Interface Surfaces for Protein-Protein Complexes

YIH-EN ANDREW BAN

Duke University Medical Center, Durham, North Carolina

HERBERT EDELSBRUNNER

Duke University, Durham, and Raindrop Geomagic, Research Triangle Park, North Carolina

AND

JOHANNES RUDOLPH

Duke University Medical Center, Durham, North Carolina

Abstract. Protein-protein interactions, which form the basis for most cellular processes, result in the formation of protein interfaces. Believing that the local shape of proteins is crucial, we take a geometric approach and present a definition of an interface surface formed by two or more proteins as a subset of their Voronoi diagram. The definition deals with the difficult and important problem of specifying interface boundaries by invoking methods used in the alpha shape representation of molecules, the discrete flow on Delaunay simplices to define pockets and reconstruct surfaces, and the assessment of the importance of topological features. We present an algorithm to construct the surface and define a hierarchy that distinguishes core and peripheral regions. This hierarchy is shown to have correlation with hot-spots in protein-protein interactions. Finally, we study the geometric and topological properties of interface surfaces and show their high degree of contortion.

Categories and Subject Descriptors: G.2.1 [**Discrete Mathematics**]: Combinatorics—*Combinatorial Algorithms*; I.3.5 [**Computer Graphics**]: Computational Geometry and Object Modeling—*Geometric Algorithms*; I.5.1 [**Pattern Recognition**]: Models—*Geometric*; J.2 [**Computer Applications**]: Physical Sciences and Engineering—*Chemistry*; *Physics*; J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and Genetics*

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Filtrations, geometric and topological algorithms, interface surfaces, protein interaction, Voronoi diagrams

A preliminary version of this article appeared in *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology*, ACM New York, 2004, pp. 205–212.

This research was supported by the National Science Foundation (NSF) grants CCR-00-86013 and EIA-99-72879, and the National Institutes of Health (NIH) grant GM61822-01.

Authors' addresses: Y.-E. A. Ban and J. Rudolph, Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, e-mail: yab@duke.edu, Rudolph@biochem.duke.edu; H. Edelsbrunner, Department of Computer Science, Duke University, Durham, NC 27708, e-mail: edels@cs.duke.edu;

1. *Introduction*

Protein-protein interactions form the basis for most cellular processes including events intimately linked to human disease such as cell division and growth. Although protein-protein interactions are ranked high on the list of unsolved problems, they remain poorly understood in regard to basic classification, specificity of recognition, and energetics of binding. A comparison can be made between the current state of the protein-protein interaction field and the field of protein structure prior to the descriptors and classifications that have now become part of the standard language. Specifically, following the definition of α-helices and β-sheets and the determination of numerous protein structures, it became possible to visualize and classify proteins into families (e.g., β-barrels or β-α-β sandwiches). These classifications have led to important insights into protein function, protein folding mechanisms, protein structure prediction, and evolutionary relationships. Even today, efforts at defining the functions of the many uncharacterized proteins in the human proteome (about 50%) rely heavily on these descriptors. In an analogous manner, descriptors of protein interfaces based on geometry (shape) and physics (forces) that allow for visualization, characterization, and classification, can be envisioned as useful to the protein-protein interaction community. For example, such studies of interfaces may reveal regions of known importance such as binding hotspots, sites where mutation of specific residues lead to significant loss in binding energy. General interfacial features to be examined include geometric characteristics such as distances, pockets, wrinkledness and physical characteristics such as electrostatics and hydrophobicity.

1.1. PRIOR WORK. Information that elucidates the driving forces and pinpoints the specificity of protein-protein interactions has been extremely difficult to obtain. The most concrete insights have come from experiments. One popular technique, known as alanine scanning mutagenesis, involves making alanine mutants for each of the interfacial residues of interest and then assaying the mutants for a change in binding affinity. Alanine scanning studies performed by Wells and collaborators [Clackson et al. 1998; Wells 1996] on the hGH/hGHbp complex have resulted in the hot-spot theory of interactions. According to this theory, although protein interaction surfaces are large and complicated, only a few specific regions of the interface are responsible for the majority of the interaction energy. Similar studies performed on other protein-protein complexes have provided evidence for the general applicability of the hot-spot theory [Castro and Anderson 1996; Schreiber and Fersht 1995; Shapiro et al. 2000]. To explain the association rate of two proteins, a theory known as electrostatic steering has been developed and appears to identify charged residues at the periphery of the protein-protein interface as a major component of long-range interactions for certain protein-protein complexes [Lee and Tidor 2001; Selzer et al. 2000; Sheinerman and Honig 2002].

Statistical studies analyzing static crystal structures of protein-protein complexes have historically provided a rough view of general features found in protein-protein interactions. Several statistical studies are available [Jones and Thornton 1997; Lo Conte et al. 1999; Xu et al. 1997] and typically have the following format: interfaces are defined by a distance threshold, the buried surface area, or a combination of the two, and statistical analyses of geometric and biochemical characteristics are performed. Results from these studies include an average buried area of protein-protein interfaces (1,600 ± 400 Å^2) and a more hydrophobic nature of interfaces in comparison to other protein surfaces. These studies have not fared well when attempting to provide deeper insights into protein-protein interactions.

The key to a computational approach for unlocking the information captured in crystal structure data is an appropriate model, which can be either physical or geometric. Most of the computational approaches to understanding protein-protein interactions have focused on energy calculations or structural analyses to predict binding hot-spots. These studies have used molecular mechanics Poisson-Boltzmann surface area calculation [Massova and Kollman 1999], simple physical models of interactions [Kortemme and Baker 2002], and evolutionary conservation [DeLano et al. 2000; Ma et al. 2003]. They differ in their emphasis on aspects of binding hot-spots, ranging from H-bonding to hydrophobic interactions to polar residues. Whereas these studies have had some success in identifying and quantifying protein binding hot-spots, they have not provided a generalizable framework for analyzing and comparing protein-protein interfaces. There are two prior approaches to constructing interface surfaces from equidistant points between the proteins, both generating smooth surfaces with ambiguous boundary definition [Gabdoulline and Wade 1996; Keil et al. 1998]. The only previous computational geometry approach to protein-protein interfaces was described by Varshney et al. [1995], giving a definition that is asymmetric and yields relatively fractured surfaces due to the use of absolute distance thresholds.

1.2. METHODS AND RESULTS. We use concepts developed in computational geometry and topology [Edelsbrunner 2001] to define interface surfaces that are symmetric and avoid fracture through the use of a relative distance threshold. The particular concepts we base our work on are the Voronoi diagram whose application to protein data has been pioneered by Richards [977], the alpha shape representation of molecules introduced in Edelsbrunner and Mücke [1994], the discrete flow on the Delaunay simplices used in the past to define pockets [Edelsbrunner et al. 1998] and to reconstruct surfaces [Edelsbrunner 2003], and the assessment of the importance of topological features as defined in Edelsbrunner et al. [2002]. Using these concepts, we deal with the difficult and important problem of specifying interface boundaries. In addition, we give a robust and efficient algorithm for constructing interface surfaces. Finally, we construct a level-of-focus hierarchy that distinguishes protected from peripheral regions. We have implemented the algorithm and use examples constructed with our software to illustrate the primarily theoretical discussions. We also use the software to analyze basic geometric and topological properties of interface surfaces and present some of our findings. A particularly tantalizing fact is the surprisingly high correlation between the protected portions of the interface surfaces and the experimentally determined hot-spot residues of protein-protein interactions. Our method and results complement and could benefit previous analyses of protein-protein interactions, particularly those that rely on discriminating general surface residues from those found at protein interfaces [Chakrabarti and Janin 2002; Ma et al. 2003].

1.3. OUTLINE. Section 2 introduces the definition of molecular interface surfaces and presents the algorithm for constructing them. Section 3 describes measures for analyzing interface surfaces and preliminary biochemical applications. Section 4 discusses extensions of this work.

2. *Definition and Algorithm*

The definition of an interface surface combines two intuitions, namely that the multichromatic part of the Voronoi diagram is the best separation between complexed molecules, and that the interesting portion of that separation is protected by a relatively tight seal. In this section, we turn the two intuitions into an unambiguous definition and an efficient algorithm.

2.1. SMOOTH INSPIRATION. The unambiguous construction of an interface surface is based on discrete data and is couched in the language of discrete geometry and combinatorial topology. We now describe an intuitive smooth process that motivates the discrete steps. Imagine a smooth map on space, $f : \mathbb{R}^3 \to \mathbb{R}$. Assuming f is generic, we gain an understanding of the function by looking at its critical points, which are minima, saddles of index one and two, and maxima. We use the critical points and their relationship to form a hierarchical partition of space. This is more easily described in one dimension less, so imagine a generic smooth map on the plane, $f : \mathbb{R}^2 \to \mathbb{R}$, whose critical points are minima, saddles, and maxima. Now flood the plane by continuously raising the water level at all locations, observing where it rises above ground as defined by f. Structurally significant events happen when we reach critical points: A minimum starts a lake, a saddles merges two lakes or forms an island, and a maximum ends an island. An alternative way of flooding the plane raises the water level without seepage. In other words, the sea rises from the global minimum of f but other minima do not automatically start a lake when the water reaches their level. Water can invade the land only by flowing over natural dams, which first happens at saddles. Once the sea reaches a saddle, it can flow into the basin on the other side, which is a recursive process. Flooding starts at the lowest point of the basin until it reaches the height of that same saddle, filling its own basins and creating its own islands in the process. In the end, the sequence of floods defines a hierarchical partition of the plane determined by the relative position and height of the saddle points. Returning to three dimensions, we obtain a similar hierarchical partition of space determined by the relative position and height of the saddles of index one.

To relate this picture with interface surfaces, let f be a generic smooth approximation of the (negative) local distance to the nearest atom or sphere in a complex of molecules. Giving up the smooth picture and translating it into a discrete construction turns out to be a major undertaking, but one that is worthwhile as it leads to a stable and extremely efficient algorithm. In the translation, we map critical points to simplices in the Delaunay triangulation: minima to tetrahedra, index-one saddles to triangles, index-two saddles to edges, and maxima to vertices. Continuous flooding without flowing over critical points translates into a retraction, which we describe as a composition of collapses. Finally, a watershed event translates into the deletion of a tetrahedron (the lowest point in the basin), a retraction (flooding of the basin), and the deletion of a triangle (the saddle causing the watershed). The retraction itself may have a complicated recursive structure mimicking the recursive sequence of watersheds. This structure is rationalized by a pairing of the critical points that mark the beginning and end of the watersheds.

2.2. SURFACES WITHOUT BOUNDARY. We begin the discrete construction by turning the first intuition on separating complexed molecules using a subcomplex of the Voronoi diagram into a technical description. Consider $\ell \geq 2$ molecules, each

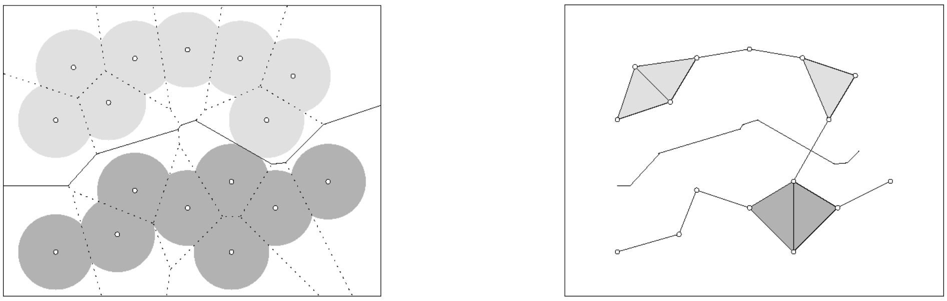

FIG. 1. Left: The solid, bi-chromatic Voronoi segments form the interface curve that separates the two collections of disks. The dotted, monochromatic segments complete the Voronoi diagram. Right: The dual complex of the union of disks and the shrunken and clipped interface curve separating the two collections.

represented by its *space-filling diagram*, which is a union of finitely many solid spheres or balls in $\mathbb{R}^3$. Denote the collections of balls by B_1 to B_ℓ and consider the set of all balls, of all collections, $B = \bigcup_{j=1}^{\ell} B_j$. We introduce the *(weighted) Voronoi diagram*, which decomposes space into convex cells, one per ball in B. Formally, $\pi_b(x) = \|x - z\|^2 - r^2$ is the *weighted square distance* of a point $x \in \mathbb{R}^3$ from a ball b with center $z \in \mathbb{R}^3$ and radius $r \in \mathbb{R}$, and the *Voronoi cell* of $b \in B$ is the set of points x for which $\pi_b(x) \leq \pi_c(x)$ for all balls $c \in B$. In the generic case, every Voronoi cell is either empty or a convex polyhedron with nonempty interior. Similarly, the intersection of two Voronoi cells is either empty or a convex polygon, that of three is either empty or a line segment, and that of four is either empty or a point. The left picture in Figure 1 illustrates this definition for two collections of disks in the plane. The *color* of a polyhedron is the index of the collection B_j that contains the generating ball. A Voronoi polygon belongs to exactly two polyhedra and is therefore either monochromatic or bichromatic. Similarly, a Voronoi segment or point can be either monochromatic or multichromatic, and the latter occurs if and only if it belongs to a bichromatic polygon. The *interface surface* $\mathbb{S} = \mathbb{S}(B_1, B_2, \ldots, B_\ell)$ consists of all multichromatic Voronoi polygons, segments and points.

For $\ell = 2$ molecules, each multichromatic segment belongs to exactly two bichromatic polygons, and each multichromatic point belongs to a topological disk formed by three or four bichromatic polygons. It follows the interface surface is a 2-manifold without boundary, and because it separates color-1 from color-2 polyhedra, it is necessarily orientable. For $\ell > 2$ molecules, we may have trichromatic segments and tri- and four-chromatic points. After removing these segments and points from $\mathbb{S}$, we get a (possibly empty) orientable 2-manifold without boundary for each pair of colors. These 2-manifolds fit together in triplets along trichromatic curves and in six-tuplets around fourchromatic points.

2.3. GROWTH AND FILTRATION. Before incorporating the second intuition into the definition, we need to understand the evolution of the space filling diagram as the balls grow. The key concept here is the filtration of dual complexes. We begin by introducing the *(weighted) Delaunay triangulation D* obtained by dualizing the Voronoi diagram: for every collection of Voronoi cells with nonempty common intersection we add the convex hull of the centers of the generating balls to D. In the assumed generic case, the convex hulls are simplices of dimension 0 to 3:

vertices, edges, triangles and tetrahedra. Similarly, we obtain the *dual complex* K of the space-filling diagram $F = \bigcup B$ by dualizing the restriction of the Voronoi diagram to the space-filling diagram: for every collection of Voronoi cells whose common intersection contains points of F we add the convex hull of the centers of the generating balls to K. The right picture in Figure 1 illustrates this definition.

Now imagine we grow the balls simultaneously in such a way that the Voronoi diagram does not change. Letting $t \in \mathbb{R}$ be time, we accomplish this by growing the ball b with center z and square radius r^2 to the ball b_t with the same center z and with square radius $r^2 + t$ at time t. (For negative time, we may have negative square radii, which correspond to imaginary radii and balls.) The growth does not affect the Voronoi diagram because it does not change the difference between any two square radii. It follows that the Delaunay triangulation does not change and the dual complex contains progressively more simplices until it eventually equals the Delaunay triangulation. We are interested in the details of this evolution. Since there are only finitely many simplices, we have only finitely many different dual complexes, which form a nested sequence interpolating between the empty complex and the Delaunay triangulation:

$$\emptyset = K_0 \subset K_1 \subset \cdots \subset K_m = D.$$

We refer to this sequence as the *filtration* of dual complexes. An elementary step in the evolution consists of adding the simplices $\tau \in K_i - K_{i-1}$ to K_{i-1}. In the generic case, this happens when the (growing) space-filling diagram encounters a new Voronoi point, segment, polygon or polyhedron. We distinguish between *critical* events in which there is only one such simplex τ, and *regular* events in which K_i differs from K_{i-1} by two or more simplices. There are three particular types of events, two critical and one regular, that are more relevant to the construction of the interface surface than the others:

Type 1. Four balls close in from all directions on a Voronoi point. This corresponds to adding a single tetrahedron to the dual complex.

Type 2. Three balls close in from all normal directions on a segment, eventually touching it at an interior point. This corresponds to adding a single triangle to the dual complex.

Type 3. Four balls close in on a Voronoi point, but they leave a gap around one of the incident segments and encounter both at the same moment. This corresponds to adding a triangle-tetrahedron pair to the dual complex.

A common representation of the filtration is the list of simplices ordered by the time they join the dual complex. Simplices that join at the same moment are ordered by dimension and remaining ties are broken arbitrarily. An algorithm for constructing this representation can be found in, Edelsbrunner and Mücke [1994], and software is publically available at http://biogeometry.duke.edu. It first computes the Delaunay triangulation, then determines the times the simplices join the dual complex, and finally sorts the Delaunay simplices by time and dimension. For biomolecular data, the number of Delaunay simplices is typically some constant times the number of balls, n, and the Delaunay triangulation can be constructed in time $O(n \log n)$, see for example Edelsbrunner [2001]. Determining the times and sorting the simplices takes again time $O(n \log n)$.

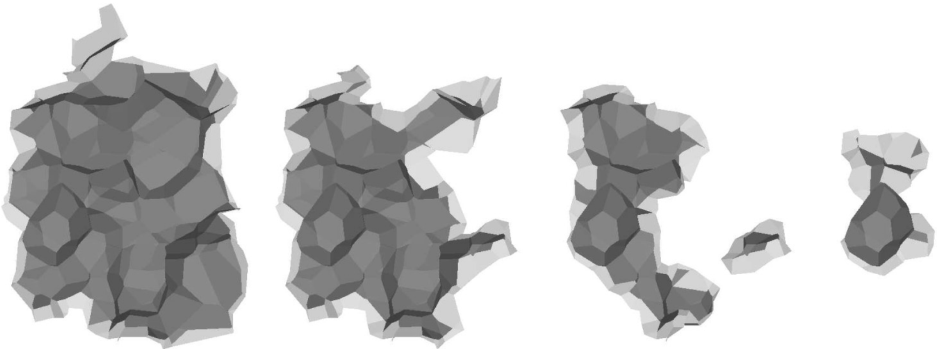

FIG. 2. Four interface surfaces in the level-of-focus hierarchy of the barnase-barstar complex. The two colors distinguish clipped polygons next to the boundary from unclipped polygons in the interior.

2.4. BOUNDARY THROUGH RETRACTION. We are now ready to incorporate the second intuition into the definition, namely that the interesting portion of the interface surface is protected by a relatively tight seal. Portions outside this seal are removed by retraction, which can be understood as a reversal of the growth process relaxed to a partial order [Edelsbrunner 2003]. We explain this by considering the filtration of dual complexes. Re-index the simplices in the corresponding sequence such that τ_{ij} is the jth new simplex in K_i. In the generic case, the simplices in $K_i - K_{i-1}$ form an abstract simplex: there are 2^k simplices τ_{ij}, for $1 \leq j \leq 2^k$, and every τ_{ij} is face of $\upsilon = \tau_{i2^k}$ and has $\sigma = \tau_{i1}$ as a face. We write $\sigma \leq \tau_{ij} \leq \upsilon$ to express the latter property. For the time being, we are only interested in regular events characterized by $k \geq 1$. Adding the $2^k > 1$ simplices to K_{i-1} does not affect its homotopy type. We note that σ is *free* in K_i, by which we mean that it is the face of a single simplex, namely of υ, but of no other simplex in K_i. We refer to the operation that deletes σ together with all simplices that contain it as a *collapse*. In our algorithm, we use only collapses for which υ is a tetrahedron. Triangles, edges and vertices that do not belong to any remaining tetrahedron are deleted as soon as they arise. We also require that a collapse deletes all and not just some simplices joining the dual complex at the same moment. We define a *retraction* as a maximal sequence of collapses. In other words, it applies collapses to a given complex until there is no further collapse possible. In the implementation of this operation, we maintain a stack of candidate pairs (σ, υ), with new pairs pushed on the stack when they appear. We may think of a retraction as the process of successively deleting sinks from an acyclic directed graph. It follows that the result of the operation is independent of the sequence in which the collapses are performed. We finally get a shrunken interface surface as a side-effect of retaining only polygons that correspond to bi-chromatic edges in the retracted complex. If this is an interior edge then we retain the entire polygon, else we clip the polygon and retain only the pieces that correspond to incident tetrahedra in the retracted complex. Figure 2 illustrates this idea by showing the retracted interface surface of two complexed proteins on the far left.

2.5. HIERARCHY THROUGH PERSISTENCE. It remains to explore the critical events characterized by $k = 0$, that is, $\upsilon = \tau_{i1}$ is the only new simplex in K_i. We use the concept of topological persistence to quantify how different υ is from being regular. Such a notion makes sense because the addition of υ to K_{i-1} either creates or destroys a topological feature and, as shown in Edelsbrunner et al. [2002], there is a unique critical matching simplex σ that earlier created what υ destroys or that will later destroy what υ creates. We call the time-lag between the addition of σ and the addition of υ the *persistence* of both. Suppose, for example, we have a critical triangle σ and a matching critical tetrahedron υ in quick succession. Then, their persistence is small, indicating that the pair is very similar to a regular event in which σ and υ would join the dual complex at the same moment. The algebraic justification of this definition is beyond the scope of this article and given in Edelsbrunner et al. [2002] along with an algorithm that generates the matching pairs in worst-case time $O(m^3)$, where m is the number of simplices in the Delaunay triangulation. Our experimental results for protein data suggest however that the running time is much better, namely $O(m)$ or similar.

We now take the shrinking process beyond the initial retraction step. Let (σ, υ) be a matching critical triangle-tetrahedron pair generated by the topological persistence algorithm. In the forward direction of the filtration, the addition of σ creates a void which is destroyed by the later addition of υ. We use the pair to define an extension of the collapse operation, which we call a *removal*: assuming σ lies on the boundary of the remaining complex, we first delete υ and then retract around υ. If the retraction reaches far enough, then σ gets deleted just because both its tetrahedra have been deleted. However, it can happen that the retraction does not reach all the way, in which case we recurse for other pairs of simplices before deleting σ. Think of the retraction from υ as creating a *dome* in the space between the molecules and the triangle σ as the *entrance* or the biggest gap in the *seal* surrounding the dome. We can now interpret the times s and u when σ and υ join the dual complex as the sizes of the entrance and the dome. We define the *seal value* of (σ, υ) as $f(s, u) = \frac{s}{u-s}$. To decide whether or not to remove σ and υ in the first place, we require that s and u are both positive and that $f(s, u)$ exceeds a positive constant threshold C_0. Since υ succeeds σ in the filtration, we have $u > s$ and therefore $f(s, u) > 0$. The seal value can be large for two reasons: because the difference in size between the dome and the entrance is small or because the entrance is large. The removal process is thus biased toward both. Note also that for $s < s' < u' < u$ we have

$$f(s, u) < f(s', u'). \tag{1}$$

This monotonicity property is important for the correctness of our algorithm because if the retraction around υ does not reach σ then this can only be because there is a triangle σ' between υ and σ that split the void created by σ before it was destroyed by υ. But then the other branch was destroyed by a tetrahedron υ' preceding υ in the filtration. In other words, $s < s' < u' < u$, where s' and u' are the times σ' and υ' join the dual complex. Inequality (1) guarantees that the simplices between υ and σ are deleted by recursive removals so that σ can eventually be deleted. The algorithm starts with the Delaunay triangulation and ends with a subcomplex that allows no further collapses or removals. Figure 3 gives some insights into the shrinking process by showing the seal values of the domes as they get deleted. The monotonically nonincreasing green graph plots the evolution of the threshold value that corresponds to the interface. The red graph above the green shows how the deletion of a dome gives access to domes with even higher seal values, which are then recursively removed. The blue and magenta graphs display the evolution of the two components of the seal function, s and $u - s$. These components help rationalize

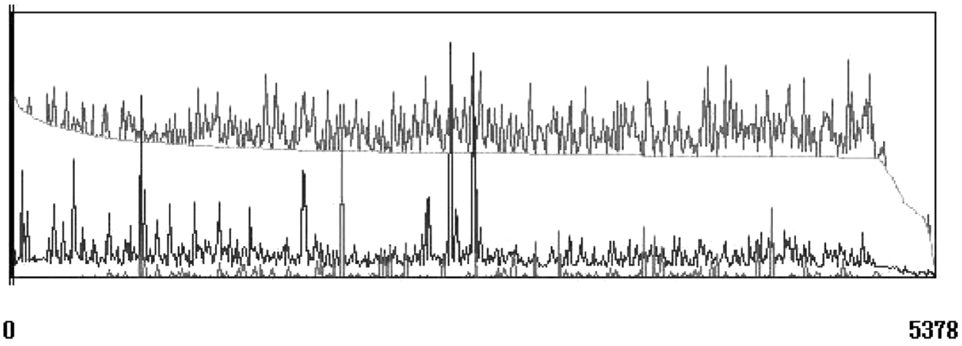

FIG. 3. Interface signatures displayed with the evolution of deletions drawn from left to right. From top to bottom: the red graph shows the seal values of the domes, the green graph tracks the minimum seal value, and the blue and magenta graphs plot s and $u - s$.

the occasional appearance of seemingly spurious specks of interface surface, which tend to have negative values of s. Such occurrences indicate clashes between the proteins and owe their existence to measurement or interpretation errors in the structure determination work. Figure 2 illustrates the result of shrinking with four pictures in the much longer nested sequence of interfaces surfaces of the barnase-barstar complex. The running time for constructing the hierarchy is constant per simplex in the Delaunay triangulation and therefore O(m).

Remark. We note here that the interface surface construction depends upon atom position and size. Structural perturbations can change the interface and cannot be avoided in the approach we take. While no theoretical guarantees are made, we observe that smaller perturbations, such as those due to experimental errors or limited conformational changes, generally only affect the interface surface boundary. There are no drastic changes on either the interior of the surface or the surface on which the boundary moves. An example of these small perturbations can be examined in the 1BRS barnase-barstar crystal structure. This structure was solved with three slightly different copies of the dimeric complex in each unit cell. Despite the differences between the three complexes, the interface surfaces are essentially indistinguishable with only a few changes at the boundary, where a small protrusion is pushed or pulled into the surfaces by about three angstroms. Our retraction process provides a means to minimize the small differences on the boundary. Slightly retracting the three copies of the barnase-barstar surfaces eliminates this protrusion, yielding essentially the same interface.

3. *Analysis*

We believe that the primary use of molecular interface surfaces, as defined in this article, will be to tease out useful information about protein-protein interactions. This can either be done directly, by studying the interface as a geometric object in its own right, or by using it as a domain on which functions expressing biochemical data are defined. An important component of these analyses is the visualization of the interface surfaces. As an illustration we provide several examples in Figure 5 to 9. All interface surfaces in this section are generated using the coordinates of proteins taken directly from crystal structure, and whose atomic radii are set to the van der Waals parameters of the AMBER95 force field [Cornell et al. 1995].

3.1. HOT-SPOTS IN PROTECTED REGIONS. The main reason for creating the level-of-focus hierarchy of the interface is its facility to distinguish protected from peripheral regions. To demonstrate the biochemical implication of this hierarchy, we show that residues which have atoms involved in the late stages of the hierarchy are somehow more critical for the interaction. We do this by constructing a simple function which we then use to distinguish hot-spot from neutral residues in the interface. Letting R be a residue, $p_0, p_1, \dots, p_k$ the polygons in the interface generated by its side-chain atoms, we define

$$h(R) = \sum_{i=0}^{k} w_i \cdot \text{area}(p_i),$$

where the weight w_i is the fraction of the interface surface that belongs to R right before the moment in time when p_i is removed. Big contributions come from large polygons removed late in the game. We predict R as a hot-spot residue if $h(R) \geq 1.68$ and as a neutral residue otherwise. This cutoff is selected to equalize the percentages of correctly predicted hot-spot and neutral residues. For a baseline comparison, we also distinguish hot-spot residues from neutral residues utilizing a measure common in protein structural studies, the area buried upon forming a complex. The *solvent accessible surface area* (*sasa*) of a protein is defined in Lee and Richards [1971], and is the area of the outer envelope of a protein represented as a union of balls whose radii are expanded by the radius (1.4 Å) of a probe sphere. The *sasa* of a residue is the contribution of that residue's atoms to the total *sasa* of the protein. Then the *buried surface area* (*bsa*) of a residue for a protein involved in a protein-protein complex is

$$bsa(R) = sasa_I(R) - sasa_C(R),$$

where $sasa_I$ is the accessible surface area of the residue in the isolated protein and $sasa_C$ is the accessible surface area of the residue in the complex. We predict a residue to be a hot-spot residue if $bsa(R) \geq 56.1$ Å^2, meaning that at least 56.1 Å^2 of the residue's surface area is buried upon complex formation. Using the alanine scanning data for the nineteen protein complexes studies in Kortemme and Baker [2002] and setting the threshold for a hot-spot residue at 2.0 kcal/mol, we employ both tests to predict hot-spot and neutral residues. With $h(R)$ we correctly identify 72.4% of the hot-spot and 72.6% of the neutral residues. This compares favorably against $bsa(R)$ which correctly identifies 65.0% of the hot-spot and 64.3% of the neutral residues. Table I indicates the hot-spot prediction accuracies of both $bsa(R)$ and $h(R)$ for individual residues grouped by type. For all residue types except valine and asparagine, $h(R)$ performs equally well or better than $bsa(R)$. Interestingly, $h(R)$ performs better than $bsa(R)$ even on tyrosine and tryptophan, two aromatic residues which tend to bury a large amount of surface area. It is important to note that we have limited both functions to side-chains, as opposed to main-chains, in an effort to be consistent with the nature of alanine scanning experiments. More precisely, we account for the area associated with each residue's interactions through its side-chain atoms, but do not directly account for the area associated with interactions through its main-chain atoms. The function $h(R)$ is similar in predictive power to the physical model of Kortemme and Baker [2002], which achieves slightly better percentages, 79% for hot-spot and 68% for neutral residues, for a threshold of 1.0 kcal/mol and worse percentages for the threshold of 2.0 kcal/mol we use.

TABLE I. HOT-SPOT PREDICTION ACCURACY FOR INDIVIDUAL RESIDUE TYPES

Residue	*bsa*(*R*)	*h*(*R*)
Aliphatic		
Valine	100%	66%
Leucine	50%	100%
Isoleucine	33%	66%
Polar		
Serine	50%	50%
Threonine	0%	100%
Asparagine	77%	77%
Glutamine	50%	100%
Aromatic		
Phenylalanine	100%	100%
Tyrosine	86%	93%
Tryptophan	60%	80%
Histidine	100%	100%
Ionizable		
Aspartic Acid	77%	77%
Glutamic Acid	25%	25%
Arginine	60%	60%
Lysine	60%	60%

Percentages indicate percent of experimentally verified hot-spots predicted correctly. No experimental hot-spot data in our database is available for unlisted residue types.

While preliminary, these results testify to the potential of the interface surface in rationalizing biochemical processes that define protein interactions. We note for example that the level-of-focus hierarchy is a geometric concept similar to the O-rings which Bogan and Thorn [1998] conjecture surround hot-spots in protein interactions, or to the protection phenomena of wrapped hydrogen bonds [Fernández and Scheraga 2003]. In this physical analogy, the seal value becomes a measure of how difficult it is to break into a dome. For example, since proteins are immersed in water, one can imagine that the seal value indicates the degree of difficulty for water to enter a dome. However, at present, we still lack an understanding of the intimate biochemical details of what the level-of-focus hierarchy captures, and are working towards this goal.

3.2. GLOBAL MEASURES. One goal of our research is the classification of interfaces into types that correspond to different kinds of protein-protein interactions. We seek global measurements that are likely to have biochemical significance. For example, it is generally acknowledged that for interfaces it is important to have a good geometric fit between the proteins. Here we focus on topological and geometric assessments of how contorted interfaces are.

We begin with topological characteristics, restricting ourselves to the relatively simple bi-chromatic case, in which the interface is a connected orientable 2-manifold with boundary. Topologically, such a manifold is completely characterized by its genus and its number of holes or boundary cycles [Massey 1967]. Most interfaces we have examined thus far have genus zero, but there are exceptions. One is the interface created by vipoxin complex, a phospholipase A_2 bound to its protein inhibitor, shown in Figure 8. It has genus three, indicating the existence of three pairs of links that lock the two proteins together, consistent with the high biochemical stability of this complex in solution. In the bi-chromatic case, having the number of holes that exceeds one is possible. For example, a portion of the Delaunay triangulation may shrink from a mono-chromatic triangle on its boundary and in this way punch a hole in the interface.

An interface can be highly contorted despite having zero genus. To get a handle on this phenomenon, we measure the average curvature and the variation from that average. A useful result in this context is the Gauss–Bonnet theorem that states the total Gaussian curvature is an invariant of a closed orientable 2-manifold, namely equal to 4π times one minus the genus [Hopf 1983]. Interfaces are not smooth so we need an equivalent piecewise linear concept. For a vertex u, we call $\theta_u = 2\pi - \sum \varphi_j$ its *angle deficiency*, where φ_j is the angle of the jth interface polygon at u. The *total angle deficiency* is the sum of angle deficiencies over all $m = \text{card}\, U$ interior vertices: $\Theta = \sum_{u \in U} \theta_u$. Following the convention from non-Euclidean geometry, we classify $\mathbb{S}$ as *elliptic*, *flat* or *hyperbolic* depending on whether Θ is positive, zero or negative. We use the average angle deficiency as a baseline and measure the root-mean-square variation as

$$W = \sqrt{\frac{1}{m} \sum_{u \in U} \left(\theta_u - \frac{\Theta}{m} \right)^2}$$

and call it the *wrinkledness* of $\mathbb{S}$. It is straightforward to compute the total angle deficiency and the wrinkledness in time proportional to the number of vertices and polygons of the interface.

To gain an intuition for several of the geometric global measures, we compute them for interface surfaces, $\mathbb{S}$, generated after the initial retraction from a set of seventy pairwise protein-protein complexes taken from Chakrabarti and Janin [2002]; see Table II. The area of $\mathbb{S}$ ranges from 397 to 2408 Å^2 with a mean of 963 Å^2. The interfacial buried surface area (*bsa*) as computed in Chakrabarti and Janin [2002] is a two-sided measure, with comparable areas ranging from 930 to 4430 Å^2 with a mean of 1906 Å^2. There is an approximate linear correlation between *bsa* and interface surface area of 1.42 with a correlation coefficient of 0.85.

The total angle deficiency results show that interfaces are contorted and span the range from hyperbolic (-4.176 radians) to elliptic (4.995 radians); see Table II. This is in clear contrast to the classical view of the protein-protein interface that has only a small (2.8 Å) mean deviation from planarity [Chakrabarti and Janin 2002; Jones and Thornton 1997]. This discrepancy in results can be explained by the planarity measure in these previous studies which first group atoms into subsets by a heuristic and then take the root-mean-square distance of each subset of atoms against their least-square planes. This generates an averaged local measure, as opposed to our global measure of total angle deficiency. In contrast to total angle deficiency, the wrinkledness has little variance with a mean value of approximately 0.2 for the set of protein–protein complexes considered. Perhaps not surprisingly, the wrinkledness notably increases when hydrogens are added into the structures (data not shown).

3.3. LOCAL MEASURES. We are interested in local measures or maps that can be used in detailed studies of protein–protein interfaces. A simple example is the weighted distance function $\varpi : \mathbb{S} \to \mathbb{R}$ that maps every point x of the interface

TABLE II. Area and Total Angle Deficiency (AD) of 70 Protein Complexes Grouped into Six Functional Categories. AD is Calculated for the Second Interface Surface in the Hierarchy

Name	Area	AD	Name	Area	AD
Protease-Inhibitor					
2ptc	575	0.436	1mct	694	3.358
1avw	1011	3.713	3tpi	643	1.166
1tgs	734	2.734	1cho	736	0.289
1acb	717	3.751	1cbw	653	0.059
1ppf	900	−0.628	1fle	546	2.088
2kai	776	1.357	1hia	847	2.781
3sgb	462	1.778	1cse	745	3.167
2sic	717	0.713	2sni	869	−0.924
1stf	718	−1.781	4cpa	672	1.228
Large protease complexes					
1bth	871	2.819	4htc	1035	1.020
1tbq	1477	−3.923	1toc	1386	−2.334
1dan	1859	2.857			
Antibody-antigen					
1jhl	638	0.609	1vfb	585	0.642
1mlc	510	0.194	3hfl	719	3.690
3hfm	825	1.528	1fbi	617	3.701
1mel	502	4.792	1dvf	775	−0.401
1nfd	904	−0.110	1ao7	866	−0.197
2jel	638	1.372	1nca	1308	−1.793
1nmb	921	−0.764	1nsn	1089	1.418
1osp	747	−2.164	1qfu	1307	−0.641
1iai	1000	0.545	1kb5	1151	−0.293
Enzyme complexes					
2pcc	580	−2.609	1gla	712	−1.150
1brs	703	−4.176	1udi	906	−0.125
1dhk	1686	−0.717	1fss	728	3.702
1ydr	783	−0.349	1dfj	1795	−0.668
G-proteins, cell cycle, signal transduction					
1a0o	397	1.647	1gua	617	1.272
1a2k	966	−0.801	1agr	1278	−0.646
1tx4	1219	−1.921	1gg2	1788	−1.921
1got	1550	−2.017	2trc	2408	3.758
1fin	1533	4.321	1aip	1639	2.705
1efu	2205	−2.961			
Miscellaneous					
1ak4	409	0.821	1igc	498	0.700
1efn	488	−1.827	1fc2	604	−0.029
1seb	1081	1.486	1atn	796	−2.109
1ycs	560	1.137	2btf	1048	0.642
1hwg	2022	4.995	1dkg	1662	1.613

surface $\mathbb{S}$ to the weighted distance from the closest ball b, which is defined as $\varpi(x) = \sqrt{\pi_b(x)}$. By construction, that ball is ambiguous since x has at least one closest ball from each color. We may visualize this map using level lines, as in Figure 4. The minima, saddles and maxima of this map are of particular interest. Thinking of ϖ as measuring the local thickness or distance between the two proteins, the minima and maxima become local extremas of interface thickness.

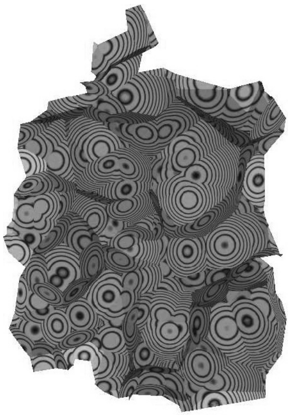

FIG. 4. Level line visualization of the weighted distance map over the interface surface of the barnase-barstar complex.

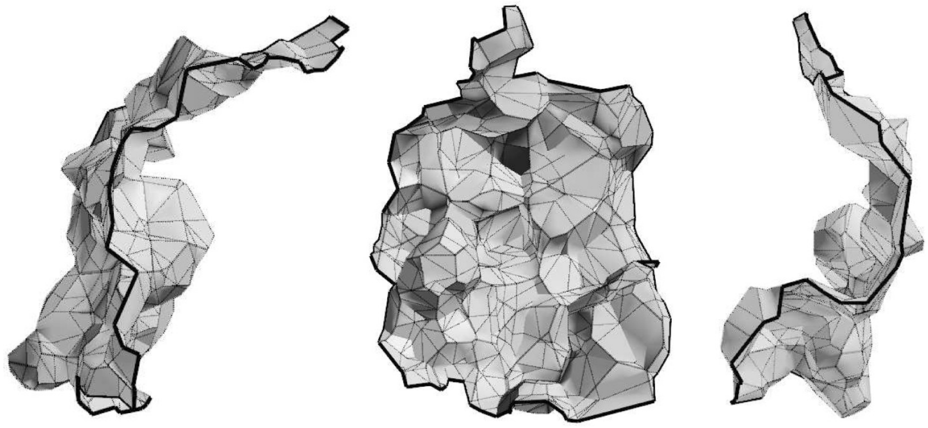

FIG. 5. Three views of the interface between Barnase and Barstar, a bacterial ribonuclease and its protein inhibitor, respectively. This experimentally well-studied complex has served as a model system for studying protein-protein interactions, in particular for characterizing binding hot-spots. The interface is somewhat smaller than average but is fairly typical in terms of shape. Generated from pdb file 1BRS.

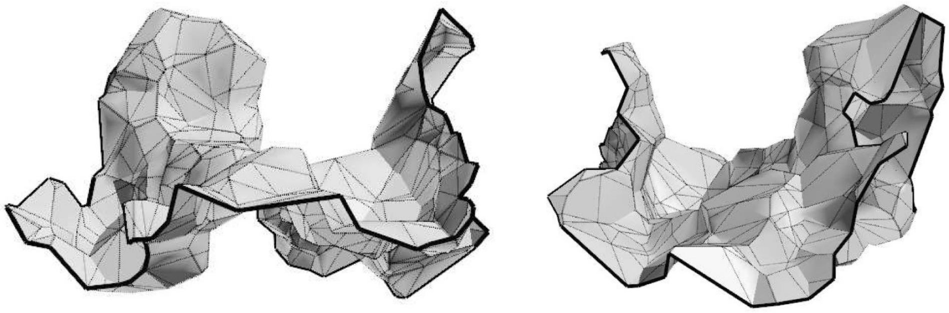

FIG. 6. Two views of the interface between colicin E9 DNase and the immunity protein IM9, a toxin produced during cell stress and its inhibitor, respectively. The affinity in the E9-IM9 complex is extremely tight (subfemtomolar). This interface is also smaller than average, but has a very prominent saddle shape. Generated from pdb file 1BXI.

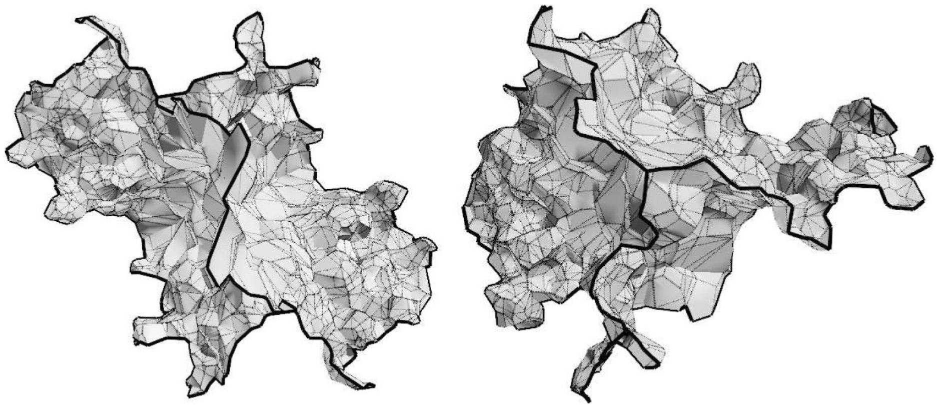

FIG. 7. Two views of the interface in human hemoglobin that demonstrate the utility of this representation for multimeric complexes. Hemoglobin consists of four separate but identical chains and the resulting interface shows the more complicated nature of a multisubunit interaction. Generated from pdb file 1A3N.

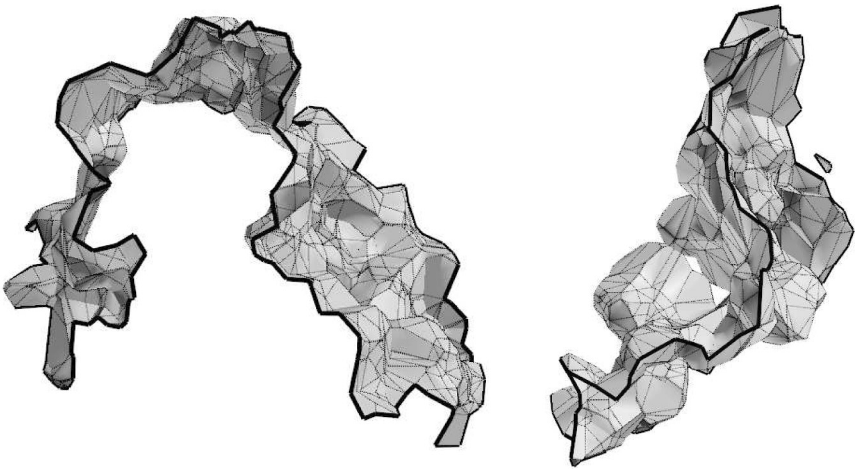

FIG. 8. On the left, the interface between human angiogenin and a placental ribonuclease inhibitor. The interaction between these proteins is extremely tight (femtomolar) and the interface exhibits both a very large surface area and an interesting overall bent shape. Generated from pdb file 1A4Y. On the right, the interface in the neurotoxic vipoxin complex from Western Sand Viper consisting of phospholipase A2 and its inhibitor. A rather unusual interface with genus 3. Generated from pdb file 1JLT.

According to smooth Morse theory [Matsumoto 2002], there are necessarily saddle points between the extrema, around which the sign of the change in local thickness changes four times. We can now explain the connection between the seal function and the local thickness map: each dome has a unique point x of locally maximum thickness, and we have $u = \varpi^2(x)$. Similarly, each seal has a point y of locally maximum thickness, which is a saddle of ϖ, and we have $s = \varpi^2(y)$. Now we just need to recall the pairing mechanism provided by topological persistence algorithm and we get the seal values as ratios $s/(u - s)$.

The method of defining continuous maps over the interface and analyzing them using ideas from Morse theory is general. We envision defining maps that express

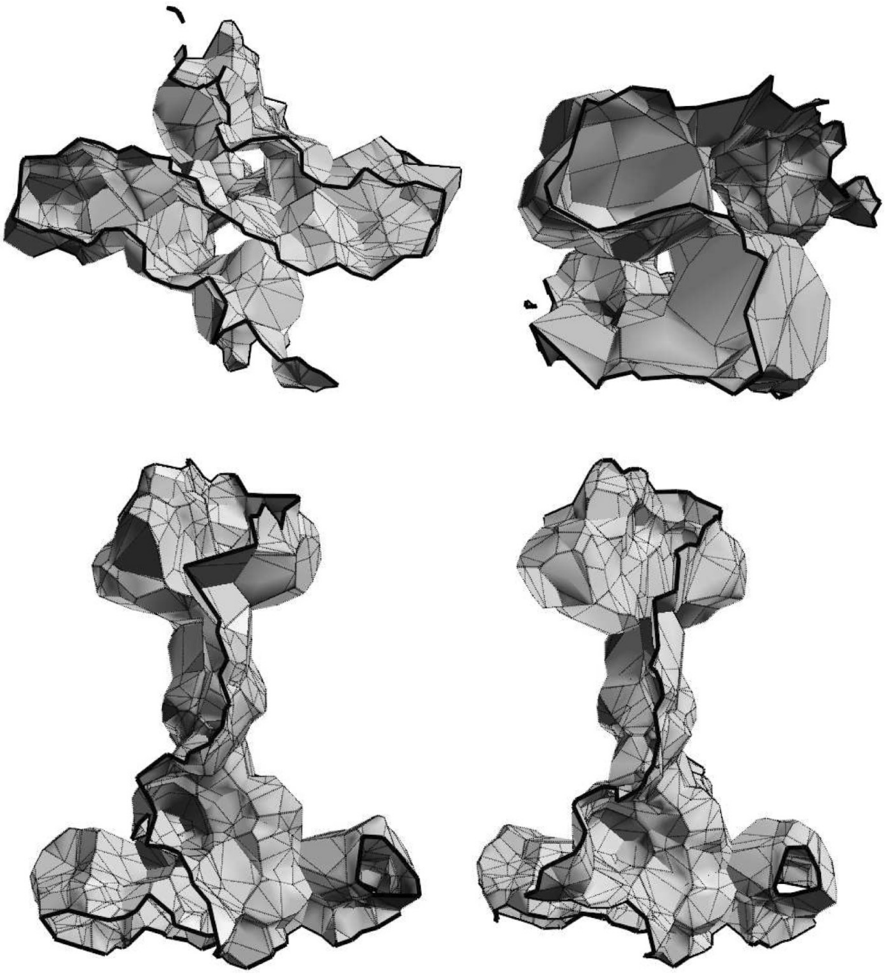

FIG. 9. Four views of the interface in HIV-1 protease, a homo-dimeric protein complex. This enzyme has been an important target for drug development against AIDS. The interface is fairly complex, in part due to the 'flaps' involved in the interaction between the two subunits. Generated from pdb file 3AID.

electrostatic and hydrophobic potentials, to name two, and to analyze them in terms of their critical points and their gradient flows. We refer to Edelsbrunner et al. [2003] for methods needed to cope with the difficulties that arise in the application of Morse theoretic ideas to piecewise linear data, and to Edelsbrunner et al. [2004] for concepts useful for comparing two or more maps defined over the same interface.

4. *Conclusion*

Given two or more proteins in complex, each represented by a space-filling diagram, we present a rigorous mathematical definition for an interface surface between them. This surface provides a more detailed view of the interaction region than traditional methods. We believe that previous studies of protein-protein interactions, instead of using fractional buried surface area or arbitrary distance cut-offs, could benefit from applying our definition in their statistical or evolutionary analyses. Taking the interface surface as an independent entity, we may study it by defining geometric and topological measures over it and map properties of both proteins on it. Additionally, we define a level-of-focus hierarchy that decomposes the interface surface into protected regions which appear to be biochemically important. This hierarchy may be studied on its own or incorporated into measures defined over the interface surface to enhance their analysis. Our novel representation of the interface surface will allow for new insights and discoveries in the study of protein-protein interactions. The generality of the interface surface definition also opens up other possibilities, such as studying water at protein-protein interfaces or internal packing of proteins. We might ask how different structural motifs within a single protein form internal surfaces, or geometrically characterize subtle structural rearrangements crucial to the functioning of proteins. Having an explicit surface representation offers practical advantages to finding such motifs and patterns. The surface can be manipulated in various ways to facilitate analysis, as we have done in our MAPS database of protein-protein interfaces and their associated mutagenesis data (http://biogeometry.duke.edu/research/docking/). In MAPS, the interface surface can be readily viewed in 3D and in a 2D representation, created by embedding the 3D surface into a disk. This flattened view allows for mapping and display of properties from each contributing protein in a complex, facilitating rapid analysis and comparisons of protein-protein interfaces. In closing, we note that although we focus on applications in protein interactions, the interface concept itself is general and there are other areas, such as nanostructures, in which interfaces arise and our geometric ideas are useful. Our software, *Ciel*, implementing these algorithms is available at http://biogeometry.duke.edu.

REFERENCES

Bogan, A. A., and Thorn, K. S. 1998. Anatomy of hot spots in protein interfaces. *J. Molec. Biol. 280*, 1–9.

Castro, M. M., and Anderson, S. 1996. Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: Effects on the kinetics and thermodynamics of binding to η-trypsin and α-chymotrypsin. *Biochemistry 35*, 11435–11446.

Chakrabarti, P., and Janin, J. 2002. Dissecting protein-protein recognition sites. *Proteins: Struct. Funct. Bioinfo. 47*, 334–343.

Clackson, T., Ultsch, M. H., Wells, J. A., and de Vos, A. M. 1998. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Molec. Biol. 277*, 1111–1128.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Amer. Chem. Soc. 117*, 5179–5197.

DeLano, W. L., Ultsch, M. H., de Vos, A. M., and Wells, J. A. 2000. Convergent solutions to binding at a protein-protein interface. *Science 287*, 1279–1283.

Edelsbrunner, H. 2001. *Geometry and Topology for Mesh Generation*. Cambridge University Press, Cambridge, England.

Edelsbrunner, H. 2003. *Surface reconstruction by wrapping finite sets in space, Discrete and Computational Geometry: The Goodman-Pollack Festschrift*. Springer-Verlag, Berlin, Germany, 379–404.

Edelsbrunner, H., Facello, M. A., and Liang, J. 1998. On the definition and the construction of pockets in macromolecules. *Disc. Appl. Math. 88*, 83–102.

Edelsbrunner, H., Harer, J., Natarajan, V., and Pascucci, V. 2004. Local and global comparison of continuous functions. In *Proceedings of IEEE Visualization 2004*, IEEE Computer Society Press, Los Alamitos, CA, 275–280.

EDELSBRUNNER, H., HARER, J., AND ZOMORODIAN, A. 2003. Hierarchical Morse complexes for piecewise linear 2-manifolds. *Disc. Comput. Geom. 30*, 87–107.

EDELSBRUNNER, H., LETSCHER, D., AND ZOMORODIAN, A. 2002. Topological persistence and simplification. *Disc. Comput. Geom. 28*, 511–533.

EDELSBRUNNER, H., AND MÜCKE, E. P. 1994. Three-dimensional alpha shapes. *ACM Trans. Graph. 13*, 43–72.

FERNÁNDEZ, A., AND SCHERAGA, H. A. 2003. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proce. Nati. Acad. Sci., USA 100*, 113–118.

GABDOULLINE, R. R., AND WADE, R. C. 1996. Analytically defined surfaces to analyze molecular interaction properties. *J. Molec. Graph. 14*, 374–375.

HOPF, H. 1983. *Differential Geometry in the Large*. Springer-Verlag, Berlin, Germany.

JONES, S., AND THORNTON, J. M. 1997. Analysis of protein-protein interaction sites using surface patches. *J. Molec. Biol. 272*, 121–132.

KEIL, M., EXNER, T., AND BRICKMAN, J. 1998. Characterisation of protein-ligand interfaces: Separating surfaces. *J. Molec. Model. 4*, 335–339.

KORTEMME, T., AND BAKER, D. 2002. A simple physical model for binding energy hot-spots in protein-protein complexes. *Proce. Nati. Acad. Sci. USA 99*, 14116–14121.

LEE, B., AND RICHARDS, F. M. 1971. The interpretation of protein structures: Estimation of Static accessibility. *J. Molecu. Biol. 55*, 379–400.

LEE, L. P., AND TIDOR, B. 2001. Barstar is electrostatically optimized for tight binding to barnase. *Nature Struc. Biol. 8*, 73–76.

LO CONTE, L., CHOTHIA, C., AND JANIN, J. 1999. The atomic structure of protein-protein recognition sites. *J. Molec. Biol. 285*, 2177–2198.

MA, B., ELKAYAM, T., WOLFSON., H., AND NUSSINOV, R. 2003. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proce. Nati. Acad. Sci. USA 100*, 5772–5777.

MASSEY, W. S. 1967. *Algebraic Topology: an Introduction*. Springer-Verlag, New York.

MASSOVA, I., AND KOLLMAN, P. A. 1999. Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J. Ameri. Chem. Soc. 121*, 8133–8143.

MATSUMOTO, Y. 2002. *An Introduction to Morse Theory*. American Mathematical Society.

RICHARDS. F. M. 1977. Areas volumes, packing and protein structures. *Ann. Rev. Biophys. Bioeng. 6*, 151–176.

SCHREIBER, G., AND FERSHT, A. R. 1995. Energetics of protein-protein interactions: Analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mole. Bio. 248*, 478–486.

SELZER, T., ALBECK, S., AND SCHREIBER, G. 2000. Rational design of faster associating and tighter binding protein complexes. *Nature Struct. Bio. 7*, 537–541.

SHAPIRO, R., RUIZ-GUTIERREZ, M., AND CHEN, C.-Z. 2000. Analysis of the interactions of human ribonuclease inhibitor with angiogenin and ribonuclease a by mutagenesis: Importance of inhibitor residues inside versus outside the C-terminal "hot-spot". *J. Mole. Bio. 302*, 497–519.

SHEINERMAN, F. B., AND HONIG, B. 2002. On the role of electrostatic interactions in the design of protein-protein interfaces. *J. Mole. Bio. 318*, 161–177.

VARSHNEY, A., BROOKS, JR., F. P., RICHARDSON, D. C., WRIGHT, W. V., AND MANOCHA, D. 1995. Definition, computing, and visualizing molecular interfaces. In *Proceedings of IEEE Visualization 1995*. IEEE Computer Society Press, Los Alamitos, CA, 36–43.

WELLS, J. A. 1996. Binding in the growth hormone receptor complex. *Proc. Nati. Acad. Sci. USA 93*, 1–6.

XU, D., TSAI, C. J., AND NUSSINOV, R. 1997. Hydrogen bonds and salt bridges across protein-protein interfaces. *Prot. Eng. 10*, 999–1012.

RECEIVED MAY 2004; REVISED MAY 2005 AND SEPTEMBER 2005; ACCEPTED FEBRUARY 2006