Proteins: Structure, Function and Bioinformatics, in press (Feb. 2008).

Fast Protein Tertiary Structure Retrieval Based on Global Surface Shape Similarity

Lee Sael[1#], Bin Li[1#], David La[2], Yi Fang[3], Karthik Ramani[3], Raif Rustamov[4] & Daisuke Kihara[1,2,5,6]*

[1] Department of Computer Science, College of Science
[2] Department of Biological Sciences, College of Science
[3]Department of Mechanical Engineering, College of Engineering
[4]Department of Mathematics, College of Science
[5] Markey Center for Structural Biology
[6] The Bindley Bioscience Center
Purdue University, West Lafayette, IN, 47907, USA

* Corresponding Author
E-mail: dkihara@purdue.edu
Tel: (765)496-2284
Fax: (765)496-1189

# These two authors have equal contribution to this work.

KEYWORDS
protein surface shape, protein structure classification, database search, structure similarity, 3D Zernike descriptor

Abstract

Characterization and identification of similar tertiary structure of proteins provides rich information for investigating function and evolution. The importance of structure similarity searches is increasing as structure databases continue to expand, partly due to the structural genomics projects. A crucial drawback of conventional protein structure comparison methods, which compare structures by their main-chain orientation or the spatial arrangement of secondary structure, is that a database search is too slow to be done in real-time. Here we introduce a global surface shape representation by 3D Zernike descriptors, which represent a protein structure compactly as a series expansion of three-dimensional functions. With this simplified representation, the search speed against a few thousand structures takes less than a minute. To investigate the agreement between surface representation defined by 3D Zernike descriptor and conventional main-chain based representation, a benchmark was performed against a protein classification generated by the CE algorithm. Despite of the different representation, 3D Zernike descriptor retrieved proteins of the same conformation defined by CE in 89.6% of the cases within the top five closest structures. The real-time protein structure search by 3D Zernike descriptor will open up new possibility of large-scale global and local protein surface shape comparison.

Introduction

The three dimensional (3D) structure, especially the surface, plays a central role in various function of proteins. For example, a group of atoms in an active site on the 3D surface of the protein which carries out the catalytic reaction of an enzyme[1]. Further, surface residues on an interface region establish physical contacts to another protein in protein-protein interactions[2,3]. Therefore, classification of the 3D structure of proteins using an appropriate representation is critical for understanding the universe of protein structure, function, and evolution[4].

The importance of characterization and comparison of protein 3D structure is further increasing recently in the context of protein function prediction[5,6,7], because a significantly increasing number of structures of unknown function have been solved in recent years by structural genomics projects[8,9,10]. Currently more than 2100 protein structures classified as “unknown function” have been deposited to the Protein Data Bank (PDB)[11], whose function are not easily assigned by conventional sequence database search methods[12,13]. To go beyond sequence data search methods, employing the 3D structure information is a reasonable and promising strategy because the evolutionary history could be better traced by using global 3D structures than from sequence alone[14,15] and similar local structure search methods could be used to identify catalytic residues involved in the same enzymatic function[16,17,18].

Several different representations have been proposed for comparing protein structures [19,20]. The most intuitive way would be to compare coordinates of corresponding residues (α carbons in the main

chain) or atoms of two proteins. Distance measurement by the root mean square deviation (RMSD) is appropriate when two proteins have the same length and have a similar overall main chain orientation[21]. When two proteins have different chain lengths, residue correspondence needs be predetermined to compute RMSD. That can be done, for example, by combining the RMSD computation by structure superimposition with the dynamic programming algorithm (DP)[22,23,24] or an iterative use of DP[25]. Also, comparing the distance map of proteins can quantify similarity of proteins based on contacts of residues[26,27]. A more coarse protein representation uses vectors that describe secondary structure segments and compares spatial arrangements of secondary structures[28].

An important point to note is that different protein structure comparison methods compare different features of protein structures. Thus distances of protein structures defined by different methods differ and consequently, database search results by different methods inevitably differ. To illustrate this, consider three programs, CE[23], SAL[22], and COSEC[28]. CE and SAL can be categorized in the same class of algorithms because both employ DP as the basis of their algorithms. However, their behavior is very different: CE first identifies similar fragment pairs of a fixed length between input two protein structures without allowing gaps in fragments, and then extends the combination of similar fragment pairs. On the other hand, SAL uses DP iteratively, allowing gaps in a structural alignment to find statistically significant matches in overall corresponding residues between two proteins. As a result, corresponding protein pairs judged as similar by CE tend to have fragments of the same secondary structure, while protein pairs found

by SAL often have very different corresponding fragments because it allows gaps in them. Now, CE and COSEC are similar with each other in the sense that both compare ungapped fragments as the basis of structure comparison. However, a large difference exists between them. CE compares fragments of two proteins in a sequential order by DP, while COSEC compares spatial arrangement of fragments of two proteins without considering sequential connectivity of fragments. Therefore, CE is more suitable for finding similarity and dissimilarity of relatively closely related proteins, while COSEC can find distantly related protein pairs which have circular permutation or domain insertions[29], which CE cannot. SAL is powerful in finding overall fold similarity of proteins which is missed by CE or COSEC. This feature of SAL is especially useful for finding template structures for protein structure prediction[22,30]. The important thing is to understand strengths and main purposes of each structure comparison algorithm, and use appropriate algorithm for questions one wants to ask.

Here, we use another representation of global structure of proteins which concerns the surface shape of proteins. A surface representation does not consider either of individual residue/atom positions or arrangement of secondary structure segments[31]. The surface of a protein has been represented in several ways, including tessellation[32,33], α-shape[34], and spherical harmonics[35]. In this study, we introduce for the first time the 3D Zernike descriptor[36] as a representation of the protein surface shape, which is based on a series expansion of a given 3D function.

The reason why we use the 3D Zernike descriptor, is because it has several strong advantages.

First, compared to conventional methods, it allows fast retrieval of protein structures. The current major structure databases, including PDB, CATH[37], and SCOP[38] only allow keyword search and browsing of precomputed classification. The DALI server[39], VAST search at NCBI[40], and eF-site database[41] allow users to search the database with a query structure, but a search often takes hours to finish. Ideally, for a routine use of protein structure comparisons against a large number of structures should be done quickly, similar to that of a BLAST search. Second, 3D Zernike descriptors are rotation invariant, *i.e.* protein structures need not be aligned for comparison. Related works, such as the multipole method[42] proposed for global protein shape comparison and an application of spherical harmonics for binding pocket and ligand comparisons by Morris *et al.*[43], need pose normalization because the methods are not rotation invariant. The multipole method uses a reference frame which is computed based on the residue C-α coordinates, and the work by Morris *et al.* poses a protein by first, second, and third moments of around the mean of surface positions. Generally speaking, pose normalization could be problematic[44] especially in comparison of protein shapes, which are almost globular and determining the principle axes may not be robust. Third, the resolution of the description of protein structures can be easily and naturally adjusted by changing the order of 3D Zernike descriptors to be considered. For example, the rough global difference of protein structures reflects the difference of the first couple of invariants which correspond to lower orders of the 3D Zernike descriptor. Moreover, other characteristics of a protein surface, such as electrostatic potentials, can be naturally incorporated into the description considering an appropriate 3D

function, which will be described elsewhere.

This manuscript is organized as follows: We first describe our implementation of 3D Zernike descriptor for protein surface shape retrieval. Then differences of the 3D Zernike descriptor and the other projection based methods are extensively discussed. Next, we report the results of our benchmark on the performance in protein structure search using a large dataset with 2432 proteins. The overall results showed a good agreement with structure comparison by the CE program[23], which compares main chain orientations of proteins, despite of the difference in view of protein shape by the two methods. We also compared 3D Zernike descriptor with another standard protein structure comparison method, DALI[26], and four other 3D object comparison methods in the computer graphics and engineering domain. Finally, differences between CE and 3D Zernike are shown, emphasizing the advantage of 3D Zernike. The effect of shape comparison at different resolution is also discussed.

Methods

Building a surface of a protein

The first step of computing 3D Zernike descriptor of a protein is to define the protein surface region in 3D space. To begin with, hetero atoms including water molecules in the PDB file of the target protein are removed. Then, the MSROLL program in Molecular Surface Package version 3.9.3[33] is used to compute the Connolly surface (triangle mesh) of the protein using default parameters. Next, the triangle

mesh is placed in a 3D cubic grid of N^3 (N = 200), compactly fitting a protein to the grid. Each voxel (a cube defined by the grid) is assigned either 1 or 0; 1 for a surface voxel which locates closer than 1.7 grid interval to any triangle defining the protein surface, and 0 otherwise. Thus, the thickness of the protein surface is 3.4 grid intervals. The inside of a protein is kept empty so that 3D Zernike descriptor focuses on capturing the surface shape of a protein.

3D Zernike descriptor

To obtain 3D Zernike descriptors, one expands a given 3D function $f(\mathbf{x})$ into a series in terms of Zernike-Canterakis basis[36] defined by the collection of functions

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

with $-l < m < l, \quad 0 \le l \le n$, and $(n-l)$ even. Here $Y_l^m(\vartheta,\varphi)$ are spherical harmonics[45]. Spherical harmonics is the angular portion of an orthogonal set of solutions to Laplace's equation, which is given by:

$$Y_l^m(\vartheta,\varphi) = N_l^m P_l^m(\cos\vartheta)e^{im\varphi} \qquad (2)$$

Here N_l^m is a normalization factor

$$N_l^m = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} \qquad (3)$$

And P_l^m is the associated Legendre functions.

$R_{nl}(r)$ are radial functions defined by Canterakis, constructed so that $Z_{nl}^m(r,\vartheta,\varphi)$ are polynomials

when written in terms of Cartesian coordinates as follows:

The conversion between spherical coordinates and Cartesian $\boldsymbol{x}$ is defined as

$$\mathbf{x} = |\mathbf{x}|\zeta = r\zeta = r(\sin\vartheta\sin\varphi, \sin\vartheta\cos\varphi, \cos\varphi)^T \tag{4}$$

Then the harmonics polynomials e_l^m are defined as

$$e_l^m(\mathbf{x}) \equiv r^l Y_l^m(\vartheta,\varphi) = r^l c_l^m \left(\frac{ix-y}{2}\right)^m z^{l-m} \sum_{\mu=0}^{\left\lfloor \frac{l-m}{2} \right\rfloor} \binom{l}{\mu}\binom{l-\mu}{m+\mu}\left(-\frac{x^2+y^2}{4z^2}\right)^\mu \tag{5}$$

where c_l^m are normalization factors:

$$c_l^m = c_l^{-m} = \frac{\sqrt{(2l+1)(l+m)!(l-m)!}}{l!} \tag{6}$$

Using the harmonics polynomials e_l^m, 3D Zernike functions (Eqn. 1) can be rewritten in Cartesian coordinates:

$$Z_{nl}^m(\mathbf{x}) = R_{nl}(r)Y_l^m(\vartheta,\varphi) = \sum_{\nu=0}^{k} q_{kl}^\nu |\mathbf{x}|^{2\nu} r^l Y_l^m(\vartheta,\varphi) = \sum_{\nu=0}^{k} q_{kl}^\nu |\mathbf{x}|^{2\nu} e_l^m(\mathbf{x}) \tag{7}$$

where $2k = n - l$ and the coefficient q_{kl}^ν are determined as follows to guarantee the orthonormality of the functions within the unit sphere,

$$q_{kl}^\nu = \frac{(-1)^k}{2^{2k}}\sqrt{\frac{2l+4k+3}{3}}\binom{2k}{k}(-1)^\nu \frac{\binom{k}{\nu}\binom{2(k+l+\nu)+1}{2k}}{\binom{k+l+\nu}{k}} \tag{8}$$

Now 3D Zernike moments of $f(\mathbf{x})$ are defined as the coefficients of the expansion in this orthonormal basis, *i.e.* by the formula

$$\Omega_{nl}^m = \tfrac{3}{4\pi}\int_{|\mathbf{x}|\le 1} f(\mathbf{x})\overline{Z}_{nl}^m(\mathbf{x})d\mathbf{x}. \tag{9}$$

To achieve rotation invariance, the moments are collected into (2*l*+1) dimensional vectors

$\Omega_{nl} = (\Omega_{nl}^{l}, \Omega_{nl}^{l-1}, \Omega_{nl}^{l-2}, \Omega_{nl}^{l-3}, ..., \Omega_{nl}^{-l})$ and define the rotationally invariant 3D Zernike descriptors F_{nl} as norms of vectors Ω_{nl}. Thus

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^{m})^2} \qquad (10)$$

Index n is called the order of the descriptor. The rotational invariance of 3D Zernike descriptors means *e.g.* that calculating F_{nl} for a protein and its rotated version would yield the same result.

In this work, binary voxelization of a protein structure is used, *i.e.* function $f(\mathbf{x})$ is defined to be 1 at the points sufficiently close (within about one voxel size) to the surface of the protein, and 0 at all other locations. The order *n* determines the resolution of the descriptor. It was shown in a previous study that the order of at most $n = 20$ (*i.e.* a total of 121 numbers to index each shape) provides sufficient accuracy[36].

Definition of distance used in this study

Now that a protein 3D structure is represented by 121 numbers, a comparison of two protein 3D structures simply results in a comparison of two series of the 121 numbers. In this study, we used three distance measures for comparing 3D Zernike descriptor of protein surface shapes. The first function is the Euclidean distance, d_E, which is the root mean square deviation of corresponding index numbers of two proteins:

$$d_E = \sqrt{\sum_{i=0}^{i=nl} (z_{Ai} - z_{Bi})^2} \qquad (11)$$

where Z_{Ai} is the ith number of 3D Zernike descriptor of protein A, *n* is the order of descriptor, and $0 \le l \le n$, and $(n-l)$ even as described above.

The second function is the Manhattan distance, d_M, which is the sum of the difference of each corresponding index numbers:

$$d_M = \sum_{i=0}^{i=nl} \left| z_{Ai} - z_{Bi} \right| \quad (12)$$

The third distance, d_c, is defined as

$$d_C = 1 - Correlation\ Coefficient\ (\mathrm{Z_A}, \mathrm{Z_B}) \quad (13)$$

Thus $d_c = 0$ when two descriptors correlates perfectly.

3D Zernike descriptor and spherical harmonic descriptors

In this section we discuss 3D Zernike descriptors mainly in comparison with the Spherical Harmonics Descriptors (SHD)[46,47], which is a popular spherical harmonics-based projection techniques used for general 3D object comparison. Projection based techniques have been used extensively in two dimensional (2D) image analysis and pattern recognition[48,49,50,51,52]. In particular, 2D Zernike moments have proved exceptionally useful for the analysis of 2D shapes arising in many areas ranging from face recognition[53], cell parts recognition[54] and optical scattering pattern recognition for identifying bacterial colonies[55]. Yeh *et al.* applied 2D Zernike moments to protein 3D structure retrieval by characterizing a structure with a set of 2D projections from 100 different directions[56]. Finally, Canterakis was able to

extend 2D Zernike polynomials and moments to 3D, introducing 3D Zernike-Canterakis polynomials[57]. Later, rotationally invariant descriptors based on Zernike-Canterakis moments were explored for 3D shape retrieval by Novotni and Klein[36], who reported improved precision-recall curves at a lower storage cost when compared toSHD.

For SHD no radial modulation is used; rather, the three-dimensional space is sampled into concentric spherical shells around the center of mass. Then a volume of a target object within each concentric sphere of a radius *r* centering at the center of mass of the object, $f_r(\vartheta,\varphi)$, is expanded in the series of spherical harmonics, $Y_l^m(\vartheta,\varphi)$:

$$f_r(\vartheta,\varphi) = \sum_l f_r^l(\vartheta,\varphi) = \sum_l \sum_{m=-l}^{l} c_{r,l}^m Y_{r,l}^m(\vartheta,\varphi) \qquad (14)$$

Spherical harmonics differs under different orientations, (ϑ,φ). However, since the L_2 norm of the function is rotation invariant, a rotation invariant signature for $f_r(\vartheta,\varphi)$ is constructed as the collection of L_2 norms of $f_r^l(\vartheta,\varphi)$ at each *l*, *i.e.* $\left\{\left\|f_r^0\right\|,\left\|f_r^1\right\|,...\right\}$. Finally, collecting the signature for each radius, *r*, will give the SHD of a protein structure. The implementation uses 32 shells, 17 descriptors per shell, making a total of 544 numbers to represent a shape.

Let us remark that 3D Zernike descriptors genuinely belong to the three-dimensional realm, while SHD are essentially a combination of two-dimensional descriptors. Indeed, notice that SHD measures similarity of objects by comparing them shell-wise. There are quite a few practical implications of this fact: (1) SHD does not capture object coherence in the radial direction, thereby incorporating less

object characteristic information[36]. For example, since the descriptors for each shell are calculated separately, the shells can be rotated independently by random angles without changing the resulting descriptors. (2) The orthonormality of the Zernike-Canterakis basis results in less information redundancy. One should notice that in SHD, descriptors coming from adjacent shells are highly correlated, making them redundant to some extent. Indeed, using 154 3D Zernike descriptors (max order 21) yields better retrieval results than using 928 SHDs (32 shells, 29 descriptors per shell) as tested on the Princeton Shape Benchmark, which is a database of general 3D objects such as airplanes and chairs[47 36]. (3) SHDs require polar sampling, which was pointed out to be problematic for the robustness of rotation invariancy[58]. Securing robustness of SHD requires a distance field based voxelization procedure where voxels are assigned continuous values between 0 and 1. On the other hand, the Zernike-Canterakis basis consists entirely of polynomials in Cartesian coordinates, thus avoiding polar sampling, and making possible to treat all voxels in the model on equal footing. In addition, 3D Zernike descriptors show optimal performance when simple binary voxelization is used[36]. Because sizeable amount of the computational time is consumed by the voxelization process, this simplicity results in faster response times for user-search engine transactions. (4) One can naturally add other protein surface properties within the 3D Zernike framework. For example, to add electrostatics, it is enough to calculate 3D Zernike descriptors of $f(x)$ set equal to the electrostatic potential value on the surface, and zero otherwise. This is not as straightforward with the SHD, because of the aforementioned robustness problem.

Comparison with the other surface shape-based structure representation methods

The 3D Zernike descriptor is compared with four other methods in terms of the performance on retrieving similar protein structures. The benchmark dataset used is described in the next section. The all four methods represent surface shape of objects, namely, the SHD[47,46], the solid angle histogram[59,60], the shape distribution[61], and the eigen value model[60]. These methods have been developed and used for recognition of 3D shapes in computer graphics and engineering domain. Below we briefly describe the idea of these methods.

To compute aforementioned SHD of a protein structure, first the protein structure is voxelized. We used the SpharmonicKit package (http://www.cs.dartmouth.edu/~geelong/sphere/) for computing SHD. The Euclidean distance was used to compare SHDs of two proteins.

The solid angle histogram (SAH) represents a distribution of local concavity and convexity of a protein structure. To obtain SAH, a protein is first voxelized. Let $K_{c,r}$ denote a set of voxels included in a sphere of a radius of *r* with the center at a voxel *c*. Then the solid angle value $SA(v_i, r)$ at a voxel v_i for a protein volume V is defined as the fraction of the intersection volume of a sphere $K_{vi,r}$ with the protein volume V relative to the volume of the sphere $K_{vi,r}$:

$$SA(v_i, r) = \frac{\left| K_{v_i,r} \cap V \right|}{\left| K_{v_i,r} \right|} \qquad (15)$$

Hence, a histogram of $SA(v_i, r)$ represents a protein structure. The solid angle histograms of two proteins

are compared by the L_1 norm as suggested in the previous work[60].

The shape distribution method describes an object as a histogram of the length of pairs of points on the surface of the object. First, a given protein is voxelized and the distance distribution if computed by randomly sampled pairs of voxels on the protein surface. We use L_2 norm to compute the distance of voxel pairs and L_1 norm to compare the similarity of two shape distributions.

The eigen value model[60] represents a given protein as a set of eigen values. The model first voxelize the protein into a 3D grid and divides the grid into cells. Then for each cell, three eigen values of the distribution of the points (voxels) are computed, resulting in total of $3p^3$ eigen values. These eigen values are registered in the according bins of a histogram, which describes the protein structure. Two histograms are compared by L_1 norm.

In a comparative study of performance of above methods on retrieval of engineering parts (*e.g.* bolts, wheels) [62], it was shown that their performance varies depending on parts. Therefore it was our curiosity that whether these methods originally developed for general objects and engineering parts can be used for protein structure search or not.

Benchmark Dataset

The benchmark dataset of protein structures consists of 2432 protein structures classified into 185 fold groups. These are a subset of structures extracted from a structure comparison results by the CE

program[23] (ftp://ftp.sdsc.edu/pub/sdsc/biology/CE/db/ ata_3_8.txt). Note that the structure representation of CE and 3D Zernike descriptor is fundamentally different: the former consider a protein structure as the spatial position of main-chain residues and the latter represents a protein structure as a surface shape. The purpose of this benchmark study is to investigate the extent of similarity between the two methods. If we observe a significant agreement between the two methods, that result suggests that 3D Zernike descriptor can be an effective tool for fast search of protein structures with a similar main-chain orientation (*i.e.* a conventional sense of protein structure similarity, which also implies evolutionary relationship) not only a similar surface shape. On the other hand, it is also expected that interesting cases that two proteins which share a similar surface shape but different main-chain orientation can be found. CE is one of standard programs for protein main-chain comparison that classifies proteins solely by geometrical aspect of proteins without consideration of evolutionary relationship as, for example, SCOP database does. Given two protein structures, CE first identifies eight residue-long fragments of a similar conformation in the two proteins by comparing corresponding distances of pair of residues within each fragment. Then, identified fragment pairs from the two proteins are combined to find larger structurally similar regions by comparing corresponding inter-fragment distances. DP is used for the calculation, thus fragment pairs are combined in a sequential order from the N-terminus to the C-terminus. Below describes the procedure we used to select the benchmark proteins.

The original CE database consists of 50,246 protein structures classified into 7,386 fold groups.

Each fold group consists of a “representing” protein, a set of “represented” proteins which satisfy several similarity criteria against the representing protein, and another set of “similar” proteins (see the README file of the database for more technical details). Starting from the CE database, first, separate fold groups are merged if the structure of their “representing” proteins is sufficiently similar, having a Z-score of 3.8 or higher by CE. The Z-score of 3.8 is recommended by the authors of the database to filter out random similarities. Next, the set of “similar” proteins are eliminated from a fold group. Then, “represented” proteins are eliminated from a fold group if the size is more than 12.5% different in length from the “representing” protein, or if the quality of the structure is not appropriate: Structures which lacks coordinates of more than ten residues, or which have only coordinates of α carbons, are removed. Small proteins that have less than a hundred residues are also eliminated. In addition, structures which have coordinates of hydrogen atoms of more than 3% of residues are filtered out, because they significantly affect surface shape of the protein. Lastly, small groups that only contain three of less “represented” proteins (and “representing” protein) are removed.

Dali protein structure comparison program

In addition, we also run the Dali algorithm[26] against the CE based benchmark dataset. Dali is another widely used protein structure comparison algorithm which is established in 1993. Dali compares two protein structures in terms of the two-dimensional distance map of the proteins. First, Dali identifies

similar sub-distance maps of two input proteins of a fixed size by comparing corresponding distances between two sub-distance maps. This step captures local regions of the two proteins which have a similar residue contact pattern. Next, the algorithm combines identified pairs of similar sub-distance maps to find significant similar structures between the two proteins. We used the standalone program of the Dali algorithm, DaliLite[63], which is available for download at http://www.ebi.ac.uk/DaliLite/ .

Benchmark Procedure

For each protein, the whole set of proteins are ranked by a given distance of the 3D Zernike descriptor. For a given distance threshold value, the sensitivity and the specificity are averaged within a group, then again averaged among all groups to give a final value in the plots (Fig. 3). The sensitivity and the specificity are defined as follows:

$$Sensitivity = TP \big/ TP + FN \quad (14)$$

$$Specificity = TP \big/ TP + FP \quad (15)$$

, where *TP*, true positive, is the number of fold group members of a query protein retrieved with a distance closer than the threshold; *FN*, false negative, is the number of the fold group members whose distance to a query protein is larger than the threshold hence missed in the search; *FP*, false positive, is the number of proteins which are not included in the fold group with the query protein but incorrectly retrieved in the search. Thus, the denominator in the Eqn. 14 for the sensitivity is the total number of all

members of the fold group. The denominator in the Eqn. 15 for the specificity is the total number of proteins retrieved above the threshold.

Results

Examples of 3D Zernike descriptor

Figure 1 shows examples of 3D Zernike descriptor of two proteins, triosephosphate isomerase (PDB code: 7tim, A chain) and interleukin-4 receptor α-chain (1iarB). Globally, 7timA has more or less a round shaped surface and 1iarB is an L-shaped structure (Fig. 1A). This apparent difference of their global surface shape is reflected by distinctive 3D Zernike descriptors shown in Figure 1B. The difference in the overall shape tends to appear in the first couple of orders of the descriptor, resulting in a relatively large Euclidean distance of 38.84, and Correlation coefficient based distance of 0.656.

Rotation Invariance

As described above and in Method section, the 3D Zernike descriptor is mathematically most noteworthy rotationally invariant. This is one of the largest advantages of 3D Zernike descriptor. However, in practice the descriptor of rotated protein structures are not exactly identical. This error is caused possibly when the protein surface shape is discretized into voxels. We found that in computing all the distance measures, *i.e.* Euclidean (Eqn. 11), Manhattan (Eqn. 12) and the correlation coefficient based

(Eqn. 13), normalizing each number in a 3D Zernike descriptor by the sum of the 121 numbers of the descriptor reduces the error the best among tested methods. Figure 2 shows an example of the variance of 3D Zernike descriptor upon rotation. Here, the two proteins used in Figure 1 are rotated to all the possible positions, and Euclidean and the correlation coefficient based distance from the original position are computed. We used these two distances because they were the top two performing functions in our protein shape search benchmark (see the next section). As for Euclidean distance (Fig. 2A), approximately 90% of the rotated structures stay within the distance of ten. In the case of the correlation coefficient based distance (Fig. 2B), approximately 90% of the rotated structures of the two proteins have less than a distance of 0.03. From this experiment, we can draw a threshold of the significance of the distance, or in the other words, determine an "invisible" range of the 3D Zernike descriptor: for example, if two proteins have an Euclidean distance of less than ten, these proteins can be considered significantly similar, or more precisely, indistinguishable from the case where the two proteins are identical but placed in a different orientation.

In order to improve the rotation invariance of the descriptor, we have tried different thickness of the surface representation, and also a continuous value assignment between 0 and 1 to surface voxels rather than the binary voxelization, but did not observe differences in the performance.

Structure Retrieval by three different methods

One of the most interesting applications of 3D Zernike descriptor is fast structure retrieval. In this experiment, we used a database derived from the CE algorithm[23] to define the structure similarity. It is important to note that the structure similarity depends on how structures are represented and compared [19,64]. CE uses combination of similar main chain fragments to compute the similarity of two protein structures. On the other hand 3D Zernike descriptor compares surface shape and Dali compares the distance maps of two proteins.

Let us first compare the CE benchmark dataset with the SCOP protein classification database[38] in order to understand the nature of protein structure classification. The unique feature of SCOP is that evolutionary relationship of proteins is also taken into account by manual curation together with protein structure similarity. Thus SCOP has been serving as an indispensable resource for elucidating relationships of protein structure and function. In Table 1A, the 185 fold groups in the CE benchmark dataset are compared with the superfamilies defined in SCOP. The number of CE fold groups which overlap with a SCOP superfamily by a certain fraction is counted. When a CE fold group overlaps with several SCOP superfamilies, the SCOP superfamily which gives the largest overlap is counted. The number of SCOP superfamilies which correspond to the CE benchmark dataset is 150, which is smaller than the number of the CE fold groups. It is found that only 75.1% of the CE fold groups correspond to one SCOP superfamily. Table 1B shows that 82.2% of the CE fold groups correspond to one SCOP fold. The overlap with SCOP folds looks larger than SCOP families because the average size of SCOP fold is

larger, thus it is more frequent that multiple CE fold groups correspond to a SCOP fold. The number of SCOP folds which correspond to the CE benchmark is 117. These results illustrate that even the widely used protein structure comparison method, CE, does not have perfect correspondence with a well established protein structure classification database, SCOP. Especially this implies that CE should be used with caution if the purpose of using CE is to investigate biological function of proteins, since only 75.1% of the CE groups agree with SCOP families. A recent work by Sierk and Pearson provides a further benchmark for protein structure comparison methods[65]. Therefore, the aim of the structure retrieval performed in this study using 3D Zernike descriptor and Dali on the CE benchmark dataset is to understand the similarity and dissimilarity of the three methods, not evaluating "accuracy" of a particular method.

Figure 3 shows the sensitivity and specificity plot of the benchmark performance on a dataset of 2432 proteins. Results of the 3D Zernike descriptor with and without pre-screening by the length of the proteins are also shown. When the pre-screening is used, a protein in the dataset is compared to a query protein only when its length is in the range of 57% to 175% of that of the query protein. The three different distance measures, namely, Euclidian, Manhattan, and the correlation coefficient-based (Eqns. 11-13) are compared. First, regardless of the pre-screening, the results are far better than random. Second, among the three distance measures, the performance of Manhattan distance is somewhat worse than the other two distance measures, but all three distance measures essentially showed similar performance.

Third, it is shown that the pre-screening is effective in improving the search performance. This is because the scale is normalized so that a structure fits in a unit sphere when computing 3D Zernike descriptor, hence the size information is lost[36].

Table 2 summarizes the search results of 3D Zernike descriptor with the length-based pre-screening is used. More than 89.0% of proteins retrieved another protein in the same CE fold group within the top five closest structures. Those successful proteins are not biased to specific types of protein folds, because the successful proteins are distributed among approximately 98% of the fold groups (Top 5, Group1). When Top 10 hits are considered, 93.1% of the proteins successfully retrieved its CE fold group member by using the Euclidean or the correlation coefficient-based distance. The search was successful for at least one protein in almost all the fold groups (99.5% by using the Euclidean distance) considering Top 10 hits. On the other hand, approximately half of the fold groups contain some members which could not retrieve its fold group member within Top 10 (Top 10, GroupsAll). These are protein structures which are judged to be similar by the main-chain orientation but not by the surface shape. Below in Figure 5 we show examples of these cases.

The structure retrieval results by DaliLite are also shown in Table 2. Interestingly, only 28.6% of proteins retrieved another protein in the same CE fold group within top five by using DaliLite. Actually the Top 5 and Top 10 results by DaliLite are only slightly better than the random retrieval.

To conclude, overall 3D Zernike descriptor showed a strong agreement in the protein structure

retrieval with CE despite of its completely different representation of protein structures. 3D Zernike descriptor agrees with CE much more than Dali (*i.e.* the DaliLite program) does with CE. The superiority of the 3D Zernike descriptor will be clearer when its performance is compared with the other shape comparison methods in Figure 4.

Next, the 3D Zernike descriptor using the correlation coefficient based distance was compared with four existing shape comparison methods, namely, the spherical harmonics descriptor (SHD), the shape distribution histogram, the solid angle histogram, and the eigen value model (Fig. 4). To our surprise, all these four methods' performance was no better than the random retrieval. This may be due to the protein surface shape being more or less globular, hence all the proteins in the benchmark dataset looked almost the same to the four shape comparison methods developed in computer graphics and engineering domain. The strikingly poor performance of the four methods in Figure 4 reminds us that these methods are originally designed to differentiate general objects, *e.g.* airplanes from cars, trees from chairs, or steering wheels from car doors. In contrast, Figure 4 clearly highlights the appropriateness of 3D Zernike descriptor's utility in the protein shape search, which was revealed to be challenging for conventional shape retrieval methods developed in computer graphics and engineering domain.

Figure 5 illustrates difference between CE and 3D Zernike descriptor. Figure 5A and 5B are protein structure pairs which are identified to be significantly similar by CE, but not by 3D Zernike. They are evolutionary closely related and thus classified into the same family in CATH and SCOP. In these two

examples, a small portion of the secondary structure elements of the protein is flipped out (figure on the right) from the mass of the protein, resulting in the change of the surface shape. Figure 5C and 5D are opposite examples. Figure 5C is a vivid example of two proteins which have a very similar surface shape but with completely different secondary structure elements, where the left structure is a β class protein and the right structure is an α class protein. Figure 5D is a protein pair with a different topology (in CATH) forming a very similar surface shape.

Taking advantage of the 3D Zernike descriptor's ability to find proteins with similar overall protein surface shape, functionally related proteins can be retrieved beyond sequence similarity and significant backbone conformation similarity. Figure 6 shows several such examples. Associated table, Table 3 gives detailed data for the proteins in Figure 6. Figure 6A is a pair of DNA topoisomerase I from human and *E. coli*. The characteristic pore of the proteins is to capture DNA double strands. The sequence identity between the two proteins is very low, and the CE only aligns 17.3% of the whole region of the two proteins. In contrast, 3D Zernike descriptor identifies the overall similar surface shape with a significant distance (compare the 3D Zernike distances with the average distance of the top hit in the benchmark, the right columns in Table 2). Figure 6B shows two DNA binding proteins. Both proteins bind to DNA with the curved U-shaped region. These two proteins have different function, but both have the characteristic surface shape which enables binding to DNA, that is captured by 3D Zernike descriptor. Figure 6C is another pair of proteins. These two proteins bind to DNA with their long tail regions. Note

that SCOP classifications of these three pairs are also different from each other. Figure 6D is a pair of subunits of membrane protein complexes. 2nwl is a submit of glutamate transporter, which is a pentamer, and 2bbh is a subunit of CorA Mg2+ transporter which is a trimer. In the both cases, the two long helices penetrate membrane and form the scaffold of the transporters. The last example, Figure 6E is a pair of transmembrane proteins. In all the cases (Fig. 6A-E), the sequence identity between the pair is below 10%, and CE only aligns partial regions of the pair. In contrast, 3D Zernike descriptor captures overall surface similarity of each pair which is required to realize their biological function with significantly close distance.

Search speed by 3D Zernike descriptor

The 3D Zernike descriptor allows rapid real-time search on the web because a protein structure is compactly represented by 121 numbers (when the order $n = 20$). If a query protein is already transformed into 3D Zernike descriptor, a search to the current benchmark dataset takes less than a second on an Intel Pentium 4 3.0 GHz processor with 2 GB of memory (Table 4). When a custom PDB file is input as the query, the following steps must be performed before database search: (1) Solvent accessible surface triangulation by Molecular Surface Package [33], (2) Surface voxelization, and (3) transformation into 3D Zernike descriptor. Taken together with the database search, this entire process takes still less than a minute. Since enlarging the database to be searched only affect to the execution time of the database

search step, a search against the entire PDB (as of August 2007) with 45,000 structures will only take a minute. The search speed can further be made faster if the database is prescreened by the length of the query protein. Note that a pairwise structure comparison by CE takes typically a couple of seconds. Thus, a database search against PDB using CE would takes more than a day.

Structure database searches using 3D Zernike descriptor can be performed through the web at: http://dragon.bio.purdue.edu/3d-surfer/. Users can search the benchmark dataset with one of the structures in the dataset (i.e. 3D Zernike descriptor of the protein is precomputed) or by uploading a custom PDB file to the server.

Resolution of the descriptors

As described above, one of the characteristics of the 3D Zernike descriptor is that the resolution of the description of shapes can be altered by changing the order of 3D Zernike descriptors. In Figure 7, two different orders (the index *n* in Eqns. 7, 9, 10), five and twenty, are used to compute similarity (Euclidean distance) of the sixteen proteins selected from different CE fold groups. Altering the order of descriptors changes the distances of proteins (*e.g.* The Euclidean distance of 1theB to 1o0eA is 28.74 in Fig. 7A, which is 13.36 in Fig. 7B). Also, the relative distance of pairs changes, which is obvious from the different topology of the two trees (Fig. 7A and 7B). When the order of five is used (Fig. 7A), an emphasis is given to describe overall shapes, such as spherical, cylinder like, or tadpole like shapes. With

the order of twenty, clusters made by using the order of five are further decomposed (Fig. 7B). To highlight the decomposition of clusters between the two trees, clusters of proteins within the Z-value of the Euclidean distance of 0.35 are shaded by the same color. The Z-score of the Euclidean distance using the order of twenty and five is computed using the average and the standard deviation of the distribution of distances of protein pairs in the CE benchmark dataset. Reducing resolution will also contribute in the search speed, because the descriptor becomes more compact. 121 numbers are used in a descriptor when the order *n* is set to twenty, and it is decreased to 12 when the order is set to five.

Discussion

Here, we have introduced 3D Zernike descriptor as a novel computationally efficient method for searching protein tertiary structures. Unlike the other existing methods for structure comparison and representation, the 3D Zernike descriptor allows an extremely rapid database search, which opens up a the possibility for a real time protein tertiary structure search on the internet. The search speed can be further increased by prescreening proteins by their length and/or by multi-resolution search using different orders of the descriptor. A search against the benchmark dataset of 2,432 proteins used in this work took only 0.46 seconds. This indicates by a simple computation that a search against the current entire PDB database with 45,000 proteins would take 18.5 seconds. A preeminent mathematical property of 3D Zernike descriptor is that it is rotation invariant. This is a significant advantage over spherical harmonics

and the multipole representation[42] which need to pose structures on a reference frame for comparison.

Because 3D Zernike descriptor concerns surface shape of proteins but not main chain orientation, in principle proteins found to be similar by 3D Zernike descriptor does not necessarily have evolutionary relationship, as illustrated in Figure 5. However, our benchmark results show that in majority of the cases 3D Zernike descriptor retrieves protein structures of the same fold (Table 2), thus demonstrates its utility in regular protein global structure database search. In a practical implementation of a tool for a real-time protein structure search, 3D Zernike descriptor could be used as a rapid primary filter, followed by an option to employ a conventional structure comparison method, such as CE, to compute main-chain similarity between a query protein against retrieved top 10 to 20 structures.

Moreover, surface shape representation made possible by 3D Zernike descriptors has numerous intriguing applications. A possible application is 3D shape matching for images by electron microscopy or electron tomography. Currently, we are developing local protein surface shape comparison and search algorithms for structure-based function annotation[66]. It would be also interesting to analyze surfaces of proteins or biological molecules with a similar function but different main-chain or molecular structure, such as binding sites of DNA-binding proteins, or proteins which display structural mimicry[67].

Biology has entered an informatics era when efficient reuse of knowledge from existing databases is crucial. In biological sequence comparison, BLAST and FASTA have enabled fast database search more than a decade ago, which revolutionized biological research. In contrast, handling of protein

3D structures is still in the realm of pairwise comparison, by which a 3D structure database search may still take hours. That would certainly render 3D structure search impractical and hinder the development of novel tools/applications based fast structure search. We believe the fast real-time 3D structure search enabled by 3D Zernike descriptors, so to speak, 3D-BLAST, will lead us to a paradigm shift in research concerning protein tertiary structure.

Acknowledgements

This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health (R01 GM075004). DK also acknowledges funding from National Science Foundation (DMS 0604776) and NIH (U24 GM077905).

Reference List

1. Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. Trends Biochem Sci 2005;30(11):622-629.

2. Winter C, Henschel A, Kim WK, Schroeder M. SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res 2006;34(Database issue):D310-D314.

3. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ. SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res 2007;35(Database issue):D580-D589.

4. Orengo CA, Thornton JM. Protein families and their evolution-a structural perspective. Annu Rev Biochem 2005;74867-900.

5. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 2006;151550-1556.

6. Hawkins T, Kihara D. Function prediction of uncharacterized proteins. J Bioinform Comput Biol 2007;5(1):1-30.

7. Hawkins T, Chitale M, Kihara D. New paradigm in protein function prediciton for large scale omics analysis. Molecular BioSystems 2008;in press.

8. Service R. Structural biology. Structural genomics, round 2. Science 2005;307(5715):1554-1558.

9. Burley SK. An overview of structural genomics. Nat Struct Biol 2000;7 Suppl932-4.

10. Zhang C, Kim SH. Overview of structural genomics: from structure to function. Curr Opin Chem Biol 2003;7(1):28-32.

11. 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(1):235-42.

12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403-10.

13. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85(8):2444-8.

14. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins.

EMBO J 1986;5(4):823-826.

15. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297(1):233-249.

16. Kinoshita K, Nakamura H. Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci 2005;14(3):711-718.

17. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the Escherichia coli genome using the sequence- to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 1998;282(4):703-11.

18. Torrance JW, Bartlett GJ, Porter CT, Thornton JM. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 2005;347(3):565-581.

19. Mizuguchi K, Go N. Seeking significance in three-dimensional protein structure comparisons. Curr Opin Struct Biol 1995;5(3):377-382.

20. 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(3):393-398.

21. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst 1978;A34827-828.

22. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003;334793-802.

23. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11(9):739-47.

24. Gerstein M, Levitt M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. Proc Int Conf Intell Syst Mol Biol 1996;459-67.

25. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 1996;266617-635.

26. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233(1):123-38.

27. Zhou X, Chou J, Wong ST. Protein structure similarity from Principle Component Correlation analysis. BMC Bioinformatics 2006;740-

28. Mizuguchi K, Go N. Comparison of spatial arrangements of secondary structural elements in proteins. Protein Eng 1995;8(4):353-62.

29. Vogel C, Morea V. Duplication, divergence and formation of novel protein topologies. Bioessays 2006;28(10):973-978.

30. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 2005;102(4):1029-1034.

31. 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(13-14):1970-1977.

32. Dupuis F, Sadoc JF, Jullien R, Angelov B, Mornon JP. Voro3D: 3D Voronoi tessellations applied to protein structures. Bioinformatics 2005;21(8):1715-1716.

33. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221(4612):709-713.

34. Edelsbrunner H, Mucke EP. 3-Dimensional Alpha-Shapes. Acm Transactions on Graphics 1994;13(1):43-72.

35. Macke TJ, Duncan BS, Goodsell DS, Olson AJ. Interactive modeling of supramolecular assemblies. J Mol Graph Model 1998;16(3):115-3.

36. 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.

37. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-108.

38. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30(1):264-7.

39. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26(1):316-9.

40. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23(3):356-69.

41. Kinoshita K, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 2003;12(8):1589-1595.

42. Gramada A, Bourne PE. Multipolar representation of protein structure. BMC Bioinformatics 2006;7242-

43. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics 2005;21(10):2347-2355.

44. Kazhdan M, Chazelle B, Dobkin D, Funkhouser T, Rusinkiewicz S. A reflective symmetry descriptor for 3D models. Algorithmica 2004;38(1):201-225.

45. Dym H, McKean H. Fourier series and integrals. 1972;

46. 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;43156-164.

47. Funkhouser T, Min P, Kazhdan M, Chen J, Halderman A, Dobkin D, Jacobs D. A search engine for 3D models. Acm Transactions on Graphics 2003;22(1):83-105.

48. Hu M-K. Visual pattern recognition by moment invariants. IRE Transactions on information theory 1962;8(2):179-187.

49. Hu M-K. Pattern recogntion by moment invariants. Proc of the IRE 1961;491428-

50. Sheng Y, Arsenault HH. Experiments on Pattern-Recognition Using Invariant Fourier-Mellin Descriptors. Journal of the Optical Society of America A-Optics Image Science and Vision 1986;3(6):771-776.

51. Casasent D, Psaltis D. Scale Invariant Optical Transform. Optical Engineering 1976;15(3):258-261.

52. Teh CH, Chin RT. On Image-Analysis by the Methods of Moments. Ieee Transactions on Pattern Analysis and Machine Intelligence 1988;10(4):496-513.

53. Foon NH, Pang Y-H, Jin ATB, Ling DNC. Efficient Method for Human Face Recognition Using Wavelet Transform and Zernike Moments. Proc Int Conf Comp Graphics, Imaging and Visualization (CGIV'04) 2004;0065-69.

54. Asadi MR, Vahedi A, Amindavar, H. Leukemia Cell Recognition with Zernike Moments of Holographic Images. Signal Processing Symposium, 2006 NORSIG 2006 Proceedings of the 7th Nordic 2006;214-217.

55. Bayraktar B, Banada PP, Hirleman ED, Bhunia AK, Robinson JP, Rajwa B. Feature extraction from light-scatter patterns of Listeria colonies for identification and classification. J Biomed Opt 2006;11(3):34006-

56. Yeh JS, Chen DY, Chen BY, Ouhyoung M. A web-based three-dimensional protein retrieval system by matching visual similarity. Bioinformatics 2005;21(13):3056-3057.

57. 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.

58. Laga H, Takahashi H, Nakajima M. Spherical Wavelet Descriptors for Content-based 3D Model Retrieval. IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06) 2006;15-

59. Connolly ML. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. Biopolymers 1986;25(7):1229-1247.

60. 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.

61. Jiantao P, Ramani K. A 3D Model Retrieval Method Using 2D Freehand Sketches. 2005;343-346.

62. Jayanti S, Kalyanaraman Y, Iyer N, Ramani K. Developing an engineering shape benchmark for CAD models. Computer-Aided Design 2006;38(9):939-953.

63. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16(6):566-567.

64. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5(7):1325-38.

65. Sierk ML, Pearson WR. Sensitivity and selectivity in protein structure comparison. Protein Sci 2004;13(3):773-785.

66. 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;in press.

67. Stebbins CE, Galan JE. Structural mimicry in bacterial virulence. Nature 2001;412(6848):701-705.

68. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 2005;5164-166.

Figure Legends

Figure 1. 3D Zernike descriptor of two example proteins. **A**, the global surface shape of the two proteins; 7timA (left) and 1iarB (right). **B**, 3D Zernike descriptor of the two proteins. Circles, 7timA; and triangles, 1iarB.

Figure 2. Variance of 3D Zernike descriptor upon rotation of proteins. All the possible rotated positions of two protein structures, 7timA and 1iarB, in three orthogonal directions with a step size of 30 degrees are computed. Histograms of distances of 3D Zernike descriptor between each of the rotated structures and the original one are shown. **A**, the Euclidean distance is used. Filled (empty) circles, the frequency (the cumulative frequency) of Euclidean distances of 7timA are plotted. Filled (empty) triangles, the frequency (the cumulative frequency) of Euclidean distances of 1iarB are plotted. **B**, the correlation coefficient is used as the distance measure. The frequency and the cumulative frequency of distances of 7timA and 1iarB are shown by solid line, dotted line, dashed line, dash-dot-dot line, respectively.

Figure 3. The sensitivity and the specificity of the benchmark dataset are plotted using three distance definitions of 3D Zernike descriptor. the Euclidean (black circles/upward triangles), the Manhattan (gray circles/dark gray squares), and the correlation

coefficient-based distance (downward triangles/light gray squares) with and without prescreening by the sequence length. When the prescreening is used, a protein in the dataset is compared to a query only when its length is in the range of 57-175% of that of the query protein. For comparison, results of a random retrieval are also plotted (black diamonds).

Figure 4. The performance of the 3D Zernike descriptor with the correlation coefficient-based distance (black solid circles) was compared with four other existing shape descriptors, the spherical harmonics descriptor (gray circles), the shape distribution histogram (black triangles), the solid angle histogram (gray triangles), and the eigen value model (black squares). See text for details of these methods. For comparison, the random retrieval is also added (gray squares).

Figure 5. Examples of protein pairs of the same main chain orientation but with a different surface shape (A, B), and pairs with a similar surface shape but with a different main chain orientation (C, D) are shown. Structural comparisons are performed by CE and 3D Zernike. The comparison between **A**, 1dz3A (response regulator SPO0A) and 1mb0A (response regulator DIVK). CE computes: RMSD=1.6 Å, Z-score=5.0, aligned/gap positions=104/3, Sequence identity (SeqID) = 26.9 %. 3D Zernike shows: Euclidean distance (d_E) = 51.21, Manhattan distance (d_M) = 438.53, Correlation coefficient (d_C) = 0.620. **B**, 1jznA

(galactose-specific C-type lectin) and 1g1qA (P-selectin lectin). CE: RMSD = 2.0 Å, Z-score = 5.9, aligned/gap positions = 115/12, SeqID = 23.5%. 3D Zernike: d_E = 52.67, d_M = 431.16, d_C = 0.602. In contrast, C and D demonstrate two instances in which 3D Zernike detect similar global surface shape of proteins with a different overall fold. **C**. 1barA (fibroblast growth factor) and 1rro (oncomodulin). CE: RMSD = 6.7Å, Z-score = 1.6, aligned/gap positions = 56/50, SeqID = 3.6 %. 3D Zernike: d_E = 12.66, d_M = 101.85, d_C =0.031. **D**. 1rypB (proteasome subunit) and 1gwz (Tyrosine phosphatase). CE: RMSD = 5.0 Å, Z-score = 2.3, aligned/gap positions = 72/70 SeqID = 9.7 %, 3D Zernike: d_E = 12.73, d_M = 108.89, d_C = 0.041.

Figure 6. Examples of protein pairs whose surface shapes are judged to be similar by 3D Zernike descriptor. Detailed data are shown in Table 3. A, 1a31 and 1cy0 (from left to right); B, 1tbp and 1t7p; C, 1b3t and 1adv; D, 2nwl and 2bbh; E, 2b2i and 2cfp. Detailed data of these protein pairs are shown in Table 3.

Figure 7. Resolution of 3D Zernike descriptor. **A**, 3D Zernike descriptors of the order of five; **B**, the order of twenty is used to construct trees representing similarity of the surface shape of sixteen proteins: 1theB, 1o0eA, 1dteA, 1aye, 1g28A, 1wbc, 1r52D, 1rxzA, 1fw8A, 2cauA, 1bas, 1ld9A, 1efwA, 1ezvC, 1yfm, and 1lwuC. The Euclidean distance is used. Proteins

within a Z-value of distance of 0.35 are grouped in a colored circle. The colors represent proteins in the same cluster in the tree constructed by using the order of five. The Z-value is calculated using the average and the standard deviation of the distribution of the Euclidean distances of proteins in the CE benchmark dataset. Phylip package [68] Fitch-Margoliash method is used to construct the trees. The length of the stems connecting two proteins represents the distance between them. The distance between 1theB to 1aye and 1theB to 1o0eA in Fig. 6A (Fig. 6B) is 15.95 (17.57) and 28.74 (13.36), respectively.

Table 1. Comparison between the CE based benchmark dataset and the SCOP database.

A. Comparison with the superfamily classification by SCOP

Overlap[a)]	**The number of CE groups (%)**[b)]
0 - 0.1	1 (0.5)
0.1 - 0.2	0
0.2 – 0.3	0
0.3 – 0.4	1 (0.5)
0.4 – 0.5	5 (2.7)
0.5 – 0.6	15 (8.1)
0.6 – 0.7	5 (2.7)
0.7 – 0.8	7 (3.8)
0.8 – 0.9	11 (5.9)
0.9 – 1.0	2 (1.1)
1.0	139 (75.1)

a) The fraction of members of a fold group in the CE based benchmark dataset that overlap with a superfamily in SCOP. When a CE fold group corresponds to multiple SCOP superfamilies, a SCOP superfamily which gives the largest overlap with the CE fold group is used to compute the fraction.

b) The percentage among the 185 CE fold groups.

B. Comparison with the fold classification by SCOP

Overlap	**The number of CE groups (%)**
0 - 0.1	1 (0.5)
0.1 - 0.2	0
0.2 – 0.3	0
0.3 – 0.4	0
0.4 – 0.5	1 (0.5)
0.5 – 0.6	10 (5.4)
0.6 – 0.7	4 (2.2)
0.7 – 0.8	8 (4.3)
0.8 – 0.9	8 (4.3)
0.9 – 1.0	2 (1.1)
1.0	152 (82.2)

Table 2. Summary of the structure retrieval using different distance metrics.

	Top1			Top5			Top10			Average rank [d]		Average distance	
	Proteins [a]	Groups1 [b]	GroupsAll [c]	Proteins	Groups1	GroupsAll	Proteins	Groups1	GroupsAll	Proteins	Groups [e]	Proteins	Groups
Euclidean	1881 (77.3%)	178 (96.2)	48 (25.9)	2179 (89.6)	182 (98.4)	72 (38.9)	2264 (93.1)	184 (99.5)	91 (49.2)	6.19	9.79	8.31	9.29
Manhattan	1846 (75.9)	177 (95.7)	41 (22.2)	2165 (89.0)	181 (97.8)	70 (37.8)	2257 (92.8)	183 (98.9)	88 (47.6)	6.33	10.07	71.92	80.04
Correlation Coefficient	1873 (77.0)	179 (96.8)	45 (24.3)	2176 (89.5)	183 (98.9)	71 (38.4)	2265 (93.1)	183 (98.9)	92 (49.7)	6.82	10.79	0.02	0.02
DaliLite [f]	307 (12.6)	49 (2.0)	0 (0.0)	696 (28.6)	85 (3.5)	1 (0.0)	897 (36.9)	104 (4.3)	1 (0.0)	108.22	183.07	19.90	24.86
Random [g]	117 (4.8)	36 (19.5)	0 (0.0)	508 (20.9)	89 (48.1)	0 (0.0)	806 (33.1)	122 (65.9)	1 (5.4)	54.2	87.17	0.09	0.14

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) A group is counted if at least one member in the group successfully retrieved another member in the group as the first position, within top 5, or top 10. In the parentheses, the percentage among all the 185 groups in the benchmark set is shown.
c) A group is counted only if all the members in the group successfully retrieved another member in the group as the first position, within top 5, or top 10.
d) The average rank and distance of the closest structure judged by the distance metric to the query.
e) The rank of proteins is first averaged within a group, then averaged across the groups.
f) DaliLite Version 2.4.4. was used. The distance d is defined as d = 100 – (the structure similarity Z-score by DaliLite).
g) A random value between 0 and 1 is assigned as the distance between the query to each protein.

Table 3. Pairs of proteins which have similar surface shape defined by 3D Zernike descriptor.

PDB ID		Function		SCOP Classification		Length (aa)		Seq. Id.(%) [a]	CE			3D Zernike [c]		
A	B	A	B	A	B	A	B		RMSD (Å)	Z-score	Aligned region (%) [b]	dE	dM	dC
1a31	1cy0	DNA topoisomerase I (human)	DNA topoisomerase I (*E. coli*)	d.163.1.2	e.10.1.1	457	534	5.8	6.3	3.9	79 (17.3)	5.58	49.9	0.001
1tbp	1t7p	TATA-binding protein	DNA polymerase	d.129.1.1	e.8.1.1	180	662	2.0	4.9	4.4	64 (35.6)	7.25	58.6	0.08
1b3t	1adv	Nuclear DNA binding protein EBNA1	Adenovirus DNA binding protein	d.58.8.1	g.51.1.1	147	287	9.0	6.7	1.6	64 (43.5)	7.65	69.4	0.28
2nwl	2bbh	Glutamate transporter	CorA Mg2+ transporter	N/A [d]	d.328.1.1	422	244	5.7	8.1	2.3	88 (36.1)	6.04	53.6	0.001
2b2i	2cfp	Ammonium transporter	Lactose permease	N/A	f.38.1.2	399	417	7.8	4.9	4.4	102 (25.6)	7.28	58.6	0.08

a) The sequence identity between the two proteins.
b) The percentage of the aligned residues relative to the shorter protein among the two.
c) The Euclidian (dE), the Manhattan (dM), and the Correlation coefficient-based (dC) distance of the 3D Zernike descriptor.
d) Not included in the current SCOP (ver. 1.73).

Table 4. Execution time (in seconds).

Grid Size[a)]	64^3 (voxels)	200^3 (voxels)
Surface triangulation[b)]	21	21
Surface voxelization	1	3
3D Zernike descriptor transformation	1	16
Database search [c)]	0.43	0.46
Total	24 (seconds)	41 (seconds)

a), the number of voxels where a protein structure is placed. b), MSROLL program in Molecular Surface Package (ver. 3.9.3) is used. c), the benchmark dataset of 2432 proteins used in the current study is searched.

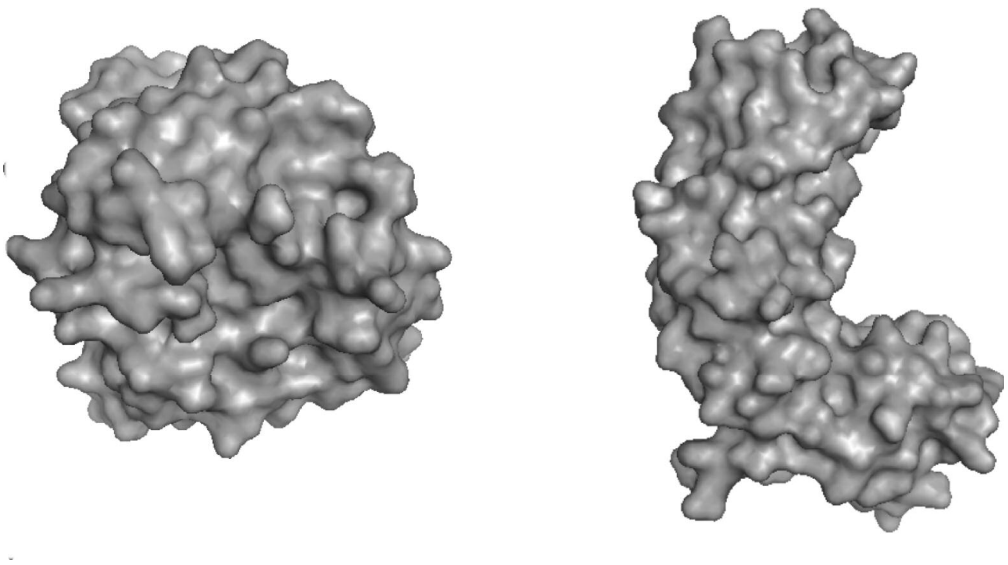

Figure 1A: The global surface shape of the two proteins, 7timA (left) and 1iarB (right).
96x53mm (600 x 600 DPI)

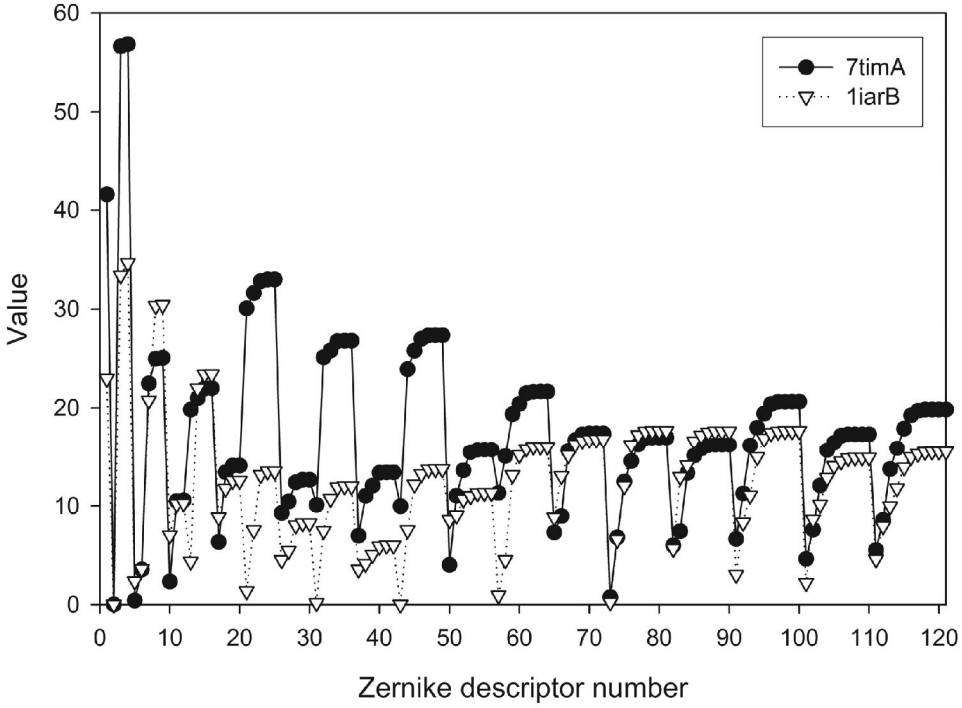

Figure 1B. 3D Zernike descriptor of the two proteins. Circles, 7timA; and triangles, 1iarB.
84x66mm (600 x 600 DPI)

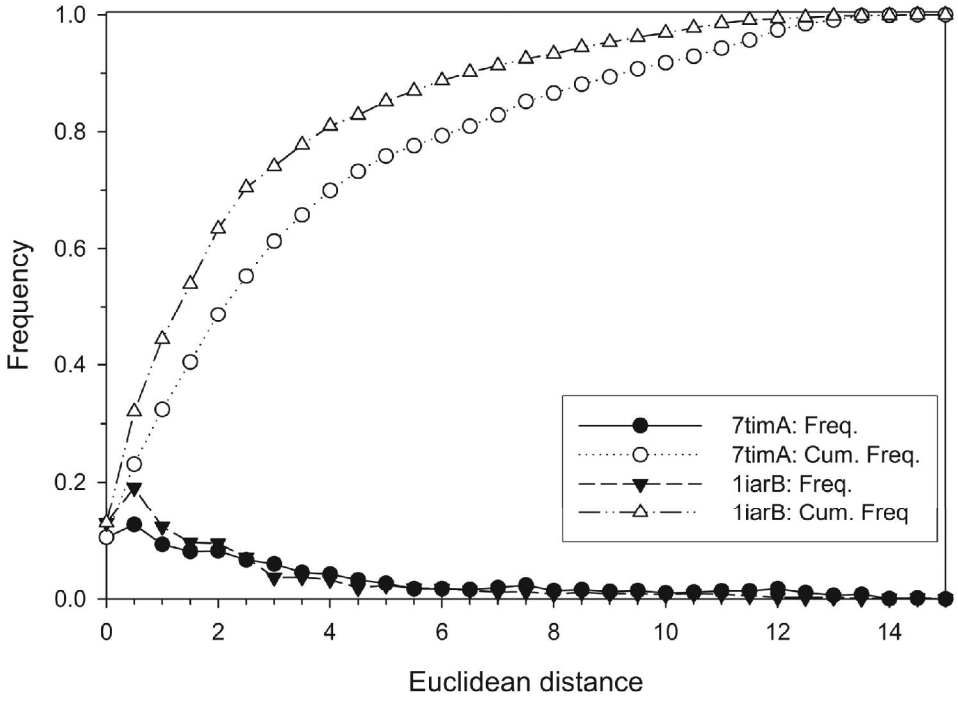

Figure 2A. Variance of 3D Zernike descriptor upon rotation of proteins. All the possible rotated positions of two protein structures, 7timA and 1iarB, in three orthogonal directions with a step size of 30 degrees are computed. Histograms of distances of 3D Zernike descriptor between each of the rotated structures and the original one are shown. A, the Euclidean distance is used. Filled (empty) circles, the frequency (the cumulative frequency) of Euclidean distances of 7timA are plotted. Filled (empty) triangles, the frequency (the cumulative frequency) of Euclidean distances of 1iarB are plotted.

85x66mm (600 x 600 DPI)

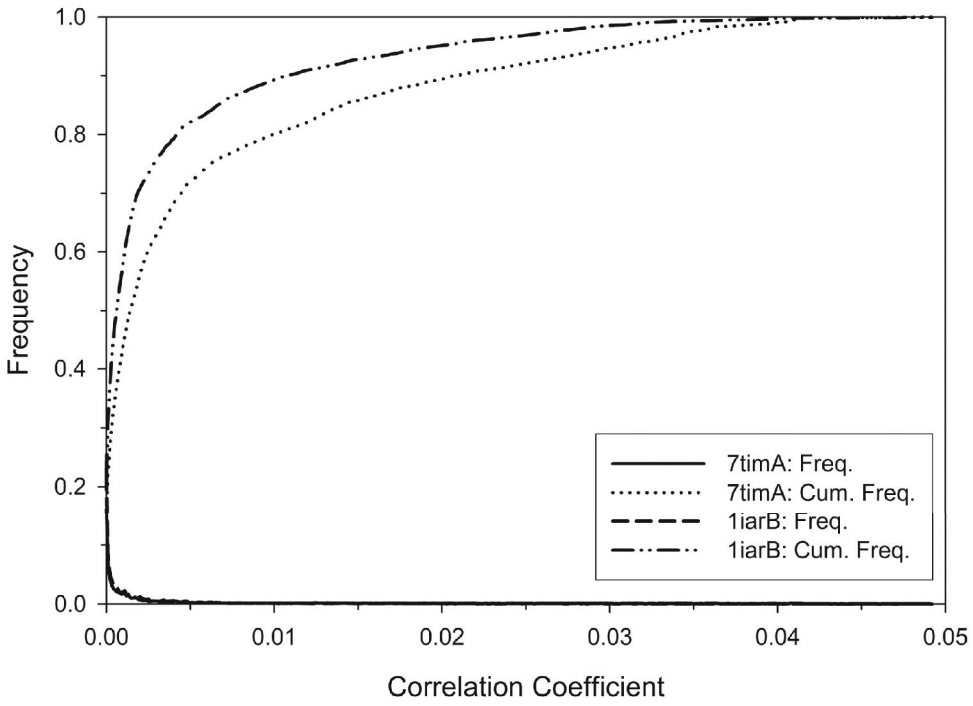

Figure 2B. Variance of 3D Zernike descriptor upon rotation of proteins. All the possible rotated positions of two protein structures, 7timA and 1iarB, in three orthogonal directions with a step size of 30 degrees are computed. Histograms of distances of 3D Zernike descriptor between each of the rotated structures and the original one are shown. Figure 2B, the correlation coefficient is used as the distance measure. The frequency and the cumulative frequency of distances of 7timA and 1iarB are shown by solid line, dotted line, dashed line, dash-dot-dot line, respectively.
85x66mm (600 x 600 DPI)

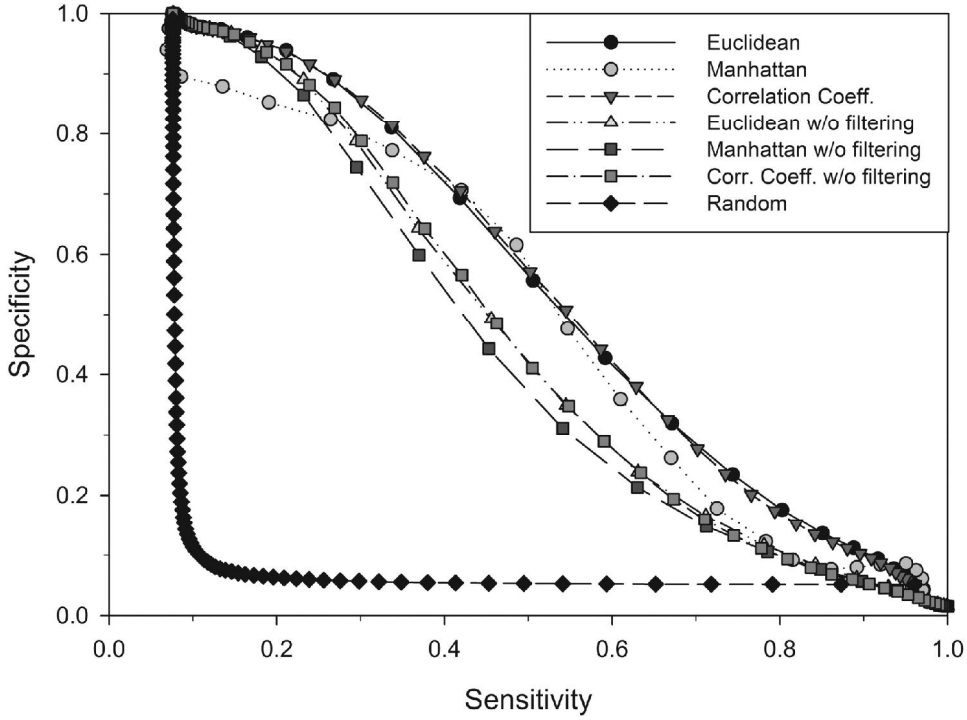

Figure 3. The sensitivity and the specificity of the benchmark dataset are plotted using three distance definitions of 3D Zernike descriptor. the Euclidean (black circles/upward triangles), the Manhattan (gray circles/dark gray squares), and the correlation coefficient-based distance (downward triangles/light gray squares) with and without prescreening by the sequence length. When the prescreening is used, two protein shapes are compared only when the length of the longer one does not exceed 125% of that of the shorter one. For comparison, results of a random retrieval are also plotted (black diamonds).

85x68mm (600 x 600 DPI)

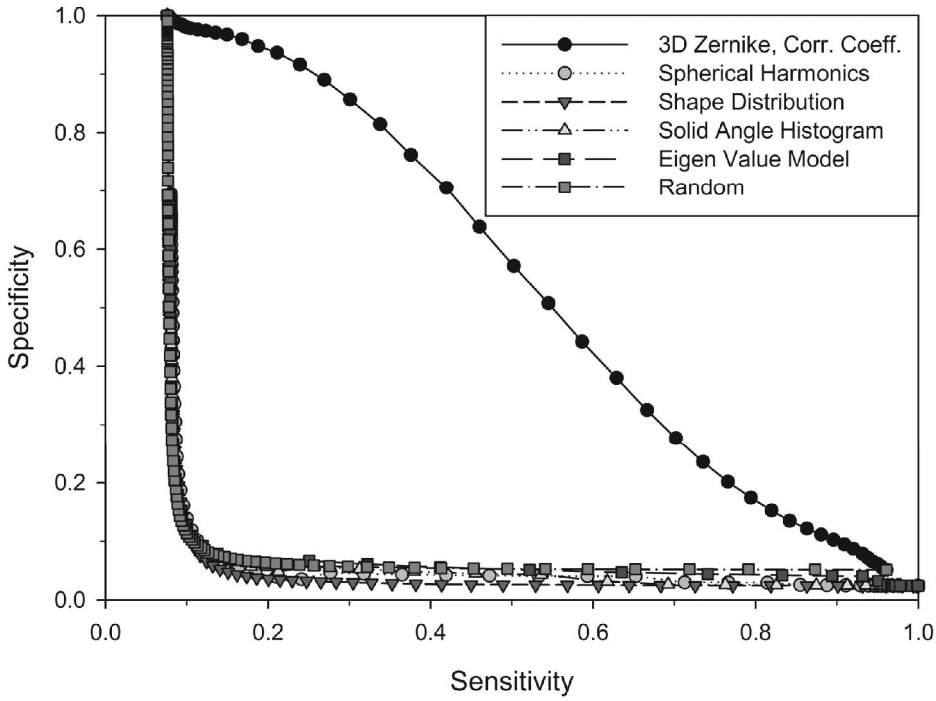

Figure 4. The performance of the 3D Zernike descriptor with the correlation coefficient-based distance (black solid circles) was compared with four other existing shape descriptors, the spherical harmonics descriptor (gray circles), the shape distribution histogram (black triangles), the solid angle histogram (gray triangles), and the eigen value model (black squares). See text for details of these methods. For comparison, the random retrieval is also added (gray squares).
82x64mm (600 x 600 DPI)

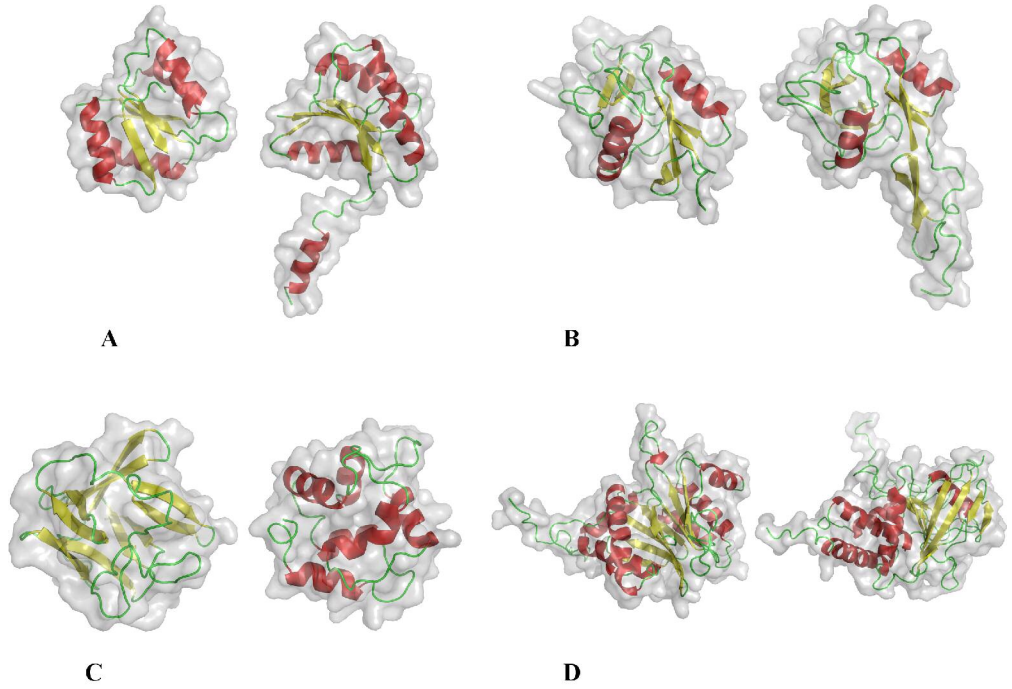

Figure 5. Examples of protein pairs of the same main chain orientation but with a different surface shape (A, B), and pairs with a similar surface shape but with a different main chain orientation (C, D) are shown. Structural comparisons are performed by CE and 3D Zernike. The comparison between A, 1dz3A (response regulator SPO0A) and 1mb0A (response regulator DIVK). CE computes: RMSD=1.6 Å, Z-score=5.0, aligned/gap positions=104/3, Sequence identity (SeqID) = 26.9 %. 3D Zernike shows: Euclidean distance (dE) = 51.21, Manhattan distance (dM) = 438.53, Correlation coefficient (dC) = 0.620. B, 1jznA (galactose-specific C-type lectin) and 1g1qA (P-selectin lectin). CE: RMSD = 2.0 Å, Z-score = 5.9, aligned/gap positions = 115/12, SeqID = 23.5%. 3D Zernike: dE = 52.67, dM = 431.16, dC = 0.602. In contrast, C and D demonstrate two instances in which 3D Zernike detect similar global surface shape of proteins with a different overall fold. C. 1barA (fibroblast growth factor) and 1rro (oncomodulin). CE: RMSD = 6.7Å, Z-score = 1.6, aligned/gap positions = 56/50, SeqID = 3.6 %. 3D Zernike: dE = 12.66, dM = 101.85, dC =0.031. D. 1rypB (proteasome subunit) and 1gwz (Tyrosine phosphatase). CE: RMSD = 5.0 Å, Z-score = 2.3, aligned/gap positions = 72/70 SeqID = 9.7 %, 3D Zernike: dE = 12.73, dM = 108.89, dC = 0.041.

99x69mm (600 x 600 DPI)

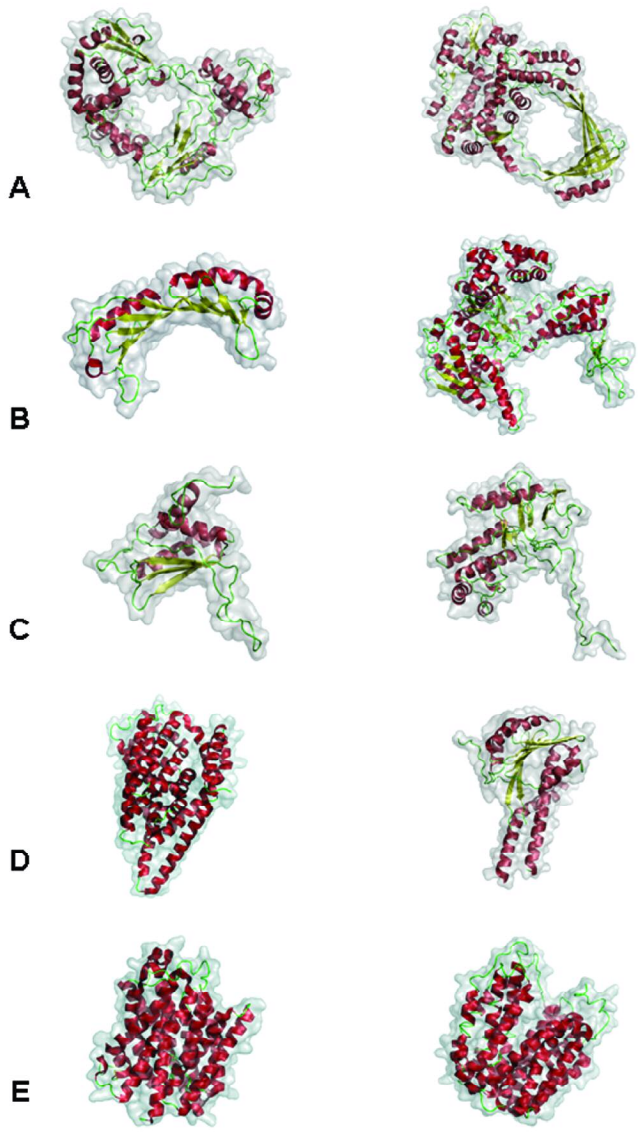

Figure 6. Examples of protein pairs whose surface shapes are judged to be similar by 3D Zernike descriptor. Detailed data are shown in Table 4. A, 1a31 and 1cy0 (from left to right); B, 1tbp and 1t7p; C, 1b3t and 1adv; D, 2nwl and 2bbh; E, 2b2i and 2cfp. Detailed data of these protein pairs are shown in Table 3.
190x254mm (600 x 600 DPI)

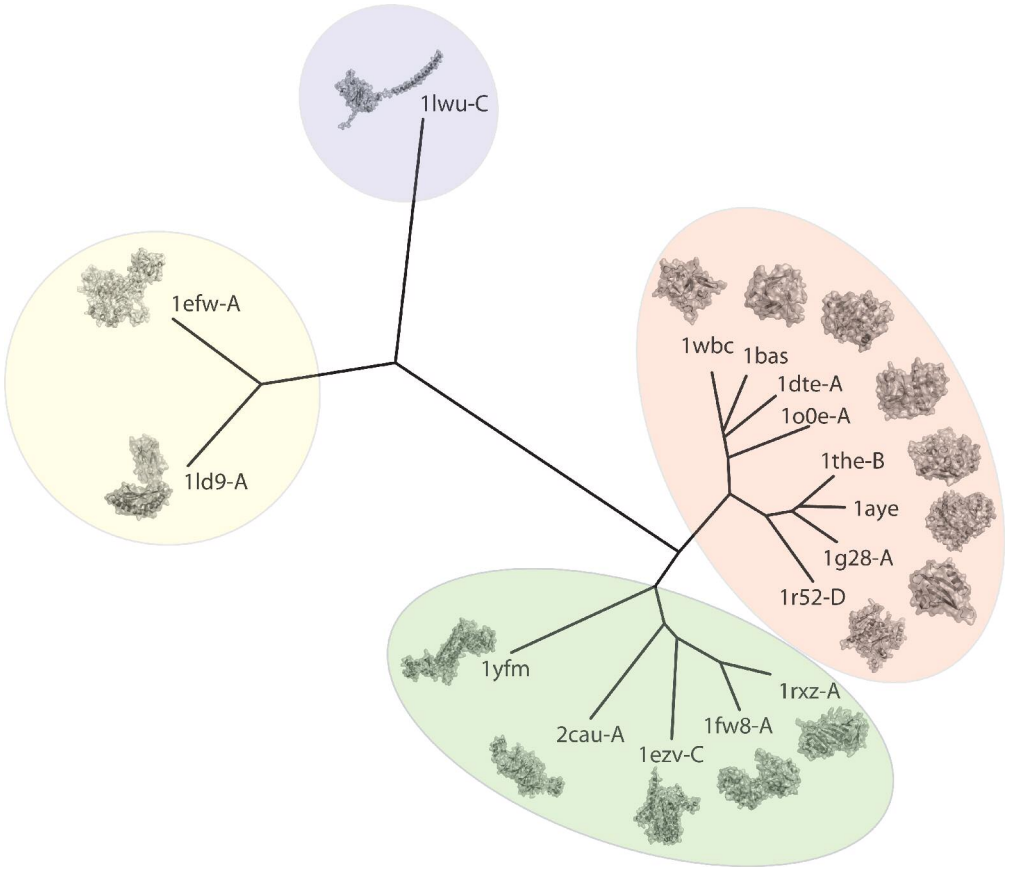

Figure 7A. Resolution of 3D Zernike descriptor. A, 3D Zernike descriptors of the order of five; B, the order of twenty is used to construct trees representing similarity of the surface shape of sixteen proteins: 1theB, 1o0eA, 1dteA, 1aye, 1g28A, 1wbc, 1r52D, 1rxzA, 1fw8A, 2cauA, 1bas, 1ld9A, 1efwA, 1ezvC, 1yfm, and 1lwuC. The Euclidean distance is used. Proteins within a Z-value of distance of 0.35 are grouped in a colored circle. The colors represent proteins in the same cluster in the tree constructed by using the order of five. The Z-value is calculated using the average and the standard deviation of the distribution of the Euclidean distances of proteins in the CE benchmark dataset. Phylip package 68 Fitch-Margoliash method is used to construct the trees. The length of the stems connecting two proteins represents the distance between them. The distance between 1theB to 1aye and 1theB to 1o0eA in Fig. 6A (Fig. 6B) is 15.95 (17.57) and 28.74 (13.36), respectively.

152x131mm (600 x 600 DPI)

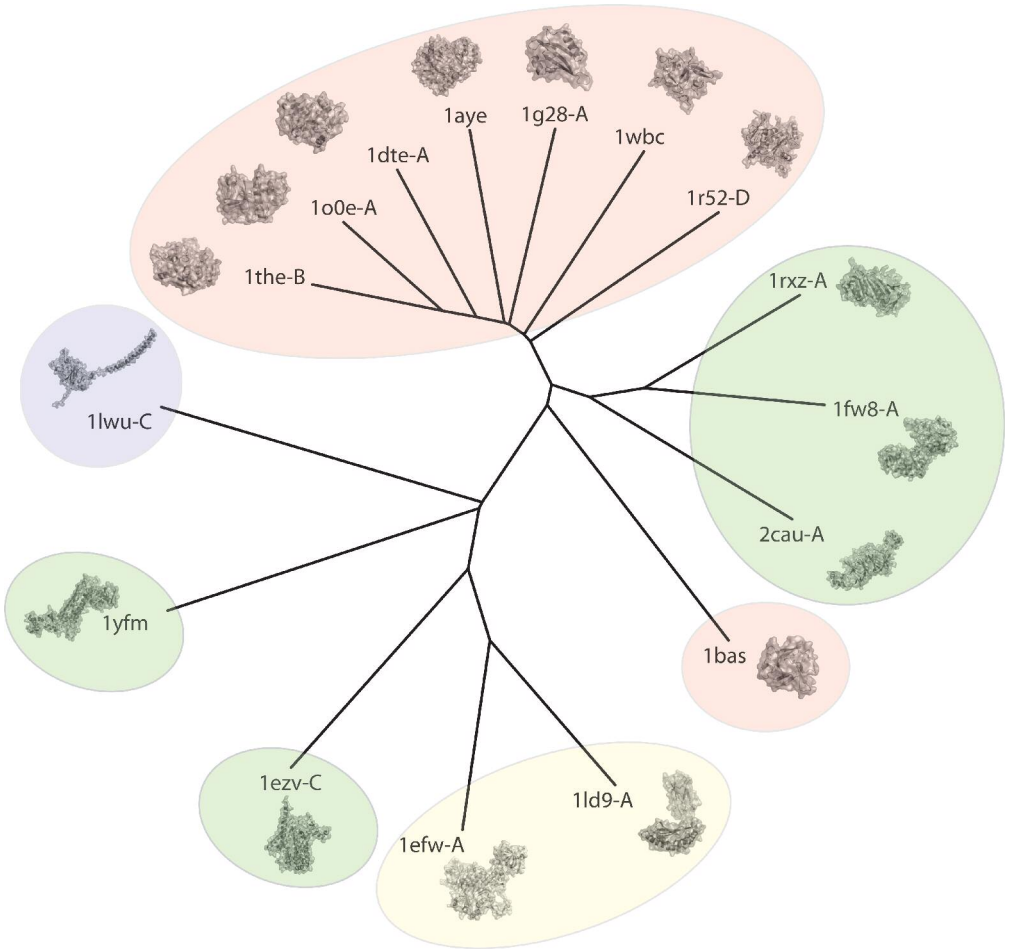

Figure 7B. Resolution of 3D Zernike descriptor. A, 3D Zernike descriptors of the order of five; B, the order of twenty is used to construct trees representing similarity of the surface shape of sixteen proteins: 1theB, 1o0eA, 1dteA, 1aye, 1g28A, 1wbc, 1r52D, 1rxzA, 1fw8A, 2cauA, 1bas, 1ld9A, 1efwA, 1ezvC, 1yfm, and 1lwuC. The Euclidean distance is used. Proteins within a Z-value of distance of 0.35 are grouped in a colored circle. The colors represent proteins in the same cluster in the tree constructed by using the order of five. The Z-value is calculated using the average and the standard deviation of the distribution of the Euclidean distances of proteins in the CE benchmark dataset. Phylip package 68 Fitch-Margoliash method is used to construct the trees. The length of the stems connecting two proteins represents the distance between them. The distance between 1theB to 1aye and 1theB to 1o0eA in Fig. 6A (Fig. 6B) is 15.95 (17.57) and 28.74 (13.36), respectively.

152x143mm (600 x 600 DPI)