Visual Analysis of Biomolecular Surfaces

Vijay Natarajan[1], Patrice Koehl[2], Yusu Wang[3], and Bernd Hamann[4]

[1] Department of Computer Science and Automation
Supercomputer Education and Research Centre
Indian Institute of Science, Bangalore, India
vijayn@csa.iisc.ernet.in
[2] Department of Computer Science
Genome Center
University of California, Davis, California, USA
koehl@cs.ucdavis.edu
[3] Department of Computer Science and Engineering
The Ohio State University, Columbus, Ohio, USA
yusu@cse.ohio-state.edu
[4] Institute for Data Analysis and Visualization
Department of Computer Science
University of California, Davis, California, USA
hamann@cs.ucdavis.edu

Summary. Surface models of biomolecules have become crucially important for the study and understanding of interaction between biomolecules and their environment. We argue for the need for a detailed understanding of biomolecular surfaces by describing several applications in computational and structural biology. We review methods used to model, represent, characterize, and visualize biomolecular surfaces focusing on the role that geometry and topology play in identifying features on the surface. These methods enable the development of efficient computational and visualization tools for studying the function of biomolecules.

1 Introduction

The molecular basis of life rests on the activity of biological macro-molecules, including nucleic acids (DNA and RNA), carbohydrates, lipids and proteins. Although each plays an essential role in life, nucleic acids and proteins are central as support of the genetic information and products of this information, respectively. A perhaps surprising finding that crystallized over the last decades is that geometric reasoning plays a major role in our attempt to understand the activities of these molecules. We address this connection between biology and geometry, focusing on hard sphere models of biomolecules. In particular, we focus on the representations of biomolecular surfaces, and their applications in computational biology.

1.1 Significance of shape

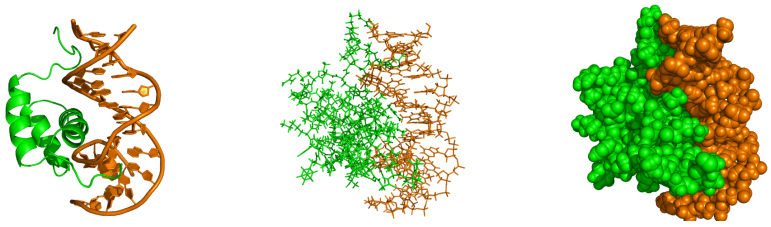
Molecular structure or shape and chemical reactivity are highly correlated as the latter depends on the positions of the nuclei and electrons within the molecule. Indeed, chemists have long used three-dimensional plastic and metal models to understand the many subtle effects of structure on reactivity and have invested in experimentally determining the structure of important molecules. The same applies to biochemistry where structural genomics projects are based on the premise that the structure of biomolecules implies their function. This premise rests on a number of specific and quantifiable correlations:

- enzymes fold into unique structures and the three-dimensional arrangement of their side-chains determines their catalytic activity;
- there is theoretical evidence that the mechanisms underlying protein complex formation depend mainly on the shapes of the biomolecules involved [1];
- the folding rate of many small proteins correlates with a gross topological parameter that quantifies the difference between distance in space and along the main-chain [2, 3, 4, 5];
- there is evidence that the geometry of a protein plays a role in defining its tolerance to mutation [6].

We note that structural biologists often refer to the 'topology' of a biomolecule when they mean the 'geometry' or 'shape' of the same. A common concrete model representing this shape is a union of balls, in which each ball corresponds to an atom. Properties of the biomolecule are then expressed in terms of properties of the union. For example, the potential active sites are detected as cavities [7, 8, 9] and the interaction with the environment is quantified through the surface area and/or volume of the union of balls [10, 11, 12]. In what follows, we discuss in detail the geometric properties of the surface of union of balls, their visualization, and their relation to the physical properties of the biomolecules they represent.

1.2 Biomolecules

Biomolecules are usually polymers of smaller subunits, whose atomic structures are known from standard chemistry. While physics and chemistry have provided significant insight into the structure of the atoms and their arrangements in small chemical structures, the focus now is set on understanding the structure and function of biomolecules, mainly nucleic acids and proteins. Our presentation of these molecules follow the general dogma in biology that states that the genetic information contained in DNA is first transcribed to RNA molecules which are then translated into proteins.



(a) (b) (c)

Fig. 1: **Visualizing protein-DNA complexes.** Homeodomains are small proteins that bind to DNA and regulate gene expression. Here we show the complex of the antennapedia homeodomain of *drosophila melanogaster* (fruit fly) and its DNA binding site [13], using three different types of visualization. The structure of this complex was determined by X-ray crystallography [13]; the coordinates are taken from the PDB file 1AHD. The protein is shown in green, and the DNA fragments in orange. (a) *Cartoon.* This representation provides a high level view of the local organization of the protein in secondary structures, shown as idealized helices. This view highlights the position of the binding site where the DNA sits. (b) *Skeletal model.* This representation uses lines to represent bonds; atom are located at their endpoints where the lines meet. It emphasizes the chemical nature of both molecules. (c) *Space-filling diagram.* Atoms are represented as balls centered at the atoms, with radii equal to the van der Waals radii of the atoms. This representation shows the tight binding between the protein and the ligand, that was not obvious from the other diagrams. Each of the representations is complementary to the others, and usually the biochemist uses all three of them when studying a protein, alone or, as illustrated here, in interaction with a ligand. All panels were drawn using Pymol (`http://www.pymol.org`)

DNA

The Deoxyribo Nucleic Acid is a long polymer built from four different building blocks, the nucleotides. The sequence in which the nucleotides are arranged contains the entire information required to describe cells and their functions. Despite this essential role in cellular functions, DNA molecules adopt surprisingly simple structures. Each nucleotide contains two parts, a backbone consisting of a deoxyribose and a phosphate, and an aromatic base, of which there are four types: adenine (A), thymine (T), guanine (G) and cytosine (C). The nucleotides are capable of being linked together to form a long chain, called a *strand*. Cells contain strands of DNA in pairs that are exact mirrors of each other. When correctly aligned, A can pair with T, G can pair with C, and the two strands form a double helix [14]. The geometry of this helix is surprisingly uniform, with only small, albeit important, structural differences between regions of different sequences. The order in which the nucleotides appear in one DNA strand defines its sequence. Some stretches of the sequence contain in-

formation that can be translated first into an RNA molecule and then into a protein. These stretches are called *genes*; the ensemble of all genes of an organism constitutes its *genome* or *genetic information*. The DNA strands can stretch for millions of nucleotides. The size of the strands vary greatly between organisms and do not necessarily reflect differences in the complexity of the organisms. For example, the wheat genome contains approximately $1.6 \cdot 10^{10}$ bases, which is close to five times the size of the human genome. For a complete list of the genomes, see `http://wit.integratedgenomics.com/GOLD/` [15].

RNA

Ribo Nucleic Acid molecules are very similar to DNA, being formed as sequences of four types of nucleotides, namely A, G, C, and uracil (U), which is a derivative of thymine. The sugar in the nucleotides of RNA is a ribose, which includes an extra oxygen compared to deoxyribose. The presence of this bulky extra oxygen prevents the formation of long and stable double helices. The single-stranded RNA can adopt a large variety of conformations, which remain difficult to predict based on its sequence. RNA molecules mainly serve as templates that are used to synthesize the active molecules, namely the proteins. The information needed to synthesize the RNA is read from the genes coded by the DNA. Interestingly, RNA is considered an essential molecule in the early steps of the origin of life. More information on the RNA world can be found in [16].

Proteins

While all biomolecules play an important part in life, there is something special about proteins, which are the products of the information contained in the genes. They are the active elements of life whose chemical activities regulate all cellular activities. As a consequence, studies of their sequence and structure occupy a central role in biology. Proteins are heteropolymer chains of amino acids, often referred to as *residues*. There are twenty types of amino acids, which share a common *backbone* and are distinguished by their chemically diverse *side-chains*, which range in size from a single hydrogen atom to large aromatic rings and can be charged or include only non-polar saturated hydrocarbons. The order in which amino acids appear defines the *primary sequence* of the protein. In its native environment, the polypeptide chain adopts a unique three-dimensional shape, referred to as the *tertiary* or *native structure* of the protein. In this structure, non-polar amino acids have a tendency to re-group and form the core of the proteins, while polar amino acids remain accessible to the solvent. The backbones are connected in sequence forming the protein *main-chain*, which frequently adopts canonical local shapes or *secondary structures*, such as α-helices and β-strands. ¿From the seminal work of Anfinsen [17], we know that the sequence fully determines

the three-dimensional structure of the protein, which itself defines its function. While the key to the decoding of the information contained in genes was found more than fifty years ago (the genetic code), we have not yet found the rules that relate a protein sequence to its structure [18, 19]. Our knowledge of protein structure therefore comes from years of experimental studies, either using X-ray crystallography or NMR spectroscopy. The first protein structures to be solved were those of hemoglobin and myoglobin [20, 21]. Currently, there are more than 37,000 protein structures in the database of biomolecular structures [22, 23]; see `http://www.rcsb.org`. More information on protein structures can be found in protein biochemistry textbooks, such as those of Branden and Tooze [24], and Creighton [25].

2 Visualizing Biomolecular Surfaces

The need for visualizing biomolecules is based on the early understanding that their shape determines their function. Early crystallographers who studied proteins and nucleic acids could not rely—as it is common nowadays—on computers and computer graphics programs for representation and analysis. They had developed a large array of finely crafted physical models that allowed them to have a feeling for these molecules. These models, usually made out of painted wood, plastic, rubber and/or metal were designed to highlight different properties of the molecule under study. In the *space-filling models*, such as those of Corey-Pauling-Koltun (CPK) [26, 27], atoms are represented as spheres, whose radii are the atoms' van der Waals radii. They provide a volumetric representation of the biomolecules, and are useful to detect cavities and pockets that are potential active sites. In the *skeletal models*, chemical bonds are represented by rods, whose junctions define the position of the atoms. These models were used for example by Kendrew and colleagues [20] in their studies of myoglobin. They are useful to the chemists by highlighting the chemical reactivity of the biomolecules and, consequently, their potential activity. With the introduction of computer graphics to structural biology, the principles of these models have been translated into software such that molecules could be visualized on the computer screen. Figure 1 shows examples of computer visualization of a protein-DNA interaction, including space-filling and skeletal representations.

Among all geometric elements of a biomolecule, its surface is probably the most important as it defines its interface, *i.e.*, its region of potential interactions with its environment. Here we limit ourselves to the definition of surface within the classical representation of biomolecules as union of balls. While other models are possible (such as a atom-based Gaussian descriptions [28], the hard-sphere model remains the most popular. Given the atom (ball) locations, the biomolecular surface can be defined in various ways. The first definition stated that the surface was simply the boundary of the union of balls. This surface, called the *van der Waals surface*, is easily computable

but not continuous [29]. The *solvent accessible surface* is the collection of points traced by the center of a probe sphere as it rolls on the van der Waals surface [30]. The accessible surface is equivalently defined as the boundary of the union of balls whose radii are expanded by the probe sphere radius. This surface is not smooth at curves where the expanded spheres meet. The *solvent excluded surface*, also called the *Connolly surface*, is defined as the surface traced by the front of a probe sphere [31]. This surface is continuous in most cases but can have cusp points and self-intersections. Figure 2 illustrates the definition of the above mentioned surfaces. Several algorithms have been proposed to construct analytic representations of the solvent excluded surface [32, 33, 34] and for triangulating the surface [35, 36].

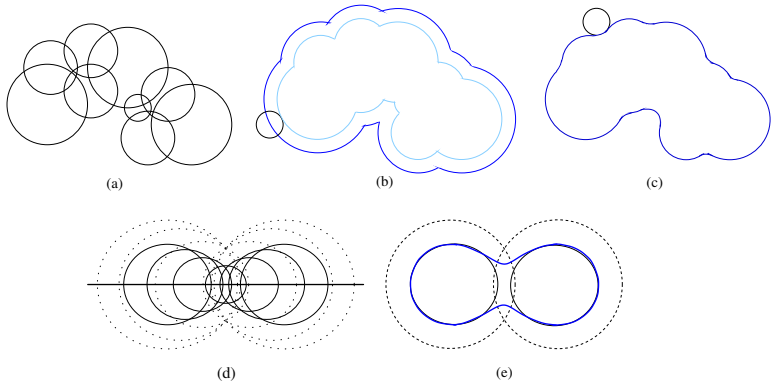


Fig. 2: (a) Each atom is modeled as a hard ball. (b) The van der Waals surface is the boundary of the union and the solvent accessible is the boundary of the union of expanded balls. Each atom is expanded by a value equal to the radius of the probe sphere. (c) The Connolly surface is traced by the front of the probe sphere. (d) The skin surface is defined as the envelope of an infinite set of spheres. Two atoms (outermost, dotted) define a family of spheres (dotted) obtained as the convex hull. Shrinking each sphere in the family results in yet another family of spheres (bold). (e) The skin surface, defined as the envelope of the shrunken family of spheres, is smooth everywhere. The radius of the two atoms is typically enlarged beforehand so that, upon shrinking, we get spheres whose radius equal the atom radius.

The *skin surface* is the envelope of families of an infinite number of evolving spheres [37]. It satisfies many desirable mathematical properties. For example, it is smooth everywhere and, although defined using an infinite number of spheres, it can be described by a finite number of quadric surface regions. Efficient algorithms have been developed recently to triangulate the skin surface [38, 39]. Besides being efficient in terms of computation time, these algorithms also generate watertight skin surface meshes (*i.e.*, without cracks) containing good quality triangles (*i.e.*, without small angles). The good quality of triangles leads to better visualizations of the surface. Watertight models

facilitate the application of volumetric analysis methods on the surface. For example, the volume occupied by the molecule, volume of voids, and their derivatives can be computed robustly given a crack-free biomolecular surface.

Multiresolution models of the mesh enables interactive visualization of the surface. Several methods have been developed to create level-of-detail representations of surface meshes [40, 41].

3 Significance of Biomolecular Surfaces

The activity of a biomolecule is encoded in its shape. Of all geometric properties of a molecule, its surface play an essential role as it delineates the region covered by the protein and therefore defines its region of interactions. Characterizing biomolecular surface therefore play an essential role for analyzing and predicting biomolecular complexes, as well as for modeling the energetics of formation of such complexes. As the surface of a molecule also defines its interface with the solvent it bathes in, the former is crucial for understanding solvation.

3.1 Solvent Models

The apparition of computers, and the rapid increase of their power has given hope that theoretical methods can play a significant role in biochemistry. Computer simulations are expected to predict molecular properties that are inaccessible to experimental probes, as well as how these properties are affected by a change in the composition of a molecular system. This has lead to a new branch in biology that works closely with structural biology and biochemistry, namely computational biology. Not surprisingly, an early and still essential focus of this new field is biomolecular dynamics [42, 43]. Soluble biomolecules adopt their stable conformation in water, and are unfolded in the gas phase. It is therefore essential to account for water in any modeling experiment. Molecular dynamics simulation that include a large number of solvent molecules are the state of the art in this field, but they are inefficient as most of the computing time is spent on updating the position of the water molecule. It should be noted that it is not always possible to account for the interaction with the solvent explicitly. For example, energy minimization of a system including both a protein and water molecules would not account for the entropy of water, which would behave like ice with respect to the protein. An alternative is to develop an approach in which the effect of the solvent is taken into account implicitly. In such implicit solvent models, the effects of water on a biomolecule is included in an effective solvation potential, $W = W_{elec} + W_{np}$, in which the first term accounts for the molecule-solvent electrostatics polarization, and the second term for the molecule-solvent van der Waals interactions and for the formation of a cavity in the solvent.

3.2 Electrostatics in implicit solvent models

Implicit solvent models reduce the solute solvent interactions to their mean-field characterization, which are expressed as a function of the solute degrees of freedom alone. They represent the solvent as a dielectric continuum that mimics the solvent-solute interactions. Many techniques have been developed to compute electrostatics energy in the context of dielectric continuum, including techniques that modify the dielectric constants in Coulomb's law, generalized Born models, and methods based on Poisson-Boltzmann equation (for a recent review, see [44]. A common element of all these techniques is that they need a good definition of the interface between the protein core and the dielectric continuum, *i.e.*, a good definition of the surface of the protein.

3.3 Non polar effects of solvent

W_{np}, the non-polar effect of water on the biomolecule is sometimes referred to as the *hydrophobic effect*. Biomolecules contain both hydrophilic and hydrophobic parts. In their folded states, the hydrophilic parts are usually at the surface where they can interact with water, and the hydrophobic parts are buried in the interior where they form a core (an "oil drop with a polar coat" [45]). In order to quantify this hydrophobic effect, Lee and Richards introduced the concept of the solvent-accessible surface [30]. They computed the accessible areas of each atom in both the folded and extended state of a protein, and found that the decrease in accessible area between the two states is greater for hydrophobic than for hydrophilic atoms. These ideas were further refined by Eisenberg and McLachlan [10], who introduced the concept of a solvation free energy, computed as a weighted sum of the accessible areas A_i of all atoms i of the biomolecule:

$$W_{np} = \sum_i \alpha_i A_i,$$

where α_i is the atomic solvation parameter. It is not clear however which surface area should be used to compute the solvation energy [46, 47, 48]. There is also some evidence that for small solute the hydrophobic term W_{np} is not proportional to the surface area [48], but rather to the solvent excluded volume of the molecule [49]. A volume-dependent solvation term was originally introduced by Gibson and Scheraga [50] as the hydration shell model. Within this debate on the exact form of the solvation energy, there is however a consensus that it depends on the geometry of the biomolecule under study. Inclusion of W_{np} in a molecular simulation therefore requires the calculation of accurate surface areas and volumes. If the simulations rely on minimization, or integrate the equations of motion, the derivatives of the solvation energy are also needed. It should be noted that calculation of the second derivatives are also of interest to study the normal modes of a biomolecule in a continuum solvent.

4 Feature-based Analysis and Visualization

To improve our understanding of a biomolecule and its functions, it is highly desirable to visualize various of its properties over its surface. Such visualization tools can for example help to emphasize important structural motifs and to reveal meaningful relations between the physiochemical information and the shape of a molecule.

One *general* framework for such visualization is as follows: the input molecular surface is considered as a *domain* $\mathbb{M}$, and one or more scalar functions $f_1, \ldots, f_k$ are defined over it, where each $f_i : \mathbb{M} \to \mathbb{R}$ is called a *descriptor function*. Such functions describe various properties that may be important, be it geometric, physiochemical, or any other type. We then visualize these descriptor functions over $\mathbb{M}$. Two key components involved here are (i) how to design meaningful descriptor functions and (ii) how to best visualize them. Below we briefly describe approaches in these two categories, focusing on topological methods[5].

4.1 Descriptor functions

Descriptors capturing physiochemical properties, such as the electrostatic potential and the local lipophilicity, are relatively easier to develop. Below we focus on molecular *shape* descriptors. In particular, most current molecular shape descriptors aim at capturing *protrusions and cavities* of the input structure, given that proteins function by interacting (binding) with other molecules, and there is a rough "lock-and-key" principle behind such binding [51] (see Figure 3 (a) for a 2D illustration).

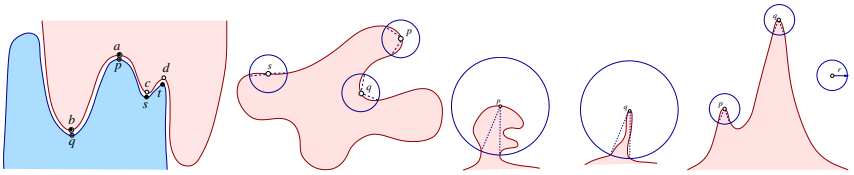
Curvature-based descriptors

The most natural choice to describe protrusions and cavities may be curvatures. A large family of molecular shape descriptors are based on curvatures. One of the most widely used one is the Connolly function [52, 53]. For any point $x \in \mathbb{M}$, consider the ball $\mathrm{B}_r(x)$ centered at x with radius r, and let $\mathrm{S}_r(x) = \partial \mathrm{B}_r(x)$ be the boundary of $\mathrm{B}_r(x)$, and S_I the portion of $\mathrm{S}_r(x)$ contained inside the surface. The *Connolly function* $f_r : \mathbb{M} \to \mathbb{R}$ is defined as (see Figure 3 (b) for a 2D illustration):

$$f_r(x) = \frac{\mathrm{Area}(\mathrm{S}_I)}{r^2}.$$

Roughly speaking, the Connolly function can be considered as an analog of the mean curvature within a fixed size neighborhood of each point [54]. A

[5] We note that both these components are widely studied in many other fields such as computer graphics, vision, and pattern recognition. We focus only on methods from the field of molecular biology.



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

Fig. 3: (a) The shape of two molecules complement each other at the interface. (b) In the 2D case, the Connolly function value is proportional to the angle spanned by the two intersection points between the circle and the curve. (c) p and q have the same Connolly function value, but the details within the neighborhood are different. (d) Using a fixed neighborhood radius is unable to distinguish the two features located at p and q.

large function value at $x \in \mathbb{M}$ means that the surface is concave around x, while a small one means that it is convex.

The Connolly function ignores the exact details of the surface contained in $\mathrm{B}_r(x)$. Hence it is insensitive to the two features pictured in Figure 3 (c). The *atomic density (AD)* function [55], $f_a : \mathbb{M} \to \mathbb{R}$, improves the Connolly function by taking a sequence of, say k, neighborhood balls around a point x with increasing radii, computing (roughly) the Connolly function with respect to each radius, and obtaining the final function value at x based on these k Connolly function values.

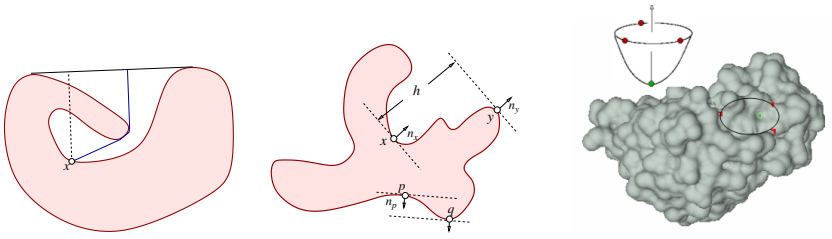
The concept of curvatures for 2-manifolds is more complicated than that of 1-manifolds — there are two principal curvatures at a given point. Several approaches have been proposed to combine these two curvatures into a single value to describe local surface features [56, 57, 58, 59, 60].

More global descriptors

The functions above are good at identifying points located at local protrusions and cavities. However, they all depend on a pre-fixed value r (the neighborhood size) — if r is small, then they may identify noise as important features; while if r is too large, then they may overlook interesting features. Furthermore, it is desirable that the function value can indicate the size (importance) of the feature that a point captures. However, none of the functions described above can measure the size of features directly (see Figure 3 (d)).

Since binding sites usually happen within cavities, it is natural to measure how *deep* a point $x \in \mathbb{M}$ is inside a cavity directly, independent of some neighborhood size. For example, a natural way to define such measures is as follows [61]. Given a molecular surface $\mathbb{M}$, let $\mathrm{CH}(\mathbb{M})$ be the convex hull of $\mathbb{M}$, and $\mathrm{Cover}(\mathbb{M}) = \mathrm{CH}(\mathbb{M}) \setminus \mathbb{M}$ intuitively covers the cavities of $\mathbb{M}$. One can

define $f_c : \mathbb{M} \to \mathbb{R}$ as $\mathrm{d}(x, \mathrm{Cover}(\mathbb{M}))$ if $x \notin \mathrm{CH}(\mathbb{M})$, and 0 otherwise; where $\mathrm{d}(x, X) = \min_{y \in X} \mathrm{d}(x, y)$ is the closest Euclidean distance from x to the set X.



(a) (b) (c)

Fig. 4: 2D illustrations of (a) $f_c(x)$ and the shortest collision-free distance, (b) elevation function (where $\mathrm{Elev}(x) = h$). (c) A local maximum of elevation function is associated with three more points to indicate the shape of a cave it captures.

This measure however ignores details inside cavities. A better measure is probably by using the shortest *collision-free* distance from a point x to the convex cover (Figure 4 (a)), which intuitively measures how difficult it is for a solvent molecule to access a specific point on the molecular surface. However, the computation of collision-free shortest path is expensive, and we are not aware of any result applying it to molecular surfaces yet. Furthermore, such concavity measures is asymmetric in measuring convexity. The *elevation function* $\mathrm{Elev} : \mathbb{M} \to \mathbb{R}$, developed in [62, 63] is independent of any pre-fixed parameter, and can measure both convexity and convexity in a meaningful manner.

Specifically, each point $x \in \mathbb{M}$ is paired with a canonical pairing partner y that shares the same normal direction n_x with x, and the function value $\mathrm{Elev}(x)$ is equal to the height difference between x and y in direction n_x. See Figure 4 (b) for a 2D illustration. The identification of the canonical pairing partners is based on a topological framework, called the *persistence algorithm*, developed by Edelsbrunner *et al.* [64]. Roughly speaking, the persistence algorithm provides a meaningful way to uniquely pair up critical points of a given function f, each pair specifies some feature with respect to f, and its *persistence* value indicates the size of this feature by measuring how long it can persist as one changes the function value locally.

The elevation function has several nice properties that make it appealing for various applications. In some sense, it finds a different and appropriate r for every point x on the surface and the function value $\mathrm{Elev}(x)$ roughly measures the depth of the cave (or protrusion) captured by x in its normal direction

(Figure 4 (b)). More interestingly, each extreme point of the elevation function is associated with pairing partners that helps the user to visualize the feature located at this point (Figure 4 (c)).

The computation of the elevation function is unfortunately costly. The function also has some discontinuity that may be undesirable when segmenting the input surface based on it. In general, there is no universally good descriptor functions. Designing meaningful and easy-to-compute molecular shape descriptors for visualization and structure characterization purposes is still in great need.

4.2 Visualizing scalar functions

To enhance the visualization of a given scalar function, one natural approach is to highlight features, such as the critical points of the input function. Another widely used technique is to segment the input surface into meaningful regions, which we focus on below.

Region-growing

Given a function $f : \mathbb{M} \to \mathbb{R}$, one family of surface segmentation methods is based on the region-growing idea [56, 65, 59]. In particular, certain *seeds* are first selected, and neighboring points are gradually clustered into the regions around these seeds. Merging and splitting operations are performed to compute regions of appropriate sizes, and/or to obtain a hierarchical representation of segmentations. For example, it is common to use the critical points of the input function as seeds. When two regions meet each other, criteria such as the size of each region and/or the difference between the function values of points from two neighboring regions decide whether to merge these two regions or not.

The advantage of such methods is that the criteria used to merge regions is not limited to the input function — such as using the size of current segment, or even combing multiple descriptor functions into the criteria. Thus they are able to create more versatile types of segments. On the other hand, the decision of merging/splitting is usually locally made and ad hoc, thus may not be optimal globally. Furthermore, various parameters control the output, such as the order of processing different regions, and it is not trivial to identify the best strategy for choosing these parameters.

Topological methods

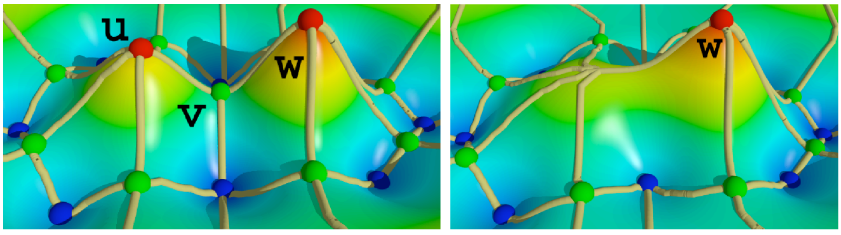
A second family of segmentation algorithms is based on Morse theory [66, 67, 68]. Such topological frameworks usually produce a hierarchical representation of segmentation easily, and are also more general — a segmentation is induced for any given function $f : \mathbb{M} \to \mathbb{R}$, with usually no parameter other than the one to specify the resolution of the segmentation.

Notation

Given a smooth 2-manifold $\mathbb{M} \subseteq \mathbb{R}^3$ and a scalar function $f : \mathbb{M} \to \mathbb{R}$, a point on $\mathbb{M}$ is *critical* if the gradient of f at this point is zero. f is a *Morse function* if none of its critical points are degenerate, that is, the Hessian matrix is non-singular for all critical points, and no two critical points have the same function value. For a Morse function defined on a 2-manifold, there are three types of critical points: minima, saddle points, and maxima. In molecular biological applications, a molecular surface is typically represented as a (triangular) mesh K, and the input scalar function $f : K \to \mathbb{R}$ over K is piecewise-linear (PL): f is given at the vertices and linearly interpolated within edges and triangles of K. The type of a critical point p of such a PL function can be determined by inspecting the *star* of p, which consists of all triangles and edges containing p [69].

Morse complex and Morse-Smale complex

An integral line of f is a maximal path on the surface $\mathbb{M}$ whose tangent vectors agree with the gradient of f at every point of the path. Integral lines have a natural origin and destination at critical points where the gradient equals zero. Grouping the integral lines based on their origin and destination results in a segmentation of the surface. The Morse-Smale (MS) complex [67] is a topological data structure that stores this segmentation (see Figure 5). A characteristic property of this segmentation is that every cell of the MS complex is monotonic (i.e, it does not contain any critical point in its interior).

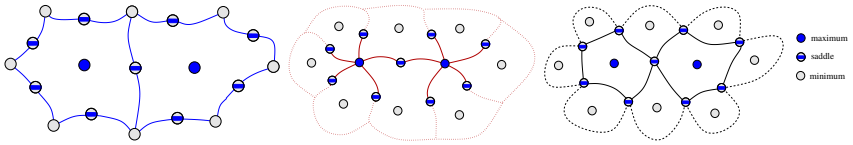


(a) (b)

Fig. 5: (a) A simple height function with two maxima surrounded by multiple local minima and its Morse-Smale complex. (b) Smoother height functions are created by canceling pairs of critical points. Canceling a saddle-maximum pair removes a topological feature. The function is modified locally by rerouting integral lines to the remaining maximum.

Alternatively, grouping integral lines based exclusively on their origin or destination results in yet another segmentation called the Morse complex (see

Figure 6). The MS complex can also be obtained as an overlay of the two Morse complexes for the set of maxima and minima, respectively.



(a) (b) (c)

Fig. 6: Morse complex computed for the set of all (a) maxima and (b) minima. (c) The Morse-Smale complex is obtained as an overlay of these two Morse complexes. Paths connecting saddles within each cell of the MS complex segment the surface into peaks and valleys.

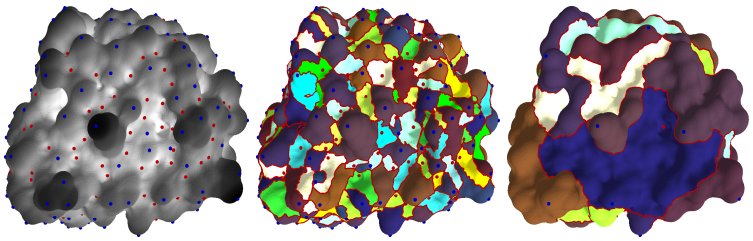
Peak-valley decomposition

The Morse complex for the set of all maxima results in a segmentation where peaks (regions around maxima) are separated (Figure 6 (a)). However, these segments extend all the way to adjoining valleys (regions around minima). The MS complex, on the other hand, does more refinement than necessary. In many cases, it is desirable to segment input surface into peaks and valleys. To this end, Natarajan *et al.* proposed an extension of the MS complex to compute such segmentations [68]. In particular, each cell in the MS complex is a *quad* containing one maximum, one minimum, and two saddles on its boundary (Figure 5 (a)). Connecting the two saddles will bisects this quad. Bisecting all quads that contain a specific maximum u, we get all regions that constitute the peak containing u. Similarly, we can obtain valleys containing a minimum v. In other words, these saddle-saddle paths describe the boundary between peaks and their adjoining valleys, and the resulting segmentation is called *peak-valley decomposition* (see Figure 6 (c)). Various criteria have been explored in [68] for the construction of such saddle-saddle paths.

Hierarchical segmentation

A major advantage of using the MS complex as a starting point for segmentation is that we can segment the surface at multiple levels of detail. A smoother Morse function can be generated from f by repeated cancellation of pairs of critical points. Each cancellation makes the function smoother by removing a topological feature. This operation can be implemented by a local modification of the gradient vector field when the two critical points are connected by a common arc in the MS complex [66]. Figure 5 (b) shows how the MS complex in Figure 5 (a) is modified after canceling a saddle-maximum pair.

The order of critical point pairs is guided by the notion of *persistence* [64], which quantifies the importance of the associated topological feature. The peak-valley decomposition can be computed at multiple levels of detail [68] by first canceling critical point pairs in the MS complex and then constructing saddle-saddle paths within the simplified complex. See Figure 7 for an example where we visualize the peak-valley segmentation of a protein molecule (chain D from the protein complex Barnase-Barstar with pdb-id 1BRS) based on the atomic density function at different levels of details. There are on-going work to use such segmentation to help to characterize protein binding sites.



(a) (b) (c)

Fig. 7: (a) The atomic density function computed for chain D of the Barnase-Barstar complex. Darker regions correspond to protrusions and lighter regions to cavities. (b) Peak-valley segmentation of the surface. (c) Coarse segmentation obtained by performing a series of critical pair cancellations.

5 Conclusions

We have provided an overview of selected methods based on well-established concepts from differential geometry, computational geometry and computational topology to characterize complex biomolecular surfaces. We have discussed how such concepts are relevant in the context of studying the complex interaction behaviors between biomolecules and their surroundings. Mathematically sound approaches to the analysis of intricate biochemical processes have become increasingly important to make progress in the study of protein folding and protein-protein interactions. Some of the most exciting and challenging questions that remain to be answered include dynamic biomolecular behavior, where surface analysis techniques like the ones discussed by us here need to be generalized substantially to support effective visualization and analysis of rapidly changing shapes.

Acknowledgments

Work done by Vijay Natarajan and Bernd Hamann was supported in part by the National Science Foundation under contracts ACI 9624034 (CAREER Award), a large Information Technology Research (ITR) grant, and the Lawrence Livermore National Laboratory under sub-contract agreement B551391. Vijay Natarajan was also supported by a faculty startup grant from the Indian Institute of Science. Patrice Koehl acknowledges support from the National Science Foundation under contract CCF-0625744. Yusu Wang was supported by DOE under grant DE-FG02-06ER25735 (Early Career Award). We thank members of the Visualization and Computer Graphics Research Group at the Institute for Data Analysis and Visualization (IDAV) at the University of California, Davis.

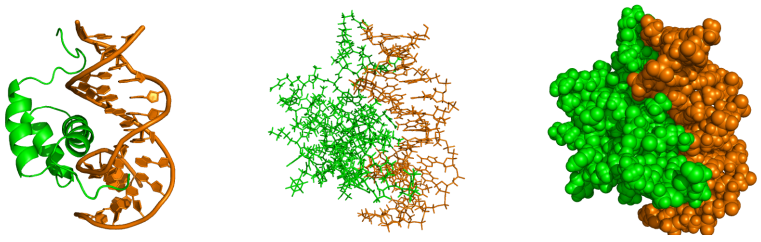
References

1. Y. Levy, P. G. Wolynes, and J. N. Onuchic. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. (USA)*, 101:511–516, 2004.
2. E. Alm and D. Baker. Prediction of protein-folding mechanisms from free energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. (USA)*, 96:11305–11310, 1999.
3. V. Muñoz and W. A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. (USA)*, 96:11311–11316, 1999.
4. K. T. Simons K. W. Plaxco and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277:985–994, 1998.
5. E. Alm, A. V. Morozov, T. Kortemme, and D. Baker. Simple physical models connect theory and experiments in protein folding kinetics. *J. Mol. Biol.*, 322:463–476, 2002.
6. P. Koehl and M. Levitt. Protein topology and stability defines the space of allowed sequences. *Proc. Natl. Acad. Sci. (USA)*, 99:1280–1285, 2002.
7. J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Prot. Sci.*, 7:1884–1897, 1998.
8. H. Edelsbrunner, M. A. Facello, and J. Liang. On the definition and construction of pockets in macromolecules. *Discrete Appl. Math.*, 88:83–102, 1998.
9. J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules. II. Inaccessible cavities in proteins. *Proteins: Struct. Func. Genet.*, 33:18–29, 1998.
10. D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature (London)*, 319:199–203, 1986.
11. T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga. Accessible surface-areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. (USA)*, 84:3086–3090, 1987.

12. J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules. I. Molecular area and volume through alpha shape. *Proteins: Struct. Func. Genet.*, 33:1–17, 1998.
13. M. Billeter, Y.Q. Qian, G. Otting, M. Muller, W. Gehring, and K. Wuthrich. Determination of the nuclear magnetic resonance solution structure of an antennapedia homeodomain-dna complex. *J. Mol. Biol.*, 234:1084–1093, 1993.
14. J. D. Watson and F. H. C. Crick. A structure for Deoxyribose Nucleic Acid. *Nature (London)*, 171:737–738, 1953.
15. A. Bernal, U. Ear, and N. Kyrpides. Genomes online database (GOLD): a monitor of genome projects world-wide. *Nucl. Acids. Res.*, 29:126–127, 2001.
16. R. F. Gesteland and J. A. Atkins. *The RNA World: the nature of modern RNA suggests a prebiotic RNA world*. Cold Spring Harbor Laboratory Press, Plainview, NY, 1993.
17. C. B. Anfinsen. Principles that govern protein folding. *Science*, 181:223–230, 1973.
18. P. Koehl and M. Levitt. A brighter future for protein structure prediction. *Nature Struct. Biol.*, 6:108–111, 1999.
19. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.
20. J. Kendrew, R. Dickerson, B. Strandberg, R. Hart, D. Davies, and D. Philips. Structure of myoglobin: a three dimensional Fourier synthesis at 2 angstrom resolution. *Nature (London)*, 185:422–427, 1960.
21. M. Perutz, M. Rossmann, A. Cullis, G. Muirhead, G. Will, and A. North. Structure of hemoglobin: a three-dimensional Fourier synthesis at 5.5 angstrom resolution, obtained by X-ray analysis. *Nature (London)*, 185:416–422, 1960.
22. F. C. Bernstein, T. F. Koetzle, G. William, D. J. Meyer, M. D. Brice, J. R. Rodgers, et al. The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.
23. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, et al. The Protein Data Bank. *Nucl. Acids. Res.*, 28:235–242, 2000.
24. C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, New York, NY, 1991.
25. T. E. Creighton. *Proteins. Structures and Molecular Principles*. Freeman, New York, NY, 1984.
26. R. B. Corey and L. Pauling. Molecular models of amino acids, peptides and proteins. *Rev. Sci. Instr.*, 24:621–627, 1953.
27. W. L. Koltun. Precision space-filling atomic models. *Biopolymers*, 3:665–679, 1965.
28. J.A. Grant and B.T. Pickup. A Gaussian description of molecular shape. *J. Phys. Chem.*, 99:3503–3510, 1995.
29. M. L. Connolly. Molecular surface: A review. *Network Science*, 1996.
30. B. Lee and F. M. Richards. Interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.
31. F. M. Richards. Areas, volumes, packing and protein structure. *Ann. Rev. Biophys. Bioeng.*, 6:151–176, 1977.
32. M. L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.
33. T. J. Richmond. Solvent accessible surface area and excluded volume in proteins. *J. Molecular Biology*, 178:63–89, 1984.

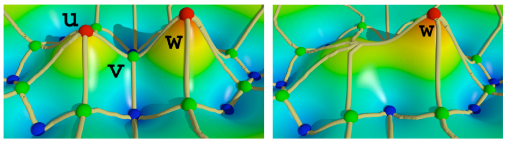
34. A. Varshney and F. P. Brooks Jr. Fast analytical computation of richard's smooth molecular surface. In *Proc. IEEE Visualization*, pages 300–307, 1993.
35. M. L. Connolly. Molecular surface triangulation. *J. Appl. Cryst.*, 18:499–505, 1985.
36. N. Akkiraju and H. Edelsbrunner. Triangulating the surface of a molecule. *Discrete Applied Mathematics*, 71:5–22, 1996.
37. H.-L Cheng, T. K. Dey, H. Edelsbrunner, and J. Sullivan. Dynamic skin triangulation. *Discrete Comput. Geom.*, 25:525–568, 2001.
38. H. L. Cheng and X. Shi. Guaranteed quality triangulation of molecular skin surfaces. In *Proc. IEEE Visualization*, pages 481–488, 2004.
39. H. L. Cheng and X. Shi. Quality mesh generation for molecular skin surfaces using restricted union of balls. In *Proc. IEEE Visualization*, pages 399–405, 2005.
40. H. Hoppe. Progressive meshes. In *ACM SIGGRAPH*, pages 99–108, 1996.
41. M. Garland. Multiresolution modeling: survey and future opportunities. In *Eurographics State of the Art Report*, 1999.
42. T. E. Cheatham and P. A. Kollman. Molecular dynamics simulation of nucleic acids. *Ann. Rev. Phys. Chem.*, 51:435–471, 2000.
43. M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Struct. Biol.*, 9:646–652, 2002.
44. P. Koehl. Electrostatics calculations: latest methodological advances. *Curr. Opin. Struct. Biol.*, 16:142–151, 2006.
45. W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.
46. R. H. Wood and P. T. Thompson. Differences between pair and bulk hydrophobic interactions. *Proc. Natl. Acad. Sci. (USA)*, 87:946–949, 1990.
47. I. Tunon, E. Silla, and J. L. Pascual-Ahuir. Molecular-surface area and hydrophobic effect. *Protein Eng.*, 5:715–716, 1992.
48. T. Simonson and A. T. Brünger. Solvation free-energies estimated from macroscopic continuum theory: an accuracy assessment. *J. Phys. Chem.*, 98:4683–4694, 1994.
49. K. Lum, D. Chandler, and J. D. Weeks. Hydrophobicity at small and large length scales. *J. Phys. Chem. B.*, 103:4570–4577, 1999.
50. K. D. Gibson and H. A. Scheraga. Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease s-peptide. *Proc. Natl. Acad. Sci. (USA)*, 58:420–427, 1967.
51. E. Fischer. Einfluss der configuration auf die wirkung derenzyme. *Ber. Dtsch. Chem. Ges.*, 27:2985–2993, 1894.
52. M. L. Connolly. Measurement of protein surface shape by solid angles. *J. Mol. Graphics*, 4:3 – 6, 1986.
53. M. L. Connolly. Shape complementarity at the hemo-globin albl subunit interface. *Biopolymers*, 25:1229–1247, 1986.
54. F. Cazals, F. Chazal, and T. Lewiner. Molecular shape analysis based upon the Morse-Smale complex and the Connolly function. In *Proc. 19th Annu. ACM Sympos. Comput. Geom.*, 2003.
55. J. C. Mitchell, R. Kerr, and L. F. Ten Eyck. Rapid atomic density measures for molecular shape characterization. *J. Mol. Graph. Model.*, 19:324–329, 2001.
56. D. A. Cosgrove, D. M. Bayada, and A. J. Johnson. A novel method of aligning molecules by local surface shape similarity. *J. Comput-Aided Mol Des*, 14:573–591, 2000.

57. B. S. Duncan and A. J. Olson. Approximation and characterization of molecular surfaces. *Biopolymers*, 33:219–229, 1993.
58. B. S. Duncan and A. J. Olson. Shape analysis of molecular surfaces. *Biopolymers*, 33:231–238, 1993.
59. T. E. Exner, M. Keil, and J. Brickmann. Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. *J. Comput. Chem.*, 23:1176–1187, 2002.
60. W. Heiden and J. Brickmann. Segmentation of protein surfaces using fuzzy logic. *J. Mol. Graphics.*, 12:106–115, 1994.
61. J. Lien and N. M. Amato. Approximate convex decomposition of polyhedra. Technical report, Technial Report TR05-001, Texas A&M University, 2005.
62. P. K. Agarwal, H. Edelsbrunner, J. Harer, and Y. Wang. Extreme elevation on a 2-manifold. In *Proc. 20th Ann. Sympos. Comput. Geom.*, pages 357–365, 2004.
63. Y. Wang, P. Agarwal, P. Brown, H. Edelsbrunner, and J. Rudulph. Fast geometric algorithm for rigid protein docking. In *Proc. 10th. Pacific Symposium on Biocomputing (PSB)*, pages 64–75, 2005.
64. H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.
65. D. Duhovny, R. Nussinov, and H. J. Wolfson. Efficient unbound docking of rigid molecules. In *WABI '02: Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, pages 185–200, 2002.
66. P. T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. A topological hierarchy for functions on triangulated surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 10(4):385–396, 2004.
67. H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete and Computational Geometry*, 30(1):87–107, 2003.
68. V. Natarajan, Y. Wang, P. Bremer, V. Pascucci, and B. Hamann. Segmenting molecular surfaces. *Computer Aided Geometric Design*, 23:495–509, 2006.
69. T. F. Banchoff. Critical points and curvature for embedded polyhedral surfaces. *American Mathematical Monthly*, 77(5):475–485, 1970.



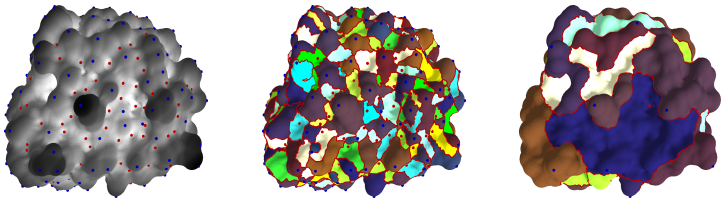
(a) (b) (c)

Fig. 8: **Visualizing protein-DNA complexes.** Complex of the antennapedia homeodomain of *drosophila melanogaster* (fruit fly) and its DNA binding site shown using three different types of visualization. The protein is shown in green, and the DNA fragments in orange. (a) The cartoon view highlights the position of the binding site where the DNA sits. (b) The skeletal model emphasizes the chemical nature of both molecules. (c) The space-filling diagram shows the tight binding between the protein and the ligand. Each representation is complementary to the others, and the biochemist uses all three of them when studying a protein.



(a) (b)

Fig. 9: (a) A simple height function with two maxima surrounded by multiple local minima and its Morse-Smale complex. (b) Smoother height functions are created by canceling pairs of critical points. Canceling a saddle-maximum pair removes a topological feature. The function is modified locally by rerouting integral lines to the remaining maximum.



(a) (b) (c)

Fig. 10: (a) The atomic density function computed for chain D of the Barnase-Barstar complex. Darker regions correspond to protrusions and lighter regions to cavities. (b) Peak-valley segmentation of the surface. (c) Coarse segmentation obtained by performing a series of critical pair cancellations.