Molecular Surface Representation Using 3D Zernike descriptors for Protein Shape Comparison and Docking

Daisuke Kihara[1,2,3*], Lee Sael[2], Rayan Chikhi[4], & Juan Esquivel-Rodriguez[2]

[1]Department of Biological Sciences, [2]Department of Computer Science, [3]Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN, 47907, USA

[4] Normale Supérieure de Cachan, Computer Science Department, 61 Avenue du President Wilson, 94235 Cachan cedex, Britanny, France

* Corresponding Author

E-mail: dkihara@purdue.edu

Tel: (765) 496-2284

Fax: (765) 496-1189

Abstract

The tertiary structures of proteins have been solved in an increasing pace in recent years. To capitalize the enormous efforts paid for accumulating the structure data, efficient and effective computational methods need to be developed for comparing, searching, and investigating interactions of protein structures. We introduce the 3D Zernike descriptor (3DZD), an emerging technique to describe molecular surfaces. The 3DZD is a series expansion of mathematical three-dimensional function, and thus a tertiary structure is represented compactly by a vector of coefficients of terms in the series. A strong advantage of the 3DZD is that it is invariant to rotation of target object to be represented. These two characteristics of the 3DZD allow rapid comparison of surface shapes, which is sufficient for real-time structure database screening. In this article, we review various applications of the 3DZD, which have been recently proposed.

Introduction

The tertiary structure of proteins provides important clue for understanding function, evolution, and interaction of the proteins and also serves as the basis for protein design. Such use of structure information is enabled by computational methods which can effectively identify similarity and dissimilarity in protein global and local structures. As the number of solved protein structures keeps increasing in a rapid pace (http://www.rcsb.org), the speed of a computational method is becoming an important factor so that the method is able to scan the entire database in a reasonable time, if not in a real-time. Conventionally, protein structures are compared in terms of their main-chain conformation [1;2], inter-residue distances [3], spatial arrangement of the secondary structures [4], and the atomic detailed structures [5]. Several review articles [6-8] provide more information about existing protein structure comparison methods to readers. Each protein representation has its own strength and has a suitable range of structure similarity and the evolutional distance, which it can capture most meaningfully.

Here, we review application of protein surface representation using moment-based descriptors, namely, the 3D Zernike descriptors (3DZD) [9;10]. The 3DZD is a mathematical series expansion of a 3D function. It has several advantages which can broaden the applicability of computational analysis of protein structures. Note that computational works handling protein surface representation are not new; related works can be found from 1970's [11;12]. It is the moment-based descriptors that are introduced to the protein bioinformatics field recently. Thus, the purpose of this article is to demonstrate the attractive features of the 3DZD, which are not equipped with most of the other surface representations, through several examples of its applications.

Before moving to the description of the 3DZD and its recent applications, we quickly mention methods for constructing and representing protein surfaces. To start with, one should define the protein surface from the PDB file [13] of the protein, which provides the Cartesian coordinates of atoms. The Connolly surface is one of the common ways to represent molecular surface [14], which rolls a probe sphere on protein surface atoms and trace the center of the probes to construct a surface. An alternative method is to overlap the 3D Gaussian function at each atom of proteins [15]. The geometry of the defined surface can be represented, for example, by a graph where each node is characterized with geometric features of that point, such as the normal vector and surface curvature [16]. The graph representation allows employing existing graph matching algorithms to make comparison between surface shapes. Other methods include the spin image method [17], where a surface point are characterized by a 2D histogram of distances to the other surface points. Alternatively, one can use a voxel representation [18], where a protein surface or the entire volume is place on a 3D grid and occupied voxels (grid points) by the protein are marked (typically with an integer of 1 and 0 otherwise). For more information, refer to some review articles [19;20].

Compared to the existing surface representations, the 3DZD has following advantageous features [21]: Since it is a series of coefficients assigned to terms in the series expansion of the 3D function (here the 3D structure of a protein surface is considered as the 3D function), the protein surface is represented very compactly as a vector of numbers. Moreover, the mathematical derivation of the 3DZD makes it rotationally invariant, that is, the orientation of a protein does not affect to its 3DZD. Therefore, time-consuming alignment of proteins is not required for comparing them and thus the 3DZD of all known proteins can be precomputed and stored. As we will discuss below, a real-time search against the entire PDB database, which has

over 100,000 chains, is made possible taking advantage of these two features [22;23]. The 3DZD can also naturally represent surface physicochemical properties, such as the electrostatic potential and the hydrophobicity, by assigning such values to the protein surface regions [24]. Lastly, by changing the order of the series expansion, the resolution of the surface representation can be easily controlled.

The next section explains the mathematical derivation of the 3DZD. The applications of the 3DZD, namely, protein global structure comparison, surface property comparison, local surface classification, binding ligand prediction by pocket shape comparison, and protein-protein docking prediction, are discussed in the subsequent sections.

3D Zernike descriptors

The 3DZD is a mathematical series expansion of 3D function, which project a 3D object to a compact representation. The mathematical foundation of the 3D Zernike moments is laid out by Canterakis [9]. Then, Novotni and Klein applied it in the form of 3D Zernike descriptor for 3D object retrieval [10]. For readers' convenience, a brief mathematical derivation of is provided. For detailed derivations and discussions, refer to the two papers [9;10] .

The first step in computing the 3DZD is deriving 3D Zernike moments. The 3D Zernike moments are series expansion of an input 3D function, $f(\mathbf{x})$, where $\mathbf{x} = (x; y; z)$, into 3D Zernike polynomial. In the case of representing a protein surface shape, $f(\mathbf{x})$, should be the description of the surface shape, which can be computed by placing the protein structure onto a 3D grid and marking voxel points with 1 where the protein surface intersects and with 0 otherwise. For representing physicochemical properties on protein surface, *e.g.* the surface electrostatic potential or the hydrophobicity, we map the value of the property instead of 1 or 0 [24].

The 3D Zernike polynomials defined on order *n,* degree *l*, and repetition *m*, are given by

$$Z_{nl}^{m}(r,\vartheta,\varphi) = R_{nl}(r)Y_{l}^{m}(\vartheta,\varphi) \quad , \qquad (1)$$

subjected to $-l < m < l$, $0 \le l \le n$, and $(n-l)$ being even. Spherical harmonics, $Y_l^m(\vartheta,\varphi)$, are functions of a set of a polar angle, ϑ , and an azimuthal angle, φ . The radial function defined by Canterakis, $R_{nl}(r)$, directly incorporates radius information, *r*, into the basis function and are constructed so that $Z_{nl}^{m}(r,\vartheta,\varphi)$ are polynomial, $Z_{nl}^{m}(\mathbf{x})$, when transformed to the Cartesian coordinates system. The 3D Zernike moments of $f(\mathbf{x})$ are defined as the coefficients of the expansion in this orthonormal basis, *i.e.* by the formula:

$$\Omega_{nl}^{m} = \frac{3}{4\pi}\int_{|\mathbf{x}|\le 1} f(\mathbf{x})\overline{Z}_{nl}^{m}(\mathbf{x})d\mathbf{x} \quad . \qquad (2)$$

After generating the 3D Zernike moments, taking the norm of 3D Zernike moments yields rotation invariance. That is, the moments are collected into (2*l*+1) dimensional vectors $\Omega_{nl} = (\Omega_{nl}^{l}, \Omega_{nl}^{l-1}, \Omega_{nl}^{l-2}, \Omega_{nl}^{l-3}, ..., \Omega_{nl}^{-l})^{T}$ and the rotational invariant 3DZD, F_{nl} , are given as the norm of vectors Ω_{nl} :

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^{m})^{2}} \quad . \qquad (3)$$

The size of the 3DZD is determined by the parameter *n* called the order, which determines the resolution of the descriptor. The 3D Zernike descriptor is a series of invariants (Eqn. 3) for each combination of *n* and *l*, where *n* ranges from 0 to the specified order. The 3DZD can be further normalized by dividing each value in the descriptor with the norm of the descriptor. This

normalization is found to reduce dependency of 3DZD to the number of voxels used to represent a protein [23].

Protein global shape comparison

The first application of the 3DZD we introduce is protein global structure comparison [23;25]. Conventional structure comparison methods, *e.g.* ones which compare protein main-chain orientation, are essentially designed to compare a pair of protein structures and thus it takes a significant time to perform a whole PDB database scan to find similar structures for a query structure. Indeed, the current PDB website (http://www.rcsb.org) only provides keyword searches and sequence-based homology searches (BLAST [26]) but does not have functionality to search proteins with structural similarity. This is very different from public sequence databases where sequence similarity searches can be performed in a real-time. The 3DZD is very suitable for fast structure database search due to the two advantages: First, because it is rotation invariant, the 3DZD for proteins in the database can be precomputed. Thus, time consuming structure alignment is not needed for comparing structures. Second, the 3DZD represents a structure as a vector of coefficients of the series expansion (concretely, a vector of 121 float numbers in our work [23]), hence comparison of structures can be done very fast, just by comparing two vectors.

We have performed a thorough study of the protein structure comparison with the surface representation using the 3DZD [23]. To investigate how well the surface representation using the 3DZD agrees with existing methods, we used the protein structure classification dataset computed by the Combinatorial Extension algorithm (CE) [2], which compares the main-chain orientation of proteins. The dataset consists of 2432 proteins which are pre-classified by CE. As

described above in the subsection of the 3DZD, the 3DZD was computed for the Connolly surface of each protein structure in the CE dataset (Fig. 1).

Comparison of two protein structures was achieved by computing the Euclidian distance between their 3DZDs. We have also used the correlation coefficient and the Manhattan distance, which gave similar results as the Euclidean distance. The results are summarized in Table 1. Overall, the 3DZD agrees well with the structure classification by CE, retrieving a protein of the same main-chain conformation (defined by CE) within the top 5 closest structure in the 89.6% of the cases. Interestingly, the agreement between the 3DZD and the CE is much higher than between CE and DALI [3] (we used the standalone DaliLite program [27]), which is another popular protein structure comparison program that compares the distance map of proteins (Table 1). The 3DZD is over 1000 times faster than DaliLite. In our original paper [23], we have also compared with four existing 3D shape comparison methods developed in the computer graphics and engineering domain for 3D object comparison, namely, the spherical harmonics descriptor [28], the shape distribution histogram [29], the solid angle histogram [30], and the Eigen-value model [30], all of which turned out to perform significantly worse than the 3DZD (Figure 4 in the paper [23]).

The agreement with the conventional main-chain comparison (CE) can be further improved by computing the 3DZD of the surface of main-chain atoms (*i.e.* the side-chain atoms are removed before computing the Connolly surface and the 3DZD) [31] (Fig. 2). As shown in Fig. 2, the 3DZDs do not differ much between the all surface representation and the main-chain surface representation. However, in terms of the database retrieval against the CE dataset, the Area Under the Curve (AUC) value of the precision-recall graph improved from 0.481 to 0.604

(Fig. 3). Random retrieval yields the AUC value of 0.017. Note that this is not a Receiver Operator Characteristic (ROC) curve, where a random retrieval has an AUC value of 0.5.

We have developed a web server, named 3D-SURFER, where users can perform real-time protein structure search against the entire PDB database [22] (http://kiharalab.org/3d-surfer). A query against the entire PDB, containing over 130,000 single chains, takes, on an average, only a couple of seconds. The web interface displays the CATH code [32] of retrieved structures and structural alignments using the CE program. In addition, geometrically interesting local features of the protein surface, such as pockets that often correspond to ligand binding sites as well as protrusions and flat regions, can also be identified using the VisGrid algorithm [18].

Low-resolution structure data comparison

The surface representation of the 3DZD can be also naturally applied for comparing low-resolution structure data from the electron microscopy (EM) [31]. In recent years, an increasing number of protein complexes have been investigated by cryo-electron microscopy. Since the EM can usually provide only low-resolution structures, it is an important task to computationally identify known structures which can fit to the EM density map [33;34]. For the representative set of 2327 proteins taken from the CE dataset, we have computed EM density map with the pdb2mrc program in the EMAN package, which simulates the EM density of protein structures [35]. Then, the isosurface of the EM density is represented by the 3DZD. As we performed for the protein structure database search, given the 3DZD of a query EM isosurface of a protein, we ranked entries in the dataset by the Euclidean distance to the query. The three examples of the searches are shown in Table 2 and Figure 4. It is shown in Figure 4 that similar EM maps are retrieved for all the three queries examined. Among the five closest EM maps retrieved, the first

two of them are indeed from the proteins of the same structure group (Table 2). The overall AUC value of the precision-recall graph was 0.489, which is similar performance to the protein structure retrieval by the 3DZD (Fig. 3).

Representing protein flexibility, surface physicochemical properties

In the application for the global surface shape comparison, binary values (1 or 0) are assigned to voxels to represent static shape of protein surfaces. Instead of binary values, decimals can be used to represent positions of atoms probabilistically to describe flexibility or uncertainty of atom positions at a certain degree [36].

In the same way, physicochemical property values, such as the electrostatic potential values or hydrophobicity values, can be mapped on the surface voxels, which can be then represented with the 3DZD [24]. Using the 3DZD, a quantitative comparison of the physicochemical properties is possible. In our work, we showed that a similarity of the surface electrostatic potential patterns of thermophilic and mesophilic proteins can be quantified and classified.

Binding ligand prediction by comparing pocket properties

The previous sections have discussed the use of the 3DZD for representing and comparing the global shape and physicochemical properties of proteins. In this section, we show an application of the 3DZD for comparing local surface shape and properties of proteins, namely, protein ligand binding pocket sites [37]. Binding ligand molecules is an important aspect of protein function and hence several methods have been developed for either detecting geometrical pockets as potential ligand binding sites of proteins [18;38;39] or for predicting the ligand molecule which

binds to a specific pocket region of a query protein [40]. This work is focusing for the latter problem. Concretely, given a pocket region in a query protein, we predict which ligand molecule binds to the pocket region by comparing the pocket with a reference set of known pocket shapes.

A binding pocket surface is defined by the Connolly surface of protein heavy atoms which locate within a certain distance to any heavy atom of the bound ligand (Fig. 5A). In addition to the 3DZD (Fig. 5B), we have also used the Pseudo Zernike moments (PZM) to represent binding pockets and compared its performance to the 3DZD. In contrast to the 3DZD, which handles a pocket as a 3D function in the space, the PZM represents a binding pocket as a spherical panoramic 2D picture from its center of gravity, to which PZM is applied. The PZM [41] are used in many pattern recognition applications to describe the shape of a 2D image. Formally, the PZM are projections of a function on a set of complete and orthogonal basis polynomials defined over the unit circle ($x^2+y^2 \leq 1$):

$$V_{n,m}(x,y) = e^{im\theta} R_{nm}(r) = e^{im\theta} \sum_{s=0}^{n-|m|} \frac{(-1)^s (2n+1-s)! \rho^{(n-s)}}{s!(n+|m|+1-s)!(n-|m|-s)!} \quad (4)$$

where $\rho = \sqrt{x^2+y^2}$, $\theta = \tan^{-1}(y/x)$, and $n \geq 0$, $|m| \leq n$. The PZM of the order n and the repetition m for a 2D image $f(x, y)$ are defined as:

$$A_{n,m} = \frac{n+1}{\pi} \int_{x^2+y^2 \leq 1} f(x,y) V^*_{n,m}(x,y) dxdy \quad (5)$$

The asterisk (*) denotes the complex conjugate.

To obtain the 2D picture of a pocket, first, a ray-casting strategy is used to represent a pocket as seen from the center of gravity. Then, a three dimensional Cartesian coordinate system is set in a way that the $\vec{x}$ axis points toward the pocket opening. $(\vec{y}, \vec{z})$ can be determined

arbitrarily as long as $(\vec{x}, \vec{y}, \vec{z})$ is orthogonal, since the PZM is rotationally invariant. Using spherical coordinates, *f(θ, ϕ)* is defined as the Euclidean distance from the center of gravity to the outermost surface of the pocket, and 0 if not intersection occurs. Figure 5C shows an example of the resulting 2D picture of a pocket, where *(x, y) = (θ, ϕ)* (Fig. 5C). Finally, the PZM is computed for the 2D picture (Fig. 5D). The surface electrostatic potential can be also represented in the same way by assigning the electrostatic potential value rather than the Euclidean distance to each direction of the 2D picture.

Since both 3DZD and PZM are vectors, the similarity of two pockets is inexpensively computed by computing the Euclidean distance between them. For predicting binding ligand for a query pocket, we used a k-nearest neighbor (k-NN) classifier with weighted voting [37]. After ranking pockets in the reference dataset by the pairwise distance to the query pocket, the k closest pockets are used to compute the score for each ligand type, *F* (ATP, NAD, heme, galactose, etc.):

$$Pocket\_score(F) = \sum_{i=1}^{k} \left(\delta_{l(i),F} \log\left(\frac{n}{i}\right) \right) \cdot \frac{\sum_{i=1}^{k} \delta_{l(i),F}}{\sum_{i=1}^{n} \delta_{l(i),F}}, \qquad (6)$$

where $l(i)$ is the ligand type of the i-th pocket, n is the total number of pockets in the reference dataset, and the indicator function $\delta_{X,Y}$ equals to 1 if X is of type Y, and is null otherwise. The first term in this scoring function assigns logarithmically decreasing weights to pockets which bind ligand *F*, as the rank *i* of the pocket goes lower. The second term balances the score by taking into account the number of pockets of type *F* in the reference dataset. The scoring function is computed for each ligand type, and the ligand with the highest *Pocket_score* is

predicted to bind to the query pocket. Using cross-validation on a test dataset consisting of 100 proteins [37], optimal parameters were estimated for the descriptors: the values w = 4.5 (resp. 0.04) and n = 4 (20) were chosen for the PZM (resp. 3DZD). The number of neighbor, k in Eqn. 6 was set to $k = 24$ for both descriptors.

In our paper [37], we examined the performance of our binding ligand prediction method, named Pocket-Surfer, on a couple of datasets including the one used previously by Kahraman *et al.* , which has 100 ligand binding pockets of nine different ligand molecules [40]. We examined the effect of different parameter values on the performance and also investigated the use of the surface electrostatic potential information together with the pocket shape information. Results were also shown for the cases when unbound pockets were used as queries. It was shown that the Pocket-surfer performs better than a similar method which employs the spherical harmonics [40].

Here, we show performance comparison of Pocket-surfer with four existing web servers, eF-Seek [42], SitesBase [43], PROSURFER[44], and XBSite2F [45]. To this end, we prepared a dataset of 118 proteins (pockets), which are commonly included among the datasets of all the five methods. This dataset includes 18 different types of ligands, such as adenosine (AND), adenosine-triphosphate (ATP), fructose 6-phosphate (F6P), flavin mononucleotide (FMN), flavin-adenine dinucleotide (FAD), and guanine (GUN). Each of the pockets in the dataset was searched against the rest, and evaluated if pockets of the same type are retrieved. The performance is evaluated with the area under curve (AUC) value defined from the receiver operating characteristic (ROC) curve and the Top-3 accuracy. The Top-3 accuracy is defined as the number of pockets for which the correct ligand is found within the three highest scoring ligands divided by the total number of pockets in the common dataset (*i.e.* 118). Results are summarized in Table 3. On this dataset, the 3DZD showed the best performance with an AUC

value of 0.86 and the top-3 accuracy of 75.0%. The PZM of the Pocket-Surfer method came the second rank in both metrics.

The PZM and the 3DZD run very fast among the methods compared. Searching against the 62,200 pockets, the size of the current PDB database, would only takes 0.7 and 6.85 seconds respectively, on a Linux machine with Pentium 4 3.0GHz CPU [37]. We also listed the execution times for the webservers in Table 3. But note that the various conditions are different among the servers, including the size of the database they have, most probably the specification of the computers, and the speed of internet connection. Particularly, it appears that the PROSURFER and SiteBase servers provide precomputed results for existing PDB entries and do not accept user uploaded structures. Therefore, the time for these two servers is probably not actual computational time for scanning the database.

Classification of local protein surfaces

In a subsequent work [46], we have further applied the local protein surface representation by the 3DZD for characterizing and classifying protein local surfaces. Unlike the previous section which only considers ligand binding pockets, local surfaces were taken from entire surface of proteins. The local surfaces are defined as the surface region within 6Å from the center of the local surface region. In total of 118,003 local surface patches were obtained from 609 representative proteins. A patch was characterized by two features, the geometric shape and the electrostatic potential, both of which were described by the 3DZD. We used the emergent self-organizing map (ESOM) [47;48], a variant of self-organizing map, for classifying the local surface patches. After obtaining initial groups of local surface patches by the ESOM, we have further clustered the patch groups, which finally resulted in 48 clusters when the surface shape is

considered and 27 clusters for the combination of the surface shape and the electrostatic potential. It was shown that surface patches of the same type were found consistently at equivalent positions at ligand binding sites of the same ligand in different proteins. Figure 6 illustrates such an example of equivalent patches which are found at heme binding sites of two different proteins.

The resulting clusters have several interesting applications. The clusters can be used as surface "alphabet", with which protein surface can be labeled and classified. Thus, a surface region, for example, a protein-docking interface or a DNA binding site, can be described as a set of surface alphabets. This description of protein surface with a set of letters would enable a variety of protein surface analyses, such as classification, function prediction, and database searches, in analogous ways to protein sequence analyses.

Protein-protein docking prediction

For the last section of this article, we discuss application of the 3DZD to protein-protein docking prediction [49]. Applications in the previous sections use the 3DZD for capturing *similarity* of global or local surface properties of proteins. In contrast, in the application for protein docking, we capture shape *complementarity* of docking interface of proteins. Since we use the3DZD to describe only surface regions of proteins treating the inner region of proteins empty, perfectly fitting interfaces of two docking proteins have identical 3DZDs (*i.e.* the 3DZDs show the Euclidean distance of 0 and the correlation coefficient of 1.0). In actual cases of protein complexes, docking interfaces of two proteins may not be perfectly complementary to each other, especially when interfaces of unbound structures are evaluated. However, our results on fifteen bound and unbound protein pairs showed that their docking interfaces (defined as a set of surface

grid points within 4.5Å to any atoms of the other protein) showed sufficient shape complementarity with 3DZD correlation coefficient values of over 0.9 in almost all the cases.

Using the 3DZD for capturing shape complementarity at docking interface, we developed a docking algorithm, LZerD (Local 3D Zernike descriptor-based Docking program) [49]. In short, LZerD uses the 3DZD and some other parameters as regional features of protein surface shape for scoring docking decoys, which are generated by geometric hashing algorithm [50]. In the following, we will present a brief description of the overall LZerD algorithm.

The first step in LZerD is to extract evenly distributed points on the protein surface at a minimum separation of 1.8Å. The geometry of each point is characterized by two features, namely, a normal vector for representing the direction of the local region and the 3DZD which is computed for a local spherical patch of 6Å radius centered at the point. Obviously, the 3DZD is used to capture the local surface shape of each surface point.

Once the shape of local regions is described, LZerD is aimed to find local regions of the two proteins (receptor and ligand proteins) that are complementary to each other and compute the pose of the two proteins so that the two regions fit. For searching poses of proteins, we apply a geometric hashing algorithm. The algorithm is divided in two stages: hashing and recognition. In the hashing phase, surface points from the ligand protein are stored in a hash table after transformed under each orthonormal coordinate frame defined using two surface point. An orthonormal coordinate frame is defined by taking a pair of surface points from the ligand protein plus a vector obtained by averaging the normal vectors of the two points. More formally, given points *a* and *b* and their corresponding normal vectors, we calculate $\vec{d} = \frac{\vec{a}_n + \vec{b}_n}{2}$ and create a Cartesian frame where point *a* is taken as the origin, $\vec{U} = \vec{AB}$ as the x-axis, $\vec{V} = \vec{AB} \times \vec{d}$ as the y-axis, and the z-axis is the cross-product of the other two axes $\vec{N} = \vec{U} \times \vec{V}$. Such reference

frames will be created for every pair of points on the ligand surface. Then, all points within 15Å of either point will be transformed to that coordinate system and stored in the hash table. Figure 7 outlines the general process of selection two points *a* and *b*, selecting all neighbor points n_i and the corresponding transformation based on the reference frame. For the docking problem, a standard hash table implementation was found to be inefficient because of the non-uniform distribution of the data in hash-space, which leads to a longer search time. Therefore, LZerD uses kd-tree, a data structure that partitions point sets recursively for efficient search. This is a specialization of a binary tree that partitions the k-dimensional space using hyper-planes. For example, for k=3 (as is the case in LZerD), the first partition corresponds to all the elements that are lower or higher than the first dimension of a node (x coordinate value), and the second partition is determined by the values for the second dimension of the node values, and so on.

Next, the recognition stage uses the previously created kd-tree to compare the shape of regions in the ligand against regions taken from the receptor proteins. Thus, the same process will be performed for every pair of points on the receptor to define reference frame, along with transformations of points. Transformed points on the receptor for each reference frame will be queried against the ligand kd-tree in order to retrieve geometrically similar points. If sufficient number of points is considered to be similar for a reference frame in terms of the local 3DZDs, normals, and point distances, then the ligand and the receptor proteins are transformed to compute the docking pose. Poses which produce too many clashes between the ligand and receptor proteins are eliminated. The docking poses are finally evaluated and ranked by a scoring function which combines terms which consider the local shape complementarity (matches of 3DZDs), directions of normals, the size of formed docking interface, and atom clashes (penalty).

Figure 8 shows the prediction results of LZerD on the 84 unbound protein complexes taken from the ZDOCK benchmark set [51]. A docking prediction for a protein complex is considered to be successful if at least one of the decoys (predicted conformations) within the top *n* ranks has an interface RMSD (the root mean square deviation computed for residues at docking interface) of 2.5Å or less. To test the performance of the scoring function of LZerD, the scoring function was applied to rerank decoys computed with the ZDOCK docking program (termed ZDOCK Reranked). Results by three existing methods, ZDOCK, PatchDock [52], and Context Shape (CS) [53] were also shown for comparison (Fig. 8). ZDOCK and LZerD clearly outperformed CS and PatchDock on these unbound complexes. LZerD and LZerD Reranked ZDOCK decoys essentially showed similar performance to ZDOCK. When compared the rank of the first correct hit for individual target, LZerD was slightly better than ZDOCK showing 33 better cases than ZDOCK prediction as opposed to 24 opposite cases. The ZDOCK Reranked showed identical performance with original ZDOCK with both having 26 better cases against each other.

Additionally, a comparison of three sample executions of these docking programs is shown in Table 4. Unlike the global and local protein shape comparison by the 3DZD, the speed of LZerD is not particularly fast because the 3DZD is used to evaluate shape complementarity of docking decoys which are generated by geometic hashing. For two cases (1D6R and 2SNI), LZerD runs faster or in a comparable time with ZDOCK. In the third case (2PCC), LZerD executes considerably slower than ZDOCK. The execution time for the geometric hashing employed in LZerD depends on the number of critical points on the protein surface. ContextShapes and PatchDock are faster due to two main reasons: the number of points and

transformations analyzed are reduced before matching shapes by using clustering and hot spot analysis, and also they employ a fast method to calculate clashes.

Conclusion

In this article, we overviewed applications of the protein surface representation using the 3DZD. Due to the rotational invariance and its compact representation, the 3DZD allows rapid real-time comparison for global and local protein surfaces, low-resolution structure data from electron microscopy or electron tomography. It also can identify complementarity of molecular surfaces, which was applied for protein docking prediction. Development of computational methods for analyzing 3D structures have been more complicated than sequence analyses tools. We believe the 3DZD will be able to play an important role for lowering the barrier for computational analyses of tertiary structures of biomolecules.

Acknowledgements

This work is supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM075004). DK also acknowledges grants from NSF (DMS0800568, EF0850009, IIS0915801).

Reference List

1. Kihara D, Skolnick J: **The PDB is a covering set of small protein structures**. *J Mol Biol* 2003, **334:**793-802.

2. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. *Protein Eng* 1998, **11:**739-47.

3. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices**. *J Mol Biol* 1993, **233:**123-38.

4. Mizuguchi K, Go N: **Comparison of spatial arrangements of secondary structural elements in proteins**. *Protein Eng* 1995, **8:**353-362.

5. Kabsch W: **A discussion of the solution for the best rotation to relate two sets of vectors**. *Acta Cryst* 1978, **A34:**827-828.

6. Mizuguchi K, Go N: **Seeking significance in three-dimensional protein structure comparisons**. *Curr Opin Struct Biol* 1995, **5:**377-382.

7. Kolodny R, Petrey D, Honig B: **Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction**. *Curr Opin Struct Biol* 2006, **16:**393-398.

8. Carugo O: **Recent progress in measuring structural similarity between proteins**. *Curr Protein Pept Sci* 2007, **8:**219-241.

9. Canterakis N: **3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition**. *Proc.11th Scandinavian Conference on Image Analysis* 1999,85-93.

10. Novotni M, Klein R: **3D Zernike descriptors for content based shape retrieval**. *ACM Symposium on Solid and Physical Modeling, Proceedings of the eighth ACM symposium on Solid modeling and applications* 2003,216-225.

11. Shrake A, Rupley JA: **Environment and exposure to solvent of protein atoms. Lysozyme and insulin**. *J Mol Biol* 1973, **79:**351-371.

12. Lee B, Richards FM: **The interpretation of protein structures: estimation of static accessibility**. *J Mol Biol* 1971, **55:**379-400.

13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28:**235-242.

14. Connolly ML: **Solvent-accessible surfaces of proteins and nucleic acids**. *Science* 1983, **221:**709-713.

15. Mitchell JC, Kerr R, Ten Eyck LF: **Rapid atomic density methods for molecular shape characterization**. *J Mol Graph Model* 2001, **19:**325-390.

16. Kinoshita K, Nakamura H: **Identification of the ligand binding sites on the molecular surface of proteins**. *Protein Sci* 2005, **14:**711-718.

17. **Identifying similar surface patches on proteins using a spin-image surface representation**. *Lecture Notes in Comp Sci* 2005, **3537:**417-428.

18. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D: **Characterization of local geometry of protein surfaces with the visibility criterion**. *Proteins* 2007, **71:**670-683.

19. Via A, Ferre F, Brannetti B, Helmer-Citterich M: **Protein surface similarities: a survey of methods to describe and compare protein surfaces**. *Cell Mol Life Sci* 2000, **57:**1970-1977.

20. Sael L, Kihara D: **Protein surface representation and comparison: New approaches in structural proteomics**. In *Biological Data Mining*. Edited by Chen J, Lonardi S. Boca Raton, Florida, USA: Chapman & Hall/CRC Press; 2009:89-109.

21. Venkatraman V, Sael L, Kihara D: **Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors**. *Cell Biochem Biophys* 2009, **54:**23-32.

22. La D, Esquivel-Rodriguez J, Venkatraman V, Li B, Sael L, Ueng S, Ahrendt S, Kihara D: **3D-SURFER: software for high-throughput protein surface comparison and analysis**. *Bioinformatics* 2009, **25:**2843-2844.

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

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

25. Mak L, Grandison S, Morris RJ: **An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison**. *J Mol Graph Model* 2007, **26:**1035-1045.

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215:**403-410.

27. Holm L, Park J: **DaliLite workbench for protein structure comparison**. *Bioinformatics* 2000, **16:**566-567.

28. Kazhdan M, Funkhouser T, Rusinkiewicz S: **Rotation invariant spherical harmonic representation of 3D shape descriptors**. *Proc.of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing* 2003, **43:**156-164.

29. Jiantao P, Ramani K: **A 3D Model Retrieval Method Using 2D Freehand Sketches**. In *Computational Science -- ICCS 2005*. Berlin/Heidelberg: Springer; 2005:343-346.

30. Kriegel H-P, Kroger P, Mashael Z, Pfeifle M, :Potke M, Seidl S: **Effective similarity search on voxelized CAD objects.** *Proc.of 8th international conference on database systems for advanced applications*. 2003,27-36.

31. Sael L, Kihara D: **Improved real-time structure search with application to low-resolution data**. *Submitted.* 2010.

32. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA: **The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies**. *Nucleic Acids Res* 2009, **37:**D310-D314.

33. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A: **Protein structure fitting and refinement guided by cryo-EM density**. *Structure* 2008, **16:**295-307.

34. Ceulemans H, Russell RB: **Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization**. *J Mol Biol* 2004, **338:**783-793.

35. Ludtke SJ, Baldwin PR, Chiu W: **EMAN: semiautomated software for high-resolution single-particle reconstructions**. *J Struct Biol* 1999, **128:**82-97.

36. Grandison S, Roberts C, Morris RJ: **The application of 3D Zernike moments for the description of "model-free" molecular structure, functional motion, and structural reliability**. *J Comput Biol* 2009, **16:**487-500.

37. Chikhi R, Sael L, Kihara D: **Real-time ligand binding pocket database search using local surface descriptors**. *Proteins* 2010, **78:**2007-2028.

38. Liang J, Edelsbrunner H, Woodward C: **Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design**. *Protein Sci* 1998, **7:**1884-1897.

39. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation**. *BMC Struct Biol* 2006, **6:**19.

40. Kahraman A, Morris RJ, Laskowski RA, Thornton JM: **Shape variation in protein binding pockets and their ligands**. *J Mol Biol* 2007, **368:**283-301.

41. Bhatia AB, Wolf E: **On the Circle Polynomials of Zernike and Related Orthogonal Sets**. *Proceedings of the Cambridge Philosophical Society* 1954, **50:**40-48.

42. Kinoshita K, Murakami Y, Nakamura H: **eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape**. *Nucleic Acids Res* 2007, **35:**W398-W402.

43. Gold ND, Jackson RM: **Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships**. *J Mol Biol* 2006, **355:**1112-1124.

44. Minai R, Matsuo Y, Onuki H, Hirota H: **Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions**. *Proteins* 2008, **72:**367-381.

45. Xiong B, Wu J, Burk D, Xue M, Jiang H, Shen J: **BSSF: a fingerprint based ultrafast binding site similarity search and function analysis server**. *BMC Bioinformatics* 2010, **11:**47.

46. Sael L, Kihara D: **Characterization and classification of local protein surfaces using self-organizing map**. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)* 2010, **1:**32-47.

47. Ultsch A: **ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM.** *Technical report.Dept.of Mathematics and Computer Sciece, Univeristy of Marburg.* 2005.

48. Ultsch A: **Maps for the visualization of high-dimensional data spaces.** *Proc.WSOM'03* 2003,225-230.

49. Venkatraman V, Yang YD, Sael L, Kihara D: **Protein-protein docking using region-based 3D Zernike descriptors**. *BMC Bioinformatics* 2009, **10:**407.

50. Wolfson H, Rigoutsos I: **Geometric hashing: an overview**. *IEEE Computational Science Engineering* 1997, **4:**10-21.

51. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z: **Protein-Protein Docking Benchmark 2.0: an update**. *Proteins* 2005, **60:**214-216.

52. Schneidman-Duhovny D, Inbar Y, Polak V, Shatsky M, Halperin I, Benyamini H, Barzilai A, Dror O, Haspel N, Nussinov R, Wolfson HJ: **Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking**. *Proteins* 2003, **52:**107-112.

53. Shentu Z, Al HM, Bystroff C, Zaki MJ: **Context shapes: Efficient complementary shape matching for protein-protein docking**. *Proteins* 2008, **70:**1056-1073.

Figure Legends

Figure 1. The steps to compute the 3DZD for protein global surface structure.

Figure 2. Example of the surface representation and the backbone representation of protein 9ldbA. The figure in the top left is the surface representation using all atom in the protein, AASurf, and figure in the top right is the backbone representation of the protein. The bottom graph shows the 3D Zernike descriptors for the two representations.

Figure 3. Precision-recall graphs of the all surface representation (AASurf) and main-chain representation (backbone) measured on the CE classification dataset. The Euclidean distance is used. The AUC values for the curves are written inside parentheses. Note that unlike the AUC value of the receiver operating characteristic (ROC), the random retrieval yields a much smaller value (in this case 0.017) in a precision-recall graph.

Figure 4. Retrieval results of simulated EM density maps. The EM density maps were simulated with EMAN2 package using a resolution of 15Å. The protein surface was generated by taking iso-surface regions of density value 10.

Figure 5. Examples of the binding pocket representation by the 3DZD and the PZM. **A**, a FAD binding site of glutathione reductase (PDB: 3grs). **B**, The 3DZD of the binding site. **C**, the 2D picture of the pocket. The color shows the distance from the center to the pocket surface. The darker, the more distance and pink shows the aperture of the pocket. **D**, the PZM of the 2D picture.

Figure 6. Examples of local surface patches. The figure illustrates the three pairs of patches locating at two heme binding sites of proteins (1dk0A, left; 1d7cA, right). The patches of the same color indicate that they are similar and belong to the same group. The numbers shows the similarity between the patch pairs, which is defined as 1 - correlation coefficient of the 3DZD of the two patches.

Figure 7. Illustration of the hashing stage. Points *a* and *b* are selected to create a new reference frame and neighbors within 15Å of the points are selected. 1), points *a* and *b* are expressed in the original coordinate system along with two neighboring points. A new coordinate system is created based on *a* and *b*, 2). Neighbors n_1 and n_2 are transformed to the new coordinate system in 3) and 4), respectively.

Figure 8. The performance of LZerD on 84 complexes in the ZDOCK Benchmark 2.0. ZDOCK reranked cases refer to results of reranking ZDOCK decoys using the LZerD's scoring function. Note that CS (Context Shapes) and Patchdock only report results for the top 3600 conformations in their corresponding papers.

Table 1. Summary of the global protein structure retrieval.

	Top1[a)]	Top5	Top10	Execution time[b)]
3DZD [c)]	1881 (77.3%)	2179 (89.6%)	2264 (93.1%)	1.46×10^{-4}(s)
DaliLite [d)]	307 (12.6%)	696 (28.6%)	897 (36.9%)	3.21 (s)
Random [e)]	117 (4.8%)	508 (20.9%)	806 (33.1%)	N/A

a) The number of query proteins which retrieved a correct member in the same group as the first position, within top 5 or top 10. In the parentheses, the percentage among all the 2432 proteins in the benchmark set is shown.

b) Average execution time for pair wise comparison (excluding time for preprocessing) in seconds. The evaluation was performed on Intel core2 CPU 6400 @ 2.13GHz processor with 5GB memory.

c) The Euclidean distance is used to compare the 3DZDs.

d) DaliLite (version 2.4.4) was used. The distance *d* is defined as *d = 100 – (the structure similarity Z-score by DaliLite)*.

e) A random value between 0 and 1 is assigned as the distance between the query to each protein.

Table 2. Examples of database search of protein EM maps.

Query	1a2aA			1bev1			1c3aA		
	Hits	Euc dist[a)]	Group [b)]	Hits	Euc dist	group	Hits	Euc dist	group
1st	1m8t-A	10.555	Y	1fpn-1	10.039	Y	1fvu-B	10.543	Y
2nd	1god-A	11.395	Y	1eah-1	10.042	Y	1iod-B	10.967	Y
3rd	1vhb-A	11.408	N	1bev-3	11.419	N	1ixx-E	11.280	Y
4th	1a3f-B	11.427	Y	1mqt-A	11.897	Y	1js9-A	12.915	N
5th	1fe5-A	11.538	Y	1wer	11.935	N	1coj-A	13.468	N

The PDB code of the top 5 hits are shown.

a) The Euclidean distance of the 3DZD between the query and the retrieved protein.

b) Y is shown for structures in the same structure group as the query. N, for otherwise.

Table 3. Performance of binding ligand prediction.

Method	Representation type	Retrieval method	ROC-AUC	Top-3 prediction accuracy (%)	Execution time [b]
Pseudo-Zernike (Pocket-surfer)	2D moments	k-NN	0.74	53.3	0.7 s
3D Zernike (Pocket-surfer)	3D moments	k-NN	0.86	75.0	6.85s
eF-Seek	Graph	Clique detection	0.49	25.0	~ 2 h
SitesBase	Geometric hashing	Geometric matching	0.60 [a]	49.3	~ 1 s [c]
PROSURFER	Fingerprinting	Pair-wise	0.57 [a]	39.6	~ 10 s [c]
XBSite2F	Fingerprinting	Pair-wise	0.55	32.8	~ 1 min

a) the AUC values for SitesBase and PROSURFER have a standard deviation of 0.02 due to incomplete ranking of dataset for some queries, see the benchmark methodology in our paper [37].

b) For the PZM and the 3DZD methods, an estimated search time for a database of 62,200 binding sites is given in the parentheses. For the other methods, we used the following webservers, since standalone programs are not available for them: eF-Seek, http://ef-site.hgc.jp/eF-seek/; SitesBase, http://www.modelling.leeds.ac.uk/sb/; PROSURFER, http://dsearch.dip.jp/top; XBSite2F, http://202.127.30.184:8080/bssf/search.jsp. For the servers, we give an order of magnitude of the execution time, as their database sizes are not identical but of the similar order to 62,200. The number of entries of the webservers is: eF-Seek, 17,500; SitesBase, 33168; PROSURFER, 48,347; Xbsite2F, 13,227.

c) It appears that the servers return pre-computed results to a limited number of queries. Thus probably this is not the actual computational time to scan the database.

Table 4. Computational time of docking prediction methods.

PDB[a)]	Length (receptor/ligand)[b)]	ContextShapes	PatchDock	ZDOCK	LZerD[c)]
1D6R	223/58	0:09:25	0:03:35	4:07:03	2:43:33 (418/999)
2SNI	275/83	0:11:08	0:06:48	3:08:39	3:33:53 (417/1057)
2PCC	341/341	0:13:48	0:16:04	4:04:16	9:14:19 (597/1291)

The execution time is shown in the hours:minutes:seconds format. All the computation is performed on a Linux computer with Intel Core i7 CPU 2.67GHz.

a) Three sample structures from the ZDOCK Benchmark 2.0.

b) Number of residues in each of the receptor and ligand structures.

c) The number of critical points for the receptor and ligand are reported in parentheses.

Figure 1

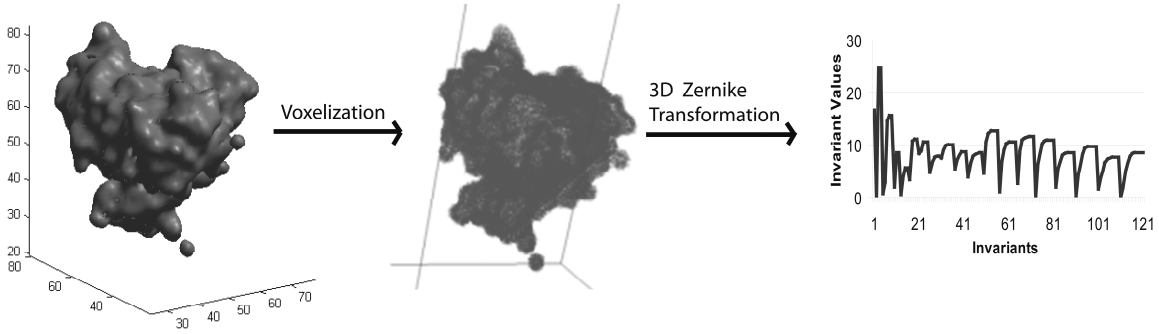

Figure 2

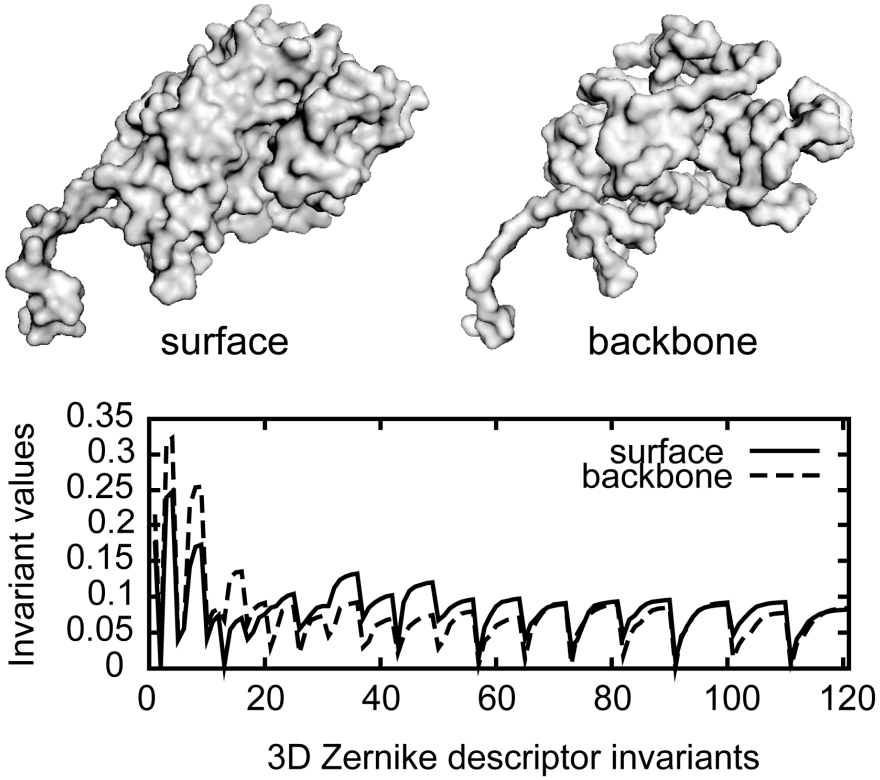

Figure 3

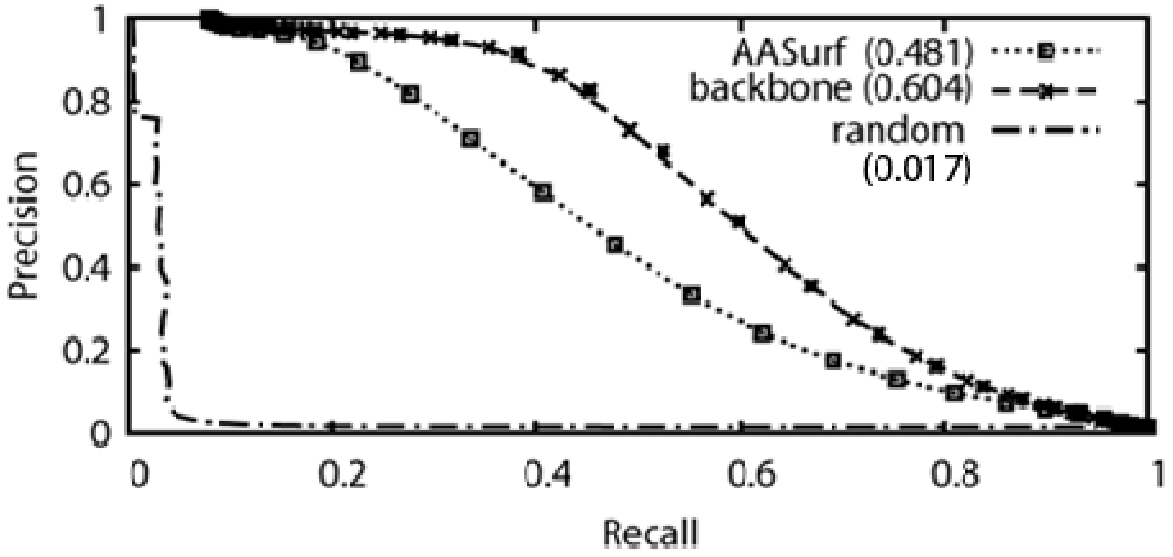

Figure 4

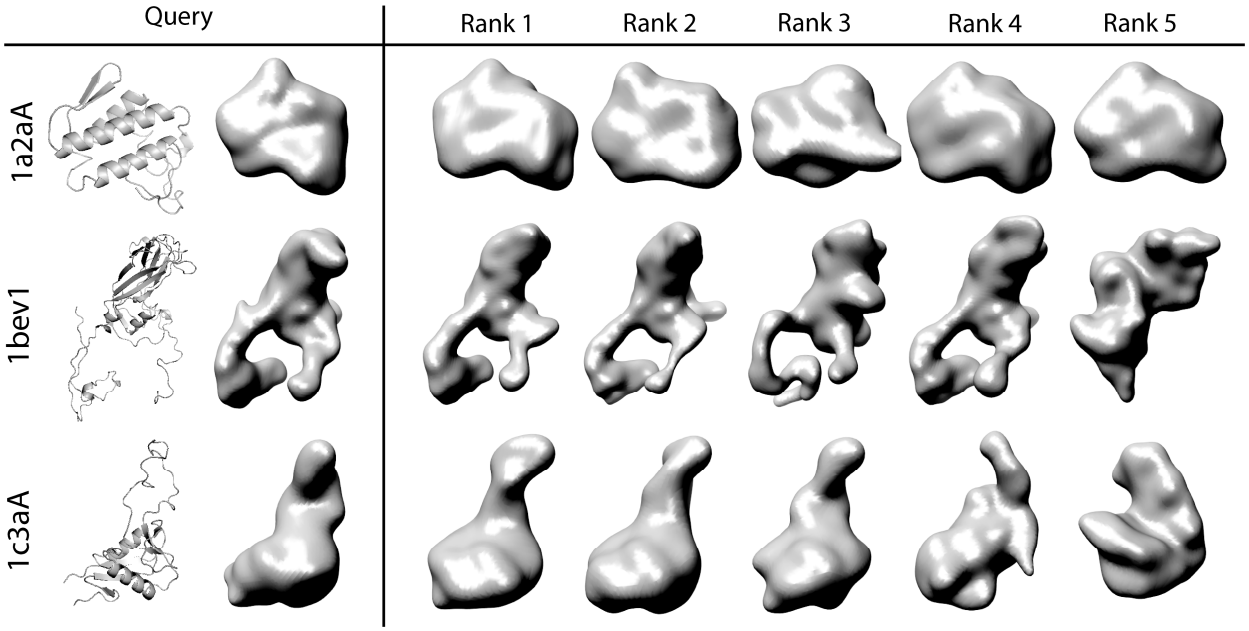

Figure 5

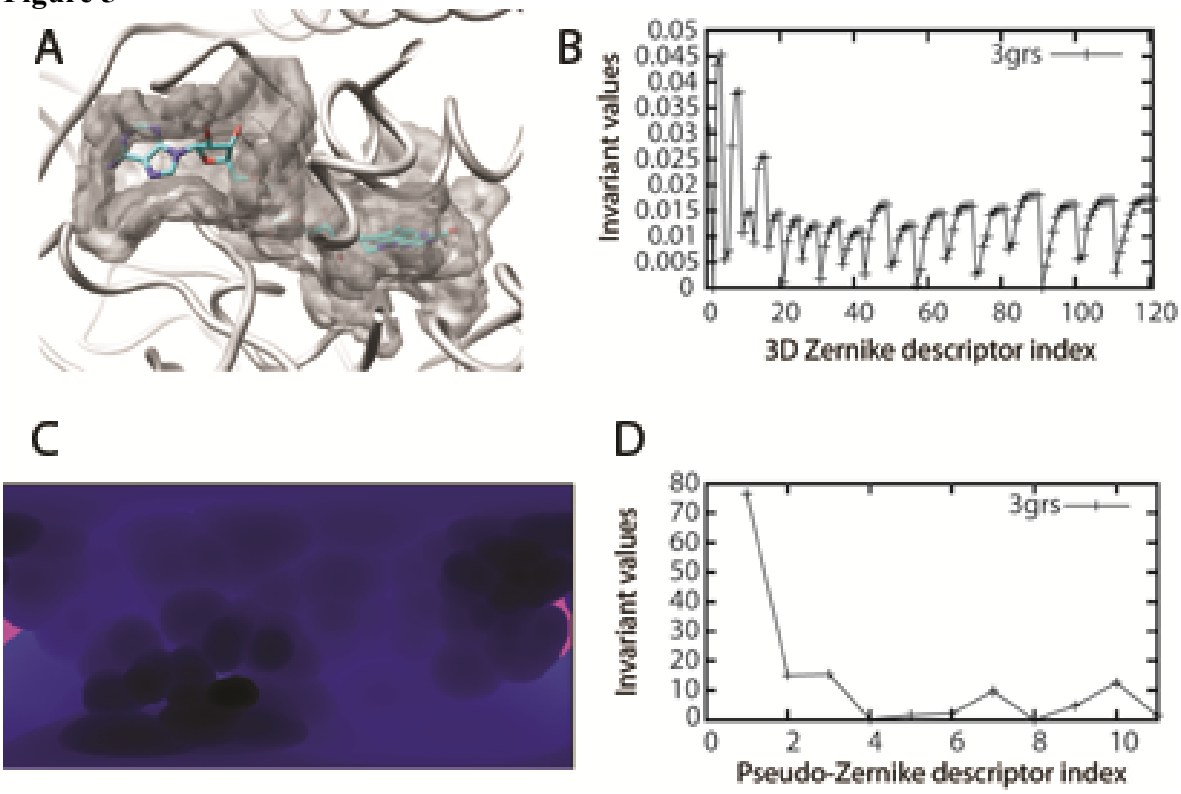

Figure 6

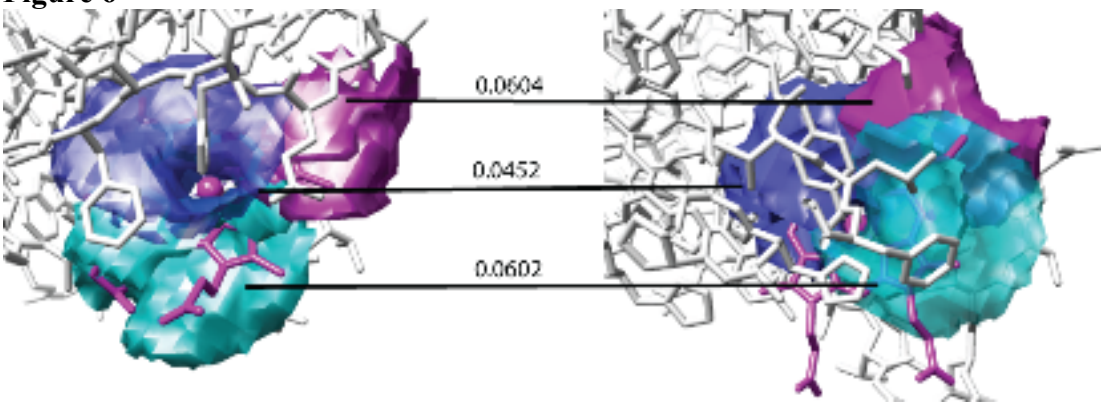

Figure 7

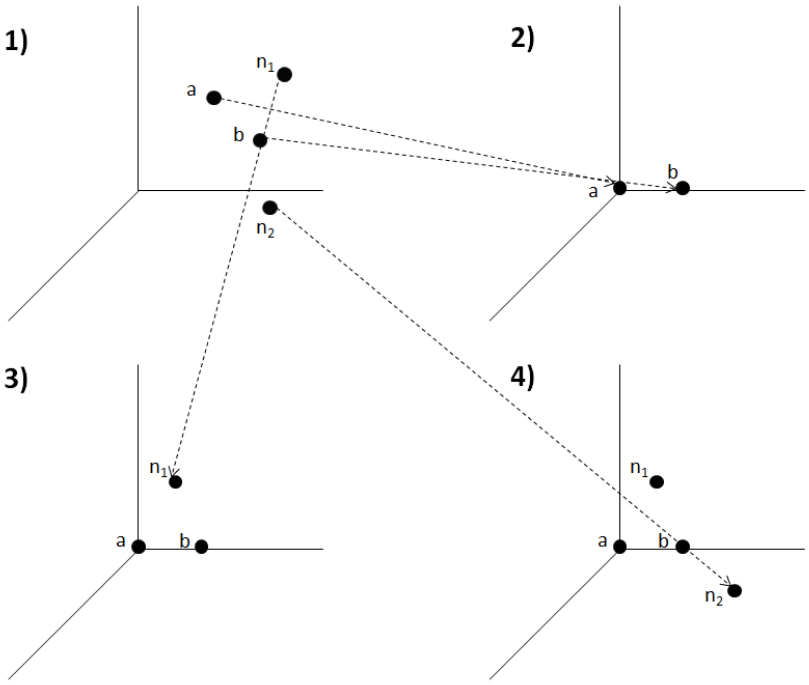

Figure 8