Cell Biochem Biophys (2009) 54:23–32
DOI 10.1007/s12013-009-9051-x

REVIEW

Potential for Protein Surface Shape Analysis Using Spherical Harmonics and 3D Zernike Descriptors

Vishwesh Venkatraman · Lee Sael · Daisuke Kihara

Published online: 12 June 2009

Abstract With structure databases expanding at a rapid rate, the task at hand is to provide reliable clues to their molecular function and to be able to do so on a large scale. This, however, requires suitable encodings of the molecular structure which are amenable to fast screening. To this end, moment-based representations provide a compact and non-redundant description of molecular shape and other associated properties. In this article, we present an overview of some commonly used representations with specific focus on two schemes namely spherical harmonics and their extension, the 3D Zernike descriptors. Key features and differences of the two are reviewed and selected applications are highlighted. We further discuss recent advances covering aspects of shape and property-based comparison at both global and local levels and demonstrate their applicability through some of our studies.

Keywords Protein structure comparison · Protein surface · Protein docking prediction · 3D Zernike descriptors · Spherical harmonics

V. Venkatraman · D. Kihara (✉)
Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA
e-mail: dkihara@purdue.edu

L. Sael · D. Kihara
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

V. Venkatraman · D. Kihara
Markey Center for Structural Biology, Purdue University, West Lafayette, IN 47907, USA

Introduction

Understanding protein structure and its relation to the various biological functions that it carries out has been a primary goal in structural biology. Based on the premise that, structurally similar proteins have similar function in many cases [1–3], various structure comparison approaches using different representations of proteins have been employed [4]. Commonly used definitions include those based on the backbone Cα positions [5, 6], the distance map [7], secondary structure elements [8], backbone torsion angles [9], and molecular surfaces [10]. The algorithms and theory used in such computational methods are tightly intertwined with the protein representations. For example, dynamic programming (DP) is a commonly used algorithm for comparing protein structures using the backbone representation [5, 11]. In other approaches, schemes based on the Monte Carlo algorithm [7], graph theory and clustering [12], and knot theory [13] have been applied. Since different representations capture different aspects of protein structures, they do not necessarily agree in the degree of structural similarity they identify. Therefore, naturally, suitable applications for each method may differ. For example, it is more appropriate to use DP-based methods that consider similarity in the protein backbone orientation [6] to compare evolutionarily closely related proteins. However, such methods will miss circularly permutated structures that are evolutionarily related [14], as they are compared sequentially from the N- to C-terminus. Methods that consider the spatial arrangement of secondary structure elements [8] have therefore been developed for dealing with these cases. Distance map-based methods [7] would be more suitable for comparing closely related structures, where their main differences may be limited to side-chain contacts. For further discussions, please refer to

some recent reviews on structure comparison methods [4, 15].

Structure databases are growing at a rapid pace, and have benefited from structural genomics projects [16–19], which have been accumulating an increasing amount of protein structures whose function has remained unknown. Searching these vast resources requires suitable encodings of the protein structure that are amenable to fast comparisons. In addition, protein structures with unknown function pose a challenging task for structure-based function prediction approaches [20–22]. In recent years, a few studies have employed surface shape for comparison [10, 23–25] to address these two issues [23]. An advantage of protein surface shape comparison is that the structural similarity can be captured across evolutionarily distant proteins [26]. This is useful for function prediction because geometrical and physicochemical properties of functional sites can be directly compared. For example, the eF-site database represents the protein as a triangulated surface mesh and employs graph matching techniques to identify common local regions between two proteins [10]. More recently, alpha shapes (a generalization of the convex hull) have been used to characterize the molecular surface as a set of contiguous patches and applied to the analysis of binding sites in proteins [27].

As speed is of importance, several shape recognition techniques make use of descriptors that capture the spatial profile of the protein as a multidimensional feature vector. Spin images [28], for example, provide a local two-dimensional description of the surface based on a reference frame defined by the associated surface points. In an another approach [29], global geometric properties of the protein are captured in the form of a probability distribution, i.e., a shape histogram, sampled from a shape function (e.g. angles, distances, areas, and volumes). As shape matching is reduced to a comparison of the histograms, it obviates the need for any feature correspondence or pre-alignment (independent of orientation).

Moment-based representations form another class of descriptors that have been used widely for pattern recognition [30]. These representations provide a compact numerical expression of the spatial features that enables rapid comparisons. Moments based on the theory of orthogonal polynomials [31], such as 2D/3D Zernike moments and Legendre moments, allow descriptors to be constructed to an arbitrary order with little or no redundancy. This feature also allows the object to be reconstructed from its moments with quality determined by the number of terms used [32].

In this review, we focus on the applications of spherical harmonics and the more recent 3D Zernike descriptors (3DZD) in biomolecular sciences. While the applications for the former have been widespread, the latter is a relatively new entrant in this field. We begin by introducing the properties of spherical harmonics and the 3DZD, while highlighting some differences and advantages in terms of their mathematical treatment. Applications for the two schemes are discussed, which range from global/local protein surface comparison to protein docking prediction.

Mathematical Foundation of Spherical Harmonics and 3DZD

Spherical Harmonic Representations in Protein Structure Analysis

Spherical harmonics [33] and their variants have gained much interest with applications in real-time rendering [34], visualization of molecular surfaces [35], electron microscopy [36], and 3D shape retrieval [37]. Mathematical properties such as orthonormality and completeness make them a suitable choice for surface modeling. The properties are, however, only valid on a unit ball and since a one-to-one mapping onto the sphere is required, it can only represent star-shaped or single-valued surfaces.

A single-valued three-dimensional (3D) surface can be parameterized in terms of spherical coordinates (θ, ϕ) as

$$\vec{x}(\theta,\phi) = r(\theta,\phi)\left[(\sin\theta\cos\phi\cdot\vec{i} + \sin\theta\sin\phi\cdot\vec{j} + \cos\theta\cdot\vec{k})\right] \tag{1}$$

where $\vec{x}$ is the set of surface (Cartesian coordinates) points, $r(\theta, \phi)$ is the distance of surface points from a chosen origin, and $\vec{i}, \vec{j}, \vec{k}$ are the unit vectors of the three perpendicular coordinate axes. This parameterization yields a bijective mapping between each surface vertex $\vec{x}$ and a pair of spherical coordinates (θ, ϕ). The surface can also be expressed as a unique linear combination of spherical harmonics basis functions, $Y_l^m(\theta,\phi)$, which for degree l order m are given by

$$r(\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} c_{l,m}Y_l^m(\theta,\phi) \tag{2}$$

The coefficients $c_{l,m}$ are uniquely determined by

$$c_{l,m} = \int_0^{2\pi}\int_0^{\pi} f(\theta,\phi)Y_l^m(\theta,\phi)^*\sin\theta\mathrm{d}\theta\mathrm{d}\phi \tag{3}$$

where * indicates complex conjugation. The basis function is defined as follows

$$Y_l^m = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}P_l^m(\cos\theta)e^{im\phi} \tag{4}$$

Here $P_l^m(\cos\theta)$ are the associated Legendre polynomials (with argument $\cos\theta$) and θ, ϕ are the spherical

coordinates. The surface is typically approximated by truncating the series of spherical harmonics coefficients to a finite number of terms. Thus, a limit L is chosen to obtain a desired level of resolution which yields $(L+1)^2$ terms.

$$r(\theta,\phi)=\sum_{l=0}^{L}\sum_{m=-l}^{l}c_{l,m}Y_l^m(\theta,\phi) \tag{5}$$

However, any rotation of $r(\theta,\ \phi)$ with respect to the coordinate system change the magnitude and/or phase of the component $c_{l,m}$, i.e., spherical harmonics are not rotationally invariant. Thus, a prealignment, which requires objects to be placed in a standard frame of reference [38], is necessary, before the objects can be compared. Translation invariance is achieved by moving the center of geometry of the object to the origin of the coordinate system. This is typically followed by a principle component analysis (PCA) step to obtain rotational invariance, but may not always yield robust normalizations and can affect descriptor performance [39]. However, the use of Wigner matrices [40] allows for a distance preserving transformation (rotation of a spherical function does not change its Euclidean norm), thus defining rotationally invariant regions [41].

Another alternative to calculating rotation invariant descriptors is to decompose the spherical function $f_l(\theta,\phi)=\sum_{m=-l}^{l}a_{lm}Y_l^m(\theta,\phi)$ as a sum of its L_2 norms of harmonics [30]. The norms of the frequency component are invariant to rotation (property of the spherical harmonics) and therefore an orientation independent descriptor F can be constructed where each component is the L_2-norm of the spherical function restricted to some frequency l:

$$F=(\|f_0(\theta,\phi)\|,\|f_1(\theta,\phi)\|,\|f_2(\theta,\phi)\|,,,,) \tag{6}$$

Using this scheme, Funkhouser et al. decomposed an object into a set of concentric spheres and then combined the rotation invariant descriptor F computed for each sphere [39]. Although this strategy can handle nonstar-like shapes, it also looses information as the representation is invariant to independent rotations of the different spherical functions representing each component. As a result, the original shape cannot be reconstructed from the calculated descriptors.

3D Zernike Descriptors

2D Zernike moments have been used in a wide range of applications in image analysis [32, 42] owing to their advantageous properties of rotation invariance, robustness to noise and small information redundancy (orthogonality of the basis functions). The extension of the 2D Zernike to its 3D counterpart was initially formulated by Canterakis [43] and later applied to shape retrieval by Novotni and Klein [44]. Their introduction into bioinformatics has, however, been quite recent with applications in protein shape comparison [26, 45] and docking [46]. A difference of the 3DZD from the spherical harmonics is the addition of a radial term, which enables 3D shapes to be modeled more precisely than spherical harmonics. The incorporation of the radial polynomial removes the star-shape requirement that affects the spherical harmonic representation.

The 3D Zernike moments are derived from a set of 3D polynomials which for a 3D object (shape function) is given by

$$Z_{nl}^m(r,\theta,\phi)=R_{nl}(r)Y_l^m(\theta,\phi) \tag{7}$$

Here Y_l^m are complex valued spherical harmonics that are orthonormal on the surface of the unit sphere and l, m, and n are integers representing the degree, order, and repetition. The radial function $R_{nl}(r)$ defined by Canterakis [43], directly incorporates radius information into the basis function and is constructed so that $Z_{nl}^m(r,\theta,\phi)$ are polynomials, when written in terms of Cartesian coordinates. The 3D Zernike moments of an object modeled by the function f(x) are defined as the coefficients of the expansion in this orthonormal basis

$$\Omega_{nl}^m=\frac{3}{4\pi}\int_{|x|\le 1}f(x)\overline{Z_{nl}^m(x)}\mathrm{d}x \tag{8}$$

As a first step, the center of gravity of the object is computed and then transformed to the origin. As these moments are not invariant under rotation, to obtain transformation invariant descriptors, i.e., the 3DZD, the norms of vectors $\|\Omega_{nl}^m\|$ are computed.

Figure 1 shows the computation process for the 3DZD. The procedure starts with the discretization of the protein molecular surface (the Connolly surface definition [47] is used) where each voxel (a cubic grid cell) records a scalar value (1 for surface and 0 otherwise). Properties other than shape such as electrostatics or hydrophobicity can also be captured by the voxelization in a similar way by assigning the values to the voxels (instead of 1 as in the surface shape). For a given order n (typically set to 10 for ligands and 20 for proteins), the 3DZD are then extracted from this voxelized structure, so that the molecule is represented by a numeric vector of length $\left(\frac{n}{2}+1\right)^2$ when n is even and $\frac{(n+1)(n+3)}{4}$ for odd values. Thus, the 3DZD yield a more compact representation, as for the same order of expansion, $(n+1)^2$ spherical harmonic coefficients are produced. Comparison of protein structures is therefore reduced to evaluating a suitable distance measure (the Euclidean distance or the correlation coefficient) between the vectors. Figure 2 shows two proteins that have dissimilar shapes which are also reflected by the difference in magnitudes of the 3DZD.

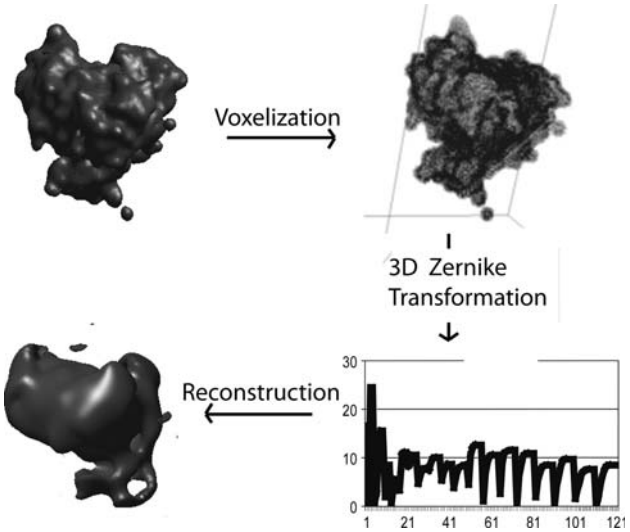

Fig. 1 Flowchart of 3DZD computation process

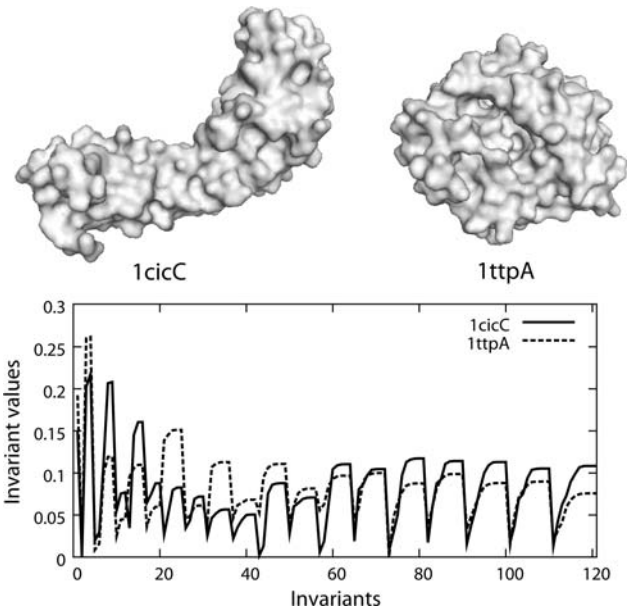

Fig. 2 Examples of the 3D Zernike descriptors for two proteins with dissimilar surface shape

Difference Between Spherical Harmonics and 3DZD

Spherical harmonics and 3DZD primarily differ in terms of the mathematical framework, with the latter having more desirable features that enhance its application to shape comparison. Firstly, by incorporating a radial function, $R_{nl}(r)$ (Eq. 7), the 3DZD can describe nonstar-like shapes (i.e., shapes which have multiple values at different distances r from the center for a given (θ, ϕ)). Spherical harmonics are restricted to star-like shapes or single-valued surfaces. Secondly, the 3DZD are rotation and translation invariant, and hence, in principle, the orientation of a structure does not affect the coefficients of the expansion. This feature is very convenient for constructing and searching a structure database as the 3DZD for the entries in the database can be precomputed. In contrast, the spherical harmonics for a given structure is dependent on its orientation. As a result, pose normalization of structures using PCA, etc., is required prior to computing the spherical harmonics for the two structures to be compared. This, however, can be problematic especially when comparing proteins as most are globular in shape, and for which the principle axes are not robustly determined. Thirdly, the 3DZD yield a more compact representation as compared to spherical harmonics, requiring fewer numbers for the same order of expansion. The rotation invariant (concentric shell scheme) harmonics named the spherical harmonics descriptors (SHD) [39] have addressed the first two issues to a certain extent. However, the size (the number of values) of the SHD is further increased as spherical harmonics is computed for each concentric sphere. Also, the harmonics of adjacent concentric cells are highly correlated. Concretely, 121 scalar values is used in the 3DZD with the order up to 20, while the SHD used 32 concentric spheres, each of which is described by 16 harmonics descriptors, resulting in 512 scalar values. Nevertheless, the 3DZD showed better shape retrieval results as tested on the Princeton Shape Benchmark, which is a database of general 3D objects, such as airplanes and chairs [44].

In summary, the 3DZD are built on the spherical harmonics and have more favorable characteristics for applications in biomolecular structure analyses. The advantages of the 3DZD over spherical harmonics have been illustrated in benchmark studies involving database searches of general 3D objects [44] and protein global structures [26].

Applications

Below, we describe applications of spherical harmonics and the 3DZD to protein and ligand molecule structure analyses. Spherical harmonics have been in use for quite some time in both bioinformatics and chemoinformatics with several works reported in literature. On the other hand, the 3DZD have been introduced to this field very recently, and their applications to several problems are still ongoing. In what follows, we first review recent research based on spherical harmonics before moving onto highlight applications of the 3DZD in protein global and local surface shape comparison and docking, most of which are done in our group.

Applications of Spherical Harmonics

Protein Structure Comparison

Gramada and Bourne [48] model protein shape quantitatively as multipolar expansion (written as a sum of spherical harmonics) terms associated with the residue C-α coordinates. Their method has the advantage of being easily extended to other residue-based properties while incorporating various levels of detail. The approach was tested on a set of kinase-like superfamily of proteins. In comparing the results with those obtained from the conventional alignment based approach, the spherical harmonic approach was found to provide a better discrimination of the families (tyrosine protein kinases, cyclin dependent kinases, etc.) within the dataset. Zhang et al. [49], on the other hand, used a voxelized representation of the protein which was normalized into a canonical coordinate system. A set of spherical harmonic coefficients was then extracted from a series of uniformly spaced concentric shells (concentric decomposition) around the center of mass of the protein. The approach was tested on a set of 37 randomly selected proteins from the PDB and pairwise comparisons using the Euclidean distance were able to retrieve structures that had significant sequence similarity with the query.

Ligand Binding Site Similarity

Molecular recognition is a key component of protein–ligand interactions and constitutes major function of proteins. The ligand recognition procedure is guided by the complementarity of physico-chemical properties of the ligand and the binding site residues of the target protein. This forms the basis for several function prediction approaches that work on the assumption that proteins with similar function may also have similar binding sites [50–56]. Kahraman et al. [57] used spherical harmonic representations to compare the shapes of the ligand and the binding pockets. The analysis was also extended to other properties such as hydrophobicity and electrostatic potentials. Based on the coefficient distances, shapes of pockets binding the same ligand were found to be more variable in comparison to the conformational variability exhibited by the ligand. In a parallel study by Morris et al. [38], the binding pockets of 40 proteins that bound four different ligand groups (adenosine triphosphate (ATP), adenosine diphosphate (ADP), heme, and steroids) were compared. While the binding pockets for steroids were found to be sufficiently similar. However, pockets of the other ligands showed significant variations in shape and therefore did not cluster together.

Protein–Protein Docking

Docking attempts to find the structure of the complex formed by two or more interacting proteins through computational means. A major challenge is to explore the six-dimensional (three translational and three rotational) space as efficiently as possible and produce orientations that are likely to mirror the native structure. Here, too, shape plays a crucial role in defining the complementarity although other factors such as hydrophobicity and electrostatics are also considered to be significant. Since flexibility adds to the complexity of the task, most approaches rely on shape to produce an initial list of candidate orientations after which further refinement is carried out [58].

Conventional approaches have typically used Fast Fourier transform (FFT) [59] that reduces the computational time of a translational search. However, there are two issues that add to the computational complexity: (1) Large grids may be required in many cases that add a significant memory overhead; (2) the FFTs have to be reevaluated for every rotational increment. The docking program HEX [60] instead uses spherical polar Fourier basis functions to represent the molecular shape and electrostatic potential of the protein. The docking search only involves rotation and translation of the initial expansion coefficients (spherical harmonic functions transform among themselves under rotation) which allows a large number of trial orientations to be tested without the use of the more expensive grids. In the blind CAPRI docking experiment [61], HEX was able to achieve a good solution (termed a hit) within the top 20 predictions in 4 out of the 7 cases [62].

Virtual Screening of Drug Molecules

A preliminary step in drug design is the identification of promising drug leads which requires large chemical databases to be searched. During the screening process, methods are typically required to exploit either shape and/or chemical property information. In addition to looking for similar molecules with similar properties, the technique should also be able to provide a suitable structural superposition. Spherical polar Fourier representations can effectively capture both steric and electrostatic properties and owing to their special rotational properties enable rapid overlap calculations [63, 64]. The approach [63, 64] was demonstrated using a small dataset of 73 common prescription drugs that were classified into 22 different categories (antibiotics, anticonvulsants, etc.). Retrieval speeds and accuracy were found to be much higher than other shape-based approaches such as the Gaussian shape overlay [65]. In addition, the clustering based on the spherical harmonic shapes was able to identify chemically meaningful groupings. Applications of these descriptors have

also been extended to virtual high-throughput screening (VHTS) [66]. A set of 1000 randomly selected drug-like decoys was used for testing and the objective was to retrieve known active ligands (vitamin D and HIV-1 protease) that were part of this database. In terms of the enrichment efficiency, the performance of the spherical harmonics approach (48%) was found to be significantly better than that of the ligand docking algorithm FRED (12%) [67].

Applications of 3DZD

Global Protein Shape Comparison

For a dataset of 2631 proteins classified into 27 classes [68], Mak et al. [41] used 3DZD to assess their performance for protein structure retrieval. The study found that the global shape comparison performed quite well (area under the ROC curve value of 0.94) although proteins with similar shapes did not tally with the evolutionary based classification.

In another study by Sael et al. [26], shape similarity was evaluated for a set of 2432 proteins (classified into 185-fold groups) culled from the CE database [6]. Here, the 3DZD-based comparison achieved around 90% agreement in retrieving proteins of the same conformation defined by CE. It was shown that this retrieval agreement with a conventional main-chain based structure comparison (i.e., CE) was significantly higher than DALI, a popular protein structure comparison method [7], and some of the other surface shape-based structure comparison approaches, including the spherical harmonics, the solid angle histogram, and the shape distribution methods. A significant advantage in using these invariants is that they are amenable to fast comparisons (searching the CE dataset with 2432 proteins took only 0.46 s). This speed is sufficiently fast for searching the whole Protein Data Bank (PDB) database with over 54,000 structures within seconds. In contrast, a pairwise structure comparison by CE typically takes a few seconds. Thus, a search against PDB by CE would take more than a day or two [26]. Several interesting examples such as DNA topoisomerase of E. coli and human and pairs of transporters were found that exemplify the case where surface shape similarity reflects the biological function of the proteins. Although, these proteins do not share significant sequence or main-chain structural identity and their similarity is only found by the surface comparison using the 3DZD.

As mentioned above, in principle, the 3DZD can be computed to represent and physicochemical properties mapped on protein surface. Sael et al. showed that surface electrostatics of proteins can be meaningfully compared by combining 3DZD computed for regions with positive and negative electrostatic potential values [45]. Figure 3 shows that thermophilic and mesophilic proteins can be distinguished by similarity of the electrostatic potential on their global surface. Examples of two families are shown: glutamate dehydrogenase and the TATA box binding protein. While both sequence and main-chain structure are unable to differentiate their thermophilic and mesophilic homologs, the 3DZD-based comparison of their electrostatic potentials provides a clear distinction with the two groups clustered at the opposite ends of the tree. For comparison, trees generated based on the sequence similarity are also shown in Fig. 3a and b. The results indicate that 3DZD are able to distinguish both the magnitude and the pattern of physicochemical properties defined on the protein surface.

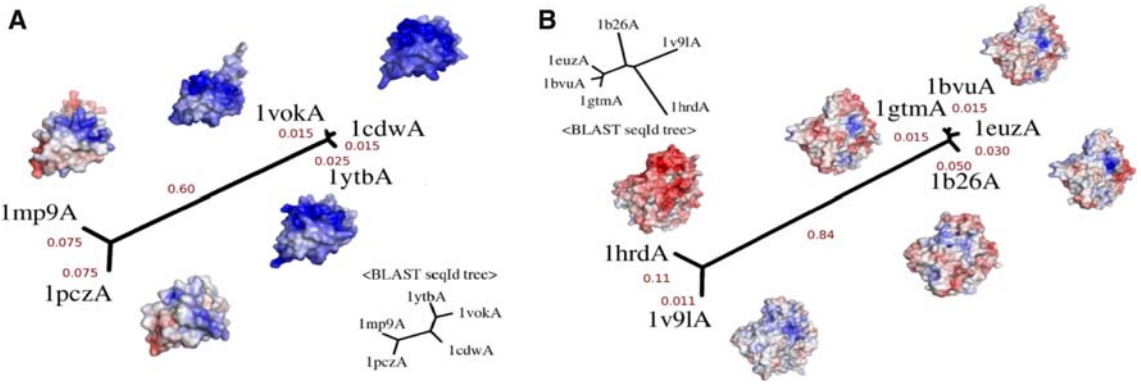

Fig. 3 Comparison of the electrostatic potentials of TATA-box binding protein (TBP) and Glutamate dehydrogenase (GDH) protein. in Trees (a) and (b) are generated using the CC distances of 3DZD (numbers on branches shown in red) of the electrostatic potential for the TBP and GDH proteins, respectively. Proteins of the TBP family include two thermophilic systems, 1mp9A (*Sulfolobus acidocaldarius*) and 1pczA (*Pyrococcus woesei*) and three mesophilic cases, 1ytbA (*Saccharomyces cerevisiae*), 1cdwA (human), and 1vokA (*Arabidopsis thaliana*). The GDH family includes 1hrdA, a mesophilic protein from Clostridium symbiosum, and the rest are thermophilic proteins: 1b26A (*Thermotoga maritima*), 1v9lA (*Pyrobaculum islandicum*), 1bvuA (*Thermococcus litoralis*), 1euzA (*Thermococcus profundus*), and 1gtmA (*Pyrococcus furiosus*). Trees representing sequence similarity are also shown

Taken together, they demonstrate the high-throughput screening ability of the descriptors as applied to global shape and physicochemical property comparison. The next section discusses their applicability to comparing local regions of proteins.

Ligand Binding Site Comparison

The 3DZD can also be used to compare local regions. The underlying idea is to segment the surface and generate 3DZD for the individual patches. For example, a molecular surface can be represented as a set of spherical patches that are centered on points of interest. Each patch captures the local shape in terms of the surface points that are contained in the sphere of arbitrary radius (say 6 Å). The number of such patches is greatly influenced by the size and shape of the protein and the number of interest points selected.

This local region-based description has been applied to the problem of identifying ligand binding interface in proteins [69]. The study is based on a subset of data that was earlier analyzed by Kahraman et al. [57]. This dataset consists of 14 proteins that bind ATP, 10 that bind flavin adenine dinuleotide (FAD), and 15 that bind nicotinamide adenine dinucleotide (NAD). For the identification of the binding sites, the following strategy is adopted:

(1) Proteins 1a0i, 1dnlA, 1e2jA that bind ATP, FMN, and NAD, respectively, were chosen as queries against the set of proteins that bind the same substrate. For example, local regions of protein 1aoi are compared with the local regions of the other ATP binding proteins. The interface regions of these proteins were therefore segmented by a sphere of a radius of 6 Å with each local segment described by corresponding 3DZD (the order of 10 was used). Typically, a protein surface is segmented into 200–1000 local regions.

(2) Local patches A and B on two proteins represented by their corresponding 3DZD (Z_A and Z_B) were then compared using the correlation (r) distance

$$CC = 1 - r(Z_A, Z_B), \quad \text{where } (-1 \le r \le 1) \tag{9}$$

(3) For each local region on the substrate binding site of the query protein, the best matching sixty local patches from other proteins are selected based on the correlation distance (CC).

(4) The chosen patches are then clustered (by the complete linkage clustering) using a distance cutoff of 6 Å.

(5) Within the top three clusters, three regions with the largest surface area are then chosen as the predicted binding sites.

(6) For each prediction, a measure of accuracy is computed. If T and P indicate the true and predicted binding sites that are composed of N_T and N_P voxels, respectively, then the accuracy can be calculated as

$$\text{RATIO} = \frac{2(N_T \cap N_P)}{N_T + N_P} \tag{10}$$

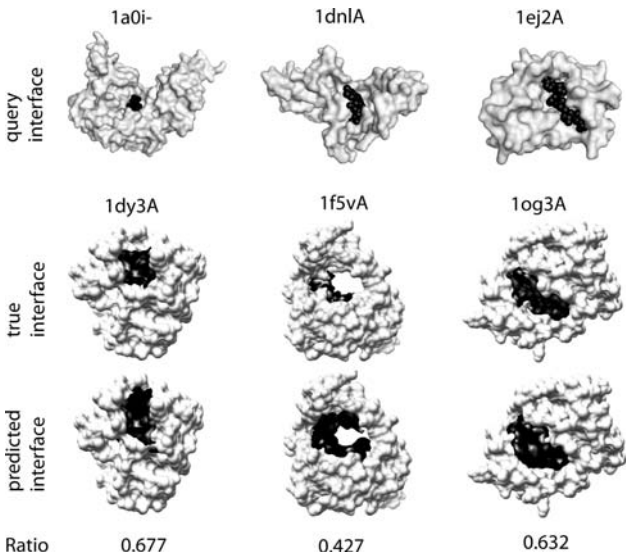

Fig. 4 Examples of local protein surface shape retrieval. The top row shows the substrate binding regions of three proteins 1a0i (ATP binding), 1dnlA (FMN), and 1e2jA (NAD) that are used as the query. The second row shows three other proteins whose interfaces are to be searched. Surface patches of 1dy3A, 1f5vA, and 1og3A are searched by patches of substrate binding regions of 1a0i, 1dnlA, and 1e2jA, respectively. The last row shows predicted binding regions of 1dy3A, 1f5vA, and 1og3A. The ratio shows the overlap between the actual and predicted binding regions

Figure 4 shows the predictions for the three cases listed above. As can be seen, this simple shape-only descriptor is able to capture the binding site region with the overlap ratio of 0.42 or above.

Local Shape Complementarity and Docking

Geometric complementarity between protein molecular surfaces (interfacial contact regions) is a widely used scheme for filtering docking conformations. Several strategies based on normals [70], grids [59], and atomic densities [71] have attempted to quantify this interaction with varying results. We have recently implemented a geometric hashing [72] based docking approach, named VDOCK, that compares local regions defined around a set of equally distributed points on the molecular surface. The shape of each such region (bounded by a 6 Å sphere centered at the point) is captured as a set of 36 3DZD numbers obtained from an order 10 expansion. Note here that in this application for protein docking prediction, we use 3DZD for capturing protein (local) surface shape complementarity rather than shape similarity. Although it may seem shape

similarity and complementarity are different, interestingly, 3DZD can also capture shape complementarity because they essentially describe contrast of distribution of values in space.

The docking method, VDOCK, looks for point correspondences that have complementary local shapes described by the correlation between their 3DZD and the orientation of the normals associated with the points. In the scoring, the amount of buried surface area and a penalty term that accounts for the undesirable clashes are also considered. The docking study was performed on 84 complexes from the ZDOCK benchmark 2.0 dataset [73]. Performance of the algorithm VDOCK is compared with the spherical harmonics-based HEX [60] and the FFT-based ZDOCK [74]. Evaluation criteria are based on the mean of the logarithm of the first ranked hit (defined as predictions whose ligand RMSD, i.e., RMSD of the backbone atoms of the unbound and bound ligand is <10 Å) and is calculated as follows:

$$\mathrm{MLR} = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\ln(\min(R_i, 1000))\right) \quad (11)$$

In Eq. 11, N is the number of complexes and R_i is the rank of the first hit. The mean rank MLR has a minimum value of 1 (all rank 1 hits) and a maximum value of 1000 when all ranked hits (if any) exceed the threshold value 1000.

Table 1 Comparison statistics for HEX (spherical harmonics), VDOCK (Zernike), and ZDOCK (FFT)

	HEX	VDOCK	ZDOCK
Mean rank	206	**164**	173
Mean ligand RMSD	7.39	**7.37**	7.55
# Cases with a better ranked hit	18	**22**	19
# HITS in top 2000	8	18	**37**

Numbers shown in bold indicate the docking program achieves a better performance for the statistics being compared

Out of the 59 cases where either program achieved at least one hit, VDOCK has a better ranked hit in 22 of the cases (see Table 1 and Fig. 5) and on the whole attains a lower mean rank and ligand RMSD.

Available Software

As far as we know, available software that uses spherical harmonics or the 3DZD for biomolecular structure analyses is very limited. The spherical harmonics-based docking and molecular superposition program HEX [60] are available as ready-to-use package (http://www.loria.fr/~ritchied/hex). For shape-based virtual screening, commercially available software PARAFIT (http://www.ceposinsilico.de) has been developed. Alternatively, one may use libraries such as SpharmonicKit (http://www.cs.dartmouth.edu/~geelong/sphere) which provide a number of algorithms (written in C) for spherical harmonic transforms. For the 3DZD, we have developed a web server, 3D-Surfer (http://dragon.bio.purdue.edu/3d-surfer/), which allows real-time structure searches of the entire PDB database [26]. Users can obtain a list of proteins in the PDB that share global surface similarity to a query protein, either specified by its PDB code or uploaded separately.

Conclusion

The surface of a molecule (protein/ligand) provides vital clues about their function. Attempts have therefore been made to condense this information into a form that is amenable to searching large structural libraries. Spherical harmonics and its extension, the 3D Zernike moments, are two such representations that characterize the shape as a unique, nonredundant set of numbers. The 3DZD score over the former in terms of rotational invariance (no prior structure alignment is needed) and compactness, thus

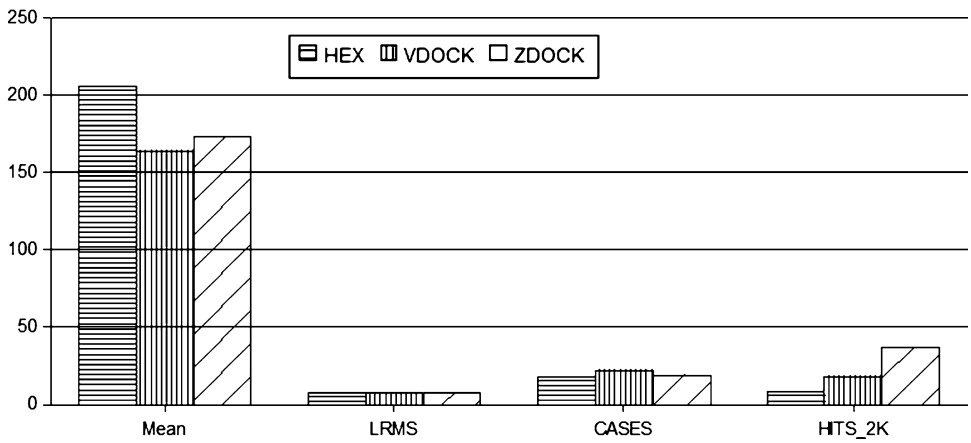

Fig. 5 Graph shows the comparison of the mean rank of the first correct hit (hit with a ligand RSMD <10Å) (Mean), mean ligand RMSD of the first correct hit (LRMS), and the number of benchmark test cases for which the program obtained a correct hit at a better rank than the other programs (CASES), and the number of cases where a correct hit is obtained within top 2000 (HITS_2K)

requiring fewer numbers to represent surface shape as compared with the spherical harmonics.

Given the succinct nature of 3DZD and the ease of comparison they offer, direct applications include similarity searching of proteins or ligands, based on shape or other criteria. Examples to this effect have already demonstrated qualities such as the speed of retrieval and their impact on cases where sequence alone has had little to offer. We have also shown the applicability of the 3D Zernike descriptors to local region matching with respect to the analysis of the binding interfaces of proteins and as measures of complementarity in protein–protein docking. In both cases, results have been quite encouraging with more applications to follow in the future.

Acknowledgments This work is supported by grants from the National Institutes of Health (R01 GM075004) and National Science Foundation (DMS0604776, DMS800568).

References

1. Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., et al. (1998). Protein folds and functions. *Structure, 6*, 875–884.
2. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., & Orengo, C. A. (2000). From structure to function: Approaches and limitations. *Nature Structural Biology, 7*(Suppl), 991–994.
3. Kihara, D., & Skolnick, J. (2004). Microbial Genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins, 55*, 464–473.
4. Mizuguchi, K., & Go, N. (1995). Seeking significance in three-dimensional protein structure comparisons. *Current Opinion in Structural Biology, 5*, 377–382.
5. Kihara, D., & Skolnick, J. (2003). The PDB is a covering set of small protein structures. *Journal of Molecular Biology, 334*, 793–802.
6. Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering, 11*, 739–747.
7. Holm, L., & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology, 233*, 123–138.
8. Mizuguchi, K., & Go, N. (1995). Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Engineering, 8*, 353–362.
9. Betancourt, M. R., & Skolnick, J. (2004). Local propensities and statistical potentials of backbone dihedral angles in proteins. *Journal of Molecular Biology, 342*, 635–649.
10. Kinoshita, K., & Nakamura, H. (2003). Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Science, 12*, 1589–1595.
11. Taylor, W. R., & Orengo, C. A. (1989). Protein structure alignment. *Journal of Molecular Biology, 208*, 1–22.
12. Lozano, M., & Escolano, F. (2006). Protein classification by matching and clustering surface graphs. Pattern Recognition, 39(4), 539–551
13. Rogen, P., & Fain, B. (2003). Automatic classification of protein structure by using Gauss integrals. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 119–124.
14. Pan, T., & Uhlenbeck, O. C. (1993). Circularly permuted DNA, RNA and proteins—A review. *Gene, 125*, 111–114.
15. Kolodny, R., Petrey, D., & Honig, B. (2006). Protein structure comparison: Implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology, 16*, 393–398.
16. Burley, S. K. (2000). An overview of structural genomics. *Nature Structural Biology, 7*(Suppl), 932–934.
17. Todd, A. E., Marsden, R. L., Thornton, J. M., & Orengo, C. A. (2005). Progress of structural genomics initiatives: An analysis of solved target structures. *Journal of Molecular Biology, 348*, 1235–1260.
18. Westbrook, J., Feng, Z., Chen, L., Yang, H., & Berman, H. M. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Research, 31*, 489–491.
19. Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., et al. (2000). Structural genomics projects in Japan. *Progress in Biophysics and Molecular Biology, 73*, 363–376.
20. Todd, A. E., Orengo, C. A., & Thornton, J. M. (2001). Evolution of function in protein superfamilies from a structural perspective. *Journal of Molecular Biology, 307*, 1113–1143.
21. Redfern, O. C., Dessailly, B., & Orengo, C. A. (2008). Exploring the structure and function paradigm. *Current Opinion in Structural Biology, 18*, 394–402.
22. Hawkins, T., & Kihara, D. (2007). Function prediction of uncharacterized proteins. *Journal of Bioinformatics and Computational Biology, 5*, 1–30.
23. Sael, L., & Kihara, D. (2009). Protein surface representation and comparison: New approaches in structural proteomics. In J. Chen & S. Lonardi (Eds.), *Biological data mining*. Boca Raton, FL, USA: Chapman & Hall/CRC Press.
24. Via, A., Ferre, F., Brannetti, B., & Helmer-Citterich, M. (2000). Protein surface similarities: A survey of methods to describe and compare protein surfaces. *Cellular and Molecular Life Sciences, 57*, 1970–1977.
25. Exner, T. E., Keil, M., & Brickmann, J. (2002). Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. *Journal of Computational Chemistry, 23*, 1176–1187.
26. Sael, L., Li, B., La, D., Fang, Y., Ramani, K., Rustamov, R., et al. (2008). Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins, 72*, 1259–1273.
27. Albou, L. P., Schwarz, B., Poch, O., Wurtz, J. M., & Moras, D. (2008). Defining and characterizing protein surface using alpha shapes. *Proteins, 76*, 1–12.
28. Bock, M. E., Cortelazzo, G. M., Ferrari, C., & Guerra, C. (2005). Identifying similar surface patches on proteins using a spin-image surface representation. pp. 417–428.
29. Ankerst, M., Kastenmuller, G., Kriegel, H.-P., & Seidl, T. (1999). 3D shape histograms for similarity search and classification in spatial databases. pp. 207–226.
30. Hu, M. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory, 8*, 179–187.
31. Teague, M. (1980). Image analysis via the general theory of moments. *Journal of the Optical Society of America, 70*, 920–930.
32. Teh, C. H., & Chin, R. T. (1988). On image-analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10*, 496–513.
33. Biedenham, L., & Louck, J. (1981). *Angular momentum in quantum physics*. Reading, MA, USA: Addison-Wesley.
34. Lisle, I., & Huang, S. (2007). Algorithms for spherical harmonic lighting. In: *GRAPHITE'07: Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia*, New York, NY, USA, ACM (pp. 235–238).

35. Max, N., & Getzoff, E. (1988). Spherical harmonic molecular surfaces. *IEEE Computer Graphics and Applications*, 8(4), 42–50.
36. Kovacs, J. A., & Wriggers, W. (2002). Fast rotational matching. *Acta Crystallographica. Section D, Biological Crystallography, 58*, 1282–1286.
37. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., et al. (2003). A search engine for 3D models. *ACM Transactions on Graphics, 22*, 83–105.
38. Morris, R. J., Najmanovich, R. J., Kahraman, A., & Thornton, J. M. (2005). Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics, 21*, 2347–2355.
39. Kazhdan, M., Funkhouser, T., & Rusinkiewicz, S. (2003). Rotation invariant spherical harmonic representation of 3D shape descriptors. In *Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on geometry processing*, vol 43 (pp. 156–164).
40. Edmonds, A. R. (1957). *Angular momentum in quantum mechanics*. Princeton, NJ: Princeton University Press.
41. Mak, L., Grandison, S., & Morris, R. J. (2007). An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *Journal of Molecular Graphics & Modelling, 26*, 1035–1045.
42. Zhang, D., & Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37(1), 1–19.
43. Canterakis, N. (1999). 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. In: *Proc. 11th scandinavian conference on image analysis* (pp. 85–93).
44. Novotni, M., & Klein, R. (2003). 3D Zernike descriptors for content based shape retrieval. In: *ACM symposium on solid and physical modeling, proceedings of the eighth ACM symposium on Solid modeling and applications* (pp. 216–225).
45. Sael, L., La, D., Li, B., Rustamov, R., & Kihara, D. (2008). Rapid comparison of properties on protein surface. *Proteins, 73*, 1–10.
46. Venkatraman, V., Yang, Y. D., Sael, L., & Kihara, D. (2009). Protein–protein docking using region-based 3D Zernike descriptors (submitted).
47. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science, 221*, 709–713.
48. Gramada, A., & Bourne, P. E. (2006). Multipolar representation of protein structure. *BMC. Bioinformatics, 7*, 242.
49. Zhang, T., Chen, W., Hu, M., & Peng, Q. (2005). A similarity computing algorithm for proteins. In *Ninth international conference on computer aided design and computer graphics (CAD-CG'05)* (pp. 168–172).
50. Binkowski, T. A., & Joachimiak, A. (2008). Protein functional surfaces: Global shape matching and local spatial alignments of ligand binding sites. *BMC Structural Biology, 8*, 45.
51. Zhang, Z., & Grigorov, M. G. (2006). Similarity networks of protein binding sites. *Proteins, 62*, 470–478.
52. Kihara, D., Sael, L., & Chikhi, R. (2009). Local surface shape-based protein function prediction using Zernike descriptors. *Biophysical Journal, 96*, 650a.
53. Chikhi, R., & Kihara, D. (2009). Real-time protein binding ligand prediction using local surface descriptors (submitted).
54. Liang, J., Edelsbrunner, H., & Woodward, C. (1998). Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science, 7*, 1884–1897.
55. Kinoshita, K., & Nakamura, H. (2005). Identification of the ligand binding sites on the molecular surface of proteins. *Protein Science, 14*, 711–718.
56. Kalidas, Y., & Chandra, N. (2008). PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. *Journal of Structural Biology, 161*, 31–42.
57. Kahraman, A., Morris, R. J., Laskowski, R. A., & Thornton, J. M. (2007). Shape variation in protein binding pockets and their ligands. *Journal of Molecular Biology, 368*, 283–301.
58. Ritchie, D. W. (2008). Recent progress and future directions in protein–protein docking. *Current Protein & Peptide Science, 9*, 1–15.
59. Chen, R., Li, L., & Weng, Z. (2003). ZDOCK: An initial-stage protein-docking algorithm. *Proteins, 52*, 80–87.
60. Ritchie, D. W., & Kemp, G. J. (2000). Protein docking using spherical polar Fourier correlations. *Proteins, 39*, 178–194.
61. Mendez, R., Leplae, R., De, M. L., & Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins, 52*, 51–67.
62. Ritchie, D. W. (2003). Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins, 52*, 98–106.
63. Mavridis, L., Hudson, B. D., & Ritchie, D. W. (2007). Toward high throughput 3D virtual screening using spherical harmonic surface representations. *Journal of Chemical Information and Modeling, 47*, 1787–1796.
64. Ritchie, D. W., & Kemp G. J. L. (1999). Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *Journal of Computational Chemistr*y, 20(4), 383–395.
65. Grant, J., Gallardo, M., & Pickup, B. (1998). A fast method of molecular shape comparison. A simple application of a Gaussian description of molecular shape. pp. 1653–1666.
66. Cai, W., Xu, J., Shao, X., Leroux, V., Beautrait, A., & Maigret, B. (2008). SHEF: A vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces. *Journal of Molecular Modeling, 14*, 393–401.
67. McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A., & Brown, F. K. (2003). Gaussian docking functions. *Biopolymers, 68*, 76–90.
68. Daras, D., Zarpalas, D., Tzovaras, D., & Strintzis, M. G. (2005). 3D shape-based techniques for protein classification. In: *International conference on image processing* (pp. 1130–1133).
69. Sael, L., & Kihara, D. (2008). Protein local surface shape comparison. Unpublished data.
70. Lawrence, M. C., & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *Journal of Molecular Biology, 234*, 946–950.
71. Mitchell, J. C., Kerr, R., & Ten Eyck, L. F. (2001). Rapid atomic density methods for molecular shape characterization. *Journal of Molecular Graphics & Modelling, 19*, 325–390.
72. Wolfson, H., & Rigoutsos, I. (1997). Geometric hashing: An overview. *Computing in Science and Engineering*, 4(4), 10–21.
73. Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., et al. (2005). Protein–protein docking benchmark 2.0: An update. *Proteins, 60*, 214–216.
74. Chen, R., & Weng, Z. (2003). A novel shape complementarity scoring function for protein–protein docking. *Proteins, 51*, 397–408.