Analytical Shape Computation of Macromolecules.

II Inaccessible Cavities in Proteins

July 10, 1996

Jie Liang[1], Herbert Edelsbrunner[2], Ping Fu[3], Pamidighantam V. Sudhakar[4] and Shankar Subramaniam[4]

Abstract. The structures of soluble proteins are well-packed, yet they contain numerous cavities. These cavities play key roles in accommodating solvents or small molecules, enabling conformational changes. From high resolution structures it is possible to locate these cavities. We have developed a precise algorithm based on alpha shapes for measuring space filling based molecular models (such as van der Waals, solvent accessible and molecular surface descriptions). We have applied this method for accurate computation of areas and volumes of cavities in several proteins. In addition, all of the atoms/residues lining the cavities are identified. We have used this method to study structure and stability of proteins. Furthermore, we have used it to locate a structural water formed proton pathway in the membrane protein bacteriorhodopsin.

[1]Department of Computer Science and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

[2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

[3]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820, USA.

[4]Beckman Institute for Advanced Science and Technology, Department of Physiology and Biophysics and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

1 Introduction

Native protein structures are well-packed and a large degree of the stabilization comes from packing interactions. Nevertheless, proteins have cavities, which either accommodate prosthetic groups or structural and functional water molecules, or they serve the function of allowing conformational flexibility. Extensive mutagenesis studies on proteins show that the alteration of cavity shapes and sizes influence both the thermodynamic stability [?, ?, ?], and in some cases the functionality of the protein [?]. It is therefore important to calculate accurately the shapes and sizes of cavities in proteins.

Lee and Richards have used a modified version of their solvent accessibility area and volume computation method to calculate cavities in myoglobin and in other proteins [?]. Using a modified version of Shrake and Rupley algorithm, Rashin *et al.* have calculated cavities in a set of 12 proteins [?]. They have analyzed their results to probe the immediate vicinity of occupied and empty cavities. They concluded that cavities lined by polar residues accommodate one or more water molecules and these water molecules form up to three hydrogen bonds with the amino acids lining the cavity. They have also estimated that the energetic cost of creating a cavity is about 1kT/10Å. More recently, Hubbard *et al.* [?, ?] have done an exhaustive analysis of cavities in 121 high resolution protein structures. They find that cavities that do not contain ordered water molecules are surrounded by hydrophobic side chains of residues in well-defined secondary structures, while cavities containing one or more water molecules are surrounded by coil regions. Experiments using water-sensitive two-dimensional heteronuclear NMR techniques show that hydrophobic cavities in proteins could contain disordered water molecules not apparent in the x-ray structural data [?].

Cavities also play an important role in protein-protein interfaces. Inter-domain contacts in proteins are often mediated by structurally well-defined interfacial water molecules isolated from bulk solvent. In addition, allosteric transformations in proteins are associated with changes in domain contacts and concomitantly with sizes and shapes of interfacial cavities. Further water-filled cavities play the role of modulating pK_a values of acidic and basic residues surrounding the cavities. In the absence of high resolution structural information that is capable of resolving all the water molecules in protein cavities, it will be extremely useful to develop accurate and fast computational methods for quantitatively calculating the shapes and sizes of these cavities.

In this paper we present an algorithm for computing accurately the location, shape, and size of internal cavities in proteins. The algorithm is based on the notions of alpha shapes and weighted Voronoi dissections of space filling based molecular models, such as the van der Waals, the solvent accessible, and and molecular surface models. The rest of the paper consists of two parts. Section 2 presents the alpha shape method and algorithmic details for computing and measuring cavities in macromolecules. Section 3 discusses applications of this method to a set of proteins and compares the results of this paper with the results in [?]. The method is also applied to study the role cavities play in the stability of the protein ribonuclease S, the conformational flexibility in the oxygen transport pathway in myoglobin, and the proton pathway in bacteriorhodopsin. Our goal is to highlight the specificity of these cavities and correlate their presence with experimental results.

2 Modeling, Computing, and Measuring Inaccessible Cavities

This section review standard geometric models of macromolecules, provides some background on alpha shapes, and finally discusses how to compute and how to measure inaccessible cavities within these models.

Geometric models. Space filling diagrams are widely used geometric models for macromolecules such as proteins and nucleic acids. They model a molecule as the union of many fused spherical balls in 3-dimensional space. Each ball represents an atom by adopting the spatial location and an appropriate radius, such as van der Waals radius of the atom. The *van der Waals* or VW model is the union of these spherical balls. Two related but different models are often used in situations where the interaction of the molecule with a solvent is considered. Model the

solvent as a sphere of appropriate radius and roll it about the van der Waals surface. The center of the solvent sphere sweeps out the surface of the *solvent accessible* or SA model, while the front of the solvent sphere defines the *molecular surface* or MS model, see see [**?**, **?**, **?**].

Recently, a combinatorial model of a molecule in terms of an associated complex has been proposed, see [**?**]. It is based on the theory of alpha shapes that has been developed in the computational geometry community [**?**]. Consider the VW model of the molecule, which is a union of balls, and the Voronoi cells of the balls. Adjacent Voronoi cells are separated by the radical plane of their balls. The Voronoi cells decompose the VW model into convex cells that are either disjoint or overlap along common boundary pieces. The *dual complex* consists of four different types of simplices: vertices, edges, triangles, and tetrahedra. The vertices are the centers of the balls. We add an edge connecting two ball centers to the complex if their two convex cells overlap along a common face. Similarly, we add a triangle spanned by three centers if their convex cells share a common edge. Finally, we add a tetrahedron spanned by four centers if their convex cells share a common point.

We refer to [**?**] for a complete and rigorous treatment of dual complexes and to the companion paper [**?**] for an intuitive description targeting biology applications. Most importantly, the dual complex faithfully represents geometric and topological properties of the molecule as represented by the VW model. Since the SA model is also a union of balls, we can repeat the same construction and obtain a dual complex representing the properties of this enlarged model.

Filter of Delaunay simplices. If we perform the construction of simplices directly for the Voronoi cells, rather than the Voronoi cells clipped to within a union of balls, we get the *Delaunay triangulation* or *complex*. The dual complex is a subset of the Delaunay complex as it is obtained by the same process but with strictly smaller cells with strictly less overlap.

Indeed, we can simultaneously grow all balls so that their Voronoi cells remain invariant and maintain the growing dual complex. This way we get a nested sequence of dual complexes, which are all subsets of the Delaunay complex. The last complex in the sequence is the Delaunay complex itself. To formalize this idea we parameterize the growth of balls by a real number α and refer to the dual complex at time α as the *α-complex*. The 0-complex is the dual complex of the VW or the SA model, depending on the choice of radii. Negative values of α correspond to shrinking the balls. Although growing the balls is a continuous process, we get a finite sequence of alpha complexes:

$$\mathcal{K}_0, \mathcal{K}_1, \ldots, \mathcal{K}_m.$$

The first complex, $\mathcal{K}_0$, is the empty complex and the last complex, $\mathcal{K}_m$, is the Delaunay complex. We use this sequence of alpha complexes to sort the simplices in the Delaunay complex. For each Delaunay simplex, σ, there is a unique *rank* $r = r(\sigma)$ with $\sigma \in \mathcal{K}_r$ and $\sigma \notin \mathcal{K}_{r-1}$. The *filter* is an ordering of the simplices,

$$\sigma_0, \sigma_1, \ldots, \sigma_n,$$

so the rank increases from left to right, that is, $r(\sigma_j) \leq r(\sigma_{j+1})$ for all j. It can happen that two simplices have the same rank in which case we order vertex before edge before triangle before tetrahedron. The relevance of the filter will become clear shortly when we discuss the computation of inaccessible cavities in a molecule.

Representing inaccessible cavities. Consider the SA model of a molecule. The complement is the part of 3-dimensional space not covered by any ball in that model. It may consists of several components, and each component except the infinitely large outside one is an *inaccessible cavity* also referred to as *void*. A void is a component of locations at which the solvent does not overlap the VW model of the molecule. The fact that the void is disconnected from the outside component means that the solvent cannot escape to infinity.

Consider the sequence of alpha complexes: $\mathcal{K}_0, \mathcal{K}_1, \ldots, \mathcal{K}_m$, and let I be the index so that $\mathcal{K}_I$ is the dual complex of the SA model. Let J be the largest index of any simplex in the filter with rank $r(\sigma_J) = I$. Hence σ_0 through σ_J belong to $\mathcal{K}_I$ and σ_{J+1} through σ_n do not belong to $\mathcal{K}_n$. We can take advantage of the fact that the simplices

in the dual complex of the SA model are readily available as a prefix of the filter. All we need is the homotopy equivalence result proved in [**?**]: for every void of the SA model there is a unique corresponding void of $\mathcal{K}_I$ which contains the SA void.

A void in $\mathcal{K}_I$ is triangulated by Delaunay simplices that are not in $\mathcal{K}_I$. We use this set to represent the void. The set consists of tetrahedra, of triangles connecting tetrahedra, and of edges connecting triangles. The set is completely determined by its tetrahedra, and a triangle or edge σ_j belongs to the set if and only if $j > J$ and σ_j is a face of at least one tetrahedron in the set.

Constructing inaccessible cavities. We describe an algorithm that constructs sets of tetrahedra, where each set corresponds to a void of $\mathcal{K}_I$. The sets are stored in a union-find system, which is a standard data structure for maintaining a system of disjoint sets. Any tetrahedron can belong to at most one set or void. The set system is manipulated through a sequence of the following types of operations:

ADD(σ): Add the tetrahedron σ as the only member of a new set to the system.

FIND(σ): Determine and return the set that contains the tetrahedron σ.

UNION(U, V): Let $U \neq V$ be two sets in the system and replace them by the union, $U \cup V$.

Specific implementations of the union-find system can be found in most algorithm texts, see e.g. [**?**]. Our implementation takes time $O(n \log n)$ for a sequence of n operations. There are other implementations that are somewhat faster, but they represent sets by trees which is less convenient than the lists used in our implementation.

We return to the main problem, namely constructing the voids. This is done by traversing the filter backwards, from σ_n down to σ_{J+1}. The tetrahedra among these simplices are collected in the sets of the union-find system. The triangles are used to trigger the union of sets. There is one special set in the system that collects all tetrahedra outside $\mathcal{K}_I$. This set does not represent any void but rather the outside component of the complement. We initialize the special set to $\{\sigma_\infty\}$, and we think of σ_∞ as the outside or complement of the Delaunay complex.

```
Initialize the set system to contain only the special set: {{σ∞}};
for j := n downto J + 1 do
    if σj is a tetrahedron then
        ADD(σj)
    elseif σj is a triangle then
        determine the two tetrahedra σ, σ' that share σj (one can be σ∞);
        U := FIND(σ); V := FIND(σ');
        if U ≠ V then UNION(U, V) endif;
    endif
endfor.
```

For the correctness of the algorithm it is important that when a triangle σ_j is encountered, its two tetrahedra, σ and σ', have already been added to the system. The construction of the filter guarantees that this is indeed always the case. After running the algorithm, each set but the special one represents a void of $\mathcal{K}_I$. Next we show how such a set of tetrahedra can be used to compute the volume and the surface area of the corresponding void in the SA model.

Computing volume. The method for computing the volume of the SA void represented by a set V of Delaunay tetrahedra is based on an inclusion-exclusion formula proved in [**?**]. We just explain the method. The volume of the SA void is

$$vol = vol_0 - vol_1 + vol_2 - vol_3,$$

where the four terms accummulate the volume of four different types of geometric objects: tetrahedra, sectors of balls, wedges of intersections of two balls, and halves of intersections of three balls.

$vol_0 = \sum vol(\sigma)$, where $vol(\sigma)$ is the volume of the tetrahedron σ and the sum extends over all $\sigma \in V$.

$vol_1 = \sum \varphi_{\nu,\sigma} \cdot vol(b_\nu)$, where $\sigma \in V$, $\nu \in \mathcal{K}_I$ is a vertex of σ, and b_ν is the ball with center ν. $vol(b_\nu)$ is the volume of b_ν and $\varphi_{\nu,\sigma}$ is the solid angle at ν inside σ. The sum extends over all $\sigma \in V$ and over all vertices $\nu \in \mathcal{K}_I$ of σ.

$vol_2 = \sum \varphi_{\varepsilon,\sigma} \cdot vol(b_\nu \cap b_\mu)$, where $\sigma \in V$, $\varepsilon \in \mathcal{K}_I$ is an edge of σ, and ν and μ are the endpoints of ε. $vol(b_\nu \cap b_\mu)$ is the volume of the intersection of the two balls, and $\varphi_{\varepsilon,\sigma}$ is the dihedral angle at ε inside σ. The sum extends over all $\sigma \in V$ and over all edges $\varepsilon \in \mathcal{K}_I$ of σ.

$vol_3 = \sum \frac{1}{2} \cdot vol(b_\nu \cap b_\mu \cap b_\lambda)$, where ν, μ, λ are vertices of a tetrahedron $\sigma \in V$ that span a triangle in $\mathcal{K}_I$. $vol(b_\nu \cap b_\mu \cap b_\lambda)$ is the volume of the triple intersection.

All angles, whether solid or dihedral, are measured in revolutions, so 0 is the empty and 1 is the full angle. The above expression for the volume of a void is fairly natural and can easily be seen to be correct if the balls lining the void overlap at most in triplets. As proved in [**?**], the expression is exact no matter how big the balls are and how many of them form common overlaps.

Computing area. The above method for volume computation extends to area. The formula for the area of the SA void represented by the set V of tetrahedra is

$$area = area_1 - area_2 + area_3.$$

Again we look at sectors of balls, wedges of intersections of two balls, and halves of intersections of three balls.

$area_1 = \sum \varphi_{\nu,\sigma} \cdot area(b_\nu)$, where the notation is the same as for vol_1, except that $area(b_\nu)$ is the surface area of ball b_ν.

$area_2 = \sum \varphi_{\varepsilon,\sigma} \cdot area(b_\nu \cap b_\mu)$, where the notation is the same as in vol_2, except that $area(b_\nu \cap b_\mu)$ denotes the surface area of the intersection of the two balls.

$area_3 = \sum \frac{1}{2} \cdot area(b_\nu \cap b_\mu \cap b_\lambda)$, where the notation is the same as in vol_3, except that $area(b_\nu \cap b_\mu \cap b_\lambda)$ is the surface area of the common intersection of the three balls.

The volume and area computations are performed by the software module VOLBL, which is part of the standard distribution of the alpha shapes software. The module works for arbitrary arrangements of spherical balls and takes no advantage of properties of molecules, such as fairly uniform ball size and limited overlap. The above formulas reduce the computation to a small collection of primitive elements. A complete description of the analytic functions used to handle the primitive elements can be found in [**?**]. Many of the functions are the same as the ones in [**?**]. The entire computation is analytical, and numerical errors are introduced only through imprecise floating-point arithmetic and mathematical function evaluation, such as the square-root and the inverse of the cosine. A recent version of VOLBL extends the volume and area computation to the MS model, albeit is some circumstances the model is ambiguous and it is not entirely clear what its volume and area should mean.

We conclude this section by mentioning that our software is robust and complete, that is, no inaccessible cavities are missed no matter how small. This is in contrast to other available software which typically has trouble with high degree of overlap, with surface arcs that form complete circles, etc., see [**?**]. 3-dimensional pictures of cavities bounded by spherical patches can be generated by software based on the structure of alpha complexes, see [**?**].

3 Results and Discussion

In the following we present the results obtained by using the alpha shape based method to compute cavities in proteins. For a brief comparison, we first compare the number and size of cavities for a set of proteins investigated

previously [?]. We then compute cavities in three other proteins where the presence of solvent and possible roles of cavities in protein function are exemplified: myoglobin where xenon binds in its hydrophobic cavities; ribonuclease (native and modified proteins) where isolated cavities and protein packing changes when proteins are modified; and bacteriorhodopsin where sites for structural water molecules may play important roles in proton pumping. All structures were taken from the Brookhaven Protein Data Bank [?], and all solvent water molecules are removed. Heteroatoms were included: the heme group in myoglobin and retinal in bacteriorhodopsin. A probe of radius 1.2 Å for water was used in our calculations, except for bacteriorhodopsin where a radius of 1.4 was used.

Inaccessible cavities and comparison with previous results. A list of proteins (the set computed by Rashin *et al.*), together with the number of cavities, total cavity SA area/volume, and total cavity MS area/volume are given in Table 3.1.

	# of Cavities		Area			Volume		
Protein	VOLBL	Rashin	SA	MS	Rashin(SA)	SA	MS	Rashin(MS)
1eca	10	9	46.4	435.8	69	6.4	252.5	401
1nxb	3	0	0.4	62.5	0	0.0	26.7	0
2act	20	21	140.4	958.2	130	35.0	611.2	449
2cha	23	26	132.2	1079.7	120	20.4	647.6	571
2lyz	12	8	58.7	498.7	53	9.0	297.2	190
2ptn	19	13	175.4	1123.0	168	31.5	702.0	494
2sn3	2	2	6.0	80.2	6	0.4	44.4	32
3cyt	8	5	3.1	211.0	2	0.1	97.1	34
3rn3	4	5	1.3	94.9	3	0.1	42.7	41
4pti	2	2	23.4	132.7	20	3.7	87.3	70
5mbn	17	23	95.8	877.0	85	14.1	503.4	391
8tln	42	30	163.4	1777.8	117	22.2	987.0	528

Table 3.1: Cavity computed from VOLBL and from Rashin *et al.* for selected proteins. Areas are in Å^2 and volumes in Å^3. SA represents the solvent accessible model and MS represents the molecular surface model. Rashin *et al.* used structures from earlier PDB entities 1sn3, 1rn3, 2mbn, and 3tln.

Rashin *et al.* used an approximate algorithm to compute cavities. The overall results in computed SA area of the two methods are comparable. However, the MS volume differ significantly. We notice that a Monte-Carlo based point-sampling method was used by Rashin *et al.* to estimate the MS volume. Pronounced difference are observed for 1eca, 2ptn and 8tln (3tln used by Rashin) molecules. In general, the VOLBL finds more cavities for proteins (except 2cha and 5mbn), indicating the accuracy and sensitivity of the method. VOLBL computed cavities show large differences from the results of Hubbard *et al.* where a probe radius of 1.4Åwas used.

Myoglobin: ligand-binding cavities. Myoglobin is an oxygen transport protein found in skeletal muscles where oxygen is stored. The prosthetic heme serves as the oxygen carrier. Diffusion and binding of oxygen to the heme iron have been the subject of numerous investigations [?, ?, ?]. The x-ray structure of myoglobin shows no direct pathway for transport of oxygen from the exterior to the heme. It has been been postulated that protein conformational changes occur concomitantly with oxygen diffusion [?]. These conformational changes are related to atomic interactions and structural changes such as residue side chain movement and cavity formation. Cavities in myoglobin have been explored experimentally through binding of the noble gas atom xenon. X-ray structural analysis shows that xenon binding cavities are surrounded by relatively non-polar residues [?]. Four known xenon binding sites are seen in the x-ray structure with varying degrees of occupancy. One site is found in a cavity near the proximal histidine and a pyrrole group of the heme [?]. A second site is located near the corner of G-H and A-B

helices close to the external surface [**?**]. These two xenon binding sites are also observed in $^{1}29$Xe NMR experiments on met-myoglobin in solution [**?**]. A third xenon binding site is located at the corner of E-F helix and H helix and is close to the surface. A fourth site is on the distal side of the heme and is close to the oxygen binding site.

Computation of cavities in myoglobin would provide insights into the locations of hydrophobic sites where ligands (such as xenon) can bind, as well as help to identify cavities that can aid in diffusion of the ligands. The cavities in myoglobin computed from alpha-shape program volbl are shown in the Figure 4.4. We obtain a total of 17 inaccessible cavities in myoglobin. Rashin *et al.* using the same-sized probe obtained 23 cavities using modified Shrake and Rupley procedure.

In our result, cavities 1, 2, 4 and 7 correspond to the xenon binding sites. The volume contained in each cavity is listed along with the identity of surrounding residues in Table 3.2 (information at atomic details is computed by our software, but these are omitted for the sake of brevity). The xenon binding site I corresponds to cavity 7, which has the highest occupancy and the lowest B factor. Site II corresponds to cavity 4, site III to cavity 1, and site IV to cavity 2. Of the four xenon binding cavities sites, site I has the smallest volume and surface area. Site II has the largest B factor (49.6 Å^2) and sites III and IV have large volumes and areas. The sizes of the computed cavities correlates well with results of the NMR experiments. Xenon atoms with slow off-rates (1×10^{-5} M/s) and sensitivity to the paramagnetic Fe atom correspond to the ones tightly bound. These are the xenon atoms in sites I and II, which have small volume, and are close to the heme prosthetic group. Xenon atoms with larger off-rates are found in the larger cavities III and IV. Residue F138 is close to site III but has no direct contact with site I. Free energy perturbation molecular dynamics simulations [**?**] suggest that apart from the immediate residues surrounding xenon site I, residue F138 also experiences large movement. This suggests that site II located between sites I and III may mediate the changes in F138.

The oxygen binding region is also apparent in our cavity computations. Cavity 5 is flanked by heme on one side, the distal histidine H64 and hydrophobic residues L29, L32, F43 and I107 on another side. Cavity 12 is flanked by residues L29, F43 and H64 (also shared by cavity 5). These cavities form a possible pathway, through which oxygen diffuses to the iron-binding site. Cavity 5 is also close to the xenon binding site IV. This suggests that the structure around the oxygen binding pocket is flexible . There are other cavities (cavities 9 and cavity 14) in addition to cavity 12 along the path towards the CD loop. These cavities could also play a role in providing a pathway for oxygen diffusion into myoglobin from the exterior.

Ribonuclease: cavities and packing interactions in proteins. Packing interactions in proteins play an important role for protein folding and protein stability. The role of packing was examined in ribonuclease S formed by the S-protein–S-peptide complex [**?**]. The S-peptide is obtained by subtilisin cleavage of the ribonuclease S. The S-peptide and S-protein associate to form native-like ribonuclease-S structure. Methionine 13 on the S-peptide packs into a hydrophobic cluster in ribonuclease S native structure. By replacing methionine in position 13 with other amino acids, Richards and coworkers investigated the role of hydrophobic interactions in maintaining the stability of the protein. The crystallographic and thermodynamic analysis of the modified proteins does not show good correlations with overall biophysical parameters such as polarizability, volume and charge. A key problem in examining the packing interactions is the presence of small cavities arising in the modified proteins, some of which accommodate one or more water molecules. Varadarajan and Richards estimated the volume changes of the cavities by subtracting mean Voronoi volumes from residue Voronoi volume at position 13 for all of the mutants. We have computed the size and shape of cavities in the native and modified S-protein-S-peptide complexes, in order to examine the changes in packing interactions arising from these small cavities. The number and sizes of cavities computed for these proteins are listed in Tables 3.3-3.9. In the wild type Rnase (`2rns`) where M is at position 13, four cavities are found. All of them are relatively small (none exceeds 23 Å^2 in MS area and 10.1 Å^3 in MS volume).

Structures of the modified proteins show significant differences in number of cavities and their sizes from those in the native protein. The modified protein M13L (`1rbh`) has five cavities, one of which (cavity 1) is in the vicinity of residue 13 (Table 3.4). This cavity also exists in the native protein. However, its size increases from 10.1 Å^3 (MS volume) in the native to 28.4 Å^3 in the modified structure. This suggests that the branched side chain of

Cavity	Area SA	Area MS	Volume SA	Volume MS	Contributing Residues
1	105.2	31.5	6.4	144.6	W7 L72 I75 L76 K79 G80 H82 A134 L135 F138 L137
2	80.1	20.1	3.2	132.5	G25 I28 L29 G65 V68 L69 L72 I107 I111 Hem130
3	41.9	9.4	1.0	67.7	W14 V17 H24 I28 L69 I111 L115
4	48.1	9.4	0.9	82.9	L104 I107 S108 I111 L135 F138 Hem130 R139
5	42.8	10.1	1.4	65.0	L29 L32 F43 H64 V68 I107 Hem130
6	33.0	5.5	0.5	65.4	L9 V10 V13 F123 A127 A130 M131
7	25.6	2.4	0.1	49.9	L89 A90 H93 L104 I142 Y146
8	25.9	4.2	0.4	44.3	W14 L72 L76 I111 M131 L135
9	21.3	1.9	0.1	42.3	A22 G25 Q26 K62 G65
10	11.4	0.3	0.0	24.9	F33 L40 L49 M55
11	14.0	0.4	0.0	29.9	L76 M131 A134 L135
12	10.8	0.2	0.0	26.5	L29 F33 F43 F46 H64
13	11.9	0.4	0.0	25.7	V13 V17 L115 H119
14	7.3	0.0	0.0	18.2	L29 F33 F46 L61
15	8.8	0.0	0.0	20.6	L11 W14 L76 K77
16	7.4	0.0	0.0	18.4	H93 I99 L104 Y146
17	8.1	0.0	0.0	20.0	E109 I112

Table 3.2: Inaccessible cavities computed by VOLBL for myoglobin `5mbn`. Areas are in Å^2 and volumes are in Å^3. SA represents solvent accessible model and MS the molecular surface model.

leucine introduces significant structural changes. On the other hand, $\Delta\Delta G$ for the M13L mutant is close to zero. It has been suggested that changes in enthalpic contributions are compensated by changes in entropic contributions [**?**]. In the modified protein M13A (`1rbc`) (see Table 3.5), cavity 1 surrounding residue 13 is present, as in the native protein. It is enlarged, due to the replacement by the small side chain in alanine. This modified protein also displays a change in cavity 5 (cavity 3 in the native protein), which is surrounded now by residues D121 and A109, in addition to residue H119 present in the native protein (the native protein has residues A4, V118 and H119 lining this cavity). Several additional cavities are also formed. Varadarajan and Richards observe the presence of a water molecule in M13A protein structure [**?**]. The modified protein M13F (`1rbe`) has seven cavities, of which three are adjacent to residue 13 (see Table 3.6). This is rather counterintuitive, since the bulkier phenylalanine replacement introduces more cavities. It has been suggested that phenylalanine substitution induces conformational changes in its vicinity [**?**]. This changes both the number of cavities as well as residue make-up of the cavity walls. A similar

Cavity	Area SA	Area MS	Volume SA	Volume MS	Contributing Residues
1	0.1	22.8	0.0	10.1	8F 47V 54V 106I 120F
2	0.0	20.4	0.0	8.6	14D 29M 33R 46F
3	0.0	22.3	0.0	8.9	4A 118V 119H
4	0.0	18.2	0.0	7.3	55Q 115Y 116V

Table 3.3: Inaccessible cavities computed by VOLBL for ribonuclease `2rns`. Areas are in Å^2 and volumes are in Å^3. SA represents the solvent accessible model and MS the molecular surface model.

	Area		Volume		
Cavity	SA	MS	SA	MS	Contributing Residues
1	5.3	48.8	0.6	28.4	F8 H12 L13 V47 V54
2	0.6	27.4	0.0	13.0	N34 L35 K37 D38
3	0.0	21.1	0.0	9.0	F8 V47 I106 F120
4	0.0	20.9	0.0	8.9	F8 V47 V54 I106
5	0.1	31.4	0.0	11.5	Q55 Y115 V116

Table 3.4: Inaccessible cavities computed by VOLBL for modified protein M13L `1rbh`. Areas are in Å^2 and volumes are in Å^3. SA represents the solvent accessible model and MS the molecular surface model.

	Area		Volume		
Cavity	SA	MS	SA	MS	Contributing Residues
1	25.2	134.9	2.6	83.5	F8 H12 A13 D14 S15 V47 E49 L51 V54
2	0.9	30.6	0.0	14.8	N34 L35 K37 D38
3	0.1	21.8	0.0	9.5	F8 E9 A13 L51
4	0.0	18.2	0.0	7.3	F8 V47 V54 I106
5	0.0	19.4	0.0	7.9	A109 H119 D121

Table 3.5: Inaccessible cavities computed by VOLBL for modified protein M13A `1rbc`. Areas are in Å^2 and volumes are in Å^3. SA represents the solvent accessible model and MS the molecular surface model.

effect is observed in M13G, where contrary to intuition, there are fewer cavities (see Table 3.7). The total cavity volume, however, is increased in this modified protein compared to the native protein.

The more isochoric substitution M13I (`1rbg`) leads to very small changes in the cavity sizes as compared to the native protein (see Table 3.8). Protein M13V (`1rbi`) has an additional cavity (volume 32.2 Å^3), although not in the immediate vicinity of residue 13 (Table 3.9). Such new cavity formation could affect protein stability.

The variable volumes of cavities formed in these modified proteins show that replacement at residue 13 leads to subtle changes in the conformation and packing of the protein structure. While relative deviations from the native structure may not be significant, modified proteins may contain cavities which differ both in size and location. Phenomenological thermodynamic analysis of protein stability due to mutations should therefore take into account these changes in cavities. The changes in cavities may also be counterintuitive to size-based arguments. This is evidenced by the fact that mutation of methionine to glycine results in elimination of a cavity. Lacking a side chain, glycine provides the flexibility necessary for a conformational change in the protein backbone.

Bacteriorhodopsin: cavities and proton pathway. A low resolution structure of the ground state of the light-driven proton pump bacteriorhodopsin (bR) has been obtained by Henderson and coworkers from electron microscopy studies [**?**]. The structure reveals the presence of a putative channel lined by a set of polar residues and the Schiff base through which a proton can be translocated across the membrane. The role played by several of the proposed channel residues has been confirmed by mutagenesis experiments [**?**, **?**, **?**, **?**, **?**, **?**]. FTIR and fluorescence labeling experiments have shown that immediate environment around a number of key residues in the putative protein transfer pathway change during the photocycle [**?**]. It has been postulated that a select number of water molecules in the pathway play a key role in the proton transport [**?**, **?**, **?**, **?**, **?**, **?**, **?**]. A quantitative study of cavities which are large enough to accommodate one or more water molecules in the bacteriorhodopsin structure would shed light on possible interactions between the water molecules and the residues in the pathway, as well as

Cavity	Area SA	Area MS	Volume SA	Volume MS	Contributing Residues
1	7.7	66.0	0.8	38.1	S21 N27 M30 K31 T36 C95 Y97
2	5.3	50.1	0.6	29.6	F8 H12 F13 V47 V54
3	0.2	25.0	0.0	11.5	F8 V47 V54 I106 F120
4	0.1	20.7	0.0	8.8	T45 I81 I106 F120 S123
5	0.1	29.4	0.0	11.7	S21 K31 T36 C95
6	0.2	23.3	0.0	10.4	D83 K98 T100
7	0.0	21.6	0.0	8.8	V54 I106 V108

Table 3.6: Inaccessible cavities computed by VOLBL for modified protein M13F `1rbe`. Areas are in Å^2 and volumes are in Å^3. SA represents the solvent accessible model and MS the molecular surface model.

Cavity	Area SA	Area MS	Volume SA	Volume MS	Contributing Residues
1	13.8	69.0	2.9	49.0	S21 N27 M30 K31 T36 C95 Y97
2	0.4	26.2	0.0	12.1	N34 L35 K37 D38
3	0.0	20.4	0.0	8.5	T45 I106 F120 S123

Table 3.7: Inaccessible cavities computed by VOLBL for modified protein M13G `1rbf`. Areas are in Å^2 and volumes are in Å^3. SA represents the solvent accessible model and MS the molecular surface model.

environmental changes around key residues during the photocycle. We compute below the number and sizes of cavities in the low resolution ground state structure of bacteriorhodopsin.

We use 1.4 Å as the probe size for a water molecule The sizes of the computed cavities along with the flanking residues are presented in Table 3.10. There are 8 cavities in bR which are inaccessible to bulk solvent and can accommodate one or more water molecules. The largest of these cavities (1 and 2) are at the cytoplasmic and extracellular ends of the protein respectively. They are capable of accommodating networks of 5-9 water molecules which can serve as the proton transfer link to the bulk electrolyte. In Figure 3.1, 3.2 and 3.3, we present a model of the protein along with the two largest cavities. As shown in the figure, the largest cavity which lines the proton channel from the cytoplasmic interior to the schiff base is encompassed between the helices C and G, with a few contacts to residues on helices B and F. The hydrophobic environment around Asp 96 in the ground state has been postulated on the basis of numerous experiments and it has been suggested that the pK_a of Asp 96 in the bR ground state is as high as 13 [**?**]. The side chain of Asp 96 along with the relatively hydrophobic side chains of Phe

Cavity	Area SA	Area MS	Volume SA	Volume MS	Contributing Residues
1	1.1	36.0	0.0	17.1	N34 L35 K37 D38 R39
2	0.7	28.2	0.0	13.6	F8 V47 V54 I106 F120
3	0.0	18.3	0.0	7.4	F8 H12 I13 V47
4	0.0	18.6	0.0	7.5	S21 K31 P93 C95
5	0.0	18.7	0.0	7.5	Q55 Y115 V116

Table 3.8: Inaccessible cavities computed by VOLBL for modified protein M13I `1rbg`. Areas are in Å^2 and volumes are in Å^3. SA represents the solvent accessible model and MS the molecular surface model.

	Area		Volume		
Cavity	SA	MS	SA	MS	Contributing Residues
1	4.8	55.7	0.4	32.1	S21 N27 M30 K31 T36 C95 Y97
2	1.3	32.1	0.1	16.4	F8 H12 V13 V47 V54
3	0.1	23.8	0.0	10.7	F8 V47 V54 I106 F120
4	0.5	26.7	0.0	12.4	N34 L35 K37 D38
5	0.0	21.8	0.0	9.4	S75 S77 M79

Table 3.9: Inaccessible cavities computed by VOLBL for modified protein M13V `1rbi`. Areas are in $Å^2$ and volumes are in $Å^3$. SA represents the solvent accessible model and MS the molecular surface model.

figure=fig/1brd.cav1.eps,height=3.0in

Figure 3.1: Cavity 1 in bacteriorhodopsin (1brd) shown in alpha shapes, which are larger than the molecular surface cavities.

figure=fig/1brd.cav2.eps,height=3.0in

Figure 3.2: Cavity 2 in bacteriorhodopsin (1brd).

figure=fig/1brd2cav.quanta.eps,height=3.0in

Figure 3.3: The alpha carbons of the seven helices and the retinal, together with residues whose side chains contribute to cavities 1 and 2.

219 and Leu 93 serve as the largest contributors to the area of the cavity, implying the presence of ordered and structured water molecules in this cavity. We display the side chains contributing to cavity 1 in Figure 3.1. Recent experiments including the structure of the M intermediate [**?**, **?**, **?**], show that during the photocycle the movement of the F helix changes the solvation environment around Asp 96. We postulate that the inaccessible cavity 1 in the ground state becomes exposed to the cytosol in this process, thus enabling the formation of a water chain that can translocate a proton.

4 Conclusions

We have presented an accurate method for computing inaccessible cavities in proteins using the alpha shape method. Molecular and solvent-accessible surface areas and volumes for the cavities can be computed using our method. In addition to computing total areas and volumes, our method also provides information about atoms and residues which contribute to the cavities. Precise knowledge computed by our method can be used for probing protein structure and stability, as well as for protein engineering. Further, the computed cavities also reveal potential water binding sites.

We have applied our method to obtain cavities in previously studied proteins. The power of the method is illustrated by three key applications. In myoglobin, our computation has located all of the cavities which bind xenon. In ribonuclease S-protein-S-peptide complex, we have analyzed the role cavities play in engineered protein stability. In bacteriorhodopsin, we show that the two largest cavities line the putative channel involved in proton transport. The precise computation of cavities can be quantitated further to analyze protein thermodynamics and this study is in progress. Effects of buried surface area/volume in multidomain proteins and protein-protein complexes is also underway.

Currently, in our software VOLBL we do not store computed area/volume for pair and triple atoms intersection. Such information would be useful for further detailed analysis of polar-polar, polar-nonpolar, and nonpolar-nonpolar interactions. Implementation of accessing such information is also in progress.

	Area		Volume		
Cavity	SA	MS	SA	MS	Contributing Residues
1	74.7	322.0	16.9	270.5	V49 T46 F219 D96 K216 L 93 L48 L92 I45 P50 G220 L100 F171 L97 L223 L174 I222 T170
2	42.6	166.6	11.6	147.3	F208 M60 Y57 D212 M209 L13 T205 R82 D85 W86
3	20.9	108.3	4.3	84.6	Y83 I119 D115 W86 RET216 M118 L87
4	17.4	104.4	3.0	80.9	W182 L181 T178 A215 RET216 L93 V177 F219
5	8.8	78.9	0.9	56.6	L97 L93 L94 T178 W182 T 90 F219
6	3.0	93.1	0.1	44.6	A81 M56 M60 D85 Y57 R82
7	3.4	63.3	0.2	39.2	L181 A215 Y185 W182 RET216 L211
8	1.2	44.7	0.0	25.8	I148 L111 T90 L94 W182
9	0.9	39.9	0.0	22.4	F153 F156 R175 T157
10	0.7	36.3	0.0	19.6	T178 L149 I148 L152 W182
11	0.1	28.8	0.0	14.5	V217 P50 K216 M20
12	0.3	32.4	0.0	17.1	V177 L174 I222 F219
13	0.1	31.2	0.0	15.6	M118 I148 T90 W182 D115
14	0.0	26.2	0.0	12.6	I52 A53 D85 T89
15	0.0	25.8	0.0	12.3	W12 N202 E9
16	0.0	26.6	0.0	12.7	L190 W138 W189
17	0.0	28.0	0.0	13.6	T90 L111 P91
18	0.0	24.7	0.0	11.5	L111 T90 P91 V112

Table 3.10: Inaccessible cavities computed by VOLBL for membrane protein bacteriorhodopsin `1brd` Areas are in Å^2 and volumes are in Å^3. SA represents the solvent accessible model and MS the molecular surface model.

figure=fig/mbncavcilab.ps,height=6.0in

Figure 4.4: Inaccessible cavities in myoglobin (5mbn). Xenon binding sites have been labeled.

Acknowledgements

The software for constructing weighted Delaunay complexes and the alpha shape filters has been written by Ernst Mücke and Michael Facello. We thank them for creating reliable and robust software so we can build on their results. Jie Liang is supported by an NSF CISE post-doctoral fellowship, grant ASC 94-04900. Herbert Edelsbrunner and and Ping Fu acknowledge support by NSF, grant ASC 92-00301, and by ONR, grant N00014-95-1-0692. Herbert Edelsbrunner also acknowledges support through the NSF Alan T. Waterman award, grant CCR 91-18874. Sharkar Subramaniam and Pamidighantam Sudharkar acknowledge support by NSF, grants ASC 89-02829 and MCB 92-19619. The authors thank NSF Meta Center Allocation for providing computational resources.