An Integrated Geometric and Topological Approach to Connecting Cavities in Biomolecules

Talha Bin Masood*
Department of Computer Science and Automation
Indian Institute of Science, Bangalore

Vijay Natarajan†
Department of Computer Science and Automation
Department of Computational and Data Science
Indian Institute of Science, Bangalore

ABSTRACT

Study of cavities and channels in molecular structure is a crucial step in understanding the function of biomolecules. Current tools and techniques for extracting these structural features are sensitive to uncertainties in atomic position and radii. In this paper, we study the problem of cavity extraction in biomolecules while taking into account such uncertainties. We propose an approach that connects user-specified cavities by computing an optimal conduit within the region occupied by the molecule. The conduit is computed using a topological representation of the occupied and empty regions and is guaranteed to satisfy well defined geometric optimality criteria. Visualization of the set of all cavities with multiple linked views serves as a useful interface for interactive extraction of stable cavities. We demonstrate the utility of the proposed method in successfully identifying biologically significant pathways between molecular cavities using several case studies.

Index Terms: G.2.2 [Discrete Mathematics]: Graph Theory; I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling; J.3 [Life and Medical Sciences]: Biology and genetics

1 INTRODUCTION

Proteins are the fundamental functional blocks of a biological system. Structural features of the protein are known to play a key role in determining the stability and function of proteins [25]. One important structural feature of protein is the set of *cavities* within it. Cavities are empty spaces within the protein. Cavities may have openings or be completely buried, and are referred to as *pockets* and *voids*, respectively. Pockets often correspond to the active site of enzymes or interacting sites for other proteins. Voids, on the other hand, are important structural features that affect the overall thermodynamic stability of the protein [14]. Filling up internal voids improves the packing of the protein thus increasing stability [12]. Given the importance of cavities in protein structure study, several algorithms and software are available to compute them given protein structure data, such as from the Protein Data Bank.

Protein structures are commonly determined from X-ray crystallography data. Inaccuracies, noise, and more generally, uncertainty in the data adversely affects existing cavity detection methods. Small inaccuracies may already cause a difference in the reported number of cavities. The atom radii are also empirically determined and hence not precisely defined values. Figure 1 shows a transmembrane protein where a cavity detection method would typically report multiple disconnected components. In many such scenarios, it is clear that a single connected cavity is the desired result. We study the problem of improved cavity extraction and describe a method that takes into consideration the uncertainty in the data. In particular, we present an integrated geometric and topological approach to connect cavities that are reported by an existing method. This approach combines the benefits of geometric measures to quantify the perturbation that is employed to connect the cavities and an efficient data structure for representing the connectivity of the empty space within the protein.

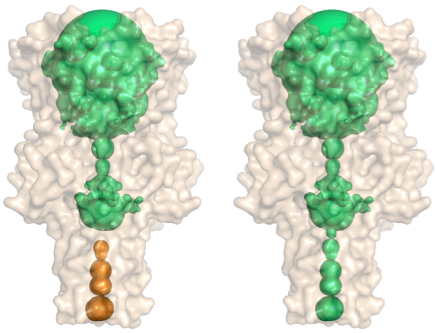


Figure 1: **Left:** The transmembrane channel through the protein 2OAR is detected as disconnected cavities. The default parameters are used to compute the cavities. **Right:** However, it can be connected by perturbing a few atoms around its bottleneck.

Numerous methods have been proposed for identification of cavities in protein molecules. These methods employ a wide variety of approaches – grid and occupancy based, graph based, model fitting, Monte Carlo simulations on solvent molecules, and Voronoi diagram based. Early tools used grid-based approaches to extract cavities [11, 15]. These methods discretized the space occupied by the protein thereby trading off accuracy in favor of computational efficiency. The notion of cavities and their classification as voids and pockets was more formally defined using the alpha shape model of a molecule proposed by Edelsbrunner et al. [9, 10] and Liang et al. [16, 17]. This enabled accurate identification of cavities and further supported the computation of geometric properties like volumes and surface areas. Tools based on the above approach are available and widely used [6, 13]. Graph based methods have also been used to identify cavities and compute their volumes [22, 27]. Given an estimate of the empty space, Varadarajan et al. describe a Monte Carlo procedure to position water molecules within to improve the accuracy of the extracted cavity [4]. Several recent methods are based on the Voronoi diagram of the atoms [18, 19, 20, 23, 24].

The above mentioned methods do not explicitly handle the adverse effects of uncertainty in the data. Some methods support user-specified parameters such as solvent radius or a growth factor but they are almost always global parameters that affect all extracted cavities. Sridharamurthy et al. [26] address the problem of uncertainty explicitly and propose a two-parameter solution that extracts so-called robust cavities. However, their solution is also global in the sense that the parameters affect all extracted cavities. Such a solution may produce undesirable results by connecting cavities that lie outside the region of interest. We propose a simple and direct approach to address the problem, where the user examines the cavities and identifies artifacts or undesirable disconnections. The user interacts with the multiple linked views provided by the visualization and specifies a pair of cavities to be connected. Our cavity

*e-mail:tbmasood@csa.iisc.ernet.in
†e-mail:vijayn@csa.iisc.ernet.in

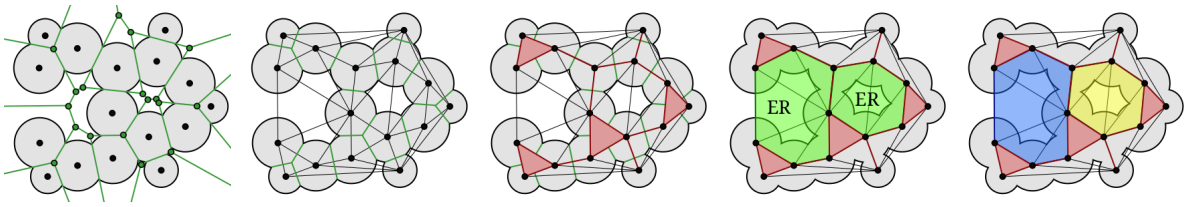


Figure 2: **(a)** Power diagram of a weighted point set in $\mathbb{R}^2$, power edges and vertices are shown in green. **(b)** The weighted Delaunay complex (black edges) is the dual of the power diagram. **(c)** The α-complex K_α for $\alpha = 0$ is shown in red. This is the dual of the intersection of power diagram and union of balls. K_α forms the occupied region (OR) of the molecule. **(d)** The empty region (ER) in green. This region is defined by Delaunay flow. The green triangles do not belong to OR and have flow towards a triangle within the molecule. The set of simplices in OR and ER form the molecular region (MR). **(e)** ER consists of two maximally connected components, called cavities, shown in blue and yellow.

connection algorithm efficiently and automatically computes an optimal conduit between the cavities. Key contributions of this paper include:

- A simple, explicit, and flexible method for extracting cavities in biomolecules from uncertain data with guaranteed bounds on the perturbation required.
- Efficient algorithms to compute a conduit between user selected cavities that satisfies well defined optimality criteria.
- Interactive visualization of cavities in a molecule with multiple linked views that facilitates identification of disconnected cavities.
- Three case studies that demonstrate the benefits of the cavity connection based method — computing ion transport channels from uncertain data, comparing cavities obtained from various mutants of a protein, and computing the migration path of carbon monoxide in myoglobin.

We evaluate the method by comparing the results with those obtained using a global parameter-driven cavity extraction method. We also note that our method may be used in conjunction with any of the Voronoi diagram based method to improve the results.

2 Biomolecule Representation

We briefly introduce the mathematical background required to define and represent the structure of biomolecules [7, 8]. Protein molecules are often modeled as union of balls. The molecule M is defined as the set $\{a_i = (p_i, r_i)\}$. Here a_i denotes a constituent atom of M modeled as a sphere with center at p_i and radius r_i.

Voronoi diagram and Delaunay triangulation

For a given set of weighted points, weighted Delaunay triangulation D is the triangulation of the input points based on proximity defined by power distance. A point p with weight w_p is equivalent to ball with radius $r_p = \sqrt{w_p}$. The power distance between a weighted point p and a point x in Euclidean space is given by $\|x-p\|^2 - w_p$. The dual of the weighted Delaunay triangulation is called weighted Voronoi diagram or power diagram [1]. The power diagram and the weighted Delaunay triangulation are shown in Figures 2(a) and 2(b), respectively.

Alpha complex

Alpha complex provides a growth model for the input spheres consistent with power distance. The growth parameter α corresponds to a radius $\sqrt{r_p^2 \pm \alpha^2}$ for a ball centered at p with radius r_p. Positive values of α correspond to growing the balls and negative values correspond to shrinking the balls. The weight of the point w_p increases or decreases by α between $-\infty$ and ∞. Note that $\alpha = 0$ corresponds to no growth. The parameter α can be varied from $-\infty$ to ∞ to obtain a filtration of triangles and tetrahedra (also called simplices) belonging to the weighted Delaunay triangulation. A sequence of α-complexes ($\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = D$) containing progressively more triangles and tetrahedra is obtained as α is increased from $-\infty$ to ∞. Figure 2(c) shows the α-complex at $\alpha = 0$ as a subset of D highlighted in red.

Cavities

For a given molecule represented as a set of balls, let D be the weighted Delaunay triangulation and $K_\alpha \subseteq D$ be the α-complex for value α. The *Delaunay flow* over D is defined as the collection of flows between adjacent tetrahedra in D going from the tetrahedron from smaller circumsphere to the larger one. Let I_{tet} denote the set of tetrahedra in D whose Delaunay flow terminate within D and I denote I_{tet} together with the corresponding faces. For a given α value, we define *molecular region* MR as $K_\alpha \cup I$. The simplices in MR can be classified into two groups based on whether they belong to the α-complex or not. The simplices in $OR = K_\alpha$ constitute the *occupied region* in the molecule, while the remaining simplices $ER = MR - OR$ capture the *empty region* in the molecule. Refer to Figure 2(d) for an illustration of ER, OR and MR. In the figure, the green region is ER while red region is OR. The union of ER and OR is the molecular region, MR. The *cavities* are defined as maximally connected subregions in ER. Let the set of all cavities be $\mathscr{C} = \{C_1, C_2, \ldots, C_k\}$, such that $ER = C_1 \cup C_2 \cup \cdots \cup C_k$ and $C_i \cap C_j = \emptyset$. The tetrahedron $t_i \in C_i$ with highest α value is selected as the *representative tetrahedron* of C_i.

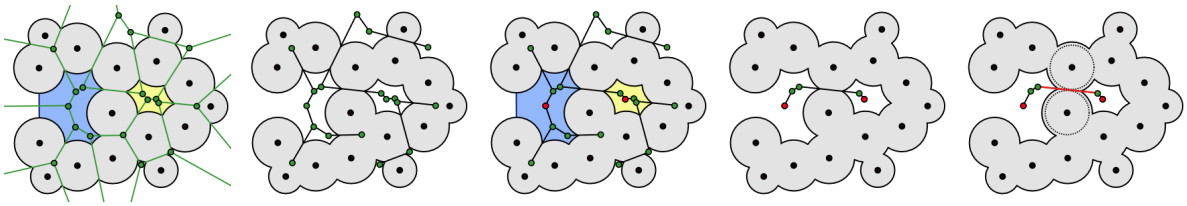
3 Cavity Connection

Cavities in a biomolecule, as defined in the previous section, are clearly sensitive to perturbations in atomic positions and radii. For example, Figure 1 illustrates a scenario where perturbing a few atoms results in detection of a single large cavity instead of disconnected cavities. Recognizing the existence of such a single connected cavity and extracting it by performing the required perturbation is an interesting and challenging problem. Current approaches to cavity extraction employ global parameters to address this problem resulting in undesired merging of multiple cavities. We aim to develop a flexible user-driven method that can improve the results of the cavity extraction algorithm by supporting the automatic computation of an optimal conduit between two given cavities.

3.1 Problem statement

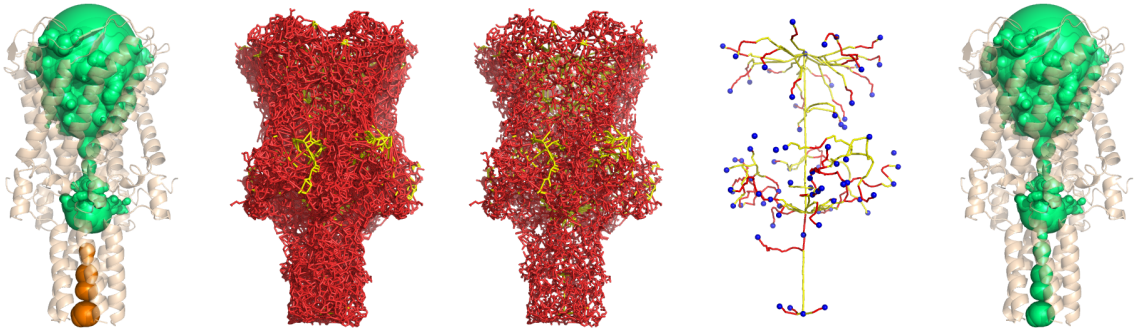
Given two disjoint cavities, the *cavity connection* problem is the computation of an optimal conduit between the cavities that (a) lies within the molecular region and (b) together with the two input cavities forms a single connected cavity after suitable perturbation of the atoms. We consider three optimality criteria that lead to different algorithms for connecting the cavities.

- BOTTLENECK : This is a min-max criterion where the objective is to minimize the maximum perturbation on an atom that



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

Figure 3: Illustration of cavity connection method based on BOTTLENECK criterion using a 2D example. **(a)** The two cavities which are required to be connected are shown in the context of the molecule shown as a set of grey disks. The G_{MR} is shown in green. **(b)** The maximum spanning tree ($MaxST$) is computed for the network. **(c)** The representative nodes of the two cavities in the $MaxST$ are colored red. **(d)** The connecting path detected between these cavities. **(e)** The only edge of the path which belongs to OR is highlighted in red. The lining atoms of this edge can be perturbed to physically connect these cavities.



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

Figure 4: Demonstration of cavity connection method applied to the protein 2OAR. **(a)** The two cavities which are required to be connected are shown in the context of the molecule shown in cartoon representation. **(b)** The complete dual graph G_{MR}. The edges which belong to OR are colored red while edges belonging to ER are colored yellow. **(c)** The $MaxST$ computed for G_{MR}. Same coloring scheme is used to identify edges in OR and ER. **(d)** The $MaxST$ is further pruned by restricting to paths connecting the cavity representatives. Here blue spheres show the cavity representatives. **(e)** Using cavity connection algorithm, the best path connecting the representative nodes of the two cavities shown in (a) is computed. The atoms are perturbed appropriately to obtain the merged cavity shown in this figure.

will result in the formation of a conduit between the cavities.

- PROXIMITY : The objective here is to minimize the number of atoms perturbed to form the conduit.
- BOTTLENECK_PROXIMITY : This is a hybrid of the above criteria. The number of perturbed atoms is minimized given an upper bound on the maximum perturbation allowed on an atom.

3.2 Cavity connection methods

We now describe efficient algorithms to connect cavities satisfying each of the above-mentioned optimality criteria.

BOTTLENECK criterion

The uncertainty in atom locations and radii determined from x-ray crystallography maps motivate the development of methods that perturb the values to extract connected cavities. The BOTTLENECK criterion aims to limit this perturbation in the atom radii to the least possible value.

Consider the dual graph G_{MR} of the tetrahedra and triangles in MR. Nodes of this dual graph correspond to the tetrahedra and the arcs correspond to the triangles. A weight is associated with each arc of G_{MR}, equal to the smallest value of α at which the corresponding triangle is inserted into the filtration. Let C_i and C_j be the two cavities that the user would like to connect. Let t_i be a representative tetrahedron belonging to the cavity C_i and t_j be the representative of C_j. Let n_i and n_j be the nodes in G_{MR} corresponding to t_i and t_j, respectively. The conduit between C_i and C_j may be represented by an alternating sequence of triangles and tetrahedra in MR and hence by a path in G_{MR}. In particular, we are interested in the path between n_i and n_j where the minimum weight over all arcs is maximized. We design a simple and efficient algorithm for computing this optimal path by recognizing that the path always lies within the maximum spanning tree of G_{MR}.

CLAIM. *The maximum spanning tree MaxST of the weighted graph G_{MR} contains a path satisfying the* BOTTLENECK *criterion for all pairs of cavities.*

Proof. Consider two nodes n_i and n_j in G_{MR}. Let P_{ij} denote an optimal path between the two nodes and a_{ij} denote the minimum weight arc within the path. We describe a proof by contradiction. Let P'_{ij} ($\neq P_{ij}$) denote the unique path between n_i and n_j in $MaxST$ and a'_{ij} denote the minimum weight arc within the path. If the weights of a'_{ij} and a_{ij} are equal then P'_{ij} is also an optimal path. If the weight of a'_{ij} is smaller than a_{ij} then we can show that a new tree may be constructed with weight greater than $MaxST$ resulting in a contradiction. Delete the arc a'_{ij} from $MaxST$. This results in two disconnected trees containing the nodes n_i and n_j respectively. Let $b_{ij} \in P_{ij}$ be an arc connecting the partitions. Clearly, the weight of b_{ij} is greater than or equal to the weight of a_{ij} and hence the weight of a'_{ij}. Replace a'_{ij} with b_{ij} in $MaxST$ to obtain a spanning tree with greater weight than $MaxST$, a contradiction. Hence, $MaxST$ always contains an optimal path. □

The conduit may be computed from the optimal path P_{ij} by perturbing the atoms lining P_{ij}. The triangles in MR that are dual to arcs in P_{ij} belong to OR and ER. We perturb the atoms incident on triangles in OR. Their radii are reduced by a value corresponding to the α-value at which the triangle is inserted into the filtration, thus establishing the connection between the cavities. Figures 3 and 4

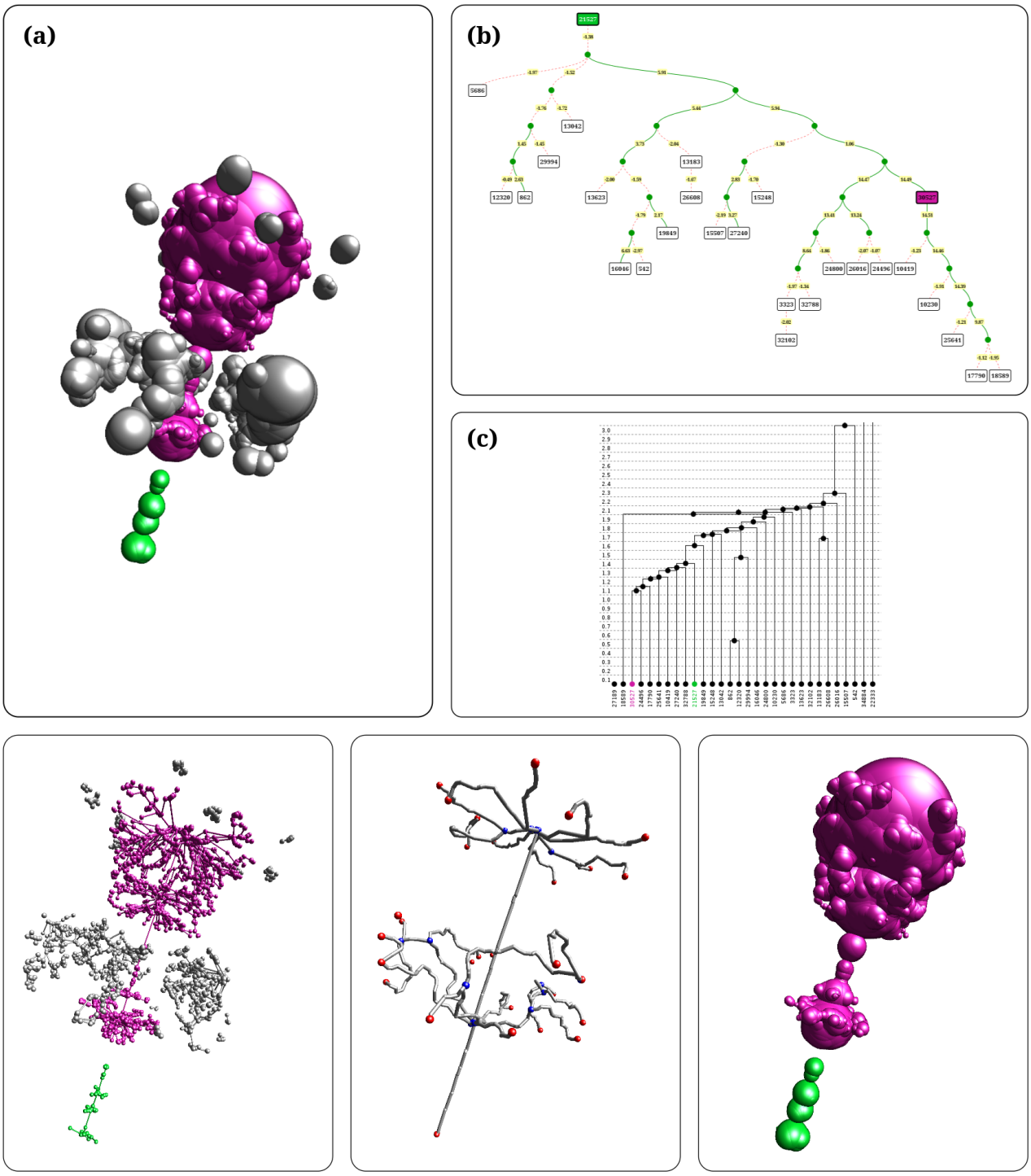


Figure 5: The three linked views of cavities in 2OAR are shown. Cavities may be selected from any of these views. **(a)** The 3D view shows the two cavities selected for connection in green and violet colors. Other cavities are shown in grey. **(b)** 2D graph visualization of the cavities. **(c)** This panel shows the cavity dendrogram in which the height is proportional to the α_{min} of the connecting path between cavities. Some additional 3D views are shown in the bottom row. From left to right: the dual graph representation, the simplified $MaxST$, and the two cavities to be connected.

illustrate this technique in 2D and 3D, respectively. A connection may also be established by perturbing the position of the atoms lining the path. However, computing such a perturbation without introducing steric clashes is a non-trivial and challenging task.

PROXIMITY criterion

Atoms that lie within the interior of the molecule are subject to greater physical constraints when compared to those lining the surface and hence less suitable for perturbation. The PROXIMITY criterion limits the number of atoms that are perturbed and hence limits the number of perturbed interior atoms.

Consider the dual graph G_{MR} as described earlier. Assign unit weight to the arcs that belong to OR and zero weight to arcs in ER. Let n_i and n_j be the representative nodes of cavities C_i and C_j in G_{MR}, respectively. The optimal conduit between C_i and C_j is represented by the shortest path P_{ij} between nodes n_i and n_j in the weighted graph G_{MR}.

The conduit corresponding to P_{ij} may be computed by selecting one atom per arc in the path and shrinking it by a value corresponding to the α-value at which the triangle is inserted into the filtration. The number of atoms perturbed is thus equal to the length of the path contained in OR.

BOTTLENECK_PROXIMITY criterion

The BOTTLENECK_PROXIMITY is a hybrid of both criteria described above. Again, the optimal conduit is represented as a path. Given cavities C_i and C_j, we first compute the path P_{ij} satisfying the BOTTLENECK criterion. Let α_{min} be the weight of the minimum weight arc in the optimal path P_{ij}. Construct a subgraph G of G_{MR} induced by arcs whose weight is greater than α_{min}. Now, construct an optimal path in G satisfying the PROXIMITY criterion. Alternatively, α_{min} may also be specified by the user instead of computing it using the BOTTLENECK criterion.

The conduit may be computed from P'_{ij} by selecting one atom per arc similar to the PROXIMITY criterion. However, the reduction in atom radius is now limited by α_{min}.

4 Visualization and Interaction

We describe three linked interactive visualizations to help the user identify important cavities and connect them based on different criteria. Figure 5 shows these three views.

3D Visualization

In this view, the cavities are shown in the context of the molecule. The cavities can be shown as union-of-balls, where each tetrahedron in the cavity is represented by its power ball whose centre is equidistant from the four atoms and radius is equal to the power distance. Alternately, we can also display the cavity in its dual graph representation, where nodes are drawn at the centre of the power ball for each tetrahedron, and edges between the nodes correspond to the common triangle face. Each cavity is given a unique color which is consistently used across different visualizations to help identify the cavity quickly. The user can pick multiple cavities for connection by simply clicking on the 3D view of the cavity. The detected connecting paths are shown as a set of cylinders in the 3D view. The $MaxST$ and the pruned $MaxST$ which connects the cavity representatives can also be visualized in this view.

2D Visualization

This view shows the abstract representation of the cavities and their connections based on BOTTLENECK criterion. We construct a pruned sub-graph of the $MaxST$ containing only the edges and nodes needed for connecting the representative nodes of all the cavities. The graph is further minimized by collapsing paths into edges. After pruning and collapsing, the graph contains the representative nodes of all the cavities and a few connecting nodes. These nodes are connected by edges, each of which represents a path in the original $MaxST$. The nodes can be colored and labeled based on different criteria. The edges are labeled by the minimum value of α in the corresponding path. This visualization is interactive and linked to other visualizations. The user can pick different cavities by selecting nodes in the graph. The connecting path is shown by highlighting the nodes and edges in the graph.

Hierarchical dendrogram

The negation of α_{min} of the optimal path P_{ij} connecting the cavities C_i and C_j can be treated as *cavity distance* measure. It can be shown cavity distance measure satisfies the non-negativity, coincidence, symmetry and triangle inequality properties. Based on this distance measure, we can cluster the cavities using hierarchical clustering and obtain the hierarchical dendrogram. This diagram shows the proximity of cavities based on BOTTLENECK criterion. It is a useful representation for showing the cavity hierarchy and may be used to identify cavities to connect. However, we do not use clustering to obtain this diagram, and instead construct this simultaneously during computation of $MaxST$.

User interaction

In addition to multiple interactive views of the cavities and their connections, the user is provided with other tools to identify important cavities. One such example is persistence based pruning of cavities. The user can interactively specify a threshold persistence value. The views are immediately updated and the cavities having less persistence than the specified threshold are removed from the three views. Another tool is automatic connection of all cavities using a perturbation below a given threshold. This is similar to what was proposed in [26].

5 Results and Discussion

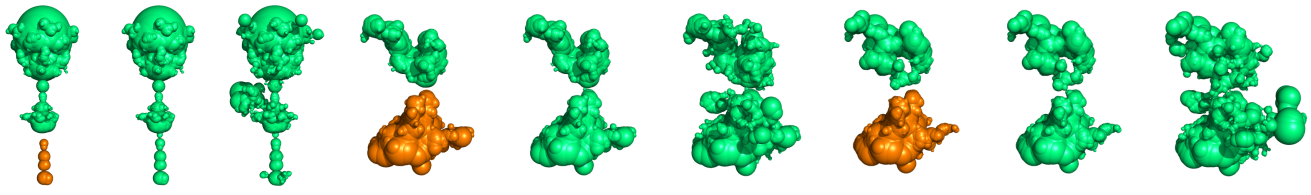
In this section, we first briefly discuss the implementation and runtime results. Then, we describe a qualitative comparison of our method with existing approaches for connecting cavities. Lastly, we demonstrate the utility of cavity connection using three example case studies. We show that our method can be used for connecting fragmented cavities to form channels. We can also study the propensity of a pathway being open based on the value of α_{min}. Using Myoglobin case study, we show that our cavity connection method can be used to reach similar conclusions as those reached after extensive molecular dynamics simulations. These case studies were carried out in collaboration with molecular biologists who are studying protein cavities and their effect on the stability of proteins. For all these examples, we use $\alpha = 0$, solvent radius of 1.4Å and BOTTLENECK criterion for cavity connection unless specified otherwise.

5.1 Runtime results

Cavity connection method is implemented as a standalone interactive software in Java 1.6 and OpenGL 3.2. The following experiments were performed on a workstation with 8 core Intel Xeon processor and 16GB of RAM. The program requires weighted Delaunay complex, alpha complex and Delaunay flow as input. We

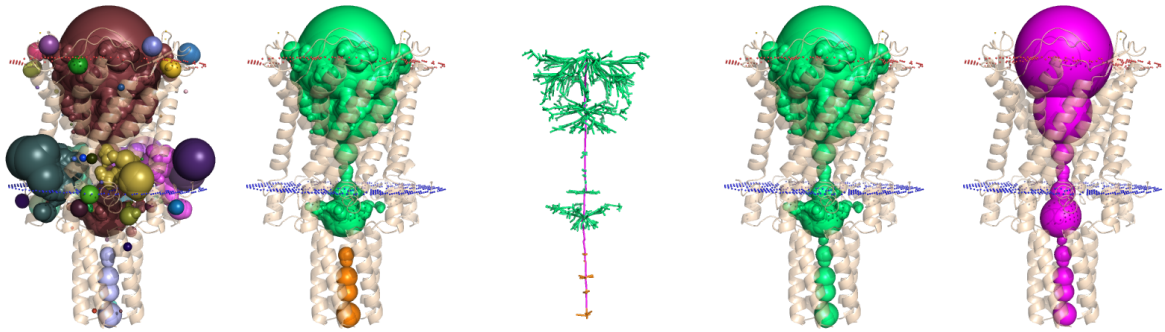
PDB id	#Atoms	Preprocess (sec)
2OAR	5772	0.357
1RHZ	4901	0.336
2YXQ	4853	0.337
2YXR	4789	0.298
1DUK	1256	0.111

Table 1: Preprocessing times for different molecules in the study.



(a) 2OAR (b) (c) (d) 2YXQ (e) (f) (g) 2YXR (h) (i)

Figure 6: Comparison of cavity connection results with RobustCavities. **(a)** The disconnected channel detected as two separate cavities (colored green and orange) in 2OAR. **(b)** The cavity connection result. **(c)** The RobustCavities result. Clearly, the volume of the merged cavity has increased by a significant amount as compared to the result obtained by our cavity connection method. **(d)–(f)** Similar result is obtained for the protein 2YXQ. **(g)–(i)** The result obtained for the protein 2YXR.



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

Figure 7: Cavity connection results for MscL transmembrane protein (PDB id: 2OAR). **(a)** All cavities detected in this protein are shown in the context of the molecule. The membrane is shown as red and blue layers. **(b)** We select two cavities at either end of the membrane for connection. **(c)** The connecting path found by the method is shown in pink. The two cavities shown in the dual graph representation for context. The maximum perturbation for the connecting path was found to be 0.4Å. **(d)** Single connected cavity after atom perturbation. **(e)** The connecting path helps identify the known ion transfer channel.

assume these are already available. Using this as input, we compute the cavities, cavity representatives, cavity attributes such as persistence, G_{MR}, $MaxST$, and pruned $MaxST$ in a preprocessing step. The preprocessing times for the five molecules we discuss in this paper are provided in Table 1. It should be noted that compared to α-complex computation time which takes a few seconds, the preprocessing time is significantly low. After preprocessing, the GUI is set-up and user can choose cavities for connection based on different criteria. The cavity connection time was observed to be in the range 2ms to 20ms for these molecules. This ensures that cavity connection can be done interactively.

5.2 Comparison

We perform qualitative comparison of our results with RobustCavities method proposed by Sridharamurthy *et al.* [26]. RobustCavities also attempts to remove inconsistencies in cavity detection by merging cavities into stable cavities using a global parameter ε. It is claimed that RobustCavities ensures that minimal change is done to the cavity volume by carefully modifying the atomic radii only for atoms lining the split triangles.

Figure 6 shows the result of our method and the RobustCavities for three protein structures. The cavity connection method causes limited perturbation to the atoms along the edges of the detected path (only edges in OR are modified). On the other hand, RobustCavities ends up connecting multiple cavities and significantly changes the volume of the merged cavity. Our method provides more flexibility and finer control over cavity connection, and does very little change to the cavity volume, a desired outcome.

It should be noted that other cavity and channel detection methods have user-defined parameters like solvent radius which can in principle be used to connect cavities. But they are global in nature, and significantly affect the cavity volume. Change in volume induced by RobustCavities is less than that induced by changing the solvent radius and extracting the cavities. Since, our method is performing better than RobustCavities, we expect similar results when compared against other methods.

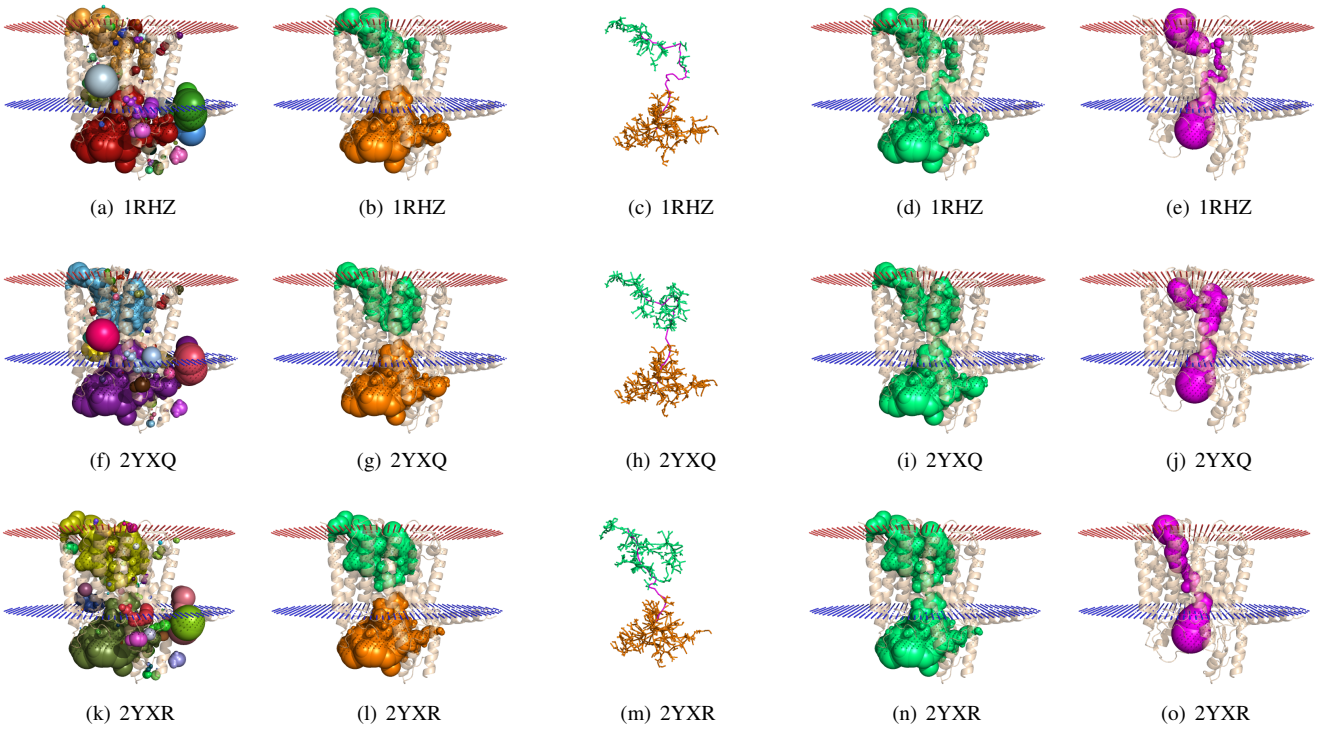
5.3 Mechanosensitive Channel of Large Conductance (MscL): Identifying a channel

This molecule has been used as a running example in this paper (Figures 1 and 4). MscL (PDB id: 2OAR) is a transmembrane ion transport channel. The transmembrane channel is detected as fragmented set of cavities instead of a single connected channel. The user selects two cavities at opposite ends of the channel (Figure 7(b)) and uses cavity connection method to find a good connecting path to merge these cavities. The α_{min} for the connecting path was found to be -1.38, which corresponds to maximum atomic perturbation of 0.4Å. Refer to Figure 7 for detailed results.

5.4 Translocase SecY: Comparing mutants

Translocase SecY is transmembrane transporter protein which forms an integral part of the translocon assembly [29]. In its wild type closed state (PDB id: 1RHZ), the plug domain of the proteins maintains a seal and prevents any leakage [21]. Half and full plug deletion mutants of this protein were created to study this protein (PDB ids: 2YXQ and 2YXR, respectively). Even after plug deletion, these mutants attain packed structures and the channel is detected as a set of fragmented cavities. However, experimental data shows that plug deletions lead to increased rate of translocation of proteins and small molecules.

We applied the BOTTLENECK criterion to connect the cavities along the channel on all the three structures. The detailed results are shown in Figure 8. The α_{min} of connecting paths were found to be -1.76, -1.41 and -1.43 for 1RHZ, 2YXQ and 2YXR, respectively. These values correspond to maximum atomic perturbations of 0.69Å, 0.42Å and 0.43Å, respectively. This clearly indicates that it is easier to open the channel in the mutants as compared to the wild-type, which supports the experimental evidence that the



(a) 1RHZ (b) 1RHZ (c) 1RHZ (d) 1RHZ (e) 1RHZ

(f) 2YXQ (g) 2YXQ (h) 2YXQ (i) 2YXQ (j) 2YXQ

(k) 2YXR (l) 2YXR (m) 2YXR (n) 2YXR (o) 2YXR

Figure 8: The results for Translocase SecY case study. **(a)** The cavities detected in the wild type protein (1RHZ). **(b)** The cavities selected for connection. **(c)** The connecting path (pink) between these cavities. The maximum perturbation for this connecting path was found to be 0.69Å. **(d)** The resulting cavity after perturbation of atoms. **(e)** The connecting path as a channel across the membrane. **(f)–(j)** Similar results for the half plug deletion mutant (2YXQ) of the protein. The maximum perturbation for the connecting path was found to be 0.42Å. **(k)–(o)** The results for the full plug deletion mutant (2YXR) of the protein. The maximum perturbation was found to be 0.43Å.

mutants are more conducive to transport of molecules through the channel.

5.5 Myoglobin: Identifying the migration path

Myoglobin (PDB id: 1DUK) functions as an oxygen storage and delivery protein in the heart and skeletal muscles. The gas molecule binds to the Fe atom present in the heme moiety buried within the protein [3]. This primary binding site where ligand carbon monoxide (CO) binds to Fe is referred to as the distal pocket (DP). Interaction of Myoglobin and Xenon (Xe) has been studied earlier and it was observed that Xe populates four pre-existing cavities in Myoglobin referred to as Xe^1 to Xe^4 [28]. Further, it has been shown in previous molecular dynamics simulation studies that CO occupies the cavities Xe^1 [5] and Xe^4 [3] for the maximum amount of time. Xe^4 is close to the distal pocket while Xe^1 is on the proximal side of the heme. The path taken by CO to migrate from distal side of heme to the proximal side and vice versa is of crucial significance for understanding the functionality of this protein.

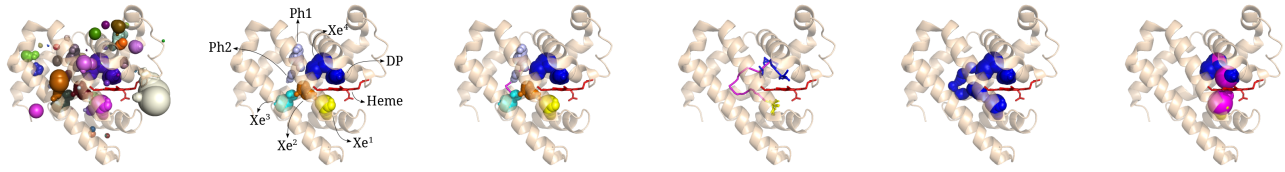
In an extended molecular dynamics simulation study by C. Bossa *et al.* on wild type sperm whale myoglobin, it was observed that over the time-scale of 80 ns, transient cavities form and collapse due to protein dynamics [2]. Two cavities which the authors labelled as Phantom 1 (Ph1) and Phantom 2 (Ph2), highlighted in Figure 9(b), were deemed important. Whereas Ph1 seemed a stable cavity existing for 98% of time, Ph2 was a transient cavity occurring 33.5% of the time over the duration of the simulation. These cavities played a crucial role in movement of CO from DP to Xe^1 since they connected the spatially distant Xe^4 and Xe^3 sites in Myoglobin. It was found that during the course of its journey CO resides in Ph2 for 0.1 ns and inhabits Ph1 for as long as 3.2 ns.

We applied BOTTLENECK criterion of cavity connection to find the connecting path between DP and Xe^1. We found that the connecting path passes through Xe^4, Ph1, Ph2, Xe^3 and Xe^2 to reach Xe^1. The α_{min} for the connecting path was found to be -1.11, which corresponds to maximum atomic perturbation of 0.25Å. This connection is shown in Figures 9(c), 9(d) and 9(e). Thus, we found that the proximal Xe^1 is connected with the distal pocket tracing a direct path through xenon binding sites around the heme moiety. This result agrees completely with the results of the extended molecular dynamics simulation performed earlier [2].

We also applied PROXIMITY criterion to find connecting path between DP and Xe^1. A short direct path with only two edges in *OR* was obtained, as shown in Figure 9(f). But, α_{min} for this connection was found to be -2.56 which corresponds to maximum perturbation of 2.16Å. This clearly suggests that direct connection between distal and proximal sides of the heme is impossible, or highly improbable. Hence, the path obtained by applying BOTTLENECK criterion is biologically significant.

6 Conclusions

Cavity detection methods suffer from unstable behavior due to uncertain nature of protein structure data. We have described a novel solution via connecting molecular cavities under different optimization criteria. We described efficient solutions using an α-complex based internal representation of the cavities and the region occupied by the molecule. The computed connection helps in quantifying the 'connection distance' between cavities. This connection distance signifies the stability of the cavity in the presence of uncertainty. A larger distance implies increased difficulty to connect the cavities. Hence, they are expected to be more stable in uncertain dynamic environments experienced by the protein. An interactive visual interface with linked views aids the user in identifying interesting cavities to connect. There is scope to improve these visualizations and the user experience further. It is important to address the problem of channel and cavity extraction in uncertain data based on sound theoretical foundations. The methods proposed in



(a) (b) (c) (d) (e) (f)

Figure 9: The results for Myoglobin case study. **(a)** The set of cavities detected in Myoglobin bound with heme. **(b)** The cavities of interest in this protein that have been studied earlier are labeled. We are interested in finding the connecting path between DP and Xe^1. **(c)** A connecting path (pink) from Xe^1 is detected that traverses through Xe^2, Xe^3, Ph2, Ph1 and Xe^4 to reach DP. This connection was suggested after extensive molecular dynamics simulations. However, we are able to detect this connection directly using the cavity connection method. The maximum perturbation required for detecting the path is found to be only 0.25Å. **(d)** The detected path (pink) along with dual graph representations of the two selected cavities. **(e)** The merged cavity (blue) formed after atom perturbation. **(f)** Using PROXIMITY criterion for finding the connecting path between DP and Xe^1 results in detection of direct path (pink) which does not pass through other Xe sites. The maximum perturbation for this path was found to be 2.16Å which suggests that direct connection between DP and Xe^1 is highly improbable.

this paper can be adapted for other Voronoi diagram based methods like CAVER [24], MOLE [23] and the techniques proposed by Lindow et al. [18] and Sridharamurthy et al. [26]. We believe that a user-driven flexible cavity connection capability would be a useful addition to these established channel and cavity detection tools.

ACKNOWLEDGEMENTS

We thank Siddharth Patel and Raghavan Vardarajan for suggesting the Translocase SecY and Myoglobin case studies and for evaluating the results. Talha Bin Masood was supported by Microsoft Corporation and Microsoft Research India under the Microsoft Research India PhD Fellowship Award. This work was partially supported by DST under Grant SR/S3/EECE/0086/2012 and the DST Center for Mathematical Biology under grant SR/S4/MS:799/12.

REFERENCES

[1] F. Aurenhammer, R. Klein, and D.-T. Lee. *Voronoi Diagrams and Delaunay Triangulations*. World Scientific, 2013.

[2] C. Bossa, M. Anselmi, D. Roccatano, A. Amadei, B. Vallone, M. Brunori, and A. Di Nola. Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophys. J.*, 86(6):3855–3862, Jun 2004.

[3] M. Brunori and Q. H. Gibson. Cavities and packing defects in the structural dynamics of myoglobin. *EMBO Rep.*, 2(8):674–679, Aug 2001.

[4] S. Chakravarty, A. Bhinge, and R. Varadarajan. A procedure for detection and quantitation of cavity volumes in proteins. *Journal of Biological Chemistry*, 277(35):31345–31353, 2002.

[5] K. Chu, J. Vojtchovsky, B. H. McMahon, R. M. Sweet, J. Berendzen, and I. Schlichting. Structure of a ligand-binding intermediate in wild-type carbonmonoxy myoglobin. *Nature*, 403(6772):921–923, Feb 2000.

[6] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research*, 34(2):W116–W118, 2006.

[7] H. Edelsbrunner. Biological applications of computational topology. In J. E. Goodman and J. O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, pages 1395–1412. CRC Press, 2004.

[8] H. Edelsbrunner. *Computational Topology. An Introduction*. Amer. Math. Soc., 2010.

[9] H. Edelsbrunner and P. Fu. Measuring space filling diagrams and voids. Technical report, UIUC-BI-MB-94-01, Beckman Inst., Univ. Illinois, Urbana, Illinois, 1994.

[10] H. Edelsbrunner and P. Koehl. The geometry of biomolecular solvation. *Combinatorial & Computational Geometry*, 52:243–275, 2005.

[11] M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.*, 15(6):359–363, Dec 1997.

[12] M. Karpusas, W. A. Baase, M. Matsumura, and B. W. Matthews. Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. *Proc. Natl. Acad. Sci. U.S.A.*, 86(21):8237–8241, Nov 1989.

[13] P. Koehl, M. Levitt, and H. Edelsbrunner. Proshape: understanding the shape of protein structures. *Software at biogeometry.duke.edu/software/proshape*, 2004.

[14] C. Lee, S.-H. Park, M.-Y. Lee, and M.-H. Yu. Regulation of protein function by native metastability. *Proceedings of the National Academy of Sciences*, 97(14):7727–7731, 2000.

[15] D. G. Levitt and L. J. Banaszak. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph*, 10(4):229–234, Dec 1992.

[16] J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules: II. inaccessible cavities in proteins. *Proteins Structure Function and Genetics*, 33(1):18–29, 1998.

[17] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities. *Protein Science*, 7(9):1884–1897, 1998.

[18] N. Lindow, D. Baum, and H. Hege. Voronoi-based extraction and visualization of molecular paths. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2025–2034, 2011.

[19] T. B. Masood, H. K. Malladi, and V. Natarajan. Facet-JFA: Faster computation of discrete voronoi diagrams. In *Proc. Indian Conference on Computer Vision Graphics and Image Processing*, ICVGIP '14, pages 20:1–20:8. ACM, 2014.

[20] T. B. Masood, S. Sandhya, N. R. Chandra, and V. Natarajan. ChExVis: a tool for molecular channel extraction and visualization. *BMC Bioinformatics*, 16:119, 2015.

[21] D. L. Minor. Puzzle plugged by protein pore plasticity. *Molecular cell*, 26(4):459–460, 2007.

[22] J. Parulek, C. Turkay, N. Reuter, and I. Viola. Implicit surfaces for interactive graph based cavity analysis of molecular simulations. In *Proc. IEEE Symp. on Biological Data Visualization (BioVis)*, pages 115–122, 2012.

[23] M. Petřek, P. Košinová, J. Koča, and M. Otyepka. MOLE: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, 15(11):1357–1363, 2007.

[24] M. Petřek, M. Otyepka, P. Banáš, P. Košinová, J. Koča, and J. Damborskỳ. Caver: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, 7(1):316, 2006.

[25] F. M. Richards. The interpretation of protein structures: Total volume, group volume distributions and packing density. *Journal of Molecular Biology*, 82(1):1 – 14, 1974.

[26] R. Sridharamurthy, H. Doraiswamy, S. Patel, R. Varadarajan, and V. Natarajan. Extraction of robust voids and pockets in proteins. In *EuroVis-Short Papers*, pages 67–71, 2013.

[27] M. S. Till and G. M. Ullmann. Mcvol-a program for calculating protein volumes and identifying cavities by a monte carlo algorithm. *Journal of molecular modeling*, 16(3):419–429, 2010.

[28] R. F. Tilton, I. D. Kuntz, and G. A. Petsko. Cavities in proteins: structure of a metmyoglobin-xenon complex solved to 1.9 A. *Biochemistry*, 23(13):2849–2857, Jun 1984.

[29] B. van den Berg, W. M. Clemons, I. Collinson, Y. Modis, E. Hartmann, S. C. Harrison, and T. A. Rapoport. X-ray structure of a protein-conducting channel. *Nature*, 427(6969):36–44, 2003.