19], and CATH [20] have been constructed using domain-based approaches. On the other hand, one of the most successful automated classifications of proteins uses concepts from knot theory to reproduce the classification provided by the CATH database with a high degree of accuracy.

This chapter is organized as follows. We review the main features of the SCOP database in Section 1. A brief description of the knot theory-based classification follows in Section 2. The drawbacks of these methods which motivated our line of research are outlined. We provide the necessary background material on alpha shapes and homology in Section 3. The definition of the neighborhood of a strand in a protein is given in detail in Section 5. We outline the algorithm used to characterize the topology of these neighborhoods in Section 6. Finally, we describe the salient features of the structural motifs that characterize these neighborhoods in Section 8.

1 The SCOP Database

The Structural Classification of Proteins (SCOP) database is a comprehensive ordering of all proteins of known structure according to their evolutionary and structural relationships. A fundamental unit of classification in this database is the protein domain. A domain is defined as an evolutionary unit observed in nature either in isolation or in more than one context in multi-domain proteins. All Protein domains are hierarchically classified into families, superfamilies, folds, and classes. The method used to construct this classification is essentially the visual inspection and comparison of structures. Any use of automatic tools in this process is aimed only at making the task manageable. The SCOP database could be considered as containing the most accurate and useful results on protein structure classification. Recent updates of the database [3] reported the introduction of integer identifiers for each node in the hierarchy (called *sunid*), and a new set of concise classification strings (called *sccs*). There is also an initiative [1] to rationalize and integrate the SCOP

information with the data about protein families housed by other prominent sequence and structural databases such as InterPro [17], CATH, and others.

The classification in SCOP is done on four hierarchical levels – family, super-family, common fold, and class. These levels embody the evolutionary and structural relationships between the domains. Proteins that have at least 30% sequence identity are classified into the same family. In addition, proteins that have lower (than 30%) sequence identity, but whose functions and structures are *very similar*, are also classified into the same family. Families, whose proteins have low sequence identities, but whose structures, and in many cases, functional features suggest that a common evolutionary origin is possible, are grouped into super-families. In the next level of hierarchy, super-families and families that have some *major* secondary structures in the *same* arrangement with the *same topological* connections are defined to a have common fold. Finally, for the convenience of users, different folds have been grouped into classes. Most folds are assigned to one of the following five structural classes based on their secondary structure composition -

1. all alpha (when the structure is mainly formed by α-helices),
2. all beta (when the structure is mainly formed by β-sheets),
3. alpha and beta (when α-helices and β-strands are largely interspersed),
4. alpha plus beta (when α-helices and β-strands are largely segregated), and
5. multi-domain (for which no homologues are known as of now).

In the latest version of SCOP (2004), the multi-domain class is further subdivided into seven classes, thus giving a total of eleven classes. The CATH database assigns to proteins a unique **C**lass, **A**rchitecture, **T**opology, and a **H**omological super-family. The methods used to achieve these assignments are similar to those employed in SCOP.

2 Knot Theory-based Classification

In 2003, Røgen and Fain [22] introduced a novel method of looking at, analyzing, and comparing protein structures that used the concepts from knot theory. The topology of a protein is captured by 30 numbers inspired by Vassiliev knot invariants. A measure for the similarity of protein shapes called the Scaled Gauss Metric (SGM) is created from these 30 numbers. The protein back-bone is analyzed as a curve in 3D space. The primary invariant calculated by the authors is the *writhing number* of the curve. This invariant essentially measures the self-linking of the curve which is the protein back-bone. The first biological applications of this measure were reported in the studies of DNA structure. It is related to the *linking number* and *twisting number* of two curves by the Călugăreanu-Fuller-White formula [24]:

$$Lk = Wr + Tw \tag{1}$$

The formula applies to a narrow closed orientable ribbon in 3D space. Here, Lk is the linking number of the two boundary curves of the ribbon, Wr is the writhing number of the central spine, and Tw is the twisting number of the two boundary curves. For a protein, the back-bone plays the role of the spine, and it is naturally oriented by the residue numbering order. Now imagine projecting the ribbon onto a 2D plane orthogonal to a randomly chosen direction. The curves defining the ribbon will seem to cross each other at certain locations in the plane of projection. Depending on the orientation of the two curve segments and the over-under relationship at each crossing, we assign a $+1$ or a -1 to the crossing. The linking number Lk counts the sum of the signed crossings between the two boundary curves, divided by two. This sum is independent of the direction of projection. The writhing number Wr counts the sum of the signed self-crossings of the ribbon's spine, now averaged over *all* projections. Finally, the twist Tw is a torsion-dependent term that measures how much one boundary curve intertwines with the other.

If we add up the unsigned individual contributions to the writhe, we obtain the (unsigned) average crossing number. A family of structural measures could be constructed using the writhe and the average crossing number as the building blocks. The authors found it sufficient to compute 30 such measures for the purpose of structural classification. Thus each protein is mapped to $\mathbb{R}^{30}$ space. Based on this mapping, the Euclidean distance between two points (or proteins) is defined as the Scaled Gauss Metric (SGM). Unlike the metrics defined by most other methods, SGM is a proper pseudo-metric - it has a zero element, it is symmetric, and most importantly, it satisfies the triangle inequality. The last property enables us to use the SGM to identify meaningful intermediate and marginal similarities, and also to distinguish between various degrees of similarity. Another desirable property of the Gauss metric is that it requires neither structural nor sequential alignment between chains, thus making the pair-wise comparison of proteins almost instantaneous. The authors used SGM to construct an automatic classification procedure for the CATH2.4 database (they essentially clustered the proteins based on the SGM). They could accurately assign more than 95% of the chains into the proper C(class), A(architecture), T(topology), and H(homologous super-family), find all new folds, and detect no false positives.

Erdmann [12] builds on the ideas of Røgen and Fain in using knot theory ideas for studying structural similarity between proteins. Supplementing the knot theory concepts with ideas from geometric convolution, the author proposes a definition of similarity based on atomic motions that preserve local back-bone topology without incurring significant errors. Similarity detection then seeks rigid body motions able to overlay pairs of substructures, each requiring a substructure-preserving motion, without necessarily requiring global structural preservation. This definition has a very broad scope - one could talk about the full rearrangement of one protein into another while preserving the global topology, or about rearrangements of sets of smaller substructures that

preserve local topology, but not the global topology, all under the same framework.

The techniques for determining structural similarity based on knot theory concepts prove to be by far the most efficient, and at the same time the most accurate method for assigning protein structure automatically. Further, the ideas presented by Erdmann could be used to develop an efficient residue-wise structural alignment scheme that might also be using the information from the structural classification. Results are awaited on this particular problem. At the same time, there are quite a few open questions that have been created by this work. There are no intuitive interpretations of the Gauss integrals except for the fundamental ones - the writhe and the average crossing number. It is also unclear how these Gauss integrals could be combined. Another question to investigate would be the relative significance of these invariants.

In spite of the amazing success in automatically classifying the CATH database with a high degree of accuracy, the Gauss integrals method has a major drawback. The problem of finding protein domains is not addressed at all. A new structure coming to SGM will not be broken into basic biologically and structurally significant pieces. From this point of view, the most desirable method for determining structural similarity would be the one that identifies protein domains using their geometric and topological properties alone, and would naturally lead to the construction of a pseudo-metric (similar to SGM), based on the definition of these domains, for measuring structural similarity. Developing a (residue-wise) structural alignment of proteins based on such a classification would be the next step. With these aims in mind, we propose the ideas for a novel characterization of tertiary structural units in proteins based on their topological and geometric properties. In the next section, we review the geometrical construction used as the framework for analyzing protein structure, and the relevant topological definitions that will be used in our analysis.

3 Alpha Shapes

An accurate representation of the protein molecule is a collection of balls, one for each atom. The equivalent picture in 2D will be the union of disks. Edelsbrunner et al. analyzed the geometry of a union of disks in 2D as early as 1983 [10]. The results for the union of balls in 3D were presented later by the author in [7]. Let B denote a finite set of balls (solid spheres) in $\mathbb{R}^3$. We specify each ball $b_i = (z_i, r_i)$ by its center $z_i \in \mathbb{R}^3$ and its radius $r_i \in \mathbb{R}$. The weighted distance of a point x from a ball b_i and is defined as the square distance from the center of the ball minus the square of the radius.

$$\pi_i(x) = \| x - z_i \|^2 - r_i^2 \tag{2}$$

The *power Voronoi cell* of a ball b_i under the power distance is the set of points that are at least as close to b_i as to any other ball in B,

$$V_i = \{x \in \mathbb{R}^3 \,|\, \pi_i(x) \leq \pi_j(x), \forall j\}. \tag{3}$$

All V_i's turn out to be convex polyhedra (see Figure 1). The dual to the power Voronoi diagram will constitute the *weighted Delaunay triangulation* of B, and is the collection of the convex hulls of the centers of those balls whose Voronoi cells have a non-empty common intersection.

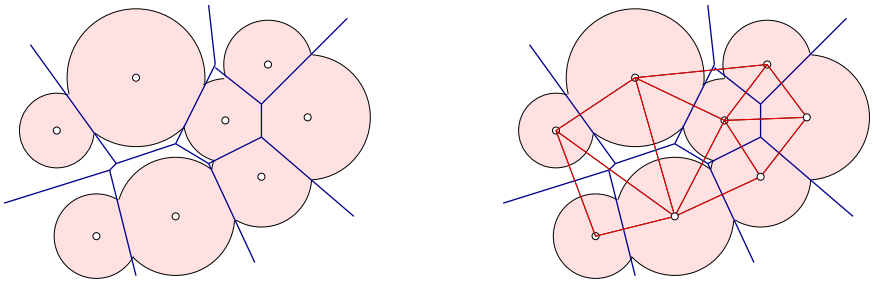

Fig. 1. Power Voronoi diagram of the disks in 2D (left) and the corresponding dual (*weighted*) Delaunay triangulation (right)

Edelsbrunner and Mücke [11] generalized the construction of the Delaunay triangulation given above to consider the dual of the power Voronoi diagram restricted to within the union of the defining balls. The Voronoi cells (3) decompose $\bigcup B$ into convex cells $R_i = \bigcup B \cap V_i = b_i \cap V_i$. The *dual complex* records the non-empty common intersection of these cells,

$$K = \{\sigma_\Lambda \,|\, \bigcap_{i \in \Lambda} R_i \neq \emptyset\}, \tag{4}$$

where Λ is a subset of the index set, and σ_Λ is the convex hull of the centers of the balls with index in Λ. Equivalently, $\sigma_\Lambda \in K$ is the common intersection of the Voronoi cells that have a non-empty intersection with the union of the balls. The *underlying space* is the set of points contained in the simplices of K, and is denoted by $|K|$. In this context, the authors refer to the underlying space as the *dual shape* of B. The concept is illustrated in 2D in Figure 2. Notable is the special case where the balls have non-empty pair-wise intersections, but have no (non-empty) triple-wise intersections. In this case, K looks like the familiar ball-and-stick diagram of a molecule. Each stick (which originally represents a covalent bond in the molecule) represents the geometric overlap between two balls.

Now consider growing the balls continuously in time and studying how their union changes. We set the weight of each ball b_i as $r_i^2 + t$ at time t and let t go from $-\infty$ to $+\infty$. Each b_i has zero weight at $t = -r_i^2$ and negative weight, and hence imaginary radius, before that time. By construction, the Voronoi cells of the balls remain unchanged. It follows that the dual complexes that

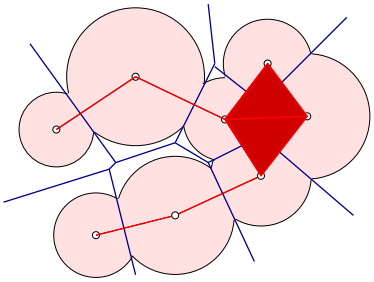

Fig. 2. The dual complex for the union of disks. The nine edges correspond to pairwise intersections and the two triangles correspond to the triple-wise intersections of the clipped Voronoi cells of the balls

arise throughout time are sub-complexes of the same Delaunay triangulation. Also, the dual complexes can only get larger in time. We use the square root $\alpha = \sqrt{t}$, as the index for time varying sets. Under this convention, with $r_i = 0$ (i.e. the ball is originally a point), the radius of the ball b_i at time t is α. Denote by B_α the collection of balls and K_α the dual complex of B_α at time $t \in \mathbb{R}$, indexed by α. We refer to K_α as the α-complex and its underlying space as the α-shape of B. For small enough (large enough negative) time, all radii are imaginary, and $\bigcup B_\alpha = \emptyset$. And for large enough time, the dual complex of B_α is equal to the Delaunay triangulation. We thus obtain a sequence of complexes that begins with the empty complex and ends with the Delaunay triangulation, $\emptyset \subseteq K_\alpha \subseteq K_\beta \subseteq D$, for every $-\infty < \alpha^2 < \beta^2 \leq +\infty$. Since there are only finitely many simplices, there are only finitely many sub-complexes of D that arise as dual complexes during the growth process. We refer to this sequence as a *filtration* of the Delaunay triangulation, $\emptyset = K^1 \subseteq \ldots \subseteq K^m = D$. We illustrate the construction by showing three complexes in the filtration of the union of the disks in the plane in Figure 3. We define a function $j(\alpha^2)$ such that $K_\alpha = K^i$ if $i = j(\alpha^2)$ in order to translate between continuous and discrete rank.

The Delaunay simplices can be sorted in the order in which they enter the dual complex. Define the *birth time* of a simplex $\sigma \in D$ as the minimum time $t = \alpha_\sigma^2$ such that $\sigma \in K_\alpha$ for all $\alpha^2 \geq t$. Thus the difference between two contiguous complexes in the filtration consists of all simplices whose birth-time coincides with the creation of the second complex,

$$K^{i+1} - K^i = \{\sigma \in D \,|\, \alpha_\sigma^2 = j^{-1}(i+1)\} \tag{5}$$

We represent the filtration by sorting the Delaunay simplices by birth time, and in case of a tie by dimension. Remaining ties are broken arbitrarily. Every dual complex K^i is a prefix of this ordering. Due to the tie breaking rule, every such prefix is a complex, even if does not coincide with a dual complex. This property of the ordering will be crucial for the algorithm that we will use to compute the connectivity of K^i.

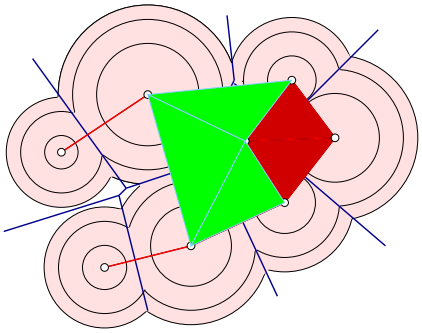

Fig. 3. Three unions of disks and the corresponding dual complexes from the filtration. The first complex consists of only the vertices (as all the balls are disjoint). The second complex is shown in red (same as the one shown in Figure 2). The simplices shown in green get added in the third complex.

4 Homology Groups

We will use homology groups as an algebraic means to study the connectivity of a topological space. The overview of the main concepts presented here follow mainly the treatment given in [9, §IV.2]. Chapter 4 in [13] presents an easy to read discussion of the same subject. Roughly speaking, for any given simplicial complex K, there is one group denoted by $H_p(K)$ in each dimension p with $0 \leq p \leq \dim K$, which measures the number of "independent p-dimensional holes" in K.

We will call a set of k-simplices a k-chain. By definition, the *sum* of two k-chains is the symmetric difference of the two sets.

$$c + d = (c \cup d) - (c \cap d)$$

We define the boundary of a simplex σ as $\partial\sigma = \{\tau \leq \sigma \mid \dim \tau = \dim \sigma - 1\}$. The *boundary* of chain is the sum of the boundaries of its simplices, $\partial c = \sum_{\sigma \in c} \partial\sigma$. Two types of chains are particularly important for us: the ones without boundary and the ones that bound. A k-cycle is a k-chain c with $\partial c = 0$. A *k-boundary* is a k-chain c for which there exists a $(k+1)$-chain d with $\partial d = c$. C_k is the set of k-chains and $(\mathsf{C}_k, +)$ is the group of k-chains. The zero of this chain group is the empty set. Let Z_k and B_k be the set of k-cycles and the set of k-boundaries respectively. Then $(\mathsf{Z}_k, +)$ is a subgroup of $(\mathsf{C}_k, +)$, and $(\mathsf{B}_k, +)$ is a subgroup of $(\mathsf{Z}_k, +)$.

The *k-th homology group* is the quotient of the k-th cycle group divided by the k-th boundary group, $\mathsf{H}_k = \mathsf{Z}_k | \mathsf{B}_k$. The size of H_k is a measure of how many k-cycles are not k-boundaries. If $\mathsf{Z}_k = \mathsf{B}_k$, then H_k is the trivial group consisting of only one element. Two k-cycles are homologous if they belong to the same homology class, $c \sim d$ if $c + d \in \mathsf{B}_k$. Equivalently, $c \sim d$ if there exists $e \in \mathsf{Z}_{k+1}$ with $d = c + \partial e$.

The most useful parameters associated with the homology groups are their ranks, which have intuitive interpretations in terms of the connectivity of the space. Given a subset S of a group G, the subgroup called the linear hull of S ($\operatorname{lin} S$) consists of all $\sum a_i x_i$, with $x_i \in S$ and $a_i \in \{0, 1\}$. A *basis* is a minimal subset S that generates the entire group, i.e. $\operatorname{lin} S = \mathsf{G}$. The *rank* of G is the cardinality of a basis. If the group is the k-th homology group of a space, $\mathsf{G} = \mathsf{H}_k$, the rank is known as the *k-th Betti number* of that space, and is denoted by $\beta_k = \operatorname{rank} \mathsf{H}_k$. Since $\mathsf{H}_k = \mathsf{Z}_k | \mathsf{B}_k$, we have

$$\operatorname{rank} \mathsf{H}_k = \operatorname{rank} \mathsf{Z}_k - \operatorname{rank} \mathsf{B}_k \tag{6}$$

In general, the 0-th Betti number (β_0) is the number of connected components. Similarly, β_1 gives the number of independent tunnels, and β_2 gives the number of independent (enclosed) voids in the space. For example, consider a torus. There is a single connected component, and hence $\beta_0 = 1$. There are two independent tunnels - one running inside the torus, and the other one is the hole in the middle. Hence $\beta_1 = 2$. There is only one independent closed void, and hence $\beta_2 = 1$. For the 2-sphere, the Betti numbers are $\beta_0 = 1, \beta_1 = 0$, and $\beta_2 = 1$. All higher Betti numbers are zero in both cases.

4.1 Persistent Homology Groups

We will study simplicial complexes for which $\beta_i = 0$ for $i \geq 2$ (details to follow in Section 5). Ideally, we would like to identify the most significant topological features of such a complex - the biggest connected components and the largest holes. At intermediate levels of growth, we would like to identify those features that are persistent – i.e. it takes a long time for them to disappear once they appear. Edelsbrunner, Letscher, and Zomorodian have formalized a notion of topological simplification within the framework of a filtration of the complex [6]. They defined the *persistence* of a non-bounding cycle as a measure of its life-time in the filtration. For each non-bounding cycle, they identify two simplices σ^i and σ^j that respectively *create* and *destroy* the non-bounding cycle in the face of the filtration. Then, the persistence of this feature is defined as $j - i - 1$. For a simplicial complex K^ℓ, the *p-persistent k-th homology group* is defined as

$$\mathsf{H}_k^{\ell,p} = \mathsf{Z}_k^\ell \,|\, (\mathsf{B}_k^{\ell+p} \cap \mathsf{Z}_k^\ell). \tag{7}$$

The p-persistent k-th Betti number of K^ℓ is the rank of this homology group - $\beta_k^{\ell,p} = \operatorname{rank} \mathsf{H}_k^{\ell,p}$.

To measure the life-time of a non-bounding cycle, we find when its homology class is created and when its class merges with the boundary group. We will defer the description of the method for computing the topological persistences till later. Now we address the main issue facing us – how to define a simplicial complex that captures the geometry of three dimensional structural units of proteins, such that we could use the tools described above to characterize its topology?

5 Definition of Neighborhood

The series of α-complexes in the filtration of the Delaunay triangulation of the protein carries all the information about the geometry of the molecule. Using the filtration as a framework, we can analyze structural units or parts that could be characterized based on their topology, hence leading to the definition of domains (or *structural motifs*). To simplify the treatment, we consider one C_α atom[4] per residue instead of looking at the all-atom model. The local structure in proteins is captured by defining the neighborhood of each C_α atom and each consecutive C_α-C_α edge as the *links* of the respective simplices in the α-complex of the protein at any α-level. We will need certain definitions to achieve this task. The notation used is the same as that given in [8].
Let K be a simplicial complex. The *closure* of a subset $L \subseteq K$ is the smallest sub-complex that contains L.

$$Cl\ L = \{\tau \in K \mid \tau \leq \sigma \in L\} \tag{8}$$

The star of a simplex τ consists of all simplices that contain τ, and the link consists of all faces of simplices in the star that do not intersect τ.

$$St\ \tau = \{\sigma \in K \mid \tau \leq \sigma\} \tag{9}$$

$$Lk\ \tau = \{\sigma \in Cl\ St\ \tau \mid \sigma \cap \tau = \emptyset\} \tag{10}$$

The star is generally not closed, but the link is always a simplicial complex. Given any α-complex K^i, we define the link of each vertex (or C_α atom) and each back-bone edge as follows [13, pg. 111].

$$\begin{aligned} Lk(v_0) = \{v_1 \mid (v_0v_1) \in K^i\}\ \cup\ \{(v_1v_2) \mid (v_0v_1v_2) \in K^i\}\ \cup \\ \{(v_1v_2v_3) \mid (v_0v_1v_2v_3) \in K^i\} \end{aligned} \tag{11}$$

$$Lk(v_0v_1) = \{v_2 \mid (v_0v_1v_2) \in K^i\} \cup \{(v_2v_3) \mid (v_0v_1v_2v_3) \in K^i\} \tag{12}$$

In words, the link of a C_α atom consists of all other C_α atoms that form an edge with it, and all other C_α-C_α edges (not necessarily consecutive) that form a triangle with it, and all other triangles of three C_α atoms that form a tetrahedron with it. The link of a back-bone edge can be interpreted similarly. In Figure 4, We illustrate these definitions in two dimensions.

Naturally, the links defined above will *grow* as the α-complex grows. We can study the connectivity of the links of C_α atoms (and C_α-C_α edges) by finding the homology groups of the links, and observe how the connectivity changes with the growth of the α-complex. We mention here that the ranks of the homology groups of the C_α and C_α-C_α links show specific patterns of variation when we run down an α-helix or a strand in a β-sheet. Such patterns can be used to characterize specific structural domains. The drawback

[4] The reader should be careful not to confuse the α used in the context of an alpha carbon atom, with the α used in the context of an α-complex.

with this approach is that the vertex and edge links provide only "local" information. We now describe how we combine a series of back-bone C_α links and C_α-C_α links to effectively capture the neighborhood of a strand, thus providing important non-local information.

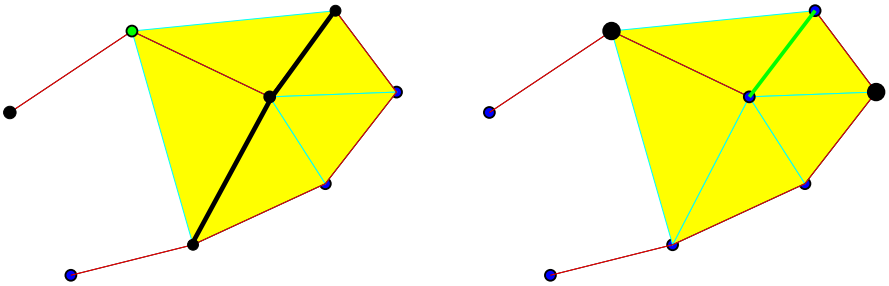

Fig. 4. Alpha-complex in 2D with the back-bone shown in red. Link (shown in black) of a residue shown in green (left figure), and that of a back-bone edge shown in green (right figure).

5.1 Link of a back-bone strand

We denote a contiguous strand of back-bone residues and the intermediate edges by $\mathcal{S}$. Formally, a strand of n residues (C_α atoms) and $n-1$ back-bone edges is defined as the sequence of vertices and edges given by $\mathcal{S} = \{v_1, e_1, v_2, \ldots, v_n\}$, where $v_{j+1} = v_j + 1$ and $e_j = (v_j, v_{j+1})$ for $1 \leq j < n$. The *link* (or the *boundary of neighborhood* to be exact) of such a strand in an α-complex K^i is defined as follows.

$$Lk(\mathcal{S}) = \left(\bigcup_{v \in \mathcal{S}} Lk(v) \right) \setminus \left(\bigcup_{v \in \mathcal{S}} St\, v \right) \tag{13}$$

where $Lk(v)$ and Stv are as defined in (11) and (9) respectively. By construction, the union of the links of all the vertices in $\mathcal{S}$ will include the links of the back-bone edges connecting them. The aim of defining the link of a strand in this way is to capture the non-trivial interactions of the strand with other parts of the protein. This is also the reason why we remove the elements of the star of each vertex in $\mathcal{S}$ from the union in (13). In practice, we are in fact even more careful in removing such "trivial" contacts (or interactions). For the strand $\mathcal{S}$ as defined above, it is natural to expect v_0 to be included in the link of $\mathcal{S}$, as the back-bone edge $e_0 = (v_0, v_1)$ will be part of the α-complex, for sufficiently large α. This observation follows from the fact that the consecutive C_α-C_α edges are among the smallest edges that appear in the Delaunay tessellation of the whole protein. Similarly, one would expect

v_{n+1} also to be included in $Lk(\mathcal{S})$. Hence, we usually modify the link of $\mathcal{S}$ as follows: $Lk(\mathcal{S}) = Lk(\mathcal{S}) \setminus \mathcal{S}'$, where $\mathcal{S}' = \{v_0, e_0, e_{n+1}, v_{n+1}\}$. We illustrate the link of a strand in two dimensions in Figure 5.

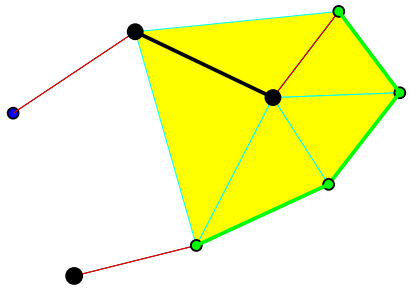

Fig. 5. Link (shown in black) of a back-bone strand (shown in green), consisting of four residues and the three intermediate edges, in the dual complex of the union of disks in 2D. Note that the link shown here is the one before we remove $\mathcal{S}'$ from the union.

The link $Lk(\mathcal{S})$ is defined for each α-complex K^i. By construction, $Lk(\mathcal{S})$ will itself be a simplicial complex. We denote the link of $\mathcal{S}$ defined for K^i by $Lk^i(\mathcal{S})$. One can naturally consider a *filtration* of the link of $\mathcal{S}$ defined for the final Delaunay triangulation D (denoted by $Lk^D(\mathcal{S})$), in the form $\emptyset = Lk^1(\mathcal{S}) \subseteq \cdots \subseteq Lk^m(\mathcal{S}) = Lk^D(\mathcal{S})$. Now we can observe the changes in the connectivity of $Lk^i(\mathcal{S})$ as the complex grows. Specific patterns in the connectivity of $Lk(\mathcal{S})$ as a function of growth (α) could be used to characterize various structural domains.

5.2 From link to "tube"

The definition of the link of a strand given in (13) efficiently captures all tertiary interactions made by the strand with other parts of the protein. The topology of the link will indeed be characteristic of these interactions. Nevertheless, there is a drawback with with this definition. Consider a strand that forms little or no contacts with the rest of the protein. The link defined in (13) will possibly be empty in this case. Isolated regions in the protein molecule where the strand merely bends on itself could typically lead to such a situation. The illustration in 2D shown in Figure 5 in fact displays such a strand. The strand of four residues (and the three intermediate back-bone edges) appears to be bending on itself. This information is not provided by the link (shown in black), as the interactions that define the link are located only at the ends of the strand.

We propose the idea of defining the neighborhood of the strand in the form of a "tube" around it. Imagine a tube that has the back-bone chain running through its center. If we thicken the tube uniformly, the parts of the tube around the strand in question would come into contact with the surface of the

tube around other parts depending on the interactions between the strand and the other parts of the protein. In the case where the strand does not form any interactions with other parts, the tube around it will also be isolated. At the same time, the tube would touch itself if the strand bends on itself, thus allowing us to characterize domains of this type. In order to implement this idea, we need to modify the representation of the simplicial complex. Consider a residue v' that is present in the link of the strand because it forms an edge $e = (v, v')$ with a vertex v in the strand. Instead of adding v' to the link, we now introduce a new vertex at the mid-point of the edge e, and add the new point to the link of the strand. We can extend this idea to simplices of higher dimensions too. In a way, we are "shrinking" the original link towards the strand to define the "tube" around it. Since the new vertex added is not in $\mathcal{S}$ or $\mathcal{S}'$, the tube might be non-empty even when the link is possibly empty due to $v' \in \mathcal{S}$ or $v' \in \mathcal{S}'$.

Formally, we perform a *barycentric subdivision* of the original alpha complex K^i. As the name suggests, we subdivide every edge in the middle. Every triangle is divided into six smaller triangles by drawing the medians. The division of tetrahedra can be understood in a similar fashion. This construction is used in the classification of closed surfaces [13, Chap. 5]. In general, given a simplicial complex $K \subset \mathbb{R}^n$, a *subdivision* of K is a simplicial complex $K^1 \subset \mathbb{R}^n$ with the property that $|K^1| = |K|$, and given $\sigma_1 \in K^1$, there exists $\sigma \in K$ such that $\sigma_1 \subset \sigma$. Thus, the simplices of K^1 are contained in the simplices of K, but K^1 and K triangulate the same subset of $\mathbb{R}^n$. The barycentric subdivision is one type of a subdivision. The barycentric subdivision of a simplex σ could also be defined as the complex obtained by adding the barycenter (or centroid) of σ as a new vertex, and connecting it to the simplices in the barycentric subdivision of the faces [9, See Exercise II.8]. Since we are working with proteins, consecutive C_α atoms in the chain form the closest interactions. All other interactions, including the tertiary contacts that we are trying to characterize, will appear in the α-complex only after the back-bone edges. Hence, we modify the generic barycentric subdivision so that the back-bone edges are not subdivided. Figure 6 illustrates the proposed barycentric subdivision for proteins in 2D.

Given the barycentrically subdivided complex K^i_{BS} of the α-complex K^i, we apply the definition of the link of strand $\mathcal{S}$ given by (13) on K^i_{BS}. The definitions of vertex and edge links (11) and (12) are also applied on K^i_{BS}. The link that results from this procedure constitutes the "tube" of $\mathcal{S}$. As illustrated in Figure 7, the tube carries all the required information, and will not be empty like the link defined earlier. As we had seen earlier, a filtration of the final tube around $\mathcal{S}$ can be maintained at each index of growth α. We can now study the topological connectivity of the tube around the strand as the complex grows. Patterns observed hence could be used to objectively define tertiary structural domains in proteins.

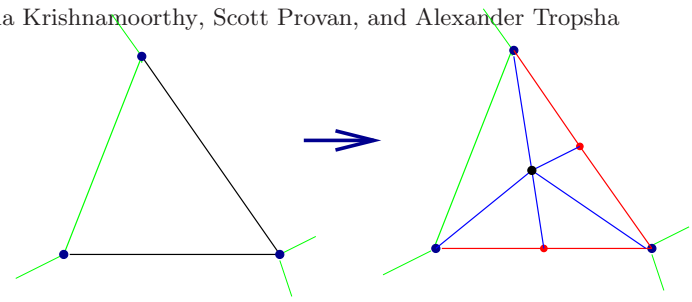

Fig. 6. Barycentric subdivision of a triangle. The edge on the left (in green) is a back-bone edge and hence is not subdivided. One new point is added in the middle of the other two edges, and two new edges are added from this new point to the original vertices in each case. A third point is added at the barycenter of the original triangle, and five new edges connecting this central point to the other points are added. Finally, the five smaller triangles are added in the interior.

5.3 A Retracted tube

A final step of geometric modification needs to be performed on the tube. Once again, strands that bend on themselves motivate the proposed change. As we have seen, the tube around the strand might touch itself at places. In this process, one or more simplices in the tube complex get identified with certain others. It is desirable if we could actually create a copy of any such simplex, and pull the copy just away from the original simplex, such that tube is not self-intersecting any more. The critical point in performing such a duplication is that we do not desire to alter the (topological) connectivity of the tube in this process. If we could achieve this goal, we would be left with an object that is topologically much nicer to handle than the original tube, but at the same time, it carries the exact connectivity information as before.

We achieve this goal by *retracting* the original tube closer towards the strand $\mathcal{S}$ in the following way: every vertex in the tube that forms an edge with a residue in $\mathcal{S}$ is retracted half-way towards the residue. This step automatically retracts every triangle in the tube that forms a tetrahedron with a

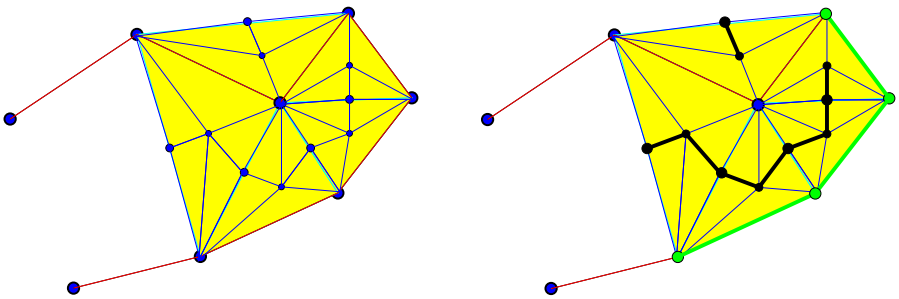

Fig. 7. Barycentric subdivision of the original complex (left) and the "tube" (shown in black) of the strand (in green), now defined on the subdivided complex.

residue in $\mathcal{S}$. The case of an edge in the tube that forms a tetrahedron with a back-bone edge in $\mathcal{S}$ is a little tricky. Retracting the former edge half-way towards the edge in the strand will generate a trapezium. We subdivide this trapezium into two triangles to make sure that the complex is triangulated. By construction, it can be seen that this trapezium will lie in a plane, and hence it can be triangulated by adding either one of its two diagonals. The advantages of this transformation are illustrated in Figure 9.

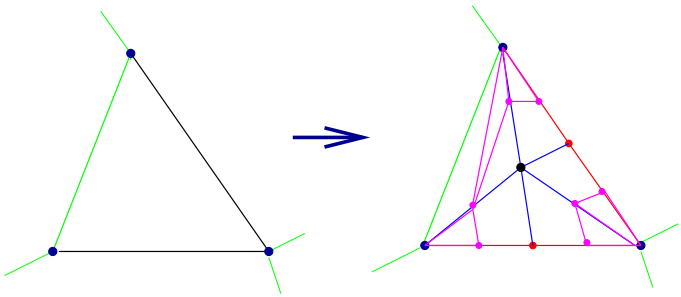

Fig. 8. Barycentric subdivision modified so that the retraction can be defined. Notice that all edges that have one vertex (end-point) as one of the residues are subdivided. Additional edges are added to triangulate any trapezium that gets added, as it happens here near the back-bone edge on the left.

With this additional simplification, the tube will always be a 2-manifold (or a surface) with boundary [15, §22], or a collection of disjoint 2-manifolds with boundary. In addition, the transformation of the tube described above can be shown to be a *strong deformation retraction* [18, §55]. These properties will be important for the method that we will employ in Sections 6 to calculate the ranks of homology groups as well as their topological persistences.

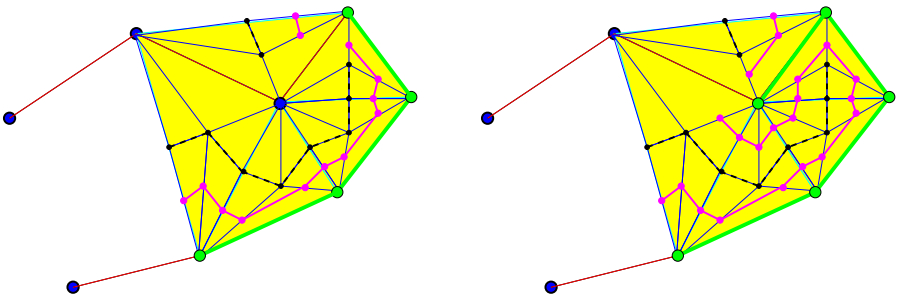

Fig. 9. Retracted "tube" of the strand with four residues (left), and the tube for the strand if we add the residue in the middle of the complex to the strand (right), both shown in magenta. The links defined earlier without retraction will be the same for both the strands (shown in black dotted lines).

6 Computing the Persistences

We now turn our attention towards the identification of the topologically persistent features of the tube complex. Our task is to pair the positive and negative simplices in the filtration of the tube complex such that each pair represents the life-time of a non-bounding cycle. For 0-cycles, we achieve this pairing while maintaining the UNION-FIND data structure [4, Chap. 22]. When a negative edge comes into the filtration, we pair it with the younger vertex (of the two) if that is unpaired yet. If not, we pair the edge with that vertex in the set to which the younger vertex belongs to, which is the oldest (i,e. has the lowest α-rank). At any point, each set is identified by the oldest unpaired vertex in that set.

The construction of the tube ensures that *all* triangles in the tube complex are negative. Also, the tube complex will be a surface (2-manifold) with boundary, or a set of disjoint surfaces with boundary. Under these conditions, we can make use of the dual relationships between the negative triangles and the positive edges that create the 1-cycles. We maintain a UNION-FIND for β_1's in the following way. The filtration is traversed in the reverse order of time. For each (negative) face, we add a SINGLE SET. Then, for each positive edge that comes in, we do the pairing just as in the case of β_0's, but taking care of the fact that the time scale is reversed now.

Edelsbrunner et al. [6] achieve the pairings in all dimensions simultaneously by performing a cycle search algorithm on a linear array, which acts similar to a hash table [4, Chap. 12]. This algorithm works for any simplicial complex (need not necessarily be a surface with boundary) and has a running time of at most $O(m^3)$, where m is the total number of sub-complexes in the filtration (or the maximum α-rank). The authors do suggest elsewhere, though, that the pairings can be achieved in near-linear time for the simpler case of surfaces, by appropriately modifying the incremental algorithm [5] for calculating Betti numbers. They have also provided the idea of using the dual graph to label faces when running the incremental algorithm. Borrowing these ideas, we achieve the pairings in almost constant time by using weighted merging for the union and path compression for find. Under these conditions, the amortized time per operation is $O(A^{-1}(m))$, where $A^{-1}(m)$ is the inverse of the Ackermann function, which grows very slowly [4, Chap. 22].

Care needs to be taken when treating the positive boundary edges while performing the union-find for the dual graph. We add a dummy dual vertex to represent the external space and assign it the dual rank of zero. A positive boundary edge in the original complex will create a dual edge that connects the dual vertex corresponding to the triangle bounded by the positive edge and the dummy external vertex. For a few of these dual edges, it will happen that the dual vertex (corresponding to the triangle) as well as the external dummy vertex will both have been paired already. We treat such edges as special cases, and record a pairing between the dual edge and the external

vertex. We call such a pairing a *forced pairing*. The number of forced pairings will be equal to the β_1 of the fully grown complex under study.

We present the details of our pairing algorithm on the following page, which is essentially a forward and a backward run of a modified incremental algorithm [5]. We assume the sequence of simplices σ^i for $0 \leq i < m$ is a filter, and the sequence of sub-complexes $K^i = \{\sigma^j | 0 \leq j \leq i\}$, for $0 \leq i < m$, is the corresponding filtration. Except in the case of forced pairings, each set will have one yet unpaired vertex, which will be the oldest in the set. The set is represented by this vertex, and the rank of the set (denoted by $r(U)$ for set U) will be the rank of this vertex. We maintain two lists of pairings - $\mathcal{P}_0$ and $\mathcal{P}_1$, for β_0 and β_1 pairs respectively. The UNION-FIND data structure supports three operations:

FIND(u): return the representative vertex of the set that contains vertex u.

UNION(u, v): substitute $U \cup V$ for U and V (represented by u and v); $r(U \cup V) = \min\{r(U), r(V)\}$.

ADD(u): add $\{u\}$ as a new singleton set

Once we have the lists of paired simplices, we can calculate the *relative persistence* of the feature represented by each pair (σ^i, σ^j) as $\lambda_{ij} = (j - i - 1)/(m - i_0 - 1)$, where m is the maximum number of α-ranks, and i_0 is the rank at which the first simplex entered the tube complex. This measure evaluates the relative life-time of each feature as compared to the entire life-time of the tube complex. We list the relative persistences of β_0's and β_1's in descending order. Choosing a cut-off value for each of these sets of relative persistences, we will be left with a fixed small number N_λ^0 of λ_{ij}^0's corresponding to the most persistent β_0's, and N_λ^1 of λ_{ij}^1's corresponding to the most persistent β_1's. We present the *persistence signature* of the structural motif represented by the tube complex of the strand $\mathcal{S}$ in question as

$$\text{Sign}(\mathcal{S}) \;=\; \{N_{\beta_0}, N_{\beta_1}; \lambda_1^0, \ldots, \lambda_{N_{\beta_0}}^0; \lambda_1^1, \ldots, \lambda_{N_{\beta_1}}^1\} \tag{14}$$

7 A Basis Set of Motifs

In order to capture all the non-local neighborhoods in a protein, we define the tube complex for a series of strands along the back-bone chain $\mathcal{S}_i$ for $i = 1, 2, \ldots$, each of length $|\mathcal{S}_i| = L$. Here, $\mathcal{S}_i = \{v_i, e_i, v_{i+1}, \ldots, v_{i+L-1}\}$. In other words, we slide a window of contiguous residues (of length L) along the back-bone chain, and study the neighborhood as defined by the tube complex for each strand. The lengths of $L = 8$ and $L = 15$ were chosen. The idea was to capture short-range as well as relatively long-range motifs. As we are going to see, the diversity of the basis motifs is higher for a higher value of L. The particular values were chosen after observing several protein structures for structural units.

Algorithm 1 Pairing Algorithm

list^2 **PAIRING**()
$\mathcal{P}_0 = \mathcal{P}_1 = \emptyset$;
for $i = 0$ to $m - 1$ do
    case σ^i is a vertex u:
        ADD(u); $r(\{u\}) = i$;
    case σ^i is an edge uv:
        $u_r = \text{FIND}(u)$; $v_r = \text{FIND}(v)$;
        if $u_r \neq v_r$
            mark σ^i as *negative*
            UNION(u_r, v_r); $\mathcal{P}_0 = \mathcal{P}_0 \cup (\arg\max\{r(u_r), r(v_r)\}, \sigma^i)$;
        else
            mark σ^i as *positive*
        endif
endfor
ADD(e); $r(\{e\}) = 0$;
for $i = m - 1$ to 0 do
    case σ^i is a triangle (dual vertex u):
        ADD(u); $r(\{u\}) = m - i$;
    case σ^i is a *positive* edge (dual edge uv, dual rank $m - i$):
        $u_r = \text{FIND}(u)$; $v_r = \text{FIND}(v)$;
        if $u_r \neq v_r$
            UNION(u_r, v_r); $\mathcal{P}_1 = \mathcal{P}_1 \cup (\arg\max\{r(u_r), r(v_r)\}, \sigma^i)$;
        else
            $\mathcal{P}_1 = \mathcal{P}_1 \cup (e, \sigma^i)$;
        endif
endfor
return($\mathcal{P}_0, \mathcal{P}_1$);

For every strand $\mathcal{S}$, we derive the persistence signature (14). The cut-off values for β_0 and β_1 persistences were chosen as $\lambda_0 = 0.43$, $\lambda_1 = 0.37$ for $L = 8$, and $\lambda_0 = 0.42$, $\lambda_1 = 0.35$ for $L = 15$. We initially observed all the relative persistences for each motif in several protein chains. The cut-offs were picked so that most significant topological features would be included in the motif, and at the same time, the total number of motifs to consider would not be too large. A diverse set of 1143 protein chains was selected. In the first run, we identified all possible recurrent motifs. A candidate motif from one of the chains was compared to all the motifs already observed (as maintained in the set of motifs) with the same number of persistent β_0 and β_1 (as denoted by N_{β_0} and N_{β_1} in the signature (14)). If each relative persistence component was within an interval of 0.12 centered at the corresponding component of one of the motifs in the set, the candidate motif was counted as an instance of the particular motif. If not, the candidate motif was added to the set as a new motif. The relative persistences for each motif was averaged over all the

instances of the same. Care was also taken to ensure that adjacent similar motifs were not double counted. In several cases, the neighborhood of the strand changes very little when we slide it by one or two residues. Thus we obtain repeated occurrences of the same motif, which in fact should be counted as a single motif that is actually longer than 15 residues. Hence we count a repeated occurrence of the same motif only if we slide the strand by at least 3 residues.

We chose a lower cut-off of 5 for the number of occurrences of a motif in the whole set of protein chains in order to include it the basis set. In case no motif with a particular $(N_{\beta_0}, N_{\beta_1})$ occurred at least 5 times, we grouped all of them with the most frequent one among them (and averaged the relative persistences). After these simplifications, we obtained a basis motif set of 361 motifs for $L = 8$ and 938 motifs for $L = 15$. We discuss the salient features of these structural motifs in the following section.

8 Features of Structural Motifs

Conventionally, one looks at the protein as being made of local units such as alpha helices and strands from beta sheets combined together in 3D arrangements. Typical units of such combinations are distinguished and given names such as a helix bundle, coiled-coil, or alpha-beta barrel. In our analysis, the motifs characterize arrangements in the neighborhood of such a unit (strand). Hence, a 15-residue portion of an alpha helix will form different motifs depending on how the remaining parts of the protein are arranged around it. The values of N_{β_0} and N_{β_1} range from 0 up to 5 for the case of $L = 8$, and from 0 up to 7 for $L = 15$. The two basis sets of motifs cover almost all possible 3D arrangements of strands of the respective lengths in proteins.

Since β_0 measures the number of connected components (section 4), it is straightforward to interpret N_{β_0} can as the number of most prominent interactions that the strand makes with the other parts of the protein. N_{β_1} gives the number of most prominent holes in the filtration of the tube complex, but it is not as easy to see how these persistent holes are created. When there are two or more adjacent dominant contacts (or interactions) of the strand with other parts, space in between two such contacts typically gives rise to a persistent hole in the tube complex. The strand bending on itself usually gives rise to holes - for example, a single turn of an alpha helix creates a hole in the tube complex. At the same time, depending on other tertiary interactions, the holes created due to the strand making helix turns might not be highly persistent. In other words, if we look at an isolated helical strand (which does not interact with any other parts of the protein), the significant holes in the tube complex are the ones due to the helix turns. We illustrate these observations by examining a few motifs in detail.

In Figure 10, we present instances of two $\{2, 2\}$ motifs. For the instance of motif 274, the strand appears to make two helix turns and then starts

to bend on itself. There is significant interaction with two helical regions and another strand lying around it. On the other hand, there seems to be two prominent interactions made by the strand in the instance of motif 287 with the beta strands on either side of it. The holes in this case are generated due to the strand bending on itself at one end. One would expect the holes in the latter case (of motif 287) to be less persistent than those in the former case, and the β_0 persistences to follow the opposite trend. The sets of persistences show these relationships clearly – for motif 274, they are {0.5096 0.4572}{0.4899 0.4060}, and for motif 287, they are {0.6655 0.4490}{0.3946 0.3681}.

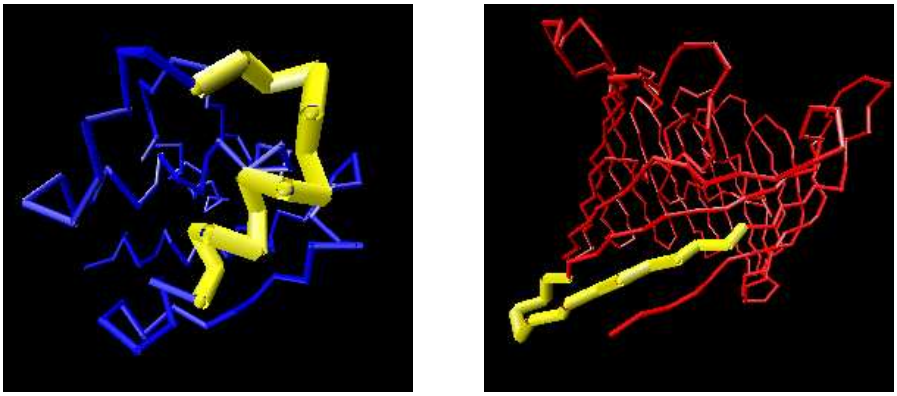

Fig. 10. Instances of $L15$ motifs 274 (in 1BKR) on the left and 287 (in 3PRN) on the right, both with $N_{\beta_0} = 2$, $N_{\beta_1} = 2$.

It typically takes a lot of structure to produce motifs with high N_{β_0} and N_{β_1}. In the same line, strands that have limited interaction with other parts usually give rise to smaller numbers. A straight strand (as opposed to a helical one) lying on the outside of a protein (thus forming limited contact with the rest of the protein) forms motifs with low N_{β_0} and N_{β_1}, as illustrated by the instance of the $L15$ motif 27 shown in Figure 11. The strand interacts with itself and does not produce any other significant contacts, thus producing a $\{1, 0\}$ motif. Similarly, the instance of $L8$ motif 162 shown in the figure produces loose interacts with three other parts of the protein, thus providing a $\{3, 0\}$ motif.

The instance of the $L15$ motif 891 shown in Figure 12 depicts the strand in the middle of several other portions of the protein, thus forming persistent interactions and holes to give a high-numbered motif ($\{5, 3\}$). A very interesting high-numbered $L15$ motif instance is also shown here - that of the $\{7, 0\}$ motif 863. The strand appears to lie far from the rest of the protein and seems to forms very little interaction. In fact, the tube complex will not become significant until the original alpha complex (of the entire protein) is

grown sufficiently. The long range interactions appear at a later point of time (in the filtration). Since the lifetime of the tube itself is short, we get large relative persistences. One could guess that there is no chance of a hole in the tube here as all the interaction lies to one side of the strand. The end result is an instance of a $\{7, 0\}$ motif.

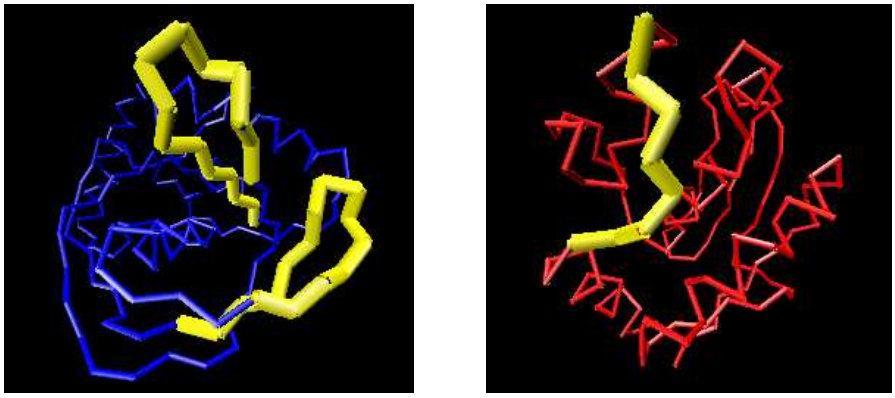

Fig. 11. Instances of $L15$ motif 27 (in 1ICJ) with $N_{\beta_0} = 1$, $N_{\beta_1} = 0$ on the left and $L8$ motif 162 (in 1KUH) with $N_{\beta_0} = 3$, $N_{\beta_1} = 0$ on the right.

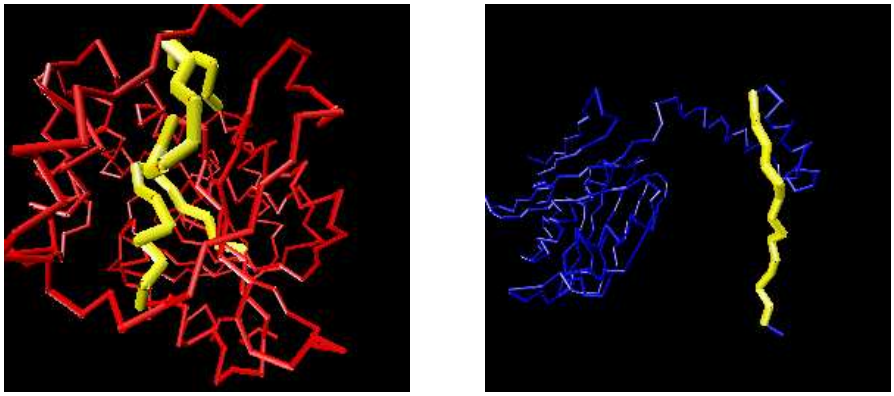

Fig. 12. Instances of $L15$ motif 891 (in 1EDE) with $N_{\beta_0} = 5$, $N_{\beta_1} = 3$ on the left and $L15$ motif 863 (in 1TMY) with $N_{\beta_0} = 7$, $N_{\beta_1} = 0$ on the right.

We have analyzed instances of all the motifs in detail. It is necessary to view the motifs in 3D so that we could rotate the protein and clearly see the different interactions involved. Visualizations are obtained using the software package called VMD. The motifs corresponding to several popular tertiary structural units (such as helix bundle) have been identified. Given the choice

of the relative persistence cut-offs and the length of the strand, we believe that these motifs provide a rigorous characterization of parts of proteins using their geometry and topology.

9 The Next Step - Classification

The structural characterization developed by us will prove useful to the biologist only when she could use the same to efficiently classify proteins, similar to the databases such as SCOP or CATH. The ultimate goal is to establish a correspondence between the structural motifs and the function of the proteins. We discuss preliminary progress made in this direction and provide ideas for achieving these goals.

The first step towards developing a classification procedure is the definition of a distance metric between two proteins based on the structural motifs that they consist of. We could collect the number of instances of each $L8$ and each $L15$ motif in a particular protein. One would also expect the overall size of the chain given by the total number of residues to be an important factor. While the counts of the individual motifs are typically single-digit numbers, the number of residues is more than 250 on an average. Hence it makes sense to divide the residue numbers by a factor of 100 and then include it with the individual counts to obtain a 1300-vector (from 938 $L15$ motifs, 361 $L8$ motifs, and the residue number) that represents each protein. We could then calculate the Euclidean distance or the 1-norm between two such vectors. We could naturally think of a way to cluster proteins using such a distance metric.

To get an idea of the accuracy of this method, we considered pairs of proteins from three different families [2] – nuclear receptor ligand-binding domains, serine proteases, and G-proteins. Each pair was distinct due to the function of the protein chains. The protein pairs considered were (2PRG,1A28), (1A0L,1AZZ), and (1EFU,5P21) from the three respective families. It was observed that there were certain motifs that occurred only in one of the families – the $L15$ motif 770 ($\{1, 6\}$) occurred only in 1A28, and the $L15$ motif 927 ($\{6, 3\}$) occurred only in 2PRG. We tried hierarchically clustering the six chains. The first family was clearly separated from the rest, but the hierarchy was not exact for the other two families. The distance metric used was the 1-norm.

As another preliminary experiment, we took a set containing ten chains from the SCOP class a (all alpha) and ten chains from the SCOP class b (all beta). A cluster analysis was able to successfully group them into two separate hierarchies, with just one chain being mis-classified. A few more similar samples (of 20 chains with 10 each from class a and class b) could be classified with an average accuracy of 80%.

In order to improve the accuracy of the method, it looks essential to add some information about how the individual motifs interact with each other. The order of the motifs along the back-bone chain could also be helpful in

making the method more efficient. Defining when two motifs are in contact might not be straightforward. One idea might be to measure the common area of intersection between the corresponding tube complexes before the retraction step. If the area is above certain threshold, we could say that the two motifs are in contact at that particular level of growth. Another piece of information that could prove discriminative might be the joint-occurrences of pairs of motifs in various proteins.

Acknowledgments

The authors would like to thank Prof. Herbert Edelsbrunner from Duke University, and Dr. Afra Zomorodian from Stanford University for providing useful comments on the material presented in this chapter.

References

1. Antonina Andreeva, Dave Howorth, Steven E. Brunner, Tim J.P. Hubbard, Cyrus Chothia, and Alexey G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32:D226–D229, 2004. Database issue.
2. Stephen A. Cammer, Charles W. Carter, Jr, and Alexander Tropsha. Identification of sequence-specific tertiary packing motifs in protein structures using Delaunay tessellation. In *Computational Methods for Macromolecules: Challenges and Applications*, volume 24 of *Lecture Notes in Computational Science and Engineering*, pages 477–494, 2000.
3. Loredana Lo Conte, Steven E. Brunner, Tim J.P. Hubbard, Cyrus Chothia, and Alexey G. Murzin. Scop database in 2002: refinements accommodate structural genomics. *Nucleic Acid Research*, 30(1):264–267, 2002.
4. T.H. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, Massachusetts, 2 edition, 2001.
5. C.J.A. Delfinado and Herbert Edelsbrunner. An incremental algorithm for betti numbers of simplicial complexes on the 3-sphere. *Comput. Aided Geom. Design*, 12:771–784, 1995.
6. H. Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.
7. Herbert Edelsbrunner. The union of balls and its dual shape. *Discrete Comput. Geom.*, 13:415–440, 1995.
8. Herbert Edelsbrunner. *Geometry and Topology for Mesh Generation*. Cambridge University Press, England, 2001.
9. Herbert Edelsbrunner. CPS 296.1: Bio-Geometric Modeling (fall 2002) – class notes, 2002. http://cs.duke.edu/education/courses/fall02/cps296.1/.
10. Herbert Edelsbrunner, D.G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory*, IT-29:551–559, 1983.
11. Herbert Edelsbrunner and Ernst P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.

12. Michael A. Erdmann. Protein similarity from knot theory and geometric convolution. Technical Report CMU-CS-03-181, School of Computer Science, Carnegie Mellon University, September 2003.
13. Peter J. Giblin. *Graphs, Surfaces and Homology.* Chapman and Hall, London, 2 edition, 1981.
14. H.M. Grindley, P.J. Artimuik, D.W. Rice, and P. Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.*, 229:707–721, 1993.
15. Michael Henle. *A Combinatorial Introduction to Topology.* W.H. Freeman and Company, San Francisco, 1979.
16. E.M. Mitchell, P.J. Artymuik, D.W. Rice, and P. Willett. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.*, 212:151–166, 1989.
17. N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, and et al. The InterPro database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, 31:315–318, 2003.
18. James R. Munkres. *Topology.* Prentice Hall, New Jersey, 2 edition, 2000.
19. Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.
20. C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.N. Swindells, and J.M. Thornton. CATH: A hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.
21. Brookhaven protein data bank. http://www.rcsb.org.
22. Peter Røgen and Boris Fain. Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci. (USA)*, 100(1):119–124, 2003.
23. W.R. Taylor and C.A. Orengo. Protein structure alignment. *J. Mol. Biol.*, 208:1–22, 1989.
24. J. White. Self-linking and the Gauss integral in higher dimensions. *American Journal of Mathematics*, 91:693–728, 1969.