Chapter 5

Protein Surface Representation and Comparison: New Approaches in Structural Proteomics

Lee Sael and Daisuke Kihara

Purdue University

5.1 Introduction 89
5.2 Evaluation Criteria 90
5.3 Surface Representation 92
5.3.1 General object representation 92
5.3.2 Protein surface definition 92
5.4 General Object Analysis Methods 93
5.4.1 Global shape analysis 93
5.4.1.1 Feature/feature distribution-based methods 93
5.4.1.2 3D coordinate centered methods 94
5.4.1.3 View-based methods 95
5.4.2 Local shape analysis 96
5.5 Protein Surface Analysis Methods 96
5.5.1 Graph-based methods 96
5.5.2 Geometric hashing 97
5.5.3 Methods using series expansion of 3D function 98
5.5.4 Several other methods 99
5.6 3D Zernike Descriptors 100
5.6.1 Characteristics of 3DZD 100
5.6.2 Steps of computing 3DZD for protein surfaces 101
5.6.3 Test results of 3DZD 102
5.7 Discussion 104
Acknowledgments 105
References 106

5.1 Introduction

The 3D shape and physicochemical properties on protein surface carry essential information for understanding function of a protein. For example, catalytic reaction of an enzyme is realized by a set of atoms on the active site. Also residues at an interface surface region establish physical contacts in protein–protein interaction. Those local surface regions which are responsible for function tend to be better conserved than other surface

regions in terms of shape and physicochemical properties, which are not detectable by conventional sequence or main-chain conformation similarity searches. Thus development of effective methods for describing and comparing protein surfaces will provide new insights into how function is realized in proteins.

Despite the importance and promise of surface-based protein characterization methods, they have not been well studied until recently partly due to its higher technical complexity. However, development of protein surface analysis methods have been highlighted recently because of its urgent need; an increasing number of structures of unknown function have been solved and accumulated by structural genomics projects in the past few years. The current protein structure database, protein data bank (PDB), contains more than 2300 structures which are categorized as "unknown function," whose function were not confidently predicted by conventional approaches. Concurrently in the computer science field, 3D object representations and searching algorithms have become a research focus in many domains, such as computer-aided design, game development, computer vision, and computational geometry. Some of developed algorithms in those domains can be readily applied to protein surface analyses. Moreover, bioinformatics resources, including databases of protein sequences and structures and classification of protein function, have been well developed. These resources enable computational analyses of the relationship of protein function and structure, facilitating development of new bioinformatics tools. Therefore now the time is ripe for extensive development of protein surface analysis methods which can provide functional annotation to proteins through surface comparison.

In the following sections, we overview evaluation criteria for methods for 3D shape analysis. Then we discuss how protein surface is defined. Next, we review 3D object analysis methods developed in the computational geometry and graphics field. What follows is a review of recent protein surface representations and comparison methods. In the last section, we introduce our recent works on surface-based fast protein structure and surface property comparison methods.

5.2 Evaluation Criteria

There are seven criteria for evaluating characteristics of general object shape and protein surface analysis methods (Tangelder and Veltkamp, 2004). In the following sections, we will refer these criteria when describing existing methods.

1. Invariance to Euclidean transformation: three Euclidean transformations, i.e., rotation, scaling, and translation, do not change the shape of

the original object. Some methods use representations that are invariant to Euclidean transformation. Although this may seem simple, it needs nontrivial mathematical derivation or often achieved by oversimplification. Euclidean transformation invariant representations are convenient in comparing objects, because their representations can be directly compared without preprocessing. When the other representations are employed, either the most similar positions of objects needed to be searched, which is time consuming, or a normalization process of object positions is needed prior to analysis.

2. Need for pose normalizing: pose normalization is to rotate, translate, and scale an 3D object to a standard position for comparison. Normalizing in terms of translation and scaling can easily be done, for example, by moving the object in a way that its center of gravity locates at the origin and then scale it into a unit sphere. However, normalization against rotation is often not robust depending on the shape of the object. Principal component analysis (PCA) is the most widely practiced method for pose normalization. PCA often fails to provide a unique robust solution when an object has symmetrical mass distribution; i.e., PCA generates many equal eigenvalues, which suggests more than one positioning of the object are possible. This is especially problematic to handling protein shapes, because they are more or less spherical.

3. Ability for partial matching: partial matching aims to find similar local regions of two objects. It is especially important in protein matching considering that a function of protein is attributed to local surface regions such as active sites and protein docking interfaces.

4. Capability of changing resolution: depending on the content of object analysis, being able to adjust the level of details of object description becomes useful. Higher resolution provides detailed information while lower resolution is more focused in the overall shape and computationally efficient.

5. Tolerance to small noises or changes: this property is also important for protein shape analysis along with the capability of changing resolution. Proteins are flexible in nature and structures are solved experimentally in different resolutions depending on the method used and experimental condition. Moreover, if a predicted structure is handled, some errors are unavoidable. Thus it is important to account for these changes by allowing resolution change or by making the method tolerant to small differences.

6. Ability to incorporate additional properties: nonshape properties such as electrostatic potential, hydrophobicity, and residue conservation are important factors in determining and analyzing function of proteins.

Thus being able to use additional properties can extend the applicability of protein surface analysis.

5.3 Surface Representation

5.3.1 General object representation

There are two types of object representations, volume- and boundary-based. Well known volume-based representations are voxels and octrees. In voxels, the volume of an object is represented by filled grid points while in octrees the object space is hierarchically subdivided. Widely used boundary-based representations include polygon mesh and point cloud. A polygon mesh is composed of nodes and edges that form triangles that are connected to completely cover the surface of an object. In point cloud, a set of (x, y, z) points on the surface are used to represent a surface.

Generally, volume-based representations are used to for experimental data, such as computed tomography scans. On the other hand, boundary-based representations are used for computer designed objects, such as ones used in computer games. Volume-based representation requires a larger space but can provide information about the interior of an object while boundary-based representation is efficient in drawing an object on the computer screen.

5.3.2 Protein surface definition

The underlying physical substance of a protein surface is the van der Waals radius of atoms of the protein. Thus an intuitive way of defining protein surface is to compute the union of boundaries of spheres of van der Waals radius of each protein atom (the van der Waals surface). Often inflated (i.e., enlarged) van der Waals radius is used for defining the surface. However, direct use of the van der Waals sphere of atoms usually leaves unoccupied spaces between atoms, making small clefts and cavities on the surface. Those small cavities, where water molecules and ions cannot enter, are negligible or often cause unnecessary noises for many applications of protein surface representation. A common way to obtain a smoother surface is to roll a probe sphere (usually of the size of a water molecule) over the van der Waals surface and to trace the center of the sphere (solvent accessible surface) or to trace the inward-facing surface of a probe sphere (solvent excluded surface or Connolly surface (Connolly, 1983)).

The other protein surface definitions include α-surface (Wang, 2001). The algorithm of α-surface connects points to construct triangle meshes, whose resolution is controlled by a parameter, α. The solvent accessible surface and the Connolly surface are also usually represented by triangle meshes. The other representations, such as point cloud and voxels are also used.

5.4 General Object Analysis Methods

This section provides a list of well known shape analysis methods in computer science and other engineering fields that are already applied or have the potential of being applied to proteins. Roughly, methods can be classified as global and local shape analysis methods.

5.4.1 Global shape analysis

Global shape descriptors represent the overall shape of objects. They can be classified into three categories, feature and feature distribution-based methods, 3D coordinate centered methods, and view-based methods.

5.4.1.1 Feature/feature distribution-based methods

Methods in this category describe an object by one or a set of features of the object, such as the volume and the area of surface. These methods are one of the earlier methods developed, and are still actively applied individually or often integrated into recent object analysis methods because of their simplicity. An advantage of these methods is that nonshape properties of objects can be easily combined with shape-based features. On the other hand, the disadvantage is that they are obviously less descriptive since an object shape is represented by a few number of features. Those features are represented as a feature vector.

Elad et al. (2000) use statistical moments, which describe the distribution of the position of vertices on a polygon mesh of an object, i.e., the center of gravity, variance, and skewness etc., as features of the object. Then the extracted moments are used as a feature vector and compared by a weighted Euclidean distance. In their method, invariance to Euclidean transformation is obtained by normalizing against the first two moments (center of gravity and variance) of the surface points. Zhang and Chen (2001) also utilize the statistical moment in addition to some other features. They propose an efficient method for computing and comparing the global features including the volume, the surface area, the volume-surface ratio, the statistical moments, and coefficients of the Fourier transform of 3D objects.

Rather than representing an object by a single value, feature distribution-based methods use a histogram of global shape features as a descriptor. An example is the shape distribution, which uses a histogram of the Euclidean distance of randomly chosen two points on the surface of an object (Osada et al., 2002). The solid angle histogram places a sphere at each representing points of an object and computes the fraction of volume of the sphere occupied by the object (Connolly, 1986).

5.4.1.2 3D coordinate centered methods

These methods directly represent 3D shapes of objects in space. There are two main approaches in this category. The first approach is to use mathematical transformation of a 3D function, which is the position of points or surface of a 3D object in the case of the shape analysis. Another approach is to compute how an object occupies the 3D space by its volume when the object is represented in voxels.

Mathematical transformations have been widely studied in 2D image processing. And the 3D version of those transformations, such as Fourier, Hough, Radon, and wavelet transformations, have been applied for 3D objects. These methods are variant to Euclidean transformations and need pose normalization prior to extraction of descriptors. More recently, spherical harmonics have been widely explored. Spherical harmonics are functions of a set of a polar angle, θ, and a colatitude angle, $\varphi : Y_l^m(\theta, \varphi)$. Since spherical harmonics form an orthonormal set of functions, a 3D function (thus, a 3D object) can be expanded as a series of spherical harmonics with a different degree l and an order m on the unit sphere.

Limitations in direct applications of spherical harmonics include its variance to Euclidean transformations and also that they can correctly capture only star like shapes, i.e., shapes that have no reentrant surfaces.

Funkhouser's group introduced a spherical harmonics-based shape descriptor, which is rotation invariant and can also be applied to nonstar like shapes (Kazhdan et al., 2003). The method first segments an object into concentric spheres and then computes spherical harmonics for each of the spheres. Since rotating a spherical function does not change its L^2 norm, combining the L^2 norm computed for each group of harmonics of the same parameters (l and m) yields a rotation invariant (thus invariant to Euclidean transformations) descriptor. Nonstar like shapes are better handled by the segmentation to concentric spheres. Application of spherical harmonics in partial matching has also been made by the same group as an extension of the spherical harmonics-based method (Funkhoser and Shilane, 2006).

The above method considers the radial information (the distance from the center of an object) by the segmentation of an object into concentric spheres. In contrast, 3D Zernike descriptors uses Zernike–Canterakis basis $Z_{nl}^m(\mathbf{x})$, which incorporates radial information into the polynomials in Cartesian coordinates $\mathbf{x} = (x, y, z)$ (Canterakis, 1999):

$$Z_{nl}^m(\mathbf{x}) = R_{nl}(r)Y_l^m(\theta, \varphi)$$

Thus, 3D Zernike descriptors are convenient to handle a 3D object described in points or voxels in Cartesian coordinates. Rotation invariance was obtained later by (Novotni and Klein, 2004) in a similar manner to what was done by (Kazhdan et al., 2003). We have applied 3D Zernike descriptors for protein surface comparison, which will be discussed in Section 5.6.

The other special functions introduced in 3D object analysis include spherical wavelets and Krawtchouk polynomials. Mathematically, spherical wavelet

descriptor (Laga et al., 2006) has two advantages over spherical harmonics. First, the level of details of description can be locally controlled. Second, sampling of points are more uniform. Weighted 3D Krawtchouk descriptor (Mademlis et al., 2006) uses polynomials of discrete variables, and thus eliminates the need for spatial discretization process. Hence no numerical approximation is involved in handling a voxelized object data. A drawback of weighted 3D Krawtchouk descriptor is again the need for pose normalization.

When objects are represented by voxels, two objects can be compared by computing the difference of distribution of the occupied voxels (volumetric difference methods). Occupied voxels of an object can be represented by a tree data structure, e.g., octree. Therefore comparison is done efficiently based on the tree representation. Volumetric difference methods are still generally slower than other global methods and still need pose normalization.

An interesting idea of representing object is to compute "energies" or cost needed to morph an object to a sphere. Then comparison is done by computing the difference in the morphing energies. A method by Leifman et al. (2003) calculates sphere projection energy as $E = \int_{dist} \vec{F}.d\vec{r}$, where *dist* is the distance between the sphere and the object surface, and $\vec{F}$ is the applied force which is assumed constant. In a method proposed by Yu et al. (2003), a feature map is used to record a local energy needed to morph an object. The object is first normalized and fitted into a unit sphere. Then local energy at each point is computed, which consists of two parts: the distance from the object surface to the bounding sphere and the number of object surfaces penetrated when a ray is shot from the sphere center. This method additionally uses Fourier transform of the feature map, which is better in tolerance to noise which may have been introduced in the pose normalization process.

5.4.1.3 View-based methods

View-based methods describe a 3D object as a set of projected 2D images of the 3D object from different viewing angles. Each image contains characteristics of the object from that angle, however, relative spatial information between images from different view points is not captured.

The most well known view-based method is light field descriptors (Chen et al., 2003). In computing the light field descriptor of a 3D object, the object is first scaled and placed into a bounding sphere. Then a light field of the object is created, which consists of 20 uniformly distributed silhouettes of the object from 10 rotational positions on the bounding sphere. Subsequently, a combination of 2D Zernike moments and Fourier transforms are used as the 2D descriptor for each silhouette. To compare descriptors of two objects, basically silhouettes of the two objects are compared exhaustively to find matches.

Ohbuchi et al. (2003) proposed another view-based technique, which captures depth of an object from each angle in addition to the 2D silhouettes. Then for each image a Fourier transform-based descriptor is generated.

5.4.2 Local shape analysis

The basis of local shape analysis is to capture geometrical feature of a local region around a given point on a surface. The curve of a local surface is described using Gaussian, mean curvature, and the shape index. Among them, the shape index has been also used for protein surface analysis. The shape index (Koenderink and van Doorn, 1992) is a single-value that ranges from -1 to 1 which measures the slope of local surface using principal curvatures. The spin image is another popular method to describe a local shape (Johnson and Hebert, 1999; de Alarc et al., 2002). A spin image is a 2D histogram of distances from a central vertex to neighboring vertices. Two distances characterize spatial relationship between the two vertices; the radial distance, α, which is defined as the perpendicular distance from the central vertex to another through the surface normal, and the axial distance, β, a signed perpendicular distance to the tangent plane of the central vertex. By definition, a spin image does not change upon rotating around the norm of a central vertex. The spin image is calculated for each vertex on a surface mesh of an object.

Using surface curvature information captured at each vertex as mentioned above, a larger surface region can be described by connecting vertices as a graph. A graph captures relative spatial information of vertices and enables partial matching of two local surfaces. However, generally speaking partial graph matching has a high complexity thus often slow for comparing large graphs. Methods for global shape analysis, such as spherical harmonics-based methods, can also be used for describing a local shape around a vertex.

5.5 Protein Surface Analysis Methods

In this section, we discuss existing methods for protein surface representation and comparison. Identifying similar global and local surface shapes of proteins has application to structure-based function prediction. Protein surface representation has been also studied in the context of protein-protein docking and protein-small ligand docking, in which case complementarity of two shapes is taken into account. We first discuss three major categories of protein surface analysis methods, namely, graph-based, geometric hashing, and methods using series expansion of 3D function.

5.5.1 Graph-based methods

Graph theoretical approaches are frequently applied for protein surface comparison since some common protein surface representations, e.g., triangular mesh, can be naturally considered as a graph. In a graph representation of a protein surface, geometrical and often physicochemical features of a local

region are assigned to each vertex and edges connecting vertices describe positional relationship of the vertices. A nice thing about graph representation is that partial matching of two protein surfaces can be done using existing algorithms in the graph theory.

The method proposed by Pickering et al. (2001) first generates a Connolly surface of the region of interest. Then for each vertex point, shape information, such as shape index and radius of curvatures, is calculated as well as biological features such as types of residues. The matching process involves finding the maximal common subgraph of two graphs representing the protein surfaces.

Kinoshita and his colleagues developed a database of protein surfaces of functional sites, named eF-site, and a method to search for the similar local surface sites in a query protein against the database (Kinoshita et al., 2002). Triangular meshes of the Connolly surface constitute a graph of a protein. Each vertex is assigned with the electrostatic potential and curvatures. To find similar local regions of two proteins, a clique detection algorithm on an association graph is used. An association graph of two graphs is formed first by creating a node for a pair of vertices, one from each protein, that have similar features. Then an edge connecting a pair of nodes is drawn when the spatial distance of the pair of original vertices belonging to the two nodes is similar. Next, the largest clique in the association graph, i.e., the largest fully connected subgraph, is selected. The selected clique is considered as the most similar part between the two protein surfaces.

SURFCOMP also uses a clique detection algorithm on an association graph (Hofbauer et al., 2004). In SURFCOMP, surface critical points are considered as vertices, which are either one the three classes, a convex, a concave, or a saddle point. Rather than using all the vertices in a Connolly surface, using critical points reduces vertices to be considered, making the method more efficient. The graphs are further simplified by several filters that compare surrounding shape, local arrangement of the critical points, and physicochemical properties.

Baldacci et al. (2006) further reduce the number of graph nodes by considering surface patches. A patch in a protein surface is a local circular region where residues included have homogeneous geometrical and physicochemical properties. The properties considered are geometrical curvature, the electrostatic potential, and hydrophobicity. Each patch contains at least ten amino acids and typically a protein surface is represented by less than ten patches. Patches are connected by edges, representing a protein by a spatial graph. They used the spatial graphs for classifying proteins by similarity of patterns of patches.

5.5.2 Geometric hashing

The Wolfson and Nussinov group applies the geometric hashing technique, which was originally developed for computer vision applications (Rosen et al., 1998). The method first extracts sparse critical points defined at the centers

of mesh faces abstracted as convex, concave, or saddle of the protein surface (Lin et al., 1994). The geometric hashing is composed of two stages, a hashing stage and a recognition (matching) stage. In the hashing stage, transformation-invariant information of protein surface shapes to be compared against (called models) is extracted and stored in a hash table. Concretely, a protein surface shape represented by critical points is placed relative to every possible admissible reference frame and their position and features are stored in the hash table. This stage can be executed off-line and the table can be reused once created. In the recognition stage, a target protein surface is placed relative to every possible reference frame and the hash table is accessed to find matching model critical points. Then a vote is registered for a pair of model and target reference frames if their critical points match. The geometric hashing allows a partial surface matching. Also a target protein can be compared with multiple proteins at the same time once they are hashed in a table. Later they also applied geometric hashing for protein-small ligand molecule docking, and protein–protein docking (Fischer et al., 1995; Halperin et al., 2002).

5.5.3 Methods using series expansion of 3D function

Using mathematical transformation has become popular in protein surface analyses as well as 3D object analysis. Here protein surface is treated as a 3D function, which is expanded in a series function. The major advantage of these methods is the compactness in description, which allows rapid real-time comparison against a large number of proteins. A series expansion is also suitable in changing resolutions of surface description. Also properties on a surface can be naturally incorporated in surface description.

An early work in this category include use of Fourier series expansion as a shape descriptor (Gerstein, 1992). Protein surfaces are superimposed and Fourier coefficients are extracted and compared at various resolutions. The author used the method to compare shape of antigen-combining sites of antibody molecules.

Thornton and her colleagues used spherical harmonics to describe the volume of ligand binding pockets of proteins (Kahraman et al., 2007). A ligand binding pocket in a protein surface is detected using the SURFNET program, which identifies a pocket by inserting spheres of a certain size. Thus a pocket is represented as overlapping spheres, which constitute the volume of the pocket. Then the spherical harmonics expansion is applied to the volumetric representation of the pocket and the coefficients are taken as the descriptor. Interesting application of their approach is direct comparison of shape of pockets and ligand molecules, which is possible because both pockets and ligands are represented as a closed volume. For comparison, shapes should be pose normalized.

Spherical harmonics has been also applied for protein–protein docking prediction (Ritchie and Kemp, 2000). By using spherical harmonics, a complete search for docking conformation over all six degrees of freedom can be

performed conveniently by rotating and translating the initial expansion coefficients.

Recently, we employed 3D Zernike function, which is an extension of spherical harmonic expansion to compare protein global surfaces (Sael et al., 2008a,b). The most favorable features of the 3D Zernike descriptor aside from its advantages originating from spherical harmonics, are its rotation invariance and applicability to nonstar-like shapes. The two advantages are worth further attention and will be described extensively in Section 5.6.

5.5.4 Several other methods

The volumetric difference method, which was originally developed for 3D object representation, has been applied for protein surface comparison. Masek et al. (1993) defines molecular "skins," which is a thin layer of voxels composing protein surface. The method compares the shape by computing the similarity of the maximum overlap between a pair of protein skins. Another volumetric difference method utilizes a genetic algorithm to find the optimal superimposition of protein surfaces or fragments of proteins (Poirrette et al., 1997). The spin image representation is also applied to identify structurally equivalent surface regions in two proteins (Bock et al., 2007).

Shentu et al. (2008) proposed a local surface structure characterization method named context shape, which considers visible directions from critical points on a protein surface. A context shape of a critical point essentially describes the visible directions from the point to a surrounding sphere of a given radius, which is not blocked by voxels occupied by the protein volume. The context shape is represented as a binary string with 1 for blocked and 0 for visible directions. They used this method to evaluate shape complementarity of two protein surfaces in protein–protein docking prediction.

Pawlowski and Godzik (2001) proposed a method which is aimed to compare physicochemical features of protein surfaces, such as electrostatic potential and hydrophobicity. Those features are mapped on a surrounding sphere of a protein and comparison is done between spheres after they are superimposed. As obvious from its design, this method does not compare shapes of proteins but only physicochemical properties, thus, it can only analyze proteins of the same structure (e.g., protein of the same family). Nevertheless this method is interesting as it can quantify the difference of properties on the surface.

There are several methods which combines surface shape information with residue or sequence information. As sequence motifs (e.g., PROSITE database) or spatial arrangements of catalytic residues (Arakaki and Skolnick, 2004) are traditionally used in function prediction in protein bioinformatics area, these methods can take advantage of accumulated knowledge of sequence-function relationship of proteins.

The SURFACE database stores a library of functionally important residues found at pocket regions of proteins (Ferrè et al., 2005). In selecting functional

sites of proteins, pockets in protein surfaces are identified by SURFNET and residues which reside in the pockets are referred to functional motif databases including PROSITE. Two local sites are compared in terms of the root means square deviation of positions of superimposed amino acids.

A method by Binkowski et al. (2003) utilizes local sequence information of binding pockets and surface shape to predict function of proteins. The local sequence of a pocket region is extracted by concatenating short sequences which compose the pocket. To compare the extracted local sequence, local sequence alignment by dynamic programming algorithm is performed.

5.6 3D Zernike Descriptors

A 3D Zernike descriptor (3DZD) is categorized as a 3D coordinate centered method using special kernel functions. Canterakis first introduced 3D Zernike moments that combine a radial function with spherical harmonics to describe objects in 3D Cartesian coordinate system (Canterakis, 1999). Novotni and Klein later applied 3D Zernike moments to construct rotation invariant descriptors of 3D objects (Novotni and Klein, 2003). 3DZD was applied to describe overall shape of small ligands by Mak et al. (2008). The first thorough applications to describe protein shape and physicochemical property on protein surface have been conducted by our group. This section summarizes our works described in two recent papers (Sael et al., 2008a,b).

5.6.1 Characteristics of 3DZD

3DZD has several significant advantages regarding the comparison of protein surface. First, it represents a protein compactly allowing fast retrieval capable for real-time database search. Second, 3DZDs are rotation invariant, that is, protein structures need not be aligned for comparison. Related works, such as spherical harmonics for binding pocket and ligand comparisons by Thornton's group (Kahraman et al., 2007), need pose normalization because the methods are not rotation invariant. Pose normalization could be problematic especially in comparison of protein shapes, which are almost globular and the principle axes are not robustly determined. Third, the resolution of the description of protein structures can be easily and naturally adjusted by changing the order of 3DZDs. The rough global difference of protein structures reflects the difference of the first couple of invariants that correspond to lower orders of the 3DZD (Sael et al., 2008b). Figure 5.1 illustrates different resolutions of a reconstructed protein surface by changing the order of 3DZD. Here the order is changed from 5 to 10, 15, 20, and 25. When a lower order is used, pear-like global surface shape of this protein is highlighted, while more description of local geometry shows up as the order becomes higher. We used

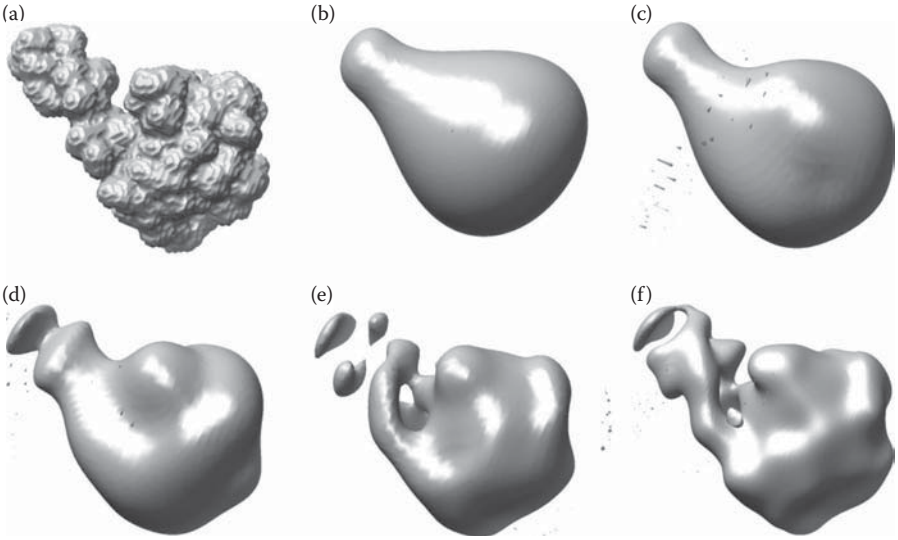

FIGURE 5.1: Resolution of 3D Zernike descriptor. (a) surface abstraction of 1ew0A, which is used as the input. (b) through (f) are reconstructed figures of 3D Zernike moments using different order from 5 up to 25 increasing with an interval of 5.

the order of 20 for our work since it yielded satisfactory results in a 3D shape retrieval benchmark by Novotni and Klein. Moreover, physicochemical properties of a protein surface, such as electrostatic potentials and hydrophobicity, can be incorporated into the description considering an appropriate 3D function (Sael et al., 2008a).

5.6.2 Steps of computing 3DZD for protein surfaces

The first steps to compute 3DZD for a protein are calculating protein surface and placing it on a cubic grid (voxelization). To represent a surface shape, each voxel is assigned 1 if it is on the surface and 0 otherwise. Real number values of other physicochemical properties can also be assigned only to the surface voxels. The resulting voxels with values are considered as the 3D function, which will be expanded in a 3DZD. Using the order of 20, a 3DZD results in 121 invariants (numbers). To convert physicochemical property which ranges from negative to positive values to 3DZD, a 3DZD for a set of voxels with a positive value assigned and those with negative value are separately computed. Then two 3DZDs are combined yielding a descriptor of 242 invariants. This is because a 3DZD recognizes the contrast of patterns of the positive and the negative value but not the value itself.

3DZDs of two proteins are compared in terms of the Euclidian distance or a correlation coefficient based distance, which is defined as 1—*correlation*

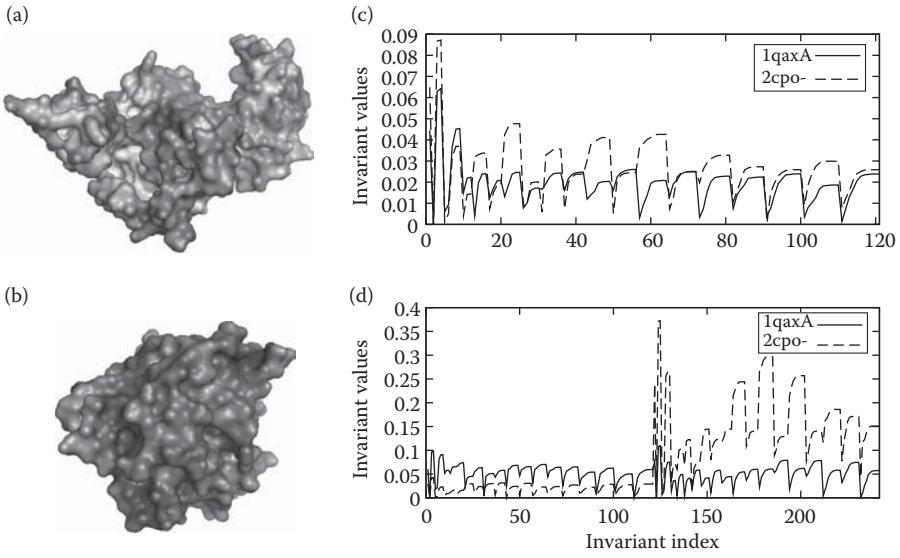

FIGURE 5.2: 3DZD of protein surface shape and electrostatic potential. Surface electrostatic potential of two proteins proteins; (a), 1qaxA; and (b), 2cpo. (c), 3DZD of surface shape; and (d), surface electrostatic potential of the two proteins is shown. The order used for both shape and electrostatic potential is 20. The number of invariants computed for shape is 121. For electrostatic potential, positive and negative regions are calculated separately forming 242 number of invariants. In (a) and (b) the gray scale ranges from -5 (black) to $+5$ (white) to represent the electrostatic potential.

coefficient. Figure 5.2 shows an example of 3DZDs of two proteins, 1qaxA and 2cpo. Figure 5.2c shows 3DZDs of the surface shape and Figure 5.2d shows 3DZDs of surface electrostatics of the two proteins. In Figure 5.2d, higher peaks at the 122th to the 242th invariants by 2cpo relative to 1qaxA indicate that 2cpo has a larger region with a negative electrostatics value.

5.6.3 Test results of 3DZD

We evaluated 3DZDs in two ways: (1) the ability to retrieve similar protein structures (Sael et al., 2008b) and (2) the ability to compare proteins in terms of their surface physicochemical properties (Sael et al., 2008a).

For the protein structure retrieval test, we prepared a dataset of 2432 protein structures, which are preclassified by another protein structure comparison method, the combinatorial extension (CE) method (Shindyalov and Bourne, 1998). CE compares two protein structures by their main-chain conformation as many other conventional protein structure comparison methods do. Despite the difference in structure representation, 3DZD retrieved proteins

of the same conformation defined by CE in 89.6% of the cases within the top five closest structures. This level of agreement with CE is the same as between CE and another commonly used protein structure comparison method, DALI. In addition to this retrieval accuracy, the strength of 3DZD is its extremely fast computational time. Computing a 3DZD for a protein takes about 37 seconds, but once it is computed, a database search against entire PDB with over 54,000 structures takes less than a minute. In contrast, typically a pairwise structure comparison by CE takes a couple of seconds. Therefore, a search against the entire PDB would take more than a day. Thus 3DZD can dramatically make a global protein structure search efficient. Even if one still wants to find proteins in a database which have the similar main-chain conformation to a query protein, 3DZD can be used as a rapid pre-screening prior to using CE.

Moreover, we found some cases where protein surface shape is indicative to functional classes of proteins. In our paper we showed such examples of pairs of DNA binding proteins and transmembrane transporters. These protein pairs have distinct surface shape similarity but does not share detectable similarity in sequence or main-chain conformation, thus their functional relevance can not be easily identified by conventional bioinformatics methods. In the case of DNA binding proteins, they have a saddle like local surface shape which is used to mount on DNA strands.

Next, we compared protein surface physicochemical properties of several protein families using 3DZDs (Sael et al., 2008a). We used globin proteins and three protein families with both thermophilic and mesophilic homologs as datasets. The globin family is known to have a conserved fold with a wide variety of function. A varied range of affinity to oxygen, different functions, and different environments where the globin proteins locate coincide with the relatively large distance of surface electrostatic potentials measured by 3DZDs. Thermophilic proteins have gained substantially higher thermal stabilities as compared to their mesophilic orthologs. And surface electrostatics has been identified as one of the major stabilization factors of thermophilic proteins. For the three protein families studied, we showed that 3DZDs successfully distinguish the thermophilic proteins from mesophilic proteins based on similarity of surface electrostatics. The sequence similarity and the main-chain conformation similarity cannot differentiate these two classes, because all of members of the families have more or less similar in sequence and structure. Since 3DZDs can be quantitatively compared, a tree can be drawn for a set of proteins based on similarity of their surface physicochemical property. This will be quite useful for studying protein function and evolution.

Further, we showed electrostatic potential of local regions of proteins can also be compared. Figure 5.3 illustrates a procedure for local surface analysis using 3DZDs. In this example, ligand binding sites of two TIM barrel proteins, 1fdjB and 1goc, are compared. TIM barrel is one of the most prevalent folds adopted by a variety of enzymes. Ligand binding sites of TIM barrel enzymes are usually located at the cleft with cluster of loops of the barrels. Reflecting

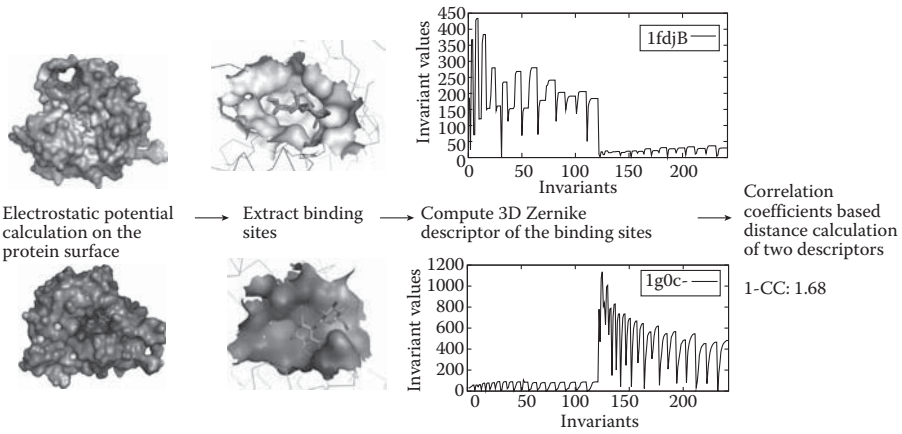

FIGURE 5.3: Local surface analysis procedure. A procedure for analyzing electrostatic potential on ligand binding region is illustrated. On the right are surface electrostatic potential of proteins 1fdjB, top, and 1g0c, bottom. Middle figure is surface electrostatic potential of extracted binding region which are the input to 3D Zernike method. The graphs are extracted 3DZD of the binding regions. The dissimilarity measures are calculated by correlation coefficients based distance of the two 3DZDs: 1.68. The gray scale ranges from -5 (black) to $+5$ (white) to represent electrostatic potential.

the nature of binding ligands, active sites show wide ranging behavior in terms of electrostatics, whose similarity can be quantified by using 3DZDs.

5.7 Discussion

In this chapter, we reviewed methods for 3D object shape analysis in the context of protein shape analysis. Some of the available methods are listed in Table 5.1.

Protein surface analysis is especially difficult because most of proteins have more or less sphere-like shape. Therefore a descriptor needs to differentiate relatively small differences. For the same reason, typical pose normalization methods, such as PCA, do not give a unique solution. Also nonshape features should be considered, such as physicochemical properties and residue conservation, as they are important to understand protein function. In addition, it is desired that computing similarity of two proteins is executed fast enough so that a database search is performed in a real-time.

To meet all these requirements, we applied 3DZD for protein surface comparison. Conventionally proteins have been long analyzed and classified in

TABLE 5.1: List of existing tools and computational resources.

	URL	Available materials
Light field descriptors (Chen et al., 2003)	http://3d.csie.ntu.edu.tw/	Source code/executable/dataset
Ef-site (Kinoshita et al., 2002)	http://ef-site.hgc.jp/eF-site/ http://ef-site.hgc.jp/eF-seek/	Database Web server
SURFACE (Ferre et al., 2005)	http://cbm.bio.uniroma2.it/surface/	Database
SURFCOMP (Hofbauer et al., 2004)	http://teachme.tuwien.ac.at/surfcomp/index.html	Toolkit/source code/executable
SURF'S UP (Pawlowski et al., 2001)	http://asia.genesilico.pl/surfs_up/	Web server
3d Zernike Server (Sael et al., 2008b)	http://dragon.bio.purdue.edu/3d-surfer/	Web server
Princeton Shape Retrieval and Analysis Group	http://shape.cs.princeton.edu/search.html	Web server

terms of their sequences and main-chain conformation. However, there are cases that these methods are not capable of detecting similarities and dissimilarities in a biologically meaningful way. In such cases, the 3DZD-based surface analysis can often do a better job than these methods, as it captures global and local protein surface shape, which is directly responsible for biological function, and also it is able to quantify similarity of physicochemical properties.

As more and more protein structures are experimentally solved, the need for effective and efficient methods for protein structure characterization increases. We expect the surface-based analysis will become a routine option for protein characterization besides sequence- and main-chain conformation-based methods. Biology has entered an informatics era when computational methods for retrieving useful knowledge from databases and reasoning using the knowledge become crucial. Various interdisciplinary approaches are essential and protein surface analysis will become one of such existing fields.

Acknowledgments

This work is supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM075004). We thank Gregg Thomas for proofreading the manuscript.

References

Arakaki, A.K. and J. Skolnick. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, 20, 2004, 1087–1096.

Baldacci, L., M. Golfarelli, A. Lumini, and S. Rizzi. Clustering techniques for protein surfaces. *Pattern Recogn.,* 39, 2006, 2370–2382.

Binkowski, T. Andrew, L. Adamian, and J. Liang. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.,* 332(2), 2003 505–526.

Bock, M. E., C. Garutti, and C. Guerra. Discovery of similar regions on protein surfaces. *J. Comput. Biol.,* 14(3), 2007, 285–299.

Canterakis, N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. In *Proceedings of the 11th Scandinavian Conference on Image Analysis*, Kangerlussuaq, Greenland, 1999, 85–93.

Chen, D. Y., M. Ouhyoung, X. P. Tian, Y. T. Shen, and M. Ouhyoung. On visual similarity based 3D model retrieval. In *Proceedings of Eurographics 2003*. Granada, Spain, 2003, 223–232.

Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic-acids. *Science,* 221, 1983, 709–713.

Connolly, M. L. Shape complementarity at the hemoglobin alpha I beta I subunit interface. *Biopolymers,* 25, 1986, 1229–1247.

de Alarc, P. A., A. D. Pascual-Montano, and J. M. Carazo. Spin Images and Neural Networks for Efficient Content-Based Retrieval in 3D Object Databases. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*. London, UK: Springer-Verlag, 2002, 225–234.

Elad, M., A. Tal, and S. Ar. Directed search in a 3d objects database using svm. Technical report, HP Laboratories, Israel, 2000.

Ferrè, Fabrizio, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich. Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinform.,* 6, 2005, 194–208.

Fischer, D., S. L. Lin, H. L. Wolfson, and R. Nussinov. A geometry-based suite of molecular docking processes. *J. Mol. Biol.,* 248(2), 1995, 459–477.

Funkhouser, T., and P. Shilane. Partial matching of 3D shapes with priority-driven search. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*. Carligari, Sardinia, Italy. ACM International Conference Proceeding Series. Eurographics Association, Aire-la-Ville, Swizerland, Vol. 256, June 26–28, 2006, 131–142.

Gerstein, M. A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites. *Acta Crystallogr.,* A48, 1992, 271–276.

Halperin, I., B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins,* 47, 2002, 409–443.

Hofbauer, C., H. Lohninger, and A. Aszódi. SURFCOMP: a novel graph-based approach to molecular surface comparison. *J. Chem. Inf. Comput. Sci.,* 44(3), 2004, 837–847.

Johnson, A. E., and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans. Pattern Anal. Mach. Intell.,* 21(5), 1999, 433–449.

Kinoshita, K., J. Furui, and H. Nakamura. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics,* 2(1), 2002, 9–22.

Kahraman, A., R. J. Morris, R. A. Laskowski, and J. M. Thornton. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.,* 368(1), 2007, 283–301.

Kazhdan, M., T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In *SGP '03: Proceedings of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*. Aire-la-Ville, Switzerland. Eurographics Association, 2003, 156–164.

Koenderink, J. J., and A. J. van Doorn. Surface shape and curvature scales. *Image Vision Comput.,* 10(8), 1992, 557–564.

Laga, H., H. Takahashi, and M. Nakajima. Spherical wavelet descriptors for content-based 3D model retrieval. In *SMI '06: Proceedings of the IEEE International Conference on Shape Modeling and Applications*, Matsushima, Japan, 2006, 75–85.

Leifman, G., S. Katz, A. Tal, and R. Meir. Signatures of 3D models for retrieval. In *4th Israel Korea Bi-National Conference on Geometric Modeling and Computer Graphics*, Tel-Aviv, Israel, 2003, 159–163.

Lin, S. L., R. Nussinov, D. Fischer, and H. J. Wolfson. Molecular surface representations by sparse critical points. *Proteins,* 18(1), 1994, 94–101.

Mademlis, A., A. Axenopoulos, P. Daras, D. Tzovaras, and M. G. Strintzis. 3D content-based search based on 3D Krawtchouk moments. In *3DPVT '06: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*. Washington, DC: IEEE Computer Society, 2006, 743–749.

Mak, L., S. Grandison, and R. J Morris. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *J. Mol. Graph. Model.,* 26(7), 2008, 1035–1045.

Masek, B. B., A. Merchant, and J. B. Matthew. Molecular skins: a new concept for quantitative shape matching of a protein with its small molecule mimics. *Proteins,* 17(2), 1993, 193–202.

Novotni, M., and R. Klein. 3D Zernike descriptors for content based shape retrieval. In *The 8th ACM Symposium on Solid Modeling and Applications*, Seattle, Washington, 2003.

Novotni, M., and R. Klein. Shape retrieval using 3D Zernike descriptors. *Computer-Aided Design 36,* 11, 2004, 1047–1062.

Ohbuchi, R., M. Nakazawa, and T. Takei. Retrieving 3D shapes based on their appearance. In *MIR '03: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*. New York, NY: ACM Press, 2003, 39–45.

Osada, R., T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. *ACM Trans. Graph.,* 21(4), 2002, 807–832.

Pawlowski, K., and A. Godzik. Surface map comparison: studying function diversity of homologous proteins. *J. Mol. Biol.,* 309, 2001, 793–800.

Pickering, S. J., A. J. Bulpitt, N. Efford, N. D. Gold, and D. R. Westhead. AI-based algorithms for protein surface comparisons. *Comput. Chem.,* 26(1), 2001, 79–84.

Poirrette, A. R., P. J. Artymiuk, D. W. Rice, and P. Willett. Comparison of protein surfaces using a genetic algorithm. *J. Comput. Aided Mol. Des.,* 11(6), 1997, 557–569.

Ritchie, D. W., and G. J. L. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins,* 39, 2000, 178–194.

Rosen, M., S. L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng.,* 11(4), 1998, 263–277.

Sael, L., D. La, B. Li, R. Rustamov, and D. Kihara. Rapid comparison of properties on protein surface. *Proteins,* 73, 2008a, 1–10.

Sael, L., B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, and D. Kihara. Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins,* 72, 2008b, 1259–1273.

Shentu, Z., M. Al Hasan, C. Bystroff, and M.J. Zaki. Context shapes: efficient complementary shape matching for protein-protein docking. *Proteins*, 70(3), 2008, 1056–1073.

Shindyalov, I. N., and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.,* 11(9), 1998, 739–747.

Tangelder, J. W. H., and R. C. Veltkamp. A survey of content based 3D shape retrieval methods. In *SMI '04: Proceedings of the Shape Modeling International 2004 (SMI'04)*. Washington, DC: IEEE Computer Society, 2004, 145–156.

Wang, X. Alpha-surface and its application to mining protein data. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, 2001, 659–662.

Yu, M., I. Atmosukarto, W. K. Leow, Z. Huang, and R. Xu. 3D model retrieval with morphing-based geometric and topological feature maps. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2003, 656–661.

Zhang, C., and T. Chen. Efficient feature extration for 2D/3D objects in mesh representation. In *Proceedings of the 2001 International Conference on Image Processing (ICIP 2001)*. Thessaloniki, Greece, 2001, October 7–10.