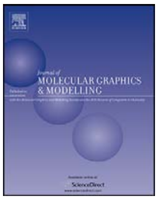
Journal of Molecular Graphics and Modelling 28 (2010) 636–649

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



Sphericity of a protein via the β-complex

Deok-Soo Kim [a,*], Jae-Kwan Kim [a], Chung-In Won [a], Chong-Min Kim [a], Joon Young Park [b], Jong Bhak [c]

[a] Department of Industrial Engineering, Hanyang University, 17 Haengdang-dong, Seongdong-gu, Seoul 133-791, South Korea
[b] Department of Industrial and Systems Engineering, Dongguk Universtiy, 3-26, Pil-Dong, Joong-Gu, Seoul 100-715, South Korea
[c] Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, 52 Eoeun-dong, Yuseong-gu, Daejeon 305-806, South Korea

ARTICLE INFO

Article history:
Received 2 September 2009
Received in revised form 30 December 2009
Accepted 4 January 2010
Available online 15 January 2010

Keywords:
Protein structure
Sphericity
Voronoi diagram
β-complex
β-shape
PDB

ABSTRACT

Molecular shape is a fundamental factor in determining the function of a molecule. As proteins tend to fold into globular shapes, the shape descriptor for protein sphericity is important in understanding molecular functions. In this paper, a definition of protein sphericity is introduced based on the recently developed geometric constructs of the β-complex and β-shape of a protein. The β-complex represents the Euclidean proximity among all the atoms in a protein, and the β-shape is the polyhedron contained within the boundary of the corresponding β-complex. Hence, the β-shape determines the proximity among the atoms on the boundary of a protein. Given the volume of a β-shape, the ratio between the surface area of a sphere with this volume and the surface area of the β-shape itself is a good measure to classify the sphericity of a protein, especially when the radius of a probe is 3.0 Å. The presented measure is invariant to translation and rotation.

1. Introduction

In molecular biology, molecular shape has long been recognized as a fundamental factor in all interactions with other molecules in the cell [1]. Due to the general consensus on the importance of morphological structure to its functions, many studies have been performed to understand the three-dimensional geometric structure of a protein. Therefore, it is often believed that a simple parameter for the overall shape descriptor of a protein would be convenient if it could be easily defined and computed. However, in this area, shape is not sufficiently well-defined and therefore an accurate measure of shape is not available. In addition, the number of atoms in molecules is usually large. For example, a protein with PDB id 1i50 has 28,279 atoms [2,3].

The shape of most proteins is classified into two categories: globular shapes, which are close to a sphere, and fibrous shapes, which look like filaments or rods. Globular proteins, often called spheroproteins, act as enzymes, hormones, transporters of other molecules, stocks of amino acids, and other roles, and they are the most interesting proteins in the design of drugs and understanding of life phenomenon. Their spherical structure is formed by hydrophobic amino acids buried in the core of a protein, whereas hydrophilic amino acids are placed on the boundary of a protein and interact with the solvent. To carry out important tasks in organisms, proteins tend to fold to a globular conformation, and thus are 5–10 kcal/mol more stable than their unfolded counterparts [4]. The entropy of an unfolded protein is large because the rotation around the bonds in the backbone and the side chains is less restricted than in a folded protein [4]. In [5], a proposal was made for a shape descriptor for proteins to meet the need for protein–protein recognition. It was also pointed out in [6] that, especially when small hydrophobic chunks of atoms assemble into an extended cluster, the driving force of the folding process consists of two parts: one proportional to the exposed surface area of the cluster, and the other proportional to the molecular volumes of the separate chunks. Fibrous proteins, often called scleroproteins, are the other class of proteins which look like a long filament or rod. This type of protein is usually an inert structural or storage protein. An example is keratin, which is tough and insoluble.

As the globular proteins are usually the targets of research, evaluating and categorizing the shape of a protein is an important task and several studies have been done. It is interesting, however, that it is difficult to find a well-accepted definition of protein sphericity in the community in spite of some previous efforts. Wootton defined the globularity as the compact-packed arrangement of the residues around a hydrophobic core of a protein [7]. Chang and Bae defined the sphericity of a protein, from the statistical mechanical perturbation theory point of view, by investigating the volumes of the protein and additional molecule used in an experiment [8]. They showed that the protein sphericity significantly affected salt-induced protein precipitation. Røgen and Bohr used the knot theory to classify protein structures based on

[*] Corresponding author. Tel.: +82 2 2220 0472; fax: +82 2 2292 0472.
E-mail address: dskim@hanyang.ac.kr (D.-S. Kim).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.01.001

global geometric shape measures such as Gauss integrals [9]. Tasylor et al. reported an algorithm to compute an ellipsoidal approximation of a protein [10]. The spherical harmonics expansion [11,12] and the Fourier description [13] were also used in spite that they were limited to star-shaped molecules. Due to the lack of a well-accepted definition of shape descriptor, the famous database such as CATH used a semi-automatic classification based on computer visualization [14]. The shape descriptor was also a fundamental building block for a database search for a similar protein of a given query protein [15]. We recommend Zhang and Lu [16] for an excellent survey of shape description techniques.

In this paper, we present a quantitative measure, which is invariant to translation and rotation, for the sphericity of protein based on computational geometric constructs called the β-complex and the β-shape. Based on the presented definition of sphericity, we also present an efficient algorithm to compute the correct sphericity of proteins which is verified through an experiment with 100 proteins selected from the Protein Data Bank [17]. We claim that the thus presented measure is a powerful, simple, global shape descriptor for molecules including proteins.

2. Computational constructs

2.1. Shape models for proteins

To devise the sphericity measure of a protein, we need a geometric model of the protein. Among others, the most popular model is the space-filling model, often called the CPK-model, where a molecule A is defined as $A = \{a_1, a_2, \cdots\}$ where $a_i = (p_i, r_i)$ is a spherical atom with a van der Waals radius r_i and a center p_i. Hence, from a geometric point of view, the hard-sphere model is simply a set of three-dimensional spheres. Fig. 1(a) shows a two-dimensional molecule consisting of six atoms. Hence, we call Fig. 1(a) a *van der Waals model*, abbreviated as *vdw-model*, of the protein.

Given a vdw-model, there are different ways to define the shape, or equivalently the boundary, of a protein when viewed from outside. The model shown in Fig. 1(b) is the boundary of the union of van der Waals atoms and is called the *vdw-boundary* of the protein. Fig. 1(c) shows a smoothed surface, called the *molecular surface*, on the vdw-model by rolling a spherical ball (shown as a black circle) around the vdw-model while the ball is contacting at least one atom. A molecular surface has been used to define the interaction characteristics of a molecule with other small molecules of a solvent where the molecule exists [18–22]. The solvent molecule is approximated by a spherical ball, called a *probe* shown as the black circle in Fig. 1(c), to simplify the computation of the surface characteristics of a molecule since it is too difficult to measure the interaction between a protein and the real molecules of the solvent.

2.2. Voronoi diagram of atoms

Suppose that $P = \{p_1, p_2, \cdots, p_n\}$ where p is a d-dimensional point. Let $\mathrm{VC}(p_i)$ denote the Voronoi cell for p_i defined as $\mathrm{VC}(p_i) = \{x \in \mathbb{R}^d | d(x, p_i) \leq d(x, p_j), i \neq j\}$ where $d(x, p)$ is the Euclidean distance between two points x and p in $\mathbb{R}^d$. Hence, $\mathrm{VC}(p_i)$ is the smallest polyhedron containing p_i in $\mathbb{R}^d$ defined by the bisectors between all pairs of points in P. Then, the Voronoi diagram $\mathrm{VD}(P)$ is defined as $\mathrm{VD}(P) = \{\mathrm{VC}(p_1), \mathrm{VC}(p_2), \cdots, \mathrm{VC}(p_n)\}$. Hence, the Voronoi diagram tessellates the space $\mathbb{R}^d$. In three-dimensional space, the boundary of a Voronoi cell consists of planar facets which are also bounded by line segments and vertices. Hence, $\mathrm{VD}(P)$ is represented as a quadruplet $\mathrm{VD}(P) = (V, E, F, C)$ where $V = \{v_1, v_2, \ldots\}, E = \{e_1, e_2, \ldots\}, F = \{f_1, f_2, \ldots\}$, and $C = \{c_1, c_2, \ldots c_n\}$ are the sets of vertices, edges, faces, and cells in the Voronoi diagram, respectively. In the Voronoi diagram, the connectivity, called the topology, among the vertices, edges, faces, and cells is appropriately stored in a data structure. The dual structure of the Voronoi diagram, the Delaunay triangulation, is a simplicial complex whose useful properties are well-known [23]. Since the conversion between the Voronoi diagram and the Delaunay triangulation takes the linear time with respect to the number of entities in the structures, their topologies are equivalent. In practice, the topology of $\mathrm{VD}(P)$ is usually stored in its dual structure called the Delaunay triangulation. The ordinary Voronoi diagram $\mathrm{VD}(P)$ is the most compact and concise representation of the proximity among the points in Euclidean space. There are many studies on its theory, algorithms, and important applications of the Voronoi diagram in various areas of science and engineering [24,23]. Efficient and robust codes for computing $\mathrm{VD}(P)$ in $\mathbb{R}^2$ and $\mathbb{R}^3$ are also available [25–27]. The Voronoi diagram has long been used in the analysis of biomolecules since Bernal's computation of molecular volume in 1959 [28,29]. Poupon presented an excellent survey about the use of Voronoi diagrams in biology [30].

Suppose that $A = \{a_1, a_2, \ldots, a_n\}$ where $a_i = (p_i, r_i)$ is a d-dimensional sphere (or an atom) with a center p_i and a radius r_i. Hence, $a_i = \{q \in \mathbb{R}^d | \|q - p_i\| \leq r_i\}$. Two atoms may intersect, whereas one cannot contain another. Let $\mathrm{VC}(a_i)$ denote a Voronoi cell for a_i defined as $\mathrm{VC}(a_i) = \{x \in \mathbb{R}^d | d(x, p_i) - r_i \leq d(x, p_j) - r_j, i \neq j\}$. Then, the Voronoi diagram $\mathrm{VD}(A)$ of the atom set A is defined as $\mathrm{VD}(A) = \{\mathrm{VC}(a_1), \mathrm{VC}(a_2), \ldots, \mathrm{VC}(a_n)\}$ which tessellates $\mathbb{R}^d$. $\mathrm{VD}(A)$ is also represented as a quadruplet $\mathrm{VD}(A) = (V^A, E^A, F^A, C^A)$ where $V^A = \{v_1^A, v_2^A, \ldots\}, E^A = \{e_1^A, e_2^A, \ldots\}, F^A = \{f_1^A, f_2^A, \ldots\}$ and $C^A = \{c_1^A, c_2^A, \ldots\}$ are the sets of vertices, edges, faces, and cells in $\mathrm{VD}(A)$, respectively. The topology among these entities is also stored in a data structure. While the topology of $\mathrm{VD}(A)$ can be directly stored in the radial edge data structure, the dual structure called the quasi-triangulation facilitates a more compact storage for the topology, as will be explained in Section 2.3. By definition, a Voronoi vertex is the center of an empty sphere tangent to the boundaries of four nearby atoms, and a Voronoi edge is a locus of points equidistant from the boundaries of three nearby atoms and is a segment of a conic curve. A Voronoi face is the mid-surface between the boundaries of two nearby atoms and is a segment of a hyperboloid of two sheets. In general, the combinatorial complexity of $\mathrm{VD}(A)$ is also quadratic, meaning that the number of vertices, edges and faces

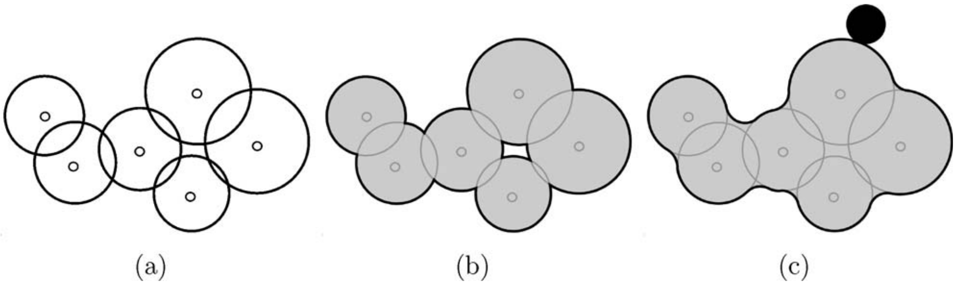


Fig. 1. Boundary models of a two-dimensional protein. (a) The vdw-model of a protein, (b) the vdw-boundary, and (c) the molecular surface with respect to the probe.

are all $O(n^2)$ in the worst case for n general spheres. However, in the world of atoms and molecules, the following are facts: (i) the radii of atoms are fixed as constants (for example, H: 1.20, C: 1.70, O: 1.52, N: 1.55, S: 1.80 [31]), and (ii) two atoms cannot get too close to each other due to the repulsive force between atoms. Due to these facts, the combinatorial complexity of the Voronoi diagram of a molecule is $O(n)$ for n atoms in the worst case. In other words, the number of faces, edges, and vertices in VD(A) for a molecule A are all linear to n. For example, the number of neighboring atoms sharing a Voronoi face with a particular atom in molecules is $O(1)$ in the worst case [22,32,33].

The robust and fast computation of the Voronoi diagram of a molecule has been a big challenge for several years in computational geometry but has only recently become practical. Readers are referred to [34,35] for more details on the properties and the algorithms.

2.3. Quasi-triangulation

Given a Voronoi diagram, it is a usual practice to store its topology in the dual structure for both storage efficiency and manipulational convenience. It is very well-known that the dual of the Voronoi diagram of points is the Delaunay triangulation, which is a simplicial complex. A simplicial complex can conveniently be stored in a compact data structure and the traversal on the topology among the simplexes in the complex is quite easy and efficient [36,37].

However, the dual structure of the Voronoi diagram of atoms, called a *quasi-triangulation*, is not always a valid triangulation. Consider a relatively small disc located in-between two larger discs in $\mathbb{R}^2$. In this case, it is possible that these three discs may define two Voronoi vertices, instead of one, as was pointed out in [38–40]. If we take the dual of the Voronoi diagram of these discs, there are two dual triangles which share two common edges. When these two dual triangles are visualized, the two triangles look identical in $\mathbb{R}^2$. Hence, the dual of this Voronoi diagram is not a valid triangulation in $\mathbb{R}^2$.

Consider a similar configuration in $\mathbb{R}^3$ such that a tiny sphere is located in the middle of three large spheres. In this case, there may be two Voronoi vertices defined from the four spheres. Then, a similar problem may exist and the dual of the Voronoi diagram of these spheres is not a valid triangulation. In this case, the two dual tetrahedra may share three common faces depending on the configuration of the four balls. It is known that a quasi-triangulation has a small number of conditions (which we call *anomalies*) which makes it a non-simplicial complex. In our previous work [41], the definition and properties of quasi-triangulation are well-described and the compact data structure to store it and facilitate an efficient traversal is provided.

We want to emphasize that it is called a quasi-triangulation because the dual of the Voronoi diagram of three-dimensional spheres consists of very few invalid tetrahedra but mostly valid dual tetrahedra. For example, the dual structures of the Voronoi diagram of atoms for the protein models in the Protein Data Bank contain only very few invalid tetrahedra.

The conversion between the Voronoi diagram and the quasi-triangulation takes $O(m)$ time in the worst case where m is the number of geometric entities in either structure. In the molecular world, $m = O(n)$ for n atoms. Hence, the topologies of the Voronoi diagram and the quasi-triangulation are equivalent from both a computational and informational point of view. For the details of quasi-triangulation and the duality between the Voronoi diagram and the quasi-triangulation, readers are referred to [41].

A Voronoi diagram in three-dimensional space is a cell structure, and it is necessary to use a non-manifold data structure such as the radial edge data structure to properly store the topology of the Voronoi diagram. On the other hand, the topology of the quasi-triangulation can be stored in a much simpler data structure called the InterWorld data structure (IWDS) [41]. IWDS consists of three arrays (i.e. a vertex array, a tetrahedron array, and a gate array), and the topology among the simplexes in the quasi-triangulation is represented by indices in the arrays. Therefore, after we compute the Voronoi diagram of spheres, we transform it to the quasi-triangulation. Note that edges and faces are not explicitly represented in IWDS [41]. Since IWDS is very compact and concise, it is useful to store the topology of the quasi-triangulation in the storage.

2.4. β-Shape and β-complex

The constructs called the β-shape and β-complex have been recently proposed, and their properties and computational algorithms are reported in [42,43]. The *β-hull* is defined in a way similar to the α-hull [44]. Consider a three-dimensional space filled with a soft matter with some spherical atoms of varying radii scattered within the matter. Carving out the matter with an omnipresent spherical eraser whose radius is β will result in a shape which we call a β-hull. Since the eraser is omnipresent, there can be interior voids as well. If the spherical eraser is the probe for a water molecule, the boundary of the β-hull is indeed the molecular surface, often called the Connolly surface [45,19]. The spherical eraser is called a *probe*.

Suppose that we have a β-hull of a set $A = \{a_1, a_2, \ldots, a_n\}$ of spherical atoms in $\mathbb{R}^3$. We can straighten the surface of the β-hull by substituting straight edges for the circular arcs and planar triangles for the spherical triangles where the vertices are the centers of the atoms contributing to the boundary of the β-hull. The straightened object bounded by planar facets is the *β-shape* of A. In this paper, we consider that the β-shape is connected to form a single component. Otherwise, each connected component of the β-shape may be handled separately. The β-shape is non-manifold and may have dangling edges, dangling triangles, and even dangling tetrahedra.

Consider again a molecule A. Suppose that a probe b of radius zero is located at a point $x \in \mathbb{R}^3$ which is not contained in any atom in A. Increase the radius of b until it touches the boundary of an atom a_i while its center is fixed at x. While keeping its tangency with a_i, increase the radius of b until it touches another atom a_j. During this radius increase process, the center of b changes. If it touches two atoms, a_i and a_j, increase the radius until it touches another atom a_k, while the tangent contacts are maintained. We increase the radius of b not until it reaches ∞ but until it reaches a predefined value of β. When a_i is touched, a vertex simplex is defined at the center of a_i. When a_i and a_j are touched, an edge simplex is also defined between the centers of two atoms in addition to another new vertex simplex corresponding to a_j. In some cases, b may touch only one atom a_i but not another atom a_j even though its radius increases to ∞. When a_i, a_j, and a_k are touched by the radius increase process, a face simplex is also defined by the three centers in addition to one more vertex and two more edge simplexes. If a_i, a_j, a_k, and a_l are touched by the radius increase process, a tetrahedral cell simplex is similarly defined.

We note the following: (i) the boundary of the β-shape is the exterior boundary of the β-complex, and (ii) the β-complex is a subset of the quasi-triangulation which lies within the boundary of the corresponding β-shape. The β-shape represents the proximity among the atoms on the boundary of a molecule and the β-complex represents the proximity among all atoms in a molecule.

According to the theory of the β-complex, a simplex in a β-complex takes one of three states: singular, regular or interior. A simplex is *singular* if it belongs to the boundary of a β-shape and does not bound any higher dimensional simplex in the corresponding β-complex. A simplex is *regular* if it belongs to the boundary of a β-shape and bounds a higher dimensional simplex in the corresponding β-complex. A simplex is *interior* if it does not belong to the boundary of a β-shape and is the intersection between neighboring higher dimensional simplexes in the corresponding β-complex. Hence, a dangling face is singular, a face separating the outside and inside of a β-shape is regular, and a face inside the boundary of a β-shape is interior. For details, refer to [46].

Above, we described the concepts of a β-shape and a β-complex using an infinite number of probes. However, we note here that their computation is not done in this way, but can be done very efficiently via the analysis of shapes of quasi-triangulation simplexes with respect to the size of atoms and the probe. The algorithm to compute the β-complex (and therefore the β-shape as well) consists of three steps and is summarized as follows: (i) the computation of the Voronoi diagram of a molecule (taking $O(n^3)$ time for general spheres and $O(n^2)$ time for molecules both in the worst case), (ii) the conversion from the Voronoi diagram to the quasi-triangulation (taking $O(n)$ time in the worst case), and (iii) the search for simplexes for the β-complex (taking $O(\log n + k)$ time in the worst case, where there are k simplexes in the β-complex). We note here that the first step, the computation of the Voronoi diagram for all proteins we tested in PDB, shows empirically $O(n)$ time for n atoms.

We want to note that the Voronoi diagram and the quasi-triangulation are used for various applications including one reported in this paper. Hence, the first two steps, the computation of the Voronoi diagram and the corresponding quasi-triangulation, can be done off-line as pre-processing and stored in the database. Therefore, the computation of the β-complex corresponding to a particular value of β requires the third step only.

Fig. 2(a) shows a two-dimensional molecule A consisting of thirteen two-dimensional atoms, and Fig. 2(b) shows the vdw-boundary of A. Note that the vdw-boundary contains two internal voids, a few shallow depressions, and a deep depression on the

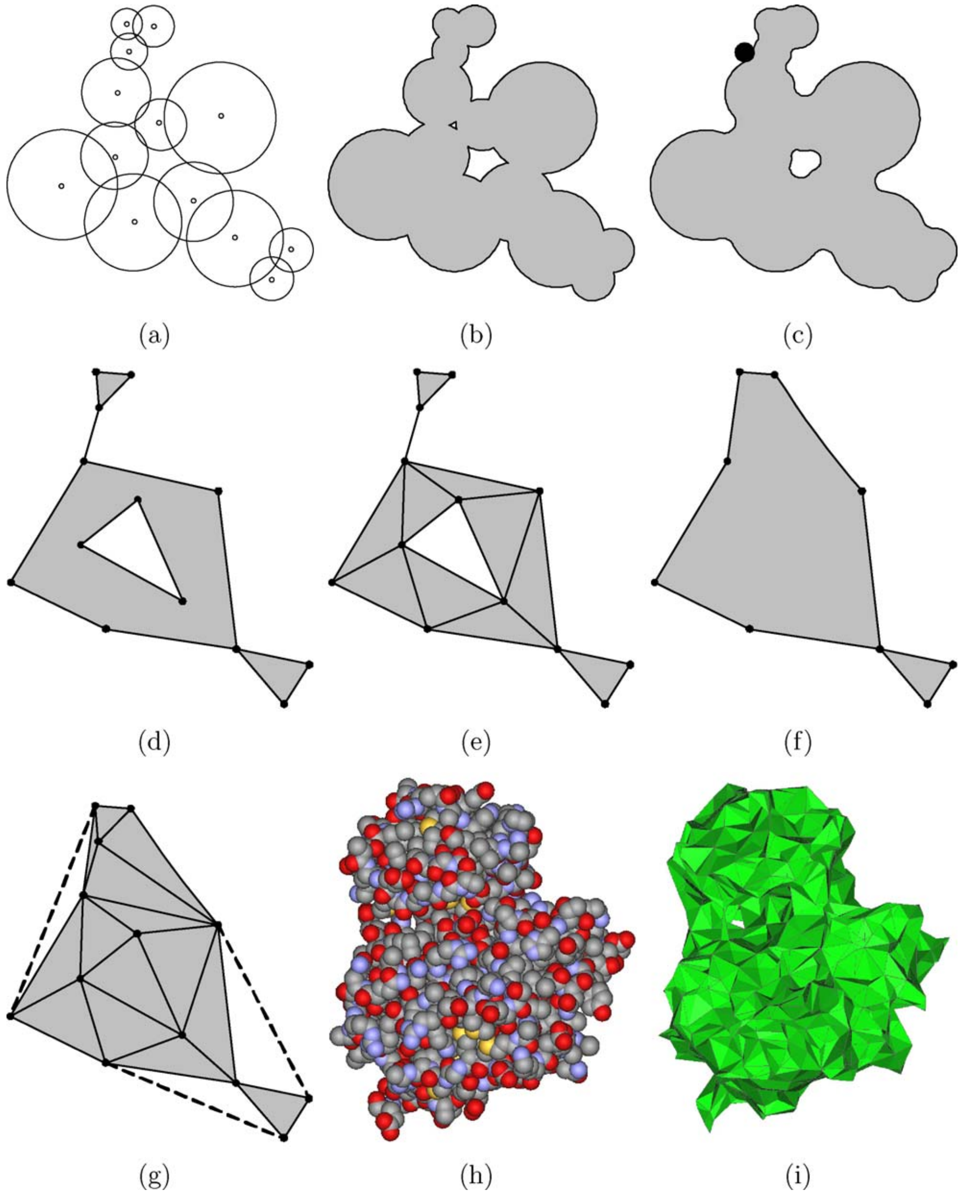


Fig. 2. Protein, the boundaries of the protein, the β-shape, and the β-complex. (a) A set of 13 atoms in $\mathbb{R}^2$, (b) the vdw-boundary, (c) the molecular surface corresponding to a small probe, (d) the corresponding β-shape, (e) the corresponding β-complex, (f) a β-shape corresponding to a larger probe, (g) the corresponding β-complex with simplexes removed from the quasi-triangulation, (h) the protein tyrosine kinase (1xba, 2068 atoms), (i) the β-shape of 1xba corresponding to $\beta = 1.42$ Å.

exterior boundary of the model. Fig. 2(c) is the molecular surface of A corresponding to a small probe (shown in the black circle). Note that only one interior void is left in Fig. 2(c), which corresponds to the larger void in the vdw-boundary of Fig. 2(b). The void corresponding to the smaller one in the vdw-boundary has disappeared. Note also that the shape of the void in the molecular surface is different from that of the vdw-boundary. Hence, the molecular surface removes some local shape characteristics which are usually unnecessarily detailed for applications such as finding drug candidates. Regions where a water molecule cannot fit may not usually interact with other meaningful chemicals. Fig. 2(d) shows the β-shape corresponding to the molecular surface of Fig. 2(c) where the shaded region is the interior of the β-shape. The β-shape in Fig. 2(d) has an interior void corresponding to the void of the molecular surface and a dangling edge corresponding to a pair of atoms that are exposed to or touched by the probe. The dangling edge corresponds to the deep depression in the external boundary of the molecular surface. The boundary of the β-shape has 13 vertices and 15 edges (12 on the exterior boundary and 3 on the interior void). Fig. 2(e) shows the β-complex corresponding to the β-shape in Fig. 2(d). Note that each vertex of the β-shape and β-complex corresponds to an atom. Fig. 2(f) and (g) shows the β-shape and β-complex corresponding to a larger probe, respectively. Note that both the dangling edge and the internal void in the β-shape of Fig. 2(d) have now disappeared. The dotted line segments in Fig. 2(g) form the quasi-triangulation of the molecule A together with the β-complex in Fig. 2(g). Fig. 2(h) shows a three-dimensional protein model, tyrosine kinase (PDB id: 1xba), consisting of 2068 atoms, and Fig. 2(i) shows the β-shape of 1xba corresponding to a probe of the radius 1.42 Å, i.e. a water molecule. The white spot in the middle of the β-shape in fact denotes a tunnel passing through the molecular surface of the protein. Note that the three-dimensional β-shape is not necessarily manifold, either. Details on the β-shape and β-complex can be found in [42,43,46].

A few important remarks: recall that the vdw-boundary may contain internal voids and have several shallow depressions. The vdw-model of protein contains the shape characteristics of the complete details of the protein. However, highly detailed information about the shape may, in fact, become noise in reasoning important shape characteristics, which may be a barrier to research on important applications such as finding drug candidates.

In its β-shape corresponding to a small probe, smaller voids disappear since the probe cannot be placed within the voids unless the probe intersects the molecule. Larger voids may remain. In the β-shape corresponding to a larger probe, larger voids may also disappear. Hence, the β-shape corresponding to an appropriate probe size removes the noisy interior and exterior details of the shape of the protein.

However, raising the size of a probe may also remove meaningful shape characteristics on the boundary of proteins. For example, if we use a probe of the size of infinity, the corresponding β-shape may be end up with the convex-hull of all atom centers. Such a β-shape does not convey many useful shape characteristics of the protein. Therefore, there is a trade-off in the choice of an appropriate probe size.

3. Methods

Let S be a sphere with a radius r in $\mathbb{R}^3$. Suppose that $Vol(S)$ and $Area(S)$ are the volume and the surface area of S given $Vol(S) = 4\pi r^3/3$ and $Area(S) = 4\pi r^2$, respectively. Let $R_S = Vol(S)/Area(S) = r/3$.

A sphere S has the minimal surface area among all shapes with an identical volume $Vol(S)$. Let X be an arbitrary three-dimensional shape where $Vol(X) = Vol(S)$. Let R_X be a ratio defined as $R_X = Vol(X)/Area(X)$. Then, $R_X \le R_S$ because $Area(X) \ge Area(S)$. Let $\rho = R_X/R_S$. Then, $0 < \rho \le 1$. The value of ρ is near 1 if X is close to a sphere and deviates farther away from 1 as the shape of X becomes further away from a sphere. Therefore, ρ can be a good measure for the sphericity of arbitrary three-dimensional shapes.

Let $Vol(vdw)$ and $Area(vdw)$ be the volume and the surface area of a protein represented in the *vdw-model*, respectively. Suppose now that $Vol(vdw)$ and $Area(vdw)$ can be efficiently computed (its computation will be explained in Section 3.1). Let R_{vdw} be the ratio defined as

$$R_{vdw} = \frac{Vol(vdw)}{Area(vdw)}. \tag{1}$$

Note that $\mathcal{C}_\beta$ and $\mathcal{S}_\beta$ are the β-complex and β-shape of a protein, respectively. Consider that the radius of a probe corresponds to a probe with a radius r_β. For example, $r_\beta = 1.42$ Å for the water molecule.

Let $Vol(\mathcal{S}_\beta)$ and $Area(\mathcal{S}_\beta)$ be the volume and boundary area of $\mathcal{S}_\beta$ that corresponds to a probe of radius r_β, respectively. Suppose that $i(X)$ denotes the *interior* of a shape X. Then, we define $Vol(\mathcal{S}_\beta) = Vol(\partial i\mathcal{S}_\beta)$ and $Area(\mathcal{S}_\beta) = Area(\partial i\mathcal{S}_\beta)$. $Vol(\mathcal{S}_\beta)$ can be computed as the summation of the volumes of the interior tetrahedral cells of $\mathcal{C}_\beta$, and $Area(\mathcal{S}_\beta)$ can be computed as the summation of the areas of the regular triangular faces of $\mathcal{S}_\beta$. Let m_{TC} be the number of tetrahedral cells in $\mathcal{C}_\beta$ and m_{TF} be the number of triangular faces in $\partial\mathcal{S}_\beta$. Then, $Vol(\mathcal{S}_\beta)$ and $Area(\mathcal{S}_\beta)$ can be computed in $O(m_{TC})$ and $O(m_{TF})$ time in the worst case. In this paper, *area* denotes the area of a boundary surface of a related shape unless otherwise stated. We call the β-shape and β-complex of a protein altogether a *β-model*, and we will use $Vol(\beta)$ and $Area(\beta)$ to denote $Vol(\mathcal{S}_\beta)$ and $Area(\mathcal{S}_\beta)$, respectively, for notational simplicity. Let a ratio R_β be defined as

$$R_\beta = \frac{Vol(\beta)}{Area(\beta)}. \tag{2}$$

Let $\mathcal{I}_X$ be a sphere where $Vol(\mathcal{I}_X) = Vol(X)$ for an arbitrary shape X. Let r_X be the radius of $\mathcal{I}_X$. We call $\mathcal{I}_X$ an *ideal sphere* of X. Hence, $\mathcal{I}_{vdw}$ and $\mathcal{I}_\beta$ are the ideal spheres of the vdw-model and the β-model of a given protein, respectively. Hence, $\mathcal{I}_{vdw}$ is the sphere with the minimum surface area among all shapes with the volume $Vol(vdw)$. $\mathcal{I}_\beta$ has the same property. Let $R^{\mathcal{I}}_{vdw}$ and $R^{\mathcal{I}}_\beta$ be defined as follows:

$$R^{\mathcal{I}}_{vdw} = \frac{Vol(\mathcal{I}_{vdw})}{Area(\mathcal{I}_{vdw})} = \frac{r_{vdw}}{3} \quad \text{and} \quad R^{\mathcal{I}}_\beta = \frac{Vol(\mathcal{I}_\beta)}{Area(\mathcal{I}_\beta)} = \frac{r_\beta}{3}. \tag{3}$$

Definition 1. Let $\rho_{vdw} = R_{vdw}/R^{\mathcal{I}}_{vdw}$ and $\rho_\beta = R_\beta/R^{\mathcal{I}}_\beta$. Then, ρ_{vdw} and ρ_β are called the sphericity of the vdw-model and β-model of a protein, respectively.

Then, it is not difficult to show the following: $\rho_{vdw} = Area(\mathcal{I}_{vdw})/Area(vdw)$ and $\rho_\beta = Area(\mathcal{I}_\beta)/Area(\beta)$. Therefore, ρ_{vdw} is the ratio of the boundary of an ideal sphere which has the same volume of the vdw-model of a protein to the boundary area of the van der Waals molecule itself. Note that ρ_{vdw} and ρ_β are dimensionless and $0 < \rho_{vdw} \le 1$ and $0 < \rho_\beta \le 1$.

To measure ρ_{vdw} and ρ_β, it is necessary to compute $Vol(vdw), Area(vdw), Vol(\beta)$, and $Area(\beta)$. Then, the ideal spheres $\mathcal{I}_{vdw}$ and $\mathcal{I}_\beta$ can be computed.

3.1. Volume and area of vdw-models

Suppose that we want to compute the area of the four overlapping disks shown in Fig. 3(a). Fig. 3(b) shows the β-complex of the four disks corresponding to $\beta = 0$. Obviously, the area of the union of the four disks is given as follows: $Area(D_1 \cup D_2 \cup D_3 \cup D_4) = \sum_i Area(D_i) - \sum_{i<j} Area(D_i \cap D_j) + \sum_{i<j<k} Area(D_i \cap D_j \cap D_k) - Area(D_1 \cap D_2 \cap D_3 \cap D_4)$, where $i, j, k \in \{1, 2, 3, 4\}$.

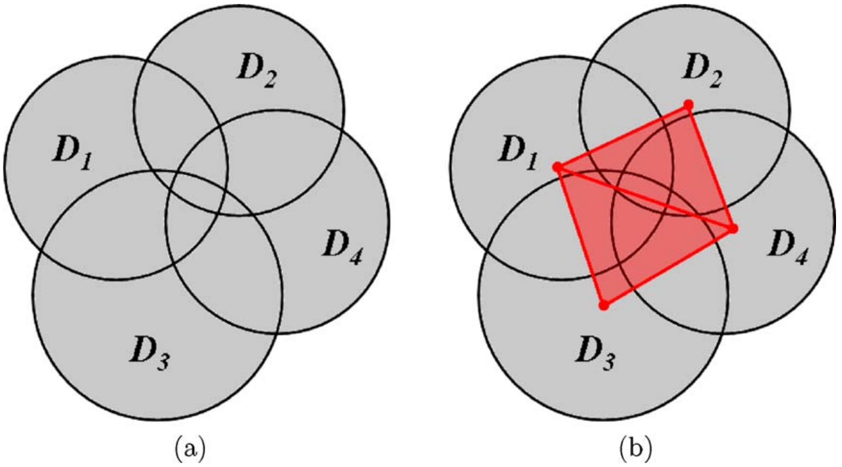


Fig. 3. Area of the four overlapping disks. (a) The configuration of the four disks, and (b) the β-complex of the given disks.

Note that the above formula seems to have 15 terms: 4 single disk terms, 6 pairwise terms, 4 triplet-wise terms, and 1 quadruplet-wise term. However, it turns out that some terms in the above formula cancel out. Since $Area(D_1 \cap D_2 \cap D_3 \cap D_4) = Area(D_1 \cap D_2 \cap D_3) + Area(D_2 \cap D_3 \cap D_4) - Area(D_2 \cap D_3)$, the above formula simplifies to the following with only 11 terms remaining: $Area(D_1 \cup D_2 \cup D_3 \cup D_4) = Area(D_1) + Area(D_2) + Area(D_3) + Area(D_4) - \{Area(D_1 \cap D_2) + Area(D_1 \cap D_3) + Area(D_1 \cap D_4) + Area(D_2 \cap D_4) + Area(D_3 \cap D_4)\} + Area(D_1 \cap D_2 \cap D_4) + Area(D_1 \cap D_3 \cap D_4)$.

The eleven terms left correspond to the simplexes in the β-complex in Fig. 3(b): Each singular term corresponds to each vertex in the β-complex; each pairwise term corresponds to each edge; and each triplet-wise term corresponds to each triangle. The terms canceled out do not have any corresponding simplexes in the β-complex. For example, the edge between the centers of D_2 and D_3, which corresponds to the dropped term $Area(D_2 \cap D_3)$, does not exist in the β-complex. Therefore, the computation of $Area(D_2 \cap D_3)$ is not necessary. Similarly, $Area(D_1 \cap D_2 \cap D_3)$, $Area(D_2 \cap D_3 \cap D_4)$, and $Area(D_1 \cap D_2 \cap D_3 \cap D_4)$ are not necessary to consider in computing the correct solution. This implies that $Area(D_1 \cup D_2 \cup D_3 \cup D_4)$ can be correctly computed by evaluating the terms corresponding to the simplexes constituting the β-complex of the four disks. A similar property holds for the cases of the three-dimensional atoms: Given a set A of three-dimensional atoms, the correct volume and the area of the boundary of the union of atoms can be computed from the beta complex of A.

Efficient computation of the correct volume and area of vdw-model has long been one of the most important research topics in computational molecular biology. There were two groups approach: analytic and non-analytic. The non-analytic approach was mostly based on Monte Carlo simulation or the enumeration of grid points of certain type, and the following studies fall in this group: Shrake and Rupley [47], Gavezzotti [48], Higo and Go [49], Karfunkel and Eyraud [50], Silla et al. [51,52], Abagyan et al. [53], and Eisenhaber et al. [54].

In the more important analytical approach, there are two technical issues that need to be addressed: (i) Formulas for the volume or area of various types of intersections among atoms, and (ii) the combinatorial structure of intersections. The first issue was studied by Richmond [55], Connolly [56], Lustig [57], and Gibson and Scheraga [58,59]. We consider the formulas reported by Gibson and Schraga [58,59] are the most thorough for the first issue. For the second issue, Kratky observed that the intersection of identically-sized disks in $\mathbb{R}^2$ can always be reduced to the signed sum of intersections of less than four disks [60]. Therefore, this observation provided a key to the reduction of the problem complexity of evaluating the union of disks. Later, Naiman and Wynn observed that the observation by Kratky could be generalized for balls with arbitrary radii in $\mathbb{R}^d$ for an arbitrary d [61]. They also observed the following: (i) the intersection among at most $d+1$ balls in $\mathbb{R}^d$ could be sufficient for getting a correct solution, and (ii) this information could be induced from a simplcial complex conveying the intersections among balls. Based on this observation, Edelsbrunner used the power diagram (or its dual structure, the regular triangulation) to derive a compact formula for the inclusion–exclusion principle among balls [62]. Since both the Voronoi diagram of spheres and the power diagram have identical information about the intersections among spherical balls, both the quasi-triangulation and the regular triangulation also have identical information about the intersections. Therefore, the simplexes in the β-complex when $\beta = 0$ can be used for the efficient, correct evaluation of the inclusion–exclusion formula presented in [62,63].

3.2. Volume and area of β-models

Let f be a triangular face in a β-complex $\mathcal{C}_\beta$ and $Area(f)$ be the area of f. Let τ be a tetrahedral cell defined by a triangle f and another vertex v. Then, the volume of τ, $Vol(\tau)$, is given by $Vol(\tau) = Area(f) \cdot h/3$ where h is the distance of v from the plane defined by f. Note that the simplexes in the β-shape and the β-complex contain all the information necessary to determine the geometry of the system of atoms. The computation of $Vol(\beta)$, and $Area(\beta)$ can be done in the linear time of the number of cell simplexes in the β-complex and the number of boundary faces in the β-shape in the worst case, respectively.

4. Results and discussions

We selected a set of 100 sample protein models as described in Table 1, which we will call a test set, with high resolutions from PDB. In the parenthesis of each entry of Table 1, the first number denotes the number of atoms in the protein model and the second number denotes the resolution of the model. Using this model set, we performed experiments to compute the volumes, surface areas, ratios, and so on. The experiments were done on a cluster computer with 118 nodes; Each node has two CPU's with 2 GB RAM and each CPU has an AMD Opteron Dual Core 2.2 GHz processor.

Table 2 shows the computation times for important geometric constructs necessary for the sphericity of eleven representative models selected from the test set (these representative models are shown in Fig. 15). In the table, column A is the time required to

Table 1
PDB ID's for the 100 selected protein models from PDB. In parenthesis, the first number denotes the number of atoms and the second number denotes the resolution.

1c26	(269, 1.70),	1d2k	(3083, 2.20),	1d4t	(915, 1.10),	1dc9	(1058, 2.10),	1dqz	(4362, 1.50)
1eai	(4540, 2.40),	1edq	(4137, 1.55),	1eqp	(3212, 1.90),	1ezg	(1106, 1.40),	1f60	(4091, 1.67)
1fa8	(2002, 1.70),	1fhl	(2598, 2.30),	1i8k	(1820, 1.80),	1iz9	(4978, 2.00),	1j27	(778, 1.70)
1jez	(4770, 2.20),	1jyh	(1256, 1.80),	1k1b	(1712, 1.90),	1l7a	(5074, 1.50),	1lbw	(3932, 2.00)
1lf1	(2348, 1.70),	1lhp	(4808, 2.10),	1lz1	(1029, 1.50),	1m0z	(4113, 1.85),	1mhn	(465, 1.80)
1mn6	(4191, 2.20),	1orj	(3847, 2.25),	1qb5	(3750, 1.90),	1qkd	(946, 1.49),	1qq1	(4110, 1.80)
1qxh	(2448, 2.20),	1r2t	(3731, 2.25),	1rav	(1952, 2.20),	1rh9	(2971, 1.50),	1sh5	(7688, 2.00)
1swh	(3446, 1.70),	1syq	(2199, 2.42),	1t45	(2642, 1.90),	1t4q	(1222, 2.10),	1t6f	(618, 1.47)
1t7n	(4776, 1.90),	1tp6	(984, 1.50),	1ugq	(3466, 2.00),	1wlg	(4312, 1.80),	1wu3	(1390, 2.15)
1x7f	(2790, 2.30),	1xg2	(3573, 1.90),	1xh3	(3186, 1.48),	1xix	(2688, 2.00),	1xqo	(2054, 1.03)
1xwg	(3519, 1.85),	1y0m	(508, 1.20),	1y2t	(2270, 1.50),	1y9u	(2387, 1.39),	1yck	(1306, 1.70)
1ym5	(2292, 2.05),	1ypf	(4547, 1.80),	1zlm	(477, 1.07),	1zpw	(663, 1.64),	1zrs	(4505, 1.50)
1zvt	(3705, 1.70),	1zx6	(448, 1.60),	2a8f	(1428, 1.35),	2ab0	(2900, 1.10),	2car	(3028, 1.09)
2cwc	(2181, 1.65),	2cwl	(4641, 1.65),	2ekc	(4054, 2.00),	2erw	(402, 1.40),	2esk	(1187, 1.36)
2et6	(4465, 2.22),	2f6l	(2509, 1.70),	2f82	(3510, 2.10),	2fn9	(4362, 1.40),	2fp8	(4772, 2.30)
2fts	(3298, 2.41),	2g7o	(544, 1.40),	2g85	(2829, 2.22),	2gas	(4840, 1.60),	2ge7	(1686, 2.00)
2ggv	(1621, 1.80),	2goi	(3054, 2.30),	2gpo	(1858, 1.95),	2guv	(2420, 1.40),	2h2r	(2167, 1.50)
2h3l	(1543, 1.00),	2i3f	(3348, 1.38),	2i49	(3115, 1.35),	2igd	(468, 1.10),	2nls	(271, 0.98)
2o37	(643, 1.25),	2o7h	(1596, 1.86),	2obi	(1330, 1.55),	2ol7	(3655, 1.35),	2op6	(1145, 1.85)
2p19	(3344, 2.10),	2yz1	(1754, 1.40),	3b7h	(597, 2.00),	3bxy	(2114, 2.00),	4eug	(1789, 1.40)

Table 2
Computational requirements for each step for ten example proteins (unit: sec). Note that the computation for each model was done at a core, and therefore the four models were processed at a node in the cluster computer.

PDB ID	#Atoms	VD	QT	β-Shape	*Vol*(β)	*Area*(β)	*Vol*(vdw)	*Area*(vdw)
		(A)	(B)	(C)	(D)	(E)	(F)	(G)
1lf1	2348	13.13	0.05	0.84	0.01	0.00	0.02	0.02
2i49	3115	17.89	0.07	1.16	0.01	0.00	0.03	0.03
1t4q	1222	6.56	0.02	0.42	0.01	0.00	0.01	0.01
2gpo	1858	10.29	0.03	0.67	0.01	0.00	0.02	0.01
2op6	1145	6.03	0.03	0.39	0.00	0.00	0.01	0.01
2yz1	1754	9.60	0.03	0.62	0.01	0.00	0.02	0.01
1zrs	4505	25.98	0.13	1.70	0.03	0.00	0.04	0.04
1iz9	4978	28.69	0.10	1.90	0.01	0.00	0.05	0.05
2h2r	2167	11.98	0.04	0.79	0.00	0.01	0.02	0.02
1eai	4540	25.39	0.12	1.70	0.00	0.02	0.04	0.04
2ol7	3655	20.32	0.07	1.35	0.01	0.00	0.03	0.04

compute the Voronoi diagram; column B is for the conversion from the Voronoi diagram to the quasi-triangulation; and column C for computing the β-shape for $\beta = 1.4$. The curves for these three statistics are shown for all proteins in the test set in Fig. 4. In Fig. 4, the horizontal axis is the number of atoms in each protein and the vertical axis is the computation time in the unit of seconds. The curve with blue rectangles is for the Voronoi diagram, the curve with red circles is for the quasi-triangulation, and the curve with green crosses is for the β-complex. Note that all the three curves show strong linearity with respect to the number of atoms in proteins.

We want to emphasize that the Voronoi diagrams of all proteins can be pre-computed and stored in a database since we use the Voronoi diagram of a protein not only for this research but also for many other studies for the protein. The Voronoi diagrams are stored in the database in the dual structure, the quasi-triangulation, since it requires only very compact memory. The Voronoi diagram and the quasi-triangulation are application independent neutral geometric constructs. Hence, the computations in the columns A and B are preprocessing steps which can be done off-line.

The columns D, E, F, and G denote the computation times of *Vol*(β), *Area*(β), *Vol*(vdw), and *Area*(vdw) for the representative proteins, respectively, and Fig. 5 shows these computation times for the whole test set. Note that the computations of all four statistics are very quick. For both the volume and area, the computational requirement for the β-model is much faster than that of the vdw-model because the formulas for the β-model contains only a few linear terms while the formulas for the vdw-model contains several non-linear terms. Note that the computation times for the volume and the area for the vdw-model are almost identical. All the four curves show strong linear behavior with respect to the number of atoms. In the course of measuring these computation times, we performed the same computation 1000 times to measure meaningful time statistics and then divided the measured time by 1000 since a single computation for a protein is too quick to be measured by the program.

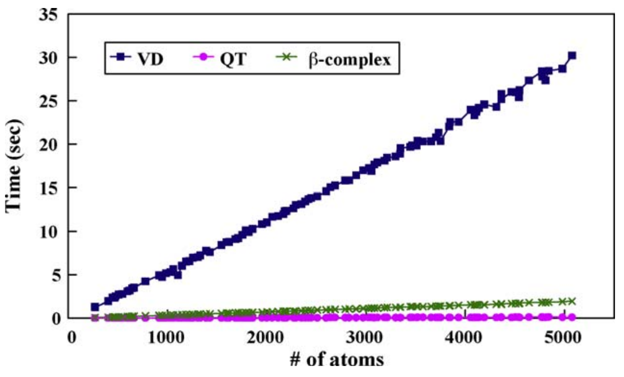


Fig. 4. Computation times of the Voronoi diagrams (rectangles), the conversion to the quasi-triangulations (crosses), and the β-complexes (circles) for all 100 test models (unit: sec).

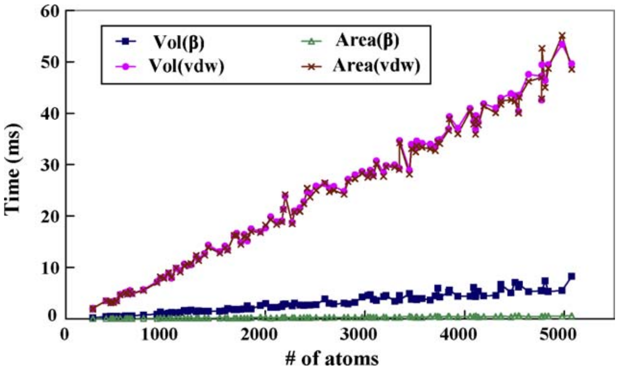


Fig. 5. Computation times of *Vol*(β), *Area*(β), *Vol*(vdw), and *Area*(vdw) for all 100 test models (unit: ms). Note that the two curves for *Vol*(vdw) and *Area*(vdw) are almost identical.

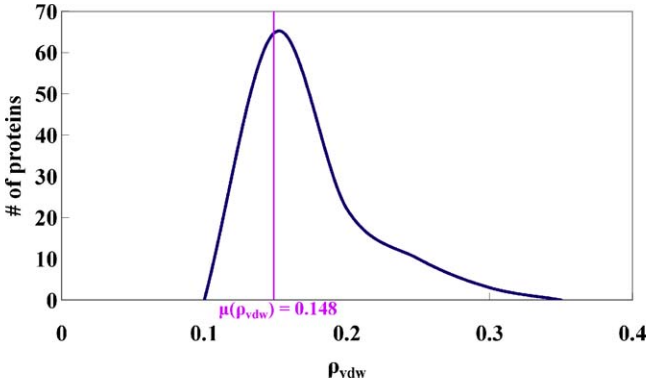


Fig. 6. Frequency distribution of 100 ρ_{vdw}'s. The average and the standard deviation of these ρ_{vdw}'s are 0.148 and 0.041, respectively.

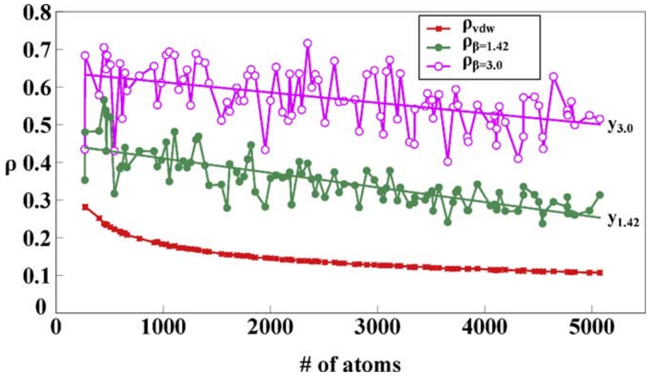


Fig. 7. Distribution of ρ_{vdw} and ρ_{β} for the 100 sample proteins (β=1.42 and 3.0). The regression lines are given as follows: $y_{1.42} = -0.000039x + 0.449$ and $y_{3.0} = -0.000028x + 0.641$.

Fig. 7 shows the distributions of the sphericity ρ_{vdw} and ρ_{β} for the whole test set where $\beta = 1.42$ and 3.0 Å. The red rectangular dots correspond to ρ_{vdw} of proteins and strongly shows an almost monotonic decreasing behavior with respect to the number of atoms in proteins, and therefore it shows a strong trend (or, bias) of ρ_{vdw} due to the number of atoms in a protein. We computed the regression curve of ρ_{vdw} as follows:

$$\rho_{vdw} = \frac{1}{(n - 96.4487)^{0.1845}} - 0.1027 \tag{4}$$

with the residual sum of squares of 0.0001719 where n is the number of atoms in proteins. It is noteworthy how extremely well ρ_{vdw} fits to a curve. Hence, we conclude that ρ_{vdw} does not discriminate the shape differences among proteins at all and therefore it is not very useful as a sphericity measure for proteins. Fig. 6 shows the distribution of ρ_{vdw}'s of 100 proteins in the test set. Note that the average and the standard deviation of these ρ_{vdw}'s are 0.148 and 0.041, respectively.

In Fig. 7, the green dots correspond to ρ_{β} where $\beta = 1.42$ and the red empty circles correspond to ρ_{β} where $\beta = 3.0$. Let $y_{1.42}$ and $y_{3.0}$ denote the regression lines corresponding to $\beta = 1.42$ and 3.0, respectively. To be specific, $y_{1.42} = -0.000039x + 0.449$ and $y_{3.0} = -0.000028x + 0.641$. Compared to ρ_{vdw} curve, the curves for ρ_{β}'s show a strong fluctuation around the corresponding regression lines. Hence, we observe that ρ_{β} reflects the shape variations among the proteins better than ρ_{vdw}. From this experiment, we conclude that ρ_{β} is a better sphericity measure than ρ_{vdw}. Note that ρ_{β} still conveys the bias due to the size of the protein itself.

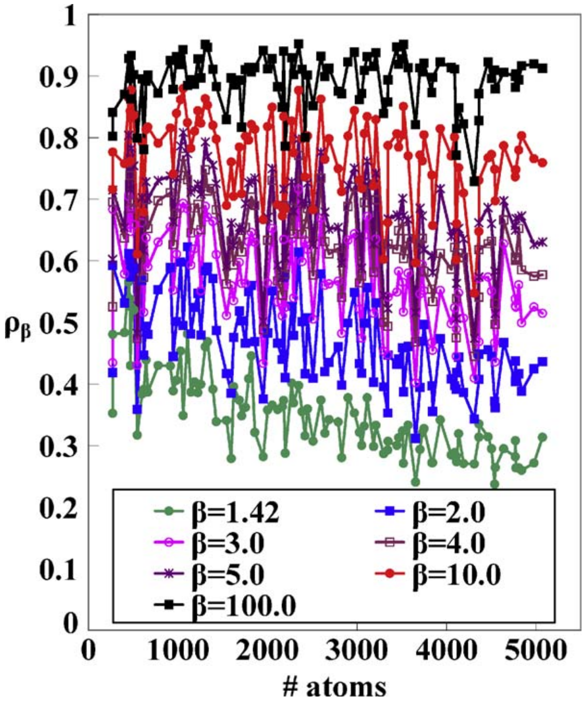


Fig. 8. Distributions of ρ_{β} for the whole test set at various β-values.

We further investigated the influence of the value of β on the distribution of ρ_{β}. Fig. 8 shows the distributions of ρ_{β} for the whole test set where $\beta = 1.42$, 2.0, 3.0, 4.0, 5.0, 10.0, and 100.0 Å from which we make two observations: first, the larger the value of β is (i.e. the larger the probe is), the higher the curve is located. This means that the average value of ρ_{β} gets larger as the value of β becomes larger. The $\mu(\rho_{\beta})$ row in Table 3 shows the average of the ρ_{β} values of the 100 proteins in the test set computed at 10 selected values of β (1.42, 2.0, 3.0, 4.0, 5.0, 7.0, 10.0, 20.0, 50.0, and 100.0 Å). Note that $\mu(\rho_{\beta})$ increases monotonically with respect to the increase of the β value. Fig. 9 shows the curves of ρ_{β} for the 11 representative proteins (1lf1, 2i49, 1t4q, 2gpo, 2op6, 2yz1, 1zrs, 1iz9, 2h2r, 1eai, and 2ol7) computed at the ten selected values of β. The horizontal axis denotes the values of β and the vertical axis denotes the value of ρ_{β}. This figure clearly shows that all eleven ρ_{β} curves show an identical pattern of monotonic increase with respect to the β value.

Second, the larger the probe is, the less bias the protein size has on the ρ_{β} curve. Fig. 10 shows the regression lines corresponding to the five ρ_{β} curves in Fig. 8, and this figure clearly shows that the regression line gets more horizontal as the β value increases. In addition to $y_{1.42}$ and $y_{3.0}$ above, $y_{2.0} = -0.000032x + 0.552$, $y_{4.0} = -0.000024x + 0.694$, $y_{5.0} = -0.000022x + 0.728$, $y_{10.0} =$

Table 3
The standard deviation of ρ_{β}'s for 10 selected values of β.

β	1.42	2.0	3.0	4.0	5.0	7.0	10.0	20.0	50.0	100.0
$\mu(\rho_{\beta})$	0.350	0.471	0.570	0.631	0.671	0.725	0.768	0.829	0.876	0.890
$\sigma(\rho_{\beta})$	0.069	0.074	0.077	0.075	0.075	0.072	0.069	0.062	0.049	0.044
σ / μ	0.198	0.157	0.135	0.118	0.112	0.100	0.090	0.075	0.056	0.050

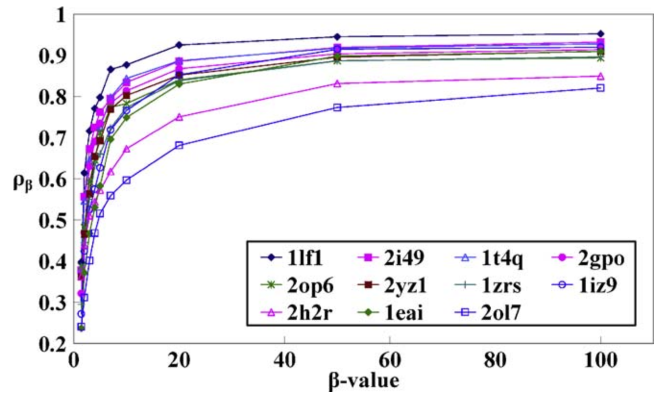


Fig. 9. Distribution of ρ_β values for some protein (1lf1, 2i49, 1t4q, 2gpo, 2op6, 2yz1, 1zrs, 1iz9, 2h2r, 1eai, and 2ol7) with respect to the value of β.

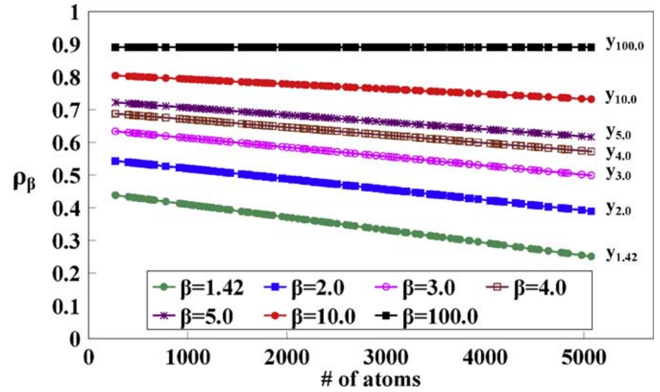


Fig. 10. Regression lines corresponding to the seven ρ_β curves in Fig. 8: $y_{1.42} = -0.000039x + 0.449$, $y_{2.0} = -0.000032x + 0.552$, $y_{3.0} = -0.000028x + 0.641$. $y_{4.0} = -0.000024x + 0.694$, $y_{5.0} = -0.000022x + 0.728$, $y_{10.0} = -0.000015x + 0.808$, and $y_{100.0} = 0.891$.

$-0.000015x + 0.808$, and $y_{100.0} = 0.891$. With this observation, it seems better to use a larger value of β to use ρ_β as the sphericity descriptor. However, it turns out that this is not always the case as there are trade-off factors with respect to the size of the proteins.

To find the optimal value of β, we studied the distribution of the standard deviation $\sigma(\rho_\beta)$ of 100 ρ_β values for the test set at each selected β value. The row $\sigma(\rho_\beta)$ of Table 3 contains $\sigma(\rho_\beta)$ values. Fig. 11 shows the curves of $\sigma(\rho_\beta)$, $\mu(\rho_\beta)$, and $\mu(\rho_\beta)/\sigma(\rho_\beta)$.

Fig. 12 shows a closer look at the behavior of $\sigma(\rho_\beta)$. The maximum value of the standard deviation occurs at $\beta = 3.0$ Å, so we can conclude that the ρ_β curve corresponding to $\beta = 3.0$ Å has the highest discrimination power for the sphericity among proteins. Note that the peak of the curve is very sharp at $\beta = 3.0$ Å and the discrimination power mostly decreases rapidly as the size of the probe gets off from $\beta = 3.0$ Å. The second peak is at 5.0 Å.

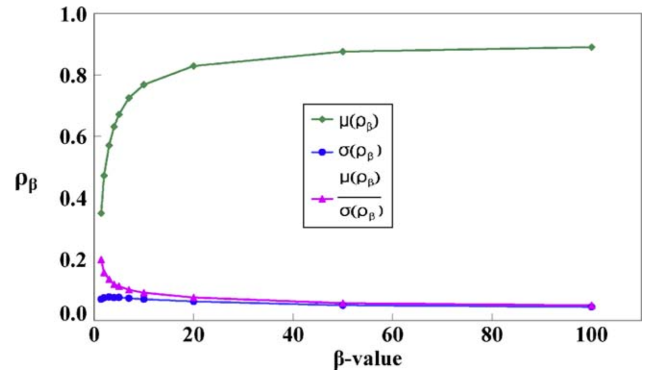


Fig. 11. Distribution of $\mu(\rho_\beta)$, $\sigma(\rho_\beta)$, and μ/σ for the whole test set at some values of β.

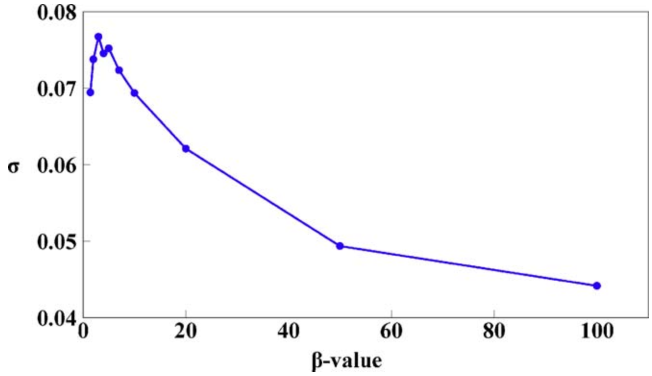


Fig. 12. Distribution of the standard deviation of the ρ_β's of all 100 test proteins with respect to β.

To verify the appropriateness of $\beta = 3.0$ Å for the sphericity of proteins, we performed an additional experiment. Fig. 13 shows the cross-plots of the ρ_β of each protein for six β values (1.42, 2.0, 4.0 and 5.0, 10.0, and 100.0 Å) with respect to $\beta = 3$ Å. Fig. 13(a) shows the distribution of the ρ_β values of each protein in the test set, where the horizontal axis denotes the ρ_β value at 1.42 Å and the vertical axis denotes the ρ_β value at 3.0 Å. For example, the protein marked in Fig. 13(a), 2cwl, has a ρ_β value of 0.295 at $\beta = 1.42$ Å and 0.628 at $\beta = 3.0$ Å. The cross-plots in Fig. 13(a) (between 1.42 and 3.0 Å), Fig. 13(b) (between 2.0 and 3.0 Å), Fig. 13(e) (between 10.0 and 3.0 Å), and Fig. 13(f) (between 100.0 and 3.0 Å) show that the corresponding ρ_β's are less correlated. On the other hand, ρ_β's between 4.0 and 3.0 Å (Fig. 13(c)), and between 5.0 and 3.0 Å (Fig. 13(d)) are more correlated.

Table 4
Statistics of eleven representative proteins selected from the 100 test proteins. The selection was made according to the ranks of ρ_β computed using $\beta = 3.0$ Å.

PDBID	#Atoms	ρ_β					Rank				
		$\beta = 1.42$	$\beta = 2.0$	$\beta = 3.0$	$\beta = 4.0$	$\beta = 5.0$	$\beta = 1.42$	$\beta = 2.0$	$\beta = 3.0$	$\beta = 4.0$	$\beta = 5.0$
1lf1	2348	0.397	0.614	0.716	0.771	0.798	21	2	1	1	3
2i49	3115	0.378	0.556	0.672	0.725	0.762	30	18	10	11	11
1t4q	1222	0.386	0.547	0.646	0.691	0.724	28	23	20	23	28
2gpo	1858	0.321	0.469	0.630	0.690	0.734	59	47	30	25	18
2op6	1145	0.387	0.481	0.593	0.645	0.711	27	40	40	43	35
2yz1	1754	0.362	0.466	0.563	0.654	0.693	38	50	50	38	42
1zrs	4505	0.295	0.445	0.550	0.632	0.660	75	59	60	52	61
1iz9	4978	0.272	0.424	0.525	0.575	0.627	91	68	70	79	77
2h2r	2167	0.363	0.440	0.511	0.544	0.573	37	62	80	86	91
1eai	4540	0.237	0.372	0.467	0.530	0.583	100	94	90	91	88
2ol7	3655	0.241	0.311	0.402	0.468	0.516	99	100	100	99	95

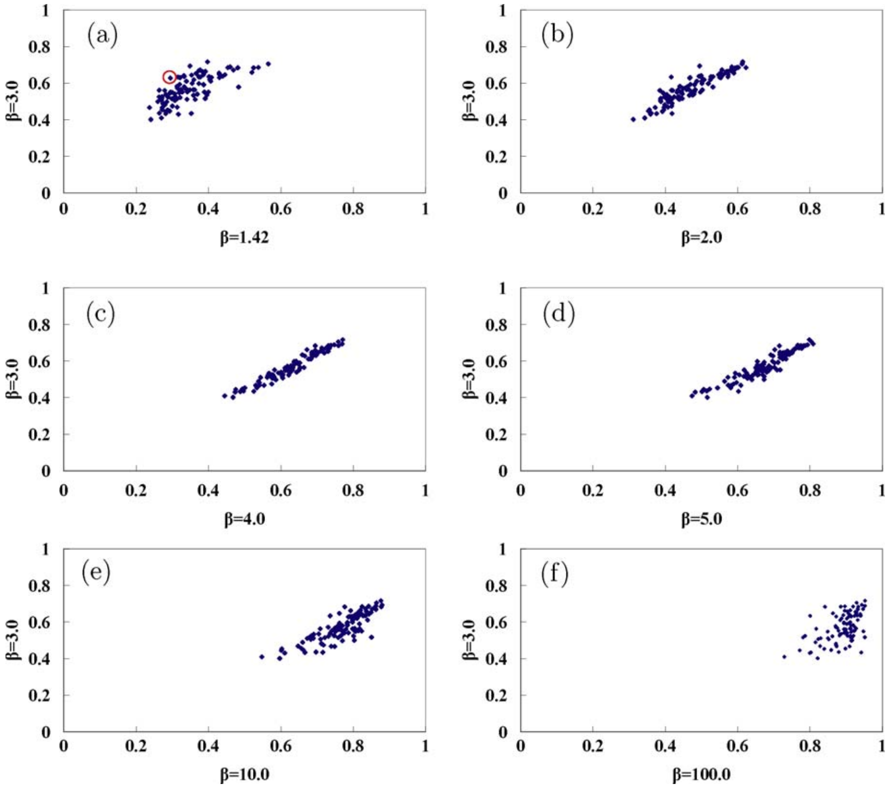


Fig. 13. Cross-plots of the ρ_β values. (a) 1.42 Å vs 3.0 Å, (b) 2.0 Å vs 3.0 Å, (c) 4.0 Å vs 3.0 Å, (d) 5.0 Å vs 3.0 Å, (e) 10.0 Å vs 3.0 Å, and (f) 100.0 Å vs 3.0 Å.

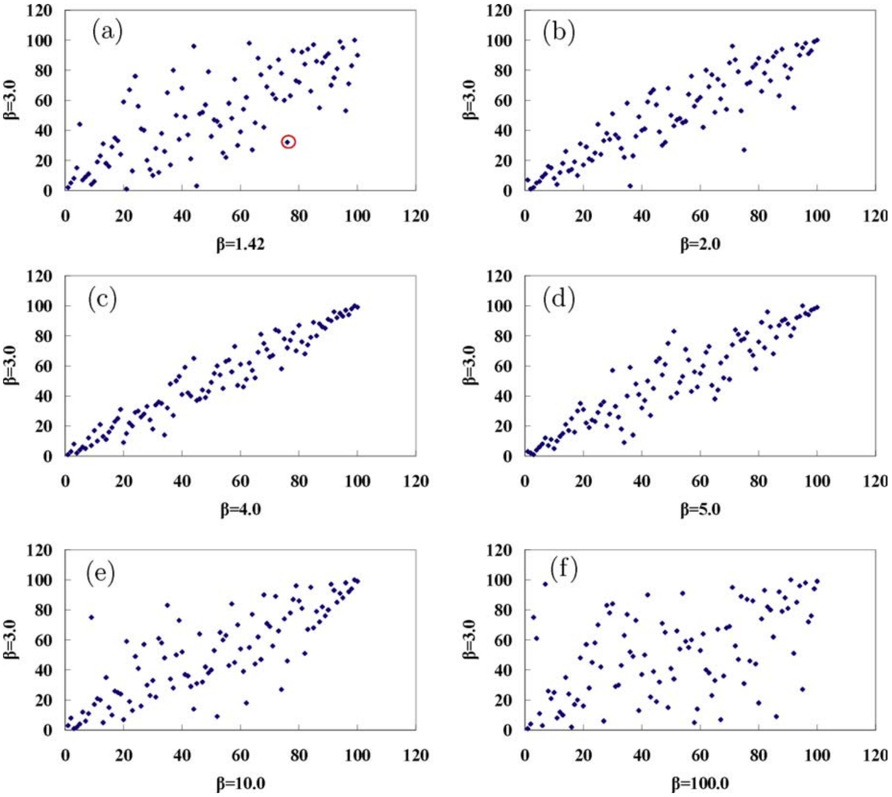


Fig. 14. Cross-plots of the protein ranks among 100 test proteins according to the ρ_β values. (a) 1.42 Å vs 3.0 Å, (b) 2.0 Å vs 3.0 Å, (c) 4.0 Å vs 3.0 Å, (d) 5.0 Å vs 3.0 Å, (e) 10.0 Å vs 3.0 Å and (f) 100.0 Å vs 3.0 Å.

Fig. 14 shows the cross-plots of the protein ranks according to the value of ρ_β among the 100 proteins. For example, the same protein 2cwl, marked in Fig. 14(a), has the rank 76 (out of the 100 proteins) in the order of the ρ_β values at $\beta = 1.42$ Å and the rank 32 at $\beta = 3.0$ Å. This figure also positively supports the previous analysis: Fig. 14(a), (b), (e), and (f) shows relatively small correlations and Fig. 14(c) and (d) shows relatively large correlations.

We interpret the least correlations in Fig. 13(a) and Fig. 14(a) as the probe of $\beta = 1.42$ Å reveals information that is too detailed about the shape and does not convey meaningful information regarding the sphericity. A local shape that a water molecule can nearly fit into on the surface of a protein may not mean much with respect to chemicals for drug candidates. Therefore, we claim that $\beta = 3.0$ Å is the optimal probe size for the sphericity measure of ρ_β for proteins.

Table 4 shows eleven selected proteins from the test set of one hundred proteins according to the ranks of ρ_β computed using $\beta = 3.0$ Å. Hence, the 3.0 Å sub-column in the *Rank* column shows $1, 10, 20, \ldots, 90, 100$. This is why we have selected the eleven representative proteins in this paper to show the detailed statistics in the experiment. Note the differences of the ranks of the eleven proteins for the different values of β. The ρ_β column shows the corresponding ρ_β values. Fig. 15 shows the eleven representative proteins from the three different orthogonal views. The order of the proteins is in the decreasing order of ranks for $\beta = 3.0$ Å. We claim that this figure validates and very strongly supports the proposition that the proposed approach using a sphericity descriptor works well.

We want to mention here that the red curve in Fig. 11 shows the coefficient of variation $\sigma(\rho_\beta)/\mu(\rho_\beta)$, which is also given in Table 3. It is known that the variation coefficient is a good measure for the variation of statistics in comparison to the magnitude of the average. However, the red curve strongly shows a monotonic decreasing pattern with its maximum at 1.42 Å. Counter to standard statistics theory, this again supports the proposition that the probe of $\beta = 1.42$ Å gives too much detailed information about the shape of proteins and is not very useful from an application point of view.

We have computed the minimum enclosing spheres for all the 100 proteins in the test set and the statistics are as follows: The average radius of all 100 minimum enclosing spheres is 38.16 Å, the radii of the minimum and the maximum enclosing spheres are 18.79 and 78.77 Å, and the standard deviation is 11.74 Å. Fig. 16(a) shows the protein 1lf1. Fig. 16(b) shows a set of spheres where each sphere is the minimum sphere enclosing each residue in 1lf1. Note that computation of such a minimum sphere enclosing a set of spheres is an NP-hard problem and good heuristics have been provided [64].

We have also computed the minimum spheres enclosing each residue and each R-group in the whole test proteins. In Fig. 17, the upper curve with red rectangles denotes the distribution of the average radius of the minimum sphere enclosing residues in the test set. For example, the value 3.343 for ALA (alanine) means that this is the average radius of the minimum enclosing spheres of all ALA residues in the 100 proteins. In Fig. 17, the lower curve with blue circles denotes the distribution of the average radius of the minimum spheres enclosing R-groups in the test set. For example, the value of ALA is 1.700. Note that, in Fig. 17, the radius of the minimum enclosing sphere of the R-group for GLY is zero. The difference between the two curves is interesting, yet was expected.

The grand average of the radii of the enclosing minimum spheres shows an interesting statistic: The grand average of the enclosing minimum spheres of all residues is 4.210 Å and its counterpart for R-groups is 2.902 Å. Note that the latter, 2.902 Å, is very close to the optimal value of measuring the sphericity, $\beta = 3.0$ Å and the standard deviation of ρ_β occurs at $\beta = 3.0$ Å.

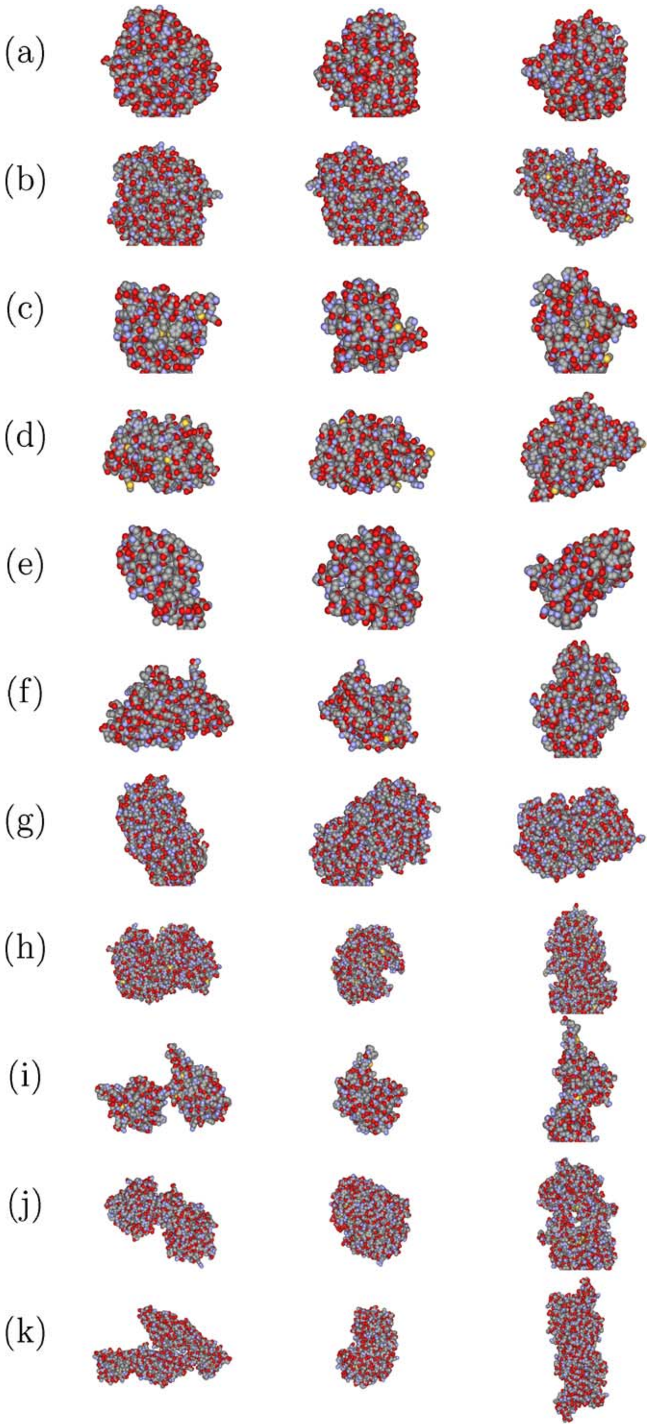


Fig. 15. Eleven representative proteins in the decreasing order of ranks for $\beta = 3.0$Å. (a) 1lf1, (b) 2i49, (c) 1t4q, (d) 2gpo, (e) 2op6, (f) 2yz1, (g) 1zrs, (h) 1iz9, (i) 2h2r, (j) 1eai, and (k) 2ol7.

Fig. 18 shows the distribution of the standard deviations of the radii of the minimum spheres enclosing each residue (the red curve) and the R-group (the blue curve) in the test set. For example, see ALA (alanine); The value of the red and blue points show the standard deviation of the radii for the enclosing sphere of ALA and the R-group of ALA, respectively. The figure shows that some residues (ARG, GLN, HIS, LYS, MET, PHE, TRP, and TYR) exist in proteins with various conformations and some residues (ALA, LEU, THR, and VAL) have relatively low variations in the conformation. The blue curve shows that the R-groups in some residues (ARG,

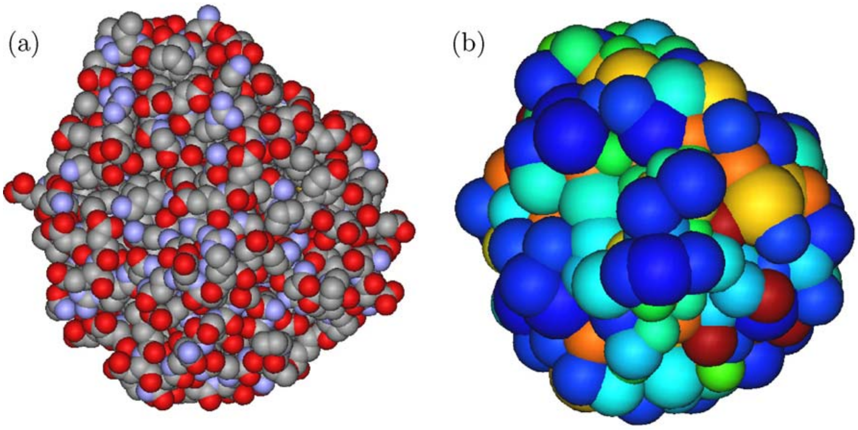


Fig. 16. A protein and its approximation model using spheres. (a) 1lf1, and (b) minimum spheres enclosing the residues.

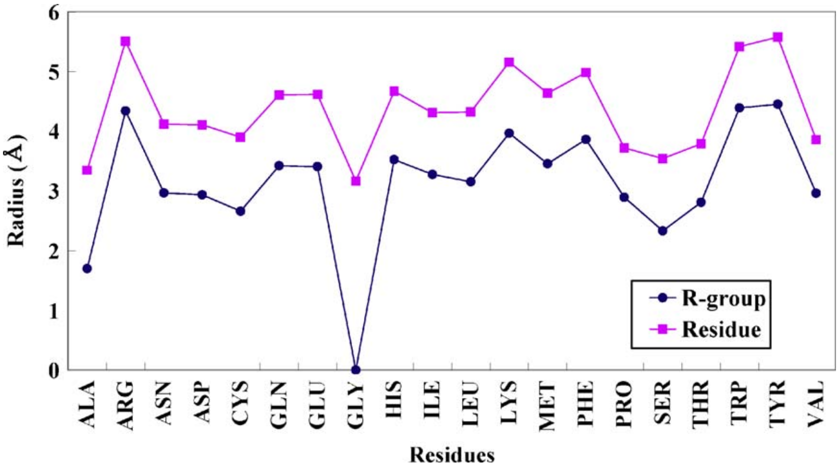


Fig. 17. The distribution of the average radii of the minimum enclosing spheres for both residues and R-groups. The red curve passing through the rectangular dots corresponds to the residues and the blue curve passing through the circular dots corresponds to the R-groups.

LYS, MET, PHE, and SER) have high conformational variation and the R-groups in some other residues (ALA, CYS, GLY, LEU, PRO, THR, TRP, and TYR) have low conformational variation. Note that the two curves show a significant discrepancy except for some residues such as ALA, ARB, ASN, LEU, LYS, MET, PHE, SET, THR, and VAL. In particular, it is interesting to note the following: as GLY has no R-group, its conformational variation is obviously null, as correctly shown in the figure; On the other hand, the R-group of ALA also has null conformational variation even though the R-group has four atoms.

The average standard deviation of the radius of the spheres for residues and R-groups are 0.212 and 0.088 Å, respectively. In the

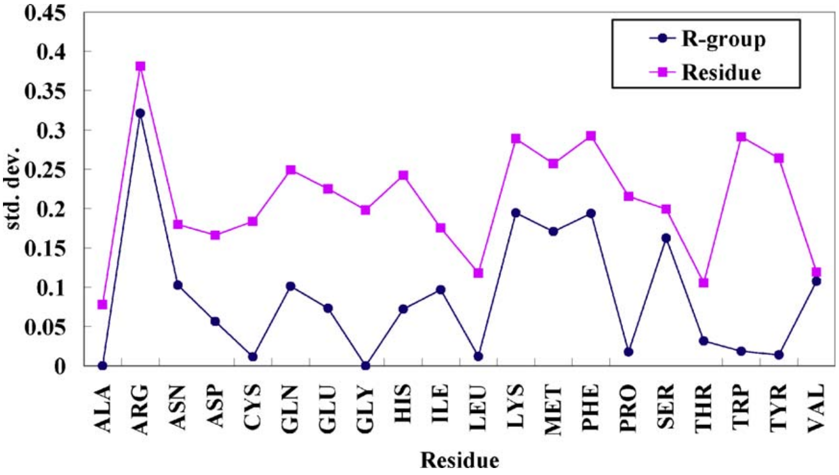


Fig. 18. Distribution of the standard deviations of the radius for the minimum enclosing spheres of both the residues and the R-groups for the 100 test proteins.

figure, TRP and TYR show the largest difference between the two curves and the reason is because TRP and TYR can have only few conformations since both TRP and TYR have an aroma-ring which constraints the motional degrees of freedom. The standard deviation has the highest value at ARG because ARG has a long side chain which causes many conformations.

5. Conclusions

The morphological features of protein are known to be important in understanding the functions of a protein. Among others, sphericity is one of the most important global shape descriptors. In this paper, we presented an approach to compute the protein sphericity using the recently developed theory of the β-complex and β-shape.

The definitions of the sphericity measures, ρ_{vdw} and ρ_{β}, were provided and experiments were performed. We find that ρ_{β} is a good measure of sphericity for proteins in both its quality and computational efficiency. We also note here that $\beta = 3.0$ Å most effectively discriminates the protein sphericity. Interestingly, 3.0 Å is approximately the average radius of the minimum enclosing spheres for all the R-groups in the test set.

Given a Voronoi diagram of a protein, the β-complex and β-shape can be computed very efficiently so that the sphericity measure can also be quickly computed. As the β-shape, corresponding to an appropriate value of β such as 3.0 Å, removes noisy tiny shape characteristics from the protein, the computed sphericity measure is more meaningful than measures based on the vdw-boundary.

The idea presented in this paper can be extended to the measures for other important shape descriptors such as cylindricity, planarity, etc. We propose that such global shape descriptors altogether can function as a fingerprint of each protein. We propose here that the β-complex and β-shape are powerful constructs for many other structural analyses of molecular structures.

The β-complex and β-shape can be conveniently used for other protein structure problems which are based on molecular shape. The problems include the computation of molecular surfaces, the void volume, the pockets, and the similarity between two proteins.

Acknowledgements

This research was supported by the Korea Science and Engineering Foundation through the National Research Lab Program funded by the Ministry of Science and Technology, Korea (No. R0A-2007-000-20048-0). The fifth author was supported by the Korea Research Foundation Grant (KRF-2007-314-D00311).

References

[1] P.G. Mezey, Shape in Chemistry: An Introduction to Molecular Shape and Topology, VCH Publishers, 1993.

[2] G. Zhang, M.G. Kazanielz, P.M. Blumberg, J.H. Hurley, Crystal structure of the Cys2 activator-binding domain of protein kinase Cδ in complex with phorbol ester, Cell 81 (1995) 917–924.

[3] P. Cramer, D.A. Bushnell, R.D. Kornberg, Structural basis of transcription: RNA polymerase II at 2.8 ångstrom resolution, Science 292 (8) (2007) 1863–1876.

[4] C.N. Pace, B.A. Shirley, M. Mcnutt, K. Gajiwala, Forces contributing to the conformational stability of proteins, The FASEB Journal 10 (1996) 75–83.

[5] G. Nicola, I.A. Vakser, A simple shape characteristic of protein–protein recognition, Structural Bioinformatics 23 (7) (2007) 789–792.

[6] D. Chandler, Interfaces and the driving force of hydrophobic assembly, Nature 437 (2005) 640–647.

[7] J.C. Wootton, Non-globular domains in protein sequences: automated segmentation using complexity measures, Computers and Chemistry 18 (3) (1994) 269–285.

[8] B.H. Chang, Y.C. Bae, Salting-out in the aqueous single-protein solution: the effect of shape factor, Biophysical Chemistry 104 (2003) 523–533.

[9] P. Røgen, H. Bohr, A new family of global protein shape descriptors, Mathematical Biosciences 182 (2003) 167–181.

[10] W. Tasylor, J. Thornton, W. Turnell, An ellipsoidal approximation of protein shape, Journal of Molecular Graphics 1 (2) (1983) 30–38.

[11] M. Kazhdan, T. Funkhouser, S. Rusinkiewicz, Rotation invariant spherical harmonic representation of 3D shape descriptors, in: Eurographics Symposium on Geometry Processing, 2003, 156–165.

[12] R.J. Morris, R.J. Najmanovich, A. Kahraman, J.M. Thornton, Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons, Bioinformatics 21 (10) (2005) 2347–2355.

[13] S.E. Leicester, J.L. Finney, R.P. Bywater, Description of molecular surface shape using fourier descriptors, Journal of Molecular Graphics 6 (1988) 104–108.

[14] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH—a hierarchic classification of protein domain structures, Structure 5 (1997) 1093–1108.

[15] D. Xu, H. Li, T. Gu, Shape representation and invariant description of protein tertiary structure in applications to shape retrieval and classification, Lecture Notes in Computer Science 4975 (2008) 556–562.

[16] D. Zhang, G. Lu, Review of shape representation and description techniques, Pattern Recognition 37 (2003) 1–19.

[17] RCSB Protein Data Bank Homepage, 2009.

[18] F.M. Richards, Areas, volumes, packing, and protein structure, Annual Review of Biophysics and Bioengineering 6 (1977) 151–176.

[19] M.L. Connolly, Analytical molecular surface calculation, Journal of Applied Crystallography 16 (1983) 548–558.

[20] M. Sanner, A.J. Olson, J.-C. Spehner, Reduced surface: an efficient way to compute molecular surfaces, Biopolymers 38 (1996) 305–320.

[21] J. Ryu, R. Park, D.-S. Kim, Molecular surfaces on proteins via beta shapes, Computer-Aided Design 39 (12) (2007) 1042–1057.

[22] A. Varshney, W.V.W. Brooks Jr., Computing smooth molecular surfaces, IEEE Computer Graphics and Applications 14 (1994) 19–25.

[23] A. Okabe, B. Boots, K. Sugihara, S.N. Chiu, Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd edition, John Wiley & Sons, Chichester, 1999.

[24] F. Aurenhammer, Voronoi diagrams—a survey of a fundamental geometric data structure, ACM Computing Surveys 23 (3) (1991) 345–405.

[25] http://home.mims.meiji.ac.jp/~sugihara/, 2009.

[26] CGAL User and Reference Manual: All Parts, Release 3.2.1, July 2006.

[27] K. Mehlhorn, S. Näher, LEDA: A Platform for Combinatorial and Geometric Computing, Cambridge University Press, 1999.

[28] J.D. Bernal, A geometrical approach to the structure of liquids, Nature 183 (4655) (1959) 141–147.

[29] J.D. Bernal, J.L. Finney, Random close-packed hard-sphere model II. Geometry of random packing of hard spheres, Discussions of the Faraday Society 43 (1967) 62–69.

[30] A. Poupon, Voronoi and Voronoi-related tessellations in studies of protein structure and interaction, Current Opinion in Structural Biology 14 (2004) 233–241.

[31] A. Bondi, van der Waals volumes and radii, Journal of Physical Chemistry 68 (1964) 441–451.

[32] M. Sanner, Modelling and applications of molecular surfaces, Ph.D. Thesis, Université de Haute-Alsace, France, 1992.

[33] D. Halperin, M.H. Overmars, Spheres, molecules, and hidden surface removal, in: Proceedings of the 10th ACM Symposium on Computational Geometry, 1994, pp. 113–122.

[34] D.-S. Kim, Y. Cho, D. Kim, Euclidean Voronoi diagram of 3D balls and its computation via tracing edges, Computer-Aided Design 37 (13) (2005) 1412–1424.

[35] D.-S. Kim, Y. Cho, D. Kim, S. Kim, J. Bhak, S.-H. Lee, Euclidean Voronoi diagrams of 3D spheres and applications to protein structure analysis, Japan Journal of Industrial and Applied Mathematics 22 (2) (2005) 251–265.

[36] J.R. Munkres, Elements of Algebraic Topology, Perseus Press, 1984.

[37] J.-D. Boissonnat, M. Yvinec, Algorithmic Geometry, Cambridge University Press, Cambridge, 1998.

[38] D.-S. Kim, D. Kim, K. Sugihara, Voronoi diagram of a circle set from Voronoi diagram of a point set. I. Topology, Computer Aided Geometric Design 18 (2001) 541–562.

[39] D.-S. Kim, D. Kim, K. Sugihara, Voronoi diagram of a circle set from Voronoi diagram of a point set. II. Geometry, Computer Aided Geometric Design 18 (2001) 563–585.

[40] M. Karavelas, M. Yvinec, Dynamic additively weighted Voronoi diagrams in 2D, in: R. Möhring, R. Raman (Eds.), Proceedings of the 10th European Symposium on Algorithms, vol. 2461 of Lecture Notes in Computer Science, Springer-Verlag, 2002, pp. 586–598.

[41] D.-S. Kim, D. Kim, Y. Cho, K. Sugihara, Quasi-triangulation and interworld data structure in three dimensions, Computer-Aided Design 38 (7) (2006) 808–819.

[42] D.-S. Kim, C.-H. Cho, D. Kim, Y. Cho, Recognition of docking sites on a protein using β-shape based on Voronoi diagram of atoms, Computer-Aided Design 38 (5) (2006) 431–443.

[43] J. Seo, Y. Cho, D. Kim, D.-S. Kim, An efficient algorithm for three-dimensional β-complex and β-shape via a quasi-triangulation, in: Proceedings of the ACM Symposium on Solid and Physical Modeling, 2007, pp. 323–328.

[44] H. Edelsbrunner, E.P. Mücke, Three-dimensional alpha shapes, ACM Transactions on Graphics 13 (1) (1994) 43–72.

[45] M.L. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (1983) 709–713.

[46] D.-S. Kim, J. Seo, D. Kim, J. Ryu, C.-H. Cho, Three-dimensional beta shapes, Computer-Aided Design 38 (11) (2006) 1179–1191.

[47] A. Shrake, J.A. Rupley, Environment and exposure to solvent of protein atoms. Lysozyme and insulin, Journal of Molecular Biology 79 (2) (1973) 351–371.

[48] A. Gavezzotti, The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state

organic reactivity, Journal of the American Chemical Society 105 (1983) 5220–5225.

[49] J. Higo, N. Go, Algorithm for rapid calculation of excluded volume of large molecules, Journal of Computational Chemistry 10 (3) (1989) 376–379.

[50] H.R. Karfunkel, V. Eyraud, An algorithm for the representation and computation of supermolecular surfaces and volumes, Journal of Computational Chemistry 10 (5) (1989) 628–634.

[51] E. Silla, F. Villar, Nilsson, J.L. Pascual-Ahuir, Tapia, molecular volumes and surfaces of biomacromolecules via GEPOL: a fast and efficient algorithm, Journal of Molecular Graphics 8 (1990) 168–172.

[52] E. Silla, I. Tunon, J.L. Pascual-Ahuir, GEPOL: an improved description of molecular surfaces. II. Computing the molecular area and volume, Journal of Computational Chemistry 12 (9) (1991) 1077–1088.

[53] R. Abagyan, M. Totrov, D. Kuznetsov, Icm: A new method for protein modeling and design: applications to docking and structure prediction from the distorted native c, Journal of Computational Chemistry 15 (3) (1994) 488–506.

[54] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, M. Scharf, The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies, Journal of Computational Chemistry 16 (3) (1995) 273–284.

[55] T.J. Richmond, Solvent accessible surface area and excluded volume in proteins. analytical equations for overlapping spheres and implications for the hydrophobic effect, Journal of Molecular Biology 178 (1) (1984) 63–89.

[56] M.L. Connolly, Computation of molecular volume, Journal of the American Chemical Society 107 (1985) 1118–1124.

[57] R. Lustig, Surface and volume of three, four, six and twelve hard fused spheres, Molecular Physics 55 (2) (1985) 305–317.

[58] K.D. Gibson, H.A. Scheraga, Volume of the intersection of three spheres of unequal size: a simplified formula, Journal of Physical Chemistry 91 (1987) 4121–4122.

[59] K.D. Gibson, H.A. Scheraga, Exact calculation of the volume and surface area of fused hard-sphere molecules with unequal atomic radii, Molecular Physics 62 (5) (1987) 1247–1265.

[60] K.W. Kratky, The area of intersection of n equal circular disks, Journal of Physics A: Mathematical and General 11 (6) (1978) 1017–1024.

[61] D.Q. Naiman, H.P. Wynn, Inclusion–exclusion bonferroni identities and inequalities for discrete tube-like problems via euler characteristics, The Annals of Statistics 20 (1) (1992) 43–76.

[62] H. Edelsbrunner, The union of balls and its dual shape, Discrete & Computational Geometry 13 (1995) 415–440.

[63] D. Attali, H. Edelsbrunner, Inclusion–exclusion formulas from independent complexes, in: Proceedings of the 21st Annual Symposium on Computational Geometry (SoCG'06), Pisa, Italy, (2006), pp. 247–254.

[64] E. Welzl, Smallest enclosing disks (balls and ellipsoids), in: Proceedings of the New Results and New Trends in Computer Science, vol. 555 of Lecture Notes in Computer Science, Springer, (1991), pp. 359–370.