On the Definition and the Construction of Pockets in Macromolecules*

Herbert Edelsbrunner †, Michael Facello ‡ and Jie Liang §

October 4, 1995

Abstract

The shape of a protein is important for its functions. This includes the location and size of identifiable regions in its complement space. We formally define pockets as regions in the complement with limited accessibility from the outside. Pockets can be efficiently constructed by an algorithm based on alpha complexes. The algorithm is implemented and applied to proteins with known three-dimensional conformations.

Keywords. Combinatorial geometry and topology, algorithms, molecular biology; molecular modeling, docking, space filling and solvent accessible models, Voronoi cells, Delaunay simplices, alpha complexes.

1 Introduction

The motivation for the work reported in this paper is the apparent difficulty to talk in mathematically concrete terms about intuitive geometric concepts sometimes referred to as 'depressions', 'canyons', 'cavities', and the like. In topology, the notions of homotopy and homology have long been used to define and study (perfect) holes of various types and dimensions. We are after a definition and study of imperfect holes, of regions people would instinctively refer to as holes although they are neither holes in the homotopical nor the homological sense.

Observations about common language reveal a great deal of confusion on what holes are. A hole in the ground is usually a depression deep or big enough so we would care about its existence. The fact we can fall into but not through it reveals it is not a hole in a topological sense. Or consider exploding a balloon by poking through its surface with a needle. The needle connects the hole holding the balloon's air with the outside. Topologically, poking a needle through the surface removes rather than creates a hole.

*This work is partially supported by the Office of Naval Research, grant N00014-95-1-0691, and by the National Science Foundation through the Alan T. Waterman award, grant CCR-9118874, and the CISE postdoctoral fellowship, NSF grant ASC-9404900.

†Department of Computer Science, University of Illinois at Urbana-Champaign, USA, and Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong.

‡Department of Computer Science and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, USA.

§Biophysics Division of the School of Life Sciences, National Center for Supercomputing Applications, and Department of Computer Science, University of Illinois at Urbana-Champaign, USA.

Pockets in proteins. The study of imperfect holes in this paper focuses on proteins and other macromolecules. The ideas are more general though and can be extended to other 3-dimensional shapes and to higher dimensions.

The functions of a protein are determined through its interaction with other molecules. Such interactions happen frequently in protected yet accessible regions of appropriate size and shape. The shape complimentary between such a protected binding site and the ligand is largely responsible for the specificity observed in protein-ligand/protein interactions. There are also the less frequent situations where the binding ligand sits in an isolated cavity/void and is completely engulfed by the protein (such as the Xe binding sites in myoglobin). For such cases, we refer to our earlier results in cavity/void identification and their area and volume measurements [9]. The above intuitive but vague description of protein binding pockets is certainly not sufficient to distinguish protected regions from unprotected ones, or to specify the precise location and extent of a protected region once it is identified. In this paper, we will formally define pockets as regions in the complement space with limited accessibility from the outside. The definition deliberately excludes shallow valleys or depressions. Although there are also binding sites of the latter type, their determination will either require a priori knowledge or an extension of the ideas described in this paper.

Intuition. The following intuition guides our formulation of an unambiguous criterion. We declare a region in the complement a pocket if it can be reached only via relatively narrow pathways: "all paths into the pocket get narrow before they get wider". This intuition can also be captured through a growth process: "a pocket becomes a void inaccessible from the outside before it disappears".

It is clear that considerations based on relative distance are required to make this intuition concrete and algorithmically useful. Such considerations are expressed in terms of Voronoi cells [18] and Delaunay simplices [5]. These are key concepts in this paper and they play a crucial role in defining, delimiting, and algorithmically constructing pockets. The algorithm is implemented and sample applications to proteins whose coordinates are available from the protein databank are given.

Outline. Section 2 discusses common sphere models of molecules and their relationship to Voronoi cells. Section 3 describes dual sets and complexes of simplices. Section 4 defines pockets based on an acyclic relation over the Delaunay tetrahedra. Section 5 presents an algorithm constructing pockets. Section 6 applies the implementation of the algorithm to a few proteins with known 3-dimensional conformations. Section 7 mentions possible extensions of this work and directions for further research.

2 Spherical Ball Models

It is common in biology to represent an atom by a spherical ball and a molecule by a union of balls. Geometric models of this type go back to Lee and Richards [14] and Richards [16]. For a fixed set of atom centers, the *space filling* model uses van der Waals radii, see e.g. [4, chapter 4], to unambiguously specify the balls and thus their union. The *solvent accessible* model increases radii to reflect accessibility for a solvent, itself modeled as a spherical ball. This section introduces the geometric terminology necessary to talk about these models and their relationship to Voronoi cells.

Distance and growth. The Euclidean distance between points $x, y \in \mathbb{R}^3$ is denoted by $|xy|$, and the *(spherical) ball* with center $z \in \mathbb{R}^3$ and radius $r \in \mathbb{R}$ is

$$b(z,r) \quad = \quad \{x \in \mathbb{R}^3 \mid |xz| \leq r\}.$$

The union of a finite set B of balls is $\bigcup B = \{x \in \mathbb{R}^3 \mid x \in b \in B\}$. The complement, $\mathbb{R}^3 - \bigcup B$, consists of one or more components. Exactly one component is unbounded and usually referred to as the *outside*. The other components are bounded and referred to as *voids* of $\bigcup B$. Figure 1 shows the union of a set of 2-dimensional balls or circular disks.

Figure 1: The union of 16 disks is connected and decomposes its complement into 1 unbounded component (the outside) and 2 bounded components (voids).

The solvent accessible model differs from the space filling model by the size of the balls; the centers are the same. This suggests we consider the union while growing the balls continuously and simultaneously. As the balls grow the union grows and the voids shrink. Which voids appear depends on the relative growth. We find it convenient to grow the balls such that the circles where two spheres meet sweep out planes.

The growth is controlled by a real parameter α^2. Formally, we choose α from $\mathbb{R}^{\frac{1}{2}}$, that is, α is either a non-negative real or it is a positive real multiple of the imaginary unit, $\sqrt{-1}$. Define $b_\alpha(z,r) = b(z, \sqrt{r^2 + \alpha^2})$ and

$$B_\alpha \quad = \quad \{b_\alpha(z,r) \mid b(z,r) \in B\}.$$

If $r^2 + \alpha^2 < 0$, the radius is imaginary and $b_\alpha = b_\alpha(z,r)$ is empty. In this case, b_α does not contribute to $\bigcup B_\alpha$ but it does influence the formation of pockets. This makes sense since we argue pockets are regions that will *become* voids in the future. Future is defined in the direction of increasing α^2, and b_α is born when α^2 passes $-r^2$.

Voronoi cells. Define the *distance* of a point $x \in \mathbb{R}^3$ from a ball $b = b(z,r)$ as $\pi_b(x) = |zx|^2 - r^2$ and note it is defined even if $r^2 < 0$. In general, $x \in b$ iff $\pi_b(x) \leq 0$. The *Voronoi cell* of $b \in B$ is

$$V_b \quad = \quad \{x \in \mathbb{R}^3 \mid \pi_b(x) \leq \pi_c(x), c \in B\}.$$

In words, V_b is the set of points x at least as close to b as to any other ball in B. Define $\mathrm{Vor}\, B = \{V_b \mid b \in B\}$. The set of points with equal distance from two balls form a plane. It follows V_b is the intersection of finitely many closed half-spaces and hence a convex polyhedron.

Figure 2: The 16 disks in figure 1 define a decomposition of $\mathbb{R}^2$ into 16 Voronoi cells.

Voronoi cells overlap at most along their boundary, and together they cover the entire space: $\mathbb{R}^3 = \bigcup \text{Vor}\, B$, see figure 2. The vertices, edges, and facets of the Voronoi cells are referred to as *Voronoi vertices*, *Voronoi edges*, and *Voronoi facets*. It is convenient to assume general position so a Voronoi edge belongs to exactly 3 Voronoi cells and a Voronoi vertex belongs to exactly 4 Voronoi cells.

Observe a point $x \in \mathbb{R}^3$ is simultaneously contained in a ball $c \in B$ and the Voronoi cell V_b of $b \neq c$ only if $\pi_b(x) \leq \pi_c(x) \leq 0$. This implies $x \in b$. In other words, $V_b \cap \bigcup B = V_b \cap b$ for every $b \in B$. The sets $R_b = V_b \cap b$ are convex and any two overlap at most along their boundary. Define $\text{Res}\, B = \{R_b \mid b \in B\}$ and note it covers the union of balls: $\bigcup B = \bigcup \text{Res}\, B$, see figure 4.

The growth process is defined so Voronoi cells do not change. Indeed, $\pi_b(x) \leq \pi_c(x)$ iff $\pi_{b_\alpha}(x) \leq \pi_{c_\alpha}(x)$, and therefore $\text{Vor}\, B = \text{Vor}\, B_\alpha$ for every $\alpha \in \mathbb{R}^{\frac{1}{2}}$. This will be important later when we take advantage of the fact the same Voronoi cells decompose every $\bigcup B_\alpha$ into convex cells.

3 Simplex Collections

The connectivity of a union of balls can be expressed by a collection of simplices that keeps track of which cells R_b overlap. This collection is used to represent the union. Similarly, sets of simplices are used to represent voids and later pockets. We begin by introducing some general terminology.

Simplicial complexes. An *abstract simplicial complex* is a finite system of sets, $\mathcal{A}$, with $X \in \mathcal{A}$ and $Y \subseteq X$ implying $Y \in \mathcal{A}$. $X \in \mathcal{A}$ is referred to as an *abstract simplex* and its *dimension* is $\dim X = \text{card}\, X - 1$. The *vertex set* is $\text{Vert}\, \mathcal{A} = \bigcup \mathcal{A}$. A *subcomplex* is an abstract simplicial complex $\mathcal{B} \subseteq \mathcal{A}$. For example, if S is any finite set, then the *nerve* of S,

$$\text{Nrv}\, S \quad = \quad \{X \subseteq S \mid \bigcap X \neq \emptyset\},$$

is an abstract simplicial complex with vertex set S. The nerve of every subset of S is a subcomplex of $\text{Nrv}\, S$. More generally, if S' is a set and $i : S' \to S$ is an injection with $a' \subseteq i(a')$ for each $a' \in S'$ then $\text{Nrv}\, S'$ is isomorphic to a subcomplex of $\text{Nrv}\, S$. Indeed, $\mathcal{B} = \{X \subseteq S \mid X = i(X'), X' \in \text{Nrv}\, S'\}$ is clearly a subcomplex of $\text{Nrv}\, S$ and isomorphic to $\text{Nrv}\, S'$.

Every abstract simplicial complex, $\mathcal{A}$, can be realized geometrically by a collection of simplices in $\mathbb{R}^d$, for some finite dimension d. The elements of $\mathrm{Vert}\,\mathcal{A}$ are represented by points, and an abstract simplex, $X \in \mathcal{A}$, is represented by the convex hull of the corresponding points. Provided d is large enough, the points can always be chosen so the convex hull is a simplex of dimension $\dim X$ and no two simplices intersect improperly. Formally, let $\iota : \mathrm{Vert}\,\mathcal{A} \rightarrow \mathbb{R}^d$ be an injection so

$$\mathrm{conv}\,\iota(X) \cap \mathrm{conv}\,\iota(Y) \quad = \quad \mathrm{conv}\,\iota(X \cap Y)$$

for all $X, Y \in \mathcal{A}$. The resulting set of simplices,

$$\mathcal{K} \quad = \quad \{\mathrm{conv}\,\iota(X) \mid X \in \mathcal{A}\},$$

is a *(geometric) simplicial complex*. The *underlying space* of $\mathcal{K}$ is $|\mathcal{K}| = \bigcup \mathcal{K}$. A *subcomplex* of $\mathcal{K}$ is a set $\{\mathrm{conv}\,\iota(X) \mid X \in \mathcal{B}\}$, $\mathcal{B}$ a subcomplex of $\mathcal{A}$.

Delaunay simplices. We form simplices by taking convex hulls of 1, 2, 3, or 4 ball centers. The collection of such simplices reflecting the overlap relation among the Voronoi cells is a complex which is now formally introduced.

Let B be a finite set of balls in $\mathbb{R}^3$, assume general position, and recall $\mathrm{Vor}\,B$ is the set of Voronoi cells. The nerve of $\mathrm{Vor}\,B$ is of course an abstract simplicial complex. It is geometrically realized by mapping each Voronoi cell to the center of the generating ball. Formally, let $\iota : \mathrm{Vor}\,B \rightarrow \mathbb{R}^3$ be defined by $\iota(V_b) = z$ if $b = b(z, r)$. The *Delaunay simplicial complex* of B is

$$\mathrm{Del}\,B \quad = \quad \{\mathrm{conv}\,\iota(X) \mid X \in \mathrm{Nrv}\,\mathrm{Vor}\,B\},$$

see figure 3. General position implies $\mathrm{Del}\,B$ is indeed a simplicial complex. The simplices

Figure 3: The Delaunay simplicial complex of the 16 disks in figure 1.

$\sigma \in \mathrm{Del}\,B$ are referred to as *Delaunay simplices*.

Consider a tetrahedron $\tau = \mathrm{conv}\,\iota(X)$ in $\mathrm{Del}\,B$. The 4 Voronoi cells in X intersect at a point $z_\tau = \bigcap X$ referred to as the *orthogonal center* of τ. Let b_1, b_2, b_3, b_4 be the balls generating the Voronoi cells in X. By construction, the distance of z_τ from the balls is the same:

$$r_\tau^2 = \pi_{b_1}(z_\tau) = \pi_{b_2}(z_\tau) = \pi_{b_3}(z_\tau) = \pi_{b_4}(z_\tau).$$

The *radius* of τ is r_τ and the *orthogonal ball* is $b_\tau = (z_\tau, r_\tau)$. The name suggests b_τ meets the b_i in some ways orthogonally. Indeed, for a point on two intersecting spheres, $x = \mathrm{bd}\,b_\tau \cap \mathrm{bd}\,b_i$, the two tangent planes passing though x meet at a right angle.

Alpha complexes. The union of balls covers only a portion of the Voronoi cells, and this portions is represented by a subcomplex of the Delaunay simplicial complex, see [12].

Recall the definitions of $R_b = V_b \cap b$ and $\mathrm{Res}\, B = \{R_b \mid b \in B\}$. The nerve of $\mathrm{Res}\, B$ is an abstract simplicial complex that can be geometrically realized by mapping cells to ball centers, the same way as before. Let $\iota : \mathrm{Res}\, B \to \mathbb{R}^3$ be defined by $\iota(R_b) = b'$. The *dual complex* of $\bigcup B$ is

$$\mathrm{Cpx} B \quad = \quad \{\mathrm{conv}\, \iota(X) \mid X \in \mathrm{Nrv}\, \mathrm{Res}\, B\},$$

see figure 4. Clearly, $\mathrm{Nrv}\, \mathrm{Res}\, B$ is isomorphic to a subcomplex of $\mathrm{Nrv}\, \mathrm{Vor}\, B$, and therefore

Figure 4: The union of disks in figure 1 is decomposed into convex cells. The dual complex connects 2 centers by an edge and 3 centers by a triangle if the corresponding cells have non-empty common intersection. The union of disks has 2 voids, each contained in a void of the dual complex.

$\mathrm{Cpx} B \subseteq \mathrm{Del}\, B$. The dual complex inherits the property of being a simplicial complex from $\mathrm{Del}\, B$.

We refer to [7] for a list of properties $\mathrm{Cpx} B$ enjoys. This includes $\mathrm{Cpx} B$ is homotopy equivalent to $\bigcup B$. More precisely, $|\mathrm{Cpx} B| \subseteq \bigcup B$ and there is a deformation retraction that takes $\bigcup B$ to $|\mathrm{Cpx} B|$. The same is true for the respective complements. More precisely, each void of $\bigcup B$ is contained in a void of $|\mathrm{Cpx} B|$ and there is a deformation retraction that takes $\mathbb{R}^3 - |\mathrm{Cpx} B|$ to $\mathbb{R}^3 - \bigcup B$.

Recall the definition of B_α, which is obtained by simultaneously growing or shrinking all balls in B. The *α-complex* of B is the dual complex of $\bigcup B_\alpha$: $\mathrm{Cpx}_\alpha B = \mathrm{Cpx} B_\alpha$. For $\alpha_1^2 \leq \alpha_2^2$ we have $b_{\alpha_1} \subseteq b_{\alpha_2}$, which implies

$$\{\emptyset\} \subseteq \mathrm{Cpx}_{\alpha_1} B \subseteq \mathrm{Cpx}_{\alpha_2} B \subseteq \mathrm{Del}\, B.$$

The bounds are tight. For sufficiently small α^2 all balls have imaginary radius and are empty, which implies $\mathrm{Cpx}_\alpha B = \{\emptyset\}$. For sufficiently large α^2 the nerves of $\mathrm{Res}\, B$ and $\mathrm{Vor}\, B = \mathrm{Vor}\, B_\alpha$ are isomorphic, which implies $\mathrm{Cpx}_\alpha B = \mathrm{Del}\, B$.

The dual set of a void. Recall a void of $\bigcup B$ is a bounded component of the complement. To be specific, let

$$\mathbb{R}^3 - \bigcup B \quad = \quad H_0 \,\dot{\cup}\, H_1 \,\dot{\cup}\, \ldots \,\dot{\cup}\, H_k$$

be the partition into maximal connected subsets. Assume H_0 is unbounded and H_1 through H_k are the voids of $\bigcup B$. As mentioned earlier, there is a deformation retraction that takes the complement of $|\,\mathrm{Cpx}\,B\,|$ to the complement of $\bigcup B$. Let

$$\mathbb{R}^3 - |\,\mathrm{Cpx}\,B\,| \quad = \quad H'_0 \,\dot{\cup}\, H'_1 \,\dot{\cup}\, \ldots \,\dot{\cup}\, H'_k$$

be the partition into components so the above mentioned deformation retraction takes H'_i to H_i, see figure 4. The voids of $|\,\mathrm{Cpx}\,B\,|$ are naturally represented by the simplices in $\mathrm{Del}\,B - \mathrm{Cpx}\,B$ that cover them. For $1 \leq i \leq k$ the *dual set* of H_i is

$$\mathcal{H}_i \quad = \quad \{\sigma \in \mathrm{Del}\,B \mid \mathrm{int}\,\sigma \subseteq H'_i\}.$$

For example, the smaller of the two voids in figure 4 has a dual set consisting of 2 triangles and 1 edge. The dual set of the larger void consists of 4 triangles and 3 edges. As shown in [7], the volume and surface area of a void H_i can be computed directly from $\mathcal{H}_i$, without explicit construction of H_i.

4 Pockets

The concept of a pocket is based on an acyclic relation over the set of Delaunay tetrahedra motivated by a continuous flow field. After defining and classifying pockets we compare them with related concepts in the literature.

Flow relation. Let T' be the set of tetrahedra in $\mathrm{Del}\,B$ and $T = T' \cup \{\tau_\infty\}$, where $\tau_\infty = \mathrm{cl}\,(\mathbb{R}^3 - |\,\mathrm{Del}\,B\,|)$ is considered a convenient dummy element. We define the *flow relation* '$\prec$' $\subseteq T \times T$ with $\tau \prec \sigma$ if

(i) τ and σ share a common triangle, φ, and

(ii) $\mathrm{int}\,\tau$ and the orthogonal center z_τ of τ lie on different sides of the plane $\mathrm{aff}\,\varphi$.

The conditions makes sense for $\sigma = \tau_\infty$ but not for $\tau = \tau_\infty$. The flow relation is acyclic because $\tau \prec \sigma$ implies $r_\tau^2 < r_\sigma^2$ or $\sigma = \tau_\infty$. In words, the radius of the orthogonal ball increases with the flow relation. This is the intuition behind the flow or vector field that motivates the definition of '$\prec$': a point flows in the direction of the closest orthogonal ball.

If $\tau \prec \sigma$ we call τ a *predecessor* of σ and σ a *successor* of τ. The set of *descendents* of τ is

$$\mathrm{des}\,\tau \quad = \quad \{\tau\} \cup \bigcup_{\tau \prec \sigma \in T} \mathrm{des}\,\sigma,$$

and the set of *ancestors* of σ is

$$\mathrm{anc}\,\sigma \quad = \quad \{\sigma\} \cup \bigcup_{\sigma \succ \tau \in T} \mathrm{anc}\,\tau.$$

$\sigma \in T$ is a *sink* if it has no successors, or equivalently $\mathrm{des}\,\sigma = \{\sigma\}$. τ_∞ is necessarily a sink. A tetrahedron $\sigma \in T'$ is a sink iff it contains its orthogonal center: $z_\sigma \in \sigma$. In general, σ cannot have more than 3 successors because z_σ can be on the other side of at most 3 of the 4 triangles bounding σ.

Sinks are important since they are responsible for the formation of voids. Indeed, if H_i is a void of $\bigcup B$ then at least one tetrahedron in $\mathcal{H}_i$ is a sink. This follows from the observation that $\tau \in \mathcal{H}_i$ and $\tau \prec \sigma$ implies $\sigma \in \mathcal{H}_i$. If $\sigma \in T$ is a sink that belongs to $\mathcal{H}_i$ then $z_\sigma \in H_i$ and $r_\tau^2 > 0$. The radii of sinks thus predict the moment in time H_i will disappear, namely when α reaches the maximum radius of any sink in $\mathcal{H}_i$. Of course, before H_i disappears it may break up into several voids, each with at least one sink.

Pockets. The combinatorial notions of closure, interior, and boundary motivate analogous combinatorial notions applicable to sets of simplices. The *closure* of a subset L of a simplicial complex $\mathcal{K}$ is $\mathrm{Cl}\, L = \{\tau \in \mathcal{K} \mid \tau \subseteq \sigma \in L\}$; it is the smallest subcomplex that contains L. The *star* of $\tau \in \mathcal{K}$ is $\mathrm{St}\, \tau = \{\sigma \in \mathcal{K} \mid \tau \subseteq \sigma\}$. $L \subseteq \mathcal{K}$ is *open* in $\mathcal{K}$ if $\mathrm{St}\, \tau \subseteq L$ for every $\tau \in L$. The *interior* of a subset $L \subseteq \mathcal{K}$ is $\mathrm{Int}\, L = \{\tau \in L \mid \mathrm{St}\, \tau \subseteq L\}$; it is the largest open set contained in L. The *boundary* of L is $\mathrm{Bd}\, L = \mathrm{Cl}\, L - \mathrm{Int}\, L$. An open set is *connected* if it cannot be partitioned into two non-empty disjoint open sets. The *components* are the maximal connected open subsets.

As mentioned earlier, the intention is to define pockets so they are generalizations of voids, possibly with connections to the outside. The relation over the tetrahedra decides which side tetrahedra belong and the divide forms the connection to the outside. More precisely, pockets consist of the Delaunay tetrahedra that do not belong to $\mathrm{Cpx} B$ and that are not ancestors of τ_∞. Define $\mathcal{P} = \mathrm{Int}\,\mathrm{Cl}\,(T - \mathrm{anc}\,\tau_\infty) - \mathrm{Cpx} B$, and let

$$\mathcal{P} \quad = \quad \mathcal{P}_1 \,\dot{\cup}\, \mathcal{P}_2 \,\dot{\cup}\, \ldots \,\dot{\cup}\, \mathcal{P}_k$$

be the partition into components. For each $1 \leq i \leq k$,

$$P_i \quad = \quad \bigcup \mathcal{P}_i - \bigcup B$$

is a *pocket* of $\bigcup B$, and $\mathcal{P}_i$ is its *dual set*. These definitions are illustrated in figure 5.

Figure 5: The 16 disks are obtained by shrinking the disks in figure 1; 3 of them have now imaginary radii. There are 2 pockets each grown from one of the voids in figure 1. Consult figures 2 and 3 to see that 5 Delaunay triangles are ancestors of τ_∞. All other triangles belong to $\mathcal{P}$ and none to the dual complex of the disk union. The component of 4 disks in the middle of the picture defines a chain of 4 vertices and 3 edges in the dual complex. This chain separates $\mathcal{P}$ into 2 components, each defining a pocket.

The above definition of pockets treats the unbounded component special and different from the voids. Sometimes this may not be appropriate and large voids are to be treated the same way as the unbounded component. This can formally be done by bounding the radii of the sinks used in the construction. For a size limit $\beta^2 \in \mathbb{R}$ define $T_\beta = \{\tau_\infty\} \cup \{\sigma \in T' \mid r_\sigma^2 > \beta^2\}$ and

$$\mathcal{P}_\beta \quad = \quad \mathrm{Int}\,\mathrm{Cl}\,(T - \textstyle\bigcup_{\sigma \in T_\beta} \mathrm{anc}\,\sigma) - \mathrm{Cpx} B.$$

As before, the subset of $\mathbb{R}^3 - \bigcup B$ covered by the interiors of the simplices in a component of $\mathcal{P}_\beta$ is a pocket, and the component is its dual set, see figure 6.

Figure 6: The upper bound on the sink radii used for the example shown excludes sinks whose orthogonal centers are not covered by the disk union in figure 1. As a result, the 2 pockets in figure 5 are reduced to 5 smaller pockets.

Mouth openings. The only type of pockets without connection to the outside are the voids. All other pockets connect to the outside at one or more places. For a pocket P_i consider the part of $\mathrm{Bd}\,\mathcal{P}_i$ not contained in $\mathrm{Cpx}\,B$. $\mathrm{Bd}\,\mathcal{P}_i$ is a simplicial complex and connectedness and components relative to $\mathrm{Bd}\,\mathcal{P}_i$ are well defined for all its open subsets. The mentioned set is indeed open in $\mathrm{Bd}\,\mathcal{P}_i$ and we let

$$\mathrm{Bd}\,\mathcal{P}_i - \mathrm{Cpx}\,B \quad = \quad \mathcal{M}_1 \,\dot{\cup}\, \mathcal{M}_2 \,\dot{\cup}\, \ldots \,\dot{\cup}\, \mathcal{M}_\ell$$

be the partition into components. The *mouths* of P_i are the sets $M_j = \bigcup \mathcal{M}_j - \bigcup B$, for $1 \leq j \leq \ell$, and their *dual sets* are the $\mathcal{M}_j$. Consider for example the two pockets in figure 5. The left and smaller pocket has 3 mouths, each defined by a single Delaunay edge. The right and bigger pocket has 4 mouths, 3 defined by a single Delaunay edge each and 1 defined by a chain of 2 Delaunay edges and 1 Delaunay vertex.

The number of mouths, ℓ, is a useful characteristic of a pocket and can be used to distinguish between different types. One would expect a pocket with different number of mouths in a protein implies different functionalities. We suggest the following terminology reflecting the resulting classification. Call a pocket a

void	if $\ell = 0$,
normal pocket	if $\ell = 1$,
simple connector	if $\ell = 2$, and
multiple connector	if $\ell \geq 3$.

In the presence of a size limit one can furthermore distinguish between connectors whose mouths connect to the same or to different components of the outside.

Related concepts. The computational biology literature contains at least 3 concepts defined as tools to study regions of limited accessibility. These are the 'molecular surface', the 'interstitial skeleton', and the 'molecular interface'. We briefly point out the similarities and differences between pockets and these concepts. The authors of this paper believe pockets are superior to all 3 concepts in terms of visual appearance, objective quantification, and wide applicability.

The molecular surface model defined by Richards [16] is a union of balls, $\bigcup B$, where gaps inaccessibly to a sphere modeling a solvent are filled. Let $MS \supseteq \bigcup B$ be the resulting object. The union of pockets is similar to albeit not the same as the difference, $MS - \bigcup B$, union all voids of MS. While pockets are defined in terms of relative distance, the criterion employed for defining molecular surface uses absolute distance, namely the radius of the solvent. Furthermore, the object obtained from MS is cluttered with tiny remains within the crevices and cusps of $\bigcup B$. Pockets do not share this visual distraction.

The interstitial skeleton defined by Connolly [3] consists of all Voronoi edges outside $\bigcup B$ and within the convex hull of the balls. A problematic feature of this concept is the lack of any possibility to clip edges inside delta regions where a depression opens up slowly towards the outside. Another disadvantage is the mess of edges that possibly attracts the eye to large pockets, but they offer little in terms of objective quantification.

The molecular interface has recently been suggested by Varshney and coauthors [17] to study the region between interacting molecules. It assumes 2 or more different molecules and consists of the points outside all molecules at distance at most ε from at least 2 of the molecules. ε is a parameter that can be chosen and adjusted. A shortcoming of this definition is its lack of dependence on any local shape characteristic. Also, it cannot be used to study depressions in a single molecule. On the other hand, pockets are easily adjusted to study the interface: compute pockets for the union of the molecules and select only the ones that touch at least 2 different molecules.

5 Algorithm

We construct pockets by growing them from sinks. We assume a pointer based data structure for $\mathrm{Del}\, B$ and a linear list that distinguishes between Delaunay simplices inside and outside an alpha complex. Both data structures are part of the alpha shape software [10], which forms the basis of our implementation. The entire software is based on exact arithmetic and the simulation of general position by infinitesimal perturbation [11]. We begin by describing the two data structures in sufficient detail to provide the context for the construction of pockets.

Simplex digraph. We refer to the pointer based data structure for $\mathrm{Del}\, B$ as the *simplex digraph*. It supports access to neighboring simplices in constant time each. Data structures with this functionality are reasonably standard and different versions have been described in the literature, see e.g. [1, 6].

The simplices of $\mathrm{Del}\, B$ are the nodes of the digraph, and they are referenced through pointers. Each simplex has direct access to its location in the linear list or filter, see below. In order to avoid a tedious discussion of the details of the simplex digraph, we stipulate functions FACES and COFACES that provide access to the neighborhood. Given a simplex $\sigma \in \mathrm{Del}\, B$ and a dimension $k < \dim \sigma$, FACES returns the k-dimensional faces:

$$\mathrm{FACES}(\sigma, k) \quad = \quad \{\tau \in \mathrm{Cl}\,\{\sigma\} \mid \dim \tau = k\}.$$

For $k > \dim \sigma$, COFACES returns the k-dimensional simplices that share σ as a face:

$$\mathrm{COFACES}(\sigma, k) \quad = \quad \{\tau \in \mathrm{St}\, \sigma \mid \dim \tau = k\}.$$

It is convenient to assume $\mathrm{COFACES}(\sigma, 3)$ includes τ_∞ if σ lies on the boundary of $|\,\mathrm{Del}\, B\,|$. We assume both functions take constant time per returned simplex.

As an example consider the problem of computing the set $N(\sigma)$ of tetrahedra adjacent to a given tetrahedron $\sigma \in \mathrm{Del}\, B$.

$N(\sigma) := \emptyset$;

```
for all φ ∈ FACES(σ, 2) do
    for both τ ∈ COFACES(φ, 3) do
        if τ ≠ σ then
            N(σ) := N(σ) ∪ {τ}
        endif
    endfor
endfor.
```

The first loop is over 4 triangles and the second over 2 tetrahedra each, so the total time for finding all adjacent tetrahedra is constant.

Filter and filtration. The Delaunay simplices are stored in the order they enter the alpha complex. We assume an array representation with constant time access via indices.

Recall the α_1-complex of B is a subcomplex of the α_2-complex if $\alpha_1^2 \leq \alpha_2^2$. It follows the infinite sequence of α^2 defines a sequence of nested complexes. Two consecutive complexes differ by one or more Delaunay simplices, and the cardinality of $\mathrm{Del}\, B$ is an upper bound on the number of complexes in the sequence. We represent the sequence by a list of simplices sorted in the order they enter. We break ties by letting vertices precede edges precede triangles precede tetrahedra. Remaining ties are broken arbitrarily. The resulting sequence of simplices,

$$\emptyset = \sigma_0, \sigma_1, \sigma_2, \ldots, \sigma_n,$$

is a *filter* of $\mathrm{Del}\, B$. The array is a representation of the filter, with pointers linking simplices to their locations in the simplex digraph. Each prefix of the filter defines a simplicial complex, $\mathcal{K}_i = \{\sigma_0, \sigma_1, \ldots, \sigma_i\}$. The resulting sequence of complexes,

$$\{\emptyset\} = \mathcal{K}_0, \mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_n = \mathrm{Del}\, B,$$

is a *filtration* of $\mathrm{Del}\, B$. For each $\alpha^2 \in \mathbb{R}$ there is an index $i(\alpha)$ with $\mathrm{Cpx}_\alpha B = \mathcal{K}_{i(\alpha)}$, but not necessarily vice versa.

Suppose we wish to construct the pockets of $\bigcup B_\alpha$, or rather their dual sets. The general idea is to traverse the latter part of the filter, from σ_{i+1} to σ_n. The algorithm is incremental, and after processing the simplices in $\mathcal{K}_j$ the data structures represent the pockets for the corresponding size limit. Each encountered tetrahedron either joins the outside, joins a set of delayed tetrahedra because it does not belong to the current set of pockets, or starts a new pocket and possibly merges some of the existing pockets into one. The delayed tetrahedra will be added at the appropriate time.

Representing pockets. The pockets are stored as sets of tetrahedra in an evolving system, Υ, represented by a *union-find* data structure. The sets in Υ are pairwise disjoint and the data structure supports the following operations:

ADD(u) :	Add $\{u\}$ as a new set to Υ.
SET(u) :	Find set $X \in \Upsilon$ with $u \in X$.
UNION(X, Y) :	Replace sets X and Y by $X \cup Y$.

A sequence of m operations takes time $\mathrm{O}(m\alpha(m))$, where $\alpha(m)$ is the extremely slowly growing inverse of Ackermann's function, see e.g. [2, chapter V]. For all practical purposes, $\alpha(m)$ can be considered a small constant.

In our application, the elements in the system are tetrahedra. Υ is initialized to $\{\{\tau_\infty\}\}$. SET(τ_∞) represents the outside and is the only set in Υ that does not represent a pocket.

Traversing the filter. The index of a simplex specifies its position in the filter. If σ_j is a tetrahedron its *depth* is

$$\begin{aligned} \mathrm{dp}\,\sigma_j &= \max\{k \mid \sigma_k \in \mathrm{des}\,\sigma_j\} \\ &= \max(\{j\} \cup \{\mathrm{dp}\,\sigma \mid \sigma_j \prec \sigma\}). \end{aligned}$$

The depth determines the minimum size limit from which moment on the tetrahedron belongs to the set of pockets. The recursive specification of depth lends itself to computing all depth values in a single traversal of the filter.

```
for j := n downto 1 do
    dp σ_j := j;
    for all τ ∈ N(σ_j) do
        if σ_j ≺ τ then
            dp σ_j := max{dp σ_j, dp τ}
        endif
    endfor
endfor.
```

Pockets are constructed by following the evolution of the ball growth. Only tetrahedra σ_j with $i(\alpha) < j \leq i(\beta)$ need to be considered, and such a σ_j belongs to $\mathcal{P}_\beta$ iff $\mathrm{dp}\,\sigma_j \leq i(\beta)$. When the traversal reaches σ_j, all tetrahedra with depth j are added to the union-find system representing the pockets. These tetrahedra are collected in an initially empty set Y_j. At the time Y_j is processed it may or may not contain σ_j.

```
for j := i(α) to i(β) do
    k := dp σ_j; Y_k := Y_k ∪ {σ_j};
    for all σ ∈ Y_j do
        ADD(σ);
        for all τ ∈ N(σ) with τ ∈ ⋃Υ do
            UNION(SET(σ), SET(τ))
        endfor
    endfor
endfor.
```

Note the test whether or not the tetrahedron τ belongs to any set in Υ that occurs in the inner `for`-loop. For $\tau = \sigma_k$ the test is equivalent to $i(\alpha) < k$ and $\mathrm{dp}\,\tau < j$.

Dual sets of pockets and mouths. The traversal constructs a pocket P as a set of tetrahedra. To compute the dual set, $\mathcal{P}$, we still need to take the closure of this set, then the interior, and remove simplices in the dual complex of $\bigcup B$. Similarly, to get the dual sets of the mouths, we need to take the boundary, remove simplices in the dual complex, and collect components. We first describe the process for pockets and then for mouths.

Let $X \in \Upsilon$ be the collection of tetrahedra defining P. The closure $\mathcal{C} = \mathrm{Cl}\,X$ is obtained by collecting all faces, with a straightforward marking mechanism to avoid duplication:

```
C := X ∪ {∅};
for all τ ∈ X do C := C ∪ FACES(τ, 2)
    ∪ FACES(τ, 1) ∪ FACES(τ, 0)
endfor.
```

To construct the interior, we use the fact a vertex or edge in $\mathcal{C}$ belongs to $\mathcal{I} = \mathrm{Int}\,\mathrm{Cl}\,X$ iff all triangles in its star belong to $\mathcal{I}$.

$\mathcal{I} := \mathcal{C} - \{\emptyset\}$;
`for` all triangles $\varphi \in \mathcal{I}$ `do`
    `for` both $\tau \in \text{COFACES}(\varphi, 3)$ `do`
        `if` $\tau \notin X$ `then` $\mathcal{I} := \mathcal{I} - \{\varphi\}$
            $- \text{FACES}(\varphi, 1) - \text{FACES}(\varphi, 0)$
        `endif`
    `endfor`
`endfor.`

The dual set of P is finally obtained by removing all simplices from $\mathcal{I}$ whose indices in the filter are less than $i + 1$.

The dual sets of the mouths M_j are the components $\mathcal{M}_j$ of $\text{Bd}\,\mathcal{P} - \text{Cpx}B$. Every boundary simplex of $\mathcal{P}$ belongs to $\mathcal{B} = \text{Bd}\,\text{Cl}\,X = \mathcal{C} - \mathcal{I}$ or to $\text{Cpx}B$ or to both. We can therefore work with $\mathcal{B}$, which can be constructed along with $\mathcal{I}$ by the above algorithm.

$\mathcal{B}$ is a 2-dimensional connected manifold because $\mathcal{I}$ is connected. This means every edge belongs to exactly 2 triangles and the star of every vertex is an alternating cycle of edges and triangles. The $\mathcal{M}_j$ are the components of $\mathcal{B} - \text{Cpx}B$. They are computed in a way analogous to the computation of the dual sets of pockets, only in one dimension lower. First, traverse the triangles $\varphi \in \mathcal{B}$ and collect the ones outside $\text{Cpx}B$ in a system represented by a union-find data structure. Whenever a triangle is added, check the 3 adjacent triangles and merge sets if they are already in the system. In the end, each set Y in the system contains the triangles of a mouth M_j. The dual set of M_j is $\mathcal{M}_j = \text{Int}\,\text{Cl}\,Y$.

6 Protein Examples

Tunnel extraction for Gramacidin A. Gramacidin A is a synthetic membrane channel and has been used as an antibiotic. It is composed of D and L amino acid residues in alternating order.

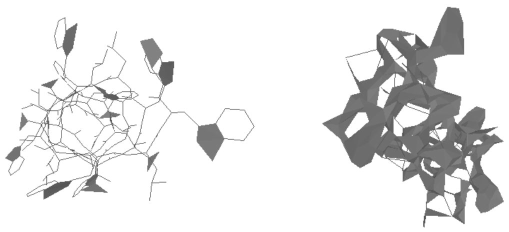

Figure 7: The alpha shape of gramacidin A reflecting the topological structure of the molecule.

Figure 7 shows the alpha complex of the molecule when $\alpha = 0$. Figure 8 shows that the tunnel of the potassium channel is extracted by the pocket construction of gramacidin A.

Pocket in HIV-I protease for inhibitor. HIV-I protease displayed in figure 9 is important for the maturation of HIV-I virus. Its complexed structure with VX–478 inhibitor has been recently solved [13]. Atoms of the HIV-I protease that are in solvent contact with the inhibitor can be identified by comparing the complexes for the bound and unbound states. The inhibitor binding site is a pocket and can be seen from the alpha complex on the left of

Figure 8: The pocket constrcuted from gramacidin A. It is a simple channel connector.

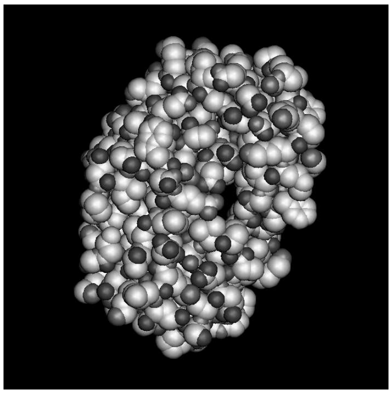

Figure 9: HIV-I protease shown in van der Waals model.

figure 10. This binding site is a simple connector with 2 mouths. The pocket of the binding site is constructed and shown on the right. Note the complementary nature of the two shapes.

Heme pocket of the myoglobin. Myoglobin is the protein that carries oxygen in muscle cells, providing the oxygen necessary for cell metabolism. Figure 11 shows the alpha complex of the apoprotein as well as the heme pocket. The heme binding site is in the pocket that can be seen from the alpha complex. The dual set of the pocket is constructed and depicted on the right hand side. Note that unlike the previous example, this is a normal pocket with a dead end.

7 Discussion and Extensions

Initial experiments have shown that the algorithm for computing pockets described in this paper cannot find shallow pockets. In systems of large molecules, shallow pockets can occur quite frequently. One possible solution to this problem is an additional parameter specifying 'steepness' or 'depth' that will add finer control over the inclusion or exclusion of the tetrahedra that flow to τ_∞.

The concept of a pocket can be applied to the complementary space of a macromolecule thus defining protrusions of the molecule. An appropriate notion of complementarity is described in [8]. The authors of this paper expect that pockets and protrusion together provide a good handle on predicting docking pairs and sites.

Figure 10: The alpha complex of HIV-I protease and the inhibitor binding pocket.

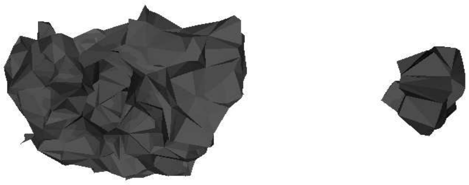

Figure 11: The alpha complex of myoglobin and the dual set of the heme binding pocket.

The notion of limited accessibility arises also in studies of shapes in other fields. For example, Miller [15] uses it to compute realistic shadings of statues. Notions of local and global accessibility are related to molecular surfaces and to pockets. The algorithmic techniques in this paper can be used to improve the performance of the algorithms in [15] by orders of magnitudes.

Acknowledgements

The authors thank Shankar Subramaniam for discussions on the relevance of pockets to the study of macromolecules, Ping Fu and Ernst Mücke for their contributions to the alpha shapes software in which the pockets software is embedded, and Siu-Wing Cheng for valuable comments on an early version of this paper.

References

[1] E. Brisson. Representing geometric structures in d dimensions: topology and order. *Discrete Comput. Geom.* **9** (1993), 387–426.

[2] T. H. Cormen, Ch. E. Leiserson and R. L. Rivest. *Introduction to Algorithms.* MIT Press, Cambridge, Mass., 1990.

[3] T. H. Connolly. Molecular interstitial skeleton. *Computer Chem.* **15** (1991), 37–45.

[4] T. E. Creighton. *Proteins. Structures and Molecular Principles.* Freeman, New York, 1984.

[5] B. Delaunay. Sur la sphère vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* **7** (1934), 793–800.

[6] D. P. Dobkin and M. J. Laszlo. Primitives for the manipulation of three-dimensional subdivisions. *Algorithmica* **4** (1989), 3–32.

[7] H. Edelsbrunner. The union of balls and its dual shape. *László Fejes Tóth Festschrift*, eds. I. Bárány and J. Pach, *Discrete Comput. Geom.* **13** (1995), 415–440.

[8] H. Edelsbrunner. Smooth surfaces for multi-scale shape representation. To appear in "Proc. 15th Conf. Software Techn. Theoret. Comput. Sci., 1995", Bangalore, India.

[9] H. Edelsbrunner, M. Facello, P. Fu and J. Liang. Measuring proteins and voids in proteins. *In* "Proc. 28th Hawaii Intern. Conf. Syst. Sci., 1995", 256–264.

[10] H. Edelsbrunner, M. Facello, P. Fu and E. P. Mücke (devs.). "Three-dimensional alpha shapes". Software developed at the Univ. Illinois at Urbana-Champaign, Illinois, 1991–95, `ftp.ncsa.uiuc.edu`.

[11] H. Edelsbrunner and E. P. Mücke. Simulation of Simplicity: a technique to cope with degenerate cases in geometric algorithms. *ACM Trans. Graphics* **9** (1990), 66–104.

[12] H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graphics* **13** (1994), 43–72.

[13] E. E. Kim, C. T. Baker, M. D. Dwyer, M. A. Murcko, B. G. Rao, R. D. Tung and M. A. Navia. Crystal structure of HIV-1 protease in complex with VX–478, a potent and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.* **117** (1995), 1181–1182.

[14] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55** (1971), 379-400.

[15] G. Miller. Efficient algorithms for local and global accessibility shading. *Computer Graphics* **28** (1994), 319–326.

[16] F. M. Richards. Areas, volumes, packing, and protein structures. *Ann. Rev. Biophys. Bioeng.* **6** (1977), 151–176.

[17] A. Varshney, F. P. Brooks, Jr., D. C. Richardson, W. V. Wright and D. Manocha. Defining, computing, and visualizing molecular interfaces. Manuscript, 1995.

[18] G. Voronoi. Nouvelles applications des paramètres continus à las théorie des formes quadratiques. *J. Reine Angew. Math.* **133** (1907), 97–178.