PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

SHORT COMMUNICATION

Rapid comparison of properties on protein surface

Lee Sael,[1] David La,[2] Bin Li,[1] Raif Rustamov,[3] and Daisuke Kihara[1,2,4,5]*

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

ABSTRACT

The mapping of physicochemical characteristics onto the surface of a protein provides crucial insights into its function and evolution. This information can be further used in the characterization and identification of similarities within protein surface regions. We propose a novel method which quantitatively compares global and local properties on the protein surface. We have tested the method on comparison of electrostatic potentials and hydrophobicity. The method is based on 3D Zernike descriptors, which provides a compact representation of a given property defined on a protein surface. Compactness and rotational invariance of this descriptor enable fast comparison suitable for database searches. The usefulness of this method is exemplified by studying several protein families including globins, thermophilic and mesophilic proteins, and active sites of TIM β/α barrel proteins. In all the cases studied, the descriptor is able to cluster proteins into functionally relevant groups. The proposed approach can also be easily extended to other surface properties. This protein surface-based approach will add a new way of viewing and comparing proteins to conventional methods, which compare proteins in terms of their primary sequence or tertiary structure.

Proteins 2008; 73:1–10.

Key words: 3D Zernike descriptor; protein surface shape; protein surface physicochemical properties; global/local surface comparison.

INTRODUCTION

The vast number of protein three-dimensional (3D) structures available in the PDB[1] provides an invaluable resource to investigate function, evolution, and the complex relationship among them. Commonly proteins are compared and classified based on their sequence[2] or tertiary structure.[3,4] However, if the main goal is the elucidation of protein function, a more intuitive approach would be to compare protein surface shapes and also physicochemical properties, such as electrostatic potential or hydrophobicity, which can be mapped on to a protein surface. This is also supported by the fact that such properties have a significant role in influencing molecular interactions.[5]

Several methods have been proposed on the comparison of physicochemical properties of proteins. The Carbo and Hodgkin indices,[6,7] which compute inner products of electrostatic potentials of two proteins, have been used to compare electron densities, electrostatic potentials, and electrostatic fields of small molecules.[6–9] Wade *et al.* have developed a method which compares the electrostatic potential on the surrounding region of a protein, which is less sensitive to small changes in protein structure.[10] Their method was used to analyze the relationship of electrostatic potentials and biological function of several proteins and ligand molecules, including plecstrin homology domain family, blue copper proteins, and proteins in the ubiquitination pathway.[10–14] Pawlowski and Godzik[15,16] introduced a method that maps pro-

Grant sponsor: National Institutes of Health; Grant numbers: R01 GM075004, U24 GM077905; Grant sponsor: National Science Foundation; Grant number: DMS 0604776.
*Correspondence to: Daisuke Kihara, Department of Biological Sciences, Purdue University, Lilly Hall, 915 West State St., West Lafayette, IN 47906. E-mail: dkihara@purdue.edu
Received 7 March 2008; Revised 14 April 2008; Accepted 9 May 2008
Published online 10 July 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22141

tein surface properties to a unit sphere. This method is advantageous in the sense that it is less sensitive to noise and that it can map multiple properties. However, as the nature of the sphere mapping, cavities on the protein surface are not properly represented. Moreover, a major drawback of these approaches is that proteins have to be prealigned using some scheme in order to determine corresponding regions. This step is often time consuming and may not always yield a unique solution especially in cases where structural similarity is low. The multiresolution attributed contour tree method compares relative positions of electrostatic potentials on protein surfaces as a tree and does not needs any prior alignment.[17] However, its performance is offset by the high time complexity owing to the tree matching and is found to be only as effective as the Carbo index.

Here, we propose a novel method for comparing properties defined on the protein surface using the 3D Zernike descriptor (3DZD).[18,19] This descriptor is a series expansion of an input 3D function with several interesting features. First, a given property can be represented by a short vector of numbers (the coefficients of terms in the series) making it amenable to fast comparison. Also, the prealignment step that is crucial to some of the methods can be avoided as the 3DZDs are rotation and translation invariant. This enables the comparison of nonhomologous proteins, where structure alignment can be unreliable. Another important feature is that the resolution of the representation can be easily altered by simply changing the order of 3DZD.

In our previous paper,[19] we have demonstrated that the 3DZD effectively captures protein surface shape similarity and is capable of searching similar protein structures from a PDB database in a real-time (http://dragon.bio.purdue.edu/3d-surfer/). We have also shown examples of proteins whose overall surface shape is similar due to their functional relevance, but do not share significant sequence or backbone conformation similarity. In this article, following the description of data set and methods, we first present a simple example of unit spheres to demonstrate that different property patterns can be discriminated by the 3DZD. We then extend this to more biological context and show that the descriptors can quantitatively compare electrostatic and hydrophobic properties of evolutionary diverse protein families.

MATERIALS AND METHODS

Data sets

The representative set consists of 184 protein structures, each of which are arbitrarily selected from different fold groups defined in a protein classification database by the Combinatorial Extension (CE) program.[4] These structures have a crystallographic resolution of 3.0 Å or higher, have no more than 10 missing residues in the structure solved, have all heavy atom positions solved, and are longer than 100 residues. Also, the structure similarity of each pair is less than a *Z*-score of 3.8 by CE.

The 43 globin structures are selected from globin family structures in the SCOP database.[20] They have less than 70% sequence identity with each other. The 19 TIM barrel proteins are selected, one from each family classified in Table 2 by Nagano *et al.*[21] A binding site of a TIM barrel protein is defined as the surface region that is closer than 3.5 Å to any atom of its ligand.

Surface representation

Protein surfaces are calculated using the Connolly molecular surface package (MSP).[22] Surface is then made discrete by placing it on a cubic grid. To represent a surface shape, each grid cell (voxel) is assigned 1 if it is on the surface and 0 otherwise. Values of other physicochemical properties, such as the electrostatic potentials, are also assigned only to the surface voxels. The electrostatic potentials are computed by APBS[23] and hydrophobicity values are taken from the eF-site database.[24] The resulting voxels with values on them are considered as a 3D function, $f(x)$, which is expanded into the 3DZD as described in the next section.

3D Zernike descriptor

3DZD is a series expansion of a 3D function, which allows a compact representation of a 3D object (i.e., a 3D function). Mathematical foundation of the 3DZD was laid by Canterakis[25] then Novotni and Klein[18] have applied it to 3D shape retrieval. Below we provide brief mathematical derivation of the 3DZD. See the two papers[18,25] for more details.

To obtain a 3DZD, given 3D function $f(\mathbf{x})$ is expanded into a series in terms of Zernike-Canterakis basis[18] defined by the collection of functions

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

with $-l < m < l$, $0 \le l \le n$, and $(n - l)$ even. Spherical harmonics,[26] $Y_l^m(\vartheta,\varphi)$, is the angular portion of an orthogonal set of solutions to Laplace's equation, which is given by:

$$Y_{i}^{m}(\vartheta,\varphi) = N_{l}^{m}P_{l}^{m}(\cos\vartheta)e^{im\varphi}, \tag{2}$$

where N_l^m is a normalization factor,

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

and P_l^m is the associated Legendre function. $R_{nl}(r)$ are radial functions defined by Canterakis, constructed so that $Z_{nl}^m(r,\vartheta,\varphi)$ are polynomials when written in terms of Cartesian coordinates. $Z_{nl}^m(r,\vartheta,\varphi)$, which are currently written in spherical coordinates, are converted into Cartesian coordinate functions $Z_{nl}^m(\mathbf{x})$ in the following three steps:

1. The conversion between spherical coordinates, (r,ϑ,φ), and Cartesian coordinates, $\mathbf{x} = (x,y,z)$, is defined as

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

2. Using Eq. (4), we define a function e_l^m in Cartesian coordinates, which is later used for rewriting the 3D Zernike function [Eq. (1)] into Cartesian coordinates. The harmonics polynomials e_l^m are defined as

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

where c_l^m are normalization factors

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

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

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

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

$$q_{kl}^v = \frac{(-1)^k}{2^{2k}}\sqrt{\frac{2l+4k+3}{3}}\binom{2k}{k}(-1)^v \times \frac{\binom{k}{v}\binom{2(k+l+v)+1}{2k}}{\binom{k+l+v}{k}}. \quad (8)$$

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

$$\Omega_{nl}^m = \frac{3}{4\pi}\int_{|\mathbf{x}|\le 1} f(\mathbf{x}) Z_{nl}^m(\mathbf{x})\,dx. \quad (9)$$

Finally, the moments are collected into $(2l + 1)$ dimensional vectors $\Omega_{nl} = (\Omega_{nl}^l, \Omega_{nl}^{l-1}, \Omega_{nl}^{l-2}, \Omega_{nl}^{l-3}, \ldots, \Omega_{nl}^{-l})$ and the rotational invariance is obtained by defining 3DZD, F_{nl}, as norms of vectors Ω_{nl}:

$$F_{nl} = \sqrt{\sum_{m=-1}^{m=1} (\Omega_{nl}^m)^2} \quad (10)$$

The parameter n is called the order of 3DZD. The order determines the resolution (i.e., the number of terms in the series expansion) of the descriptor. As stated earlier, n defines the range of l, and a 3DZD is a series of invariants [Eq. (10)] for each pair of n and l, where n ranges from 0 to the specified order. For example, n ranges from 0 to 20 for a 3DZD of an order of 20. We use $n = 20$, which yields a total of 121 invariants, because it is shown to provide sufficient accuracy in a previous works of shape comparison.[18] The rotational invariance of 3DZD means that calculating F_{nl} for a protein and its rotated versions would yield the same descriptor.

As for the surface electrostatic potentials, 3DZD is computed separately for the pattern of positive values and for the negative values and later combined in the following way: First, voxels with a positive electrostatic potential value are kept but all the other voxels with a negative electrostatic potential value are reset with a value of zero. Then 3DZD of the pattern of the positive values in the cubic grid is computed. Next, similarly, voxels with a negative electrostatic potential value are kept but all the other voxels are reset with a value of zero. Then 3DZD of the pattern of the negative values is computed. Then, the two 3DZDs, one for voxels with a positive value and another one for voxels with a negative value are combined, yielding a descriptor with 2×121–242 invariants. This is because Eq. (10) does not differentiate positive and negative values, but only a pattern of non-zero values in the 3D space, as will be seen in Figure 1. Finally, we normalize numbers in a descriptor by the norm of the descriptor. This normalization is found to reduce dependency of 3DZD on the number of voxels used to represent a protein.

Similarity measures

Two distances are used to compare 3DZDs. The Euclidean distance (EUC) is defined as

$$\mathrm{EUC} = \sqrt{\sum_{i=1}^{i=N} (z_{Ai} - z_{Bi})^2} \quad (11)$$

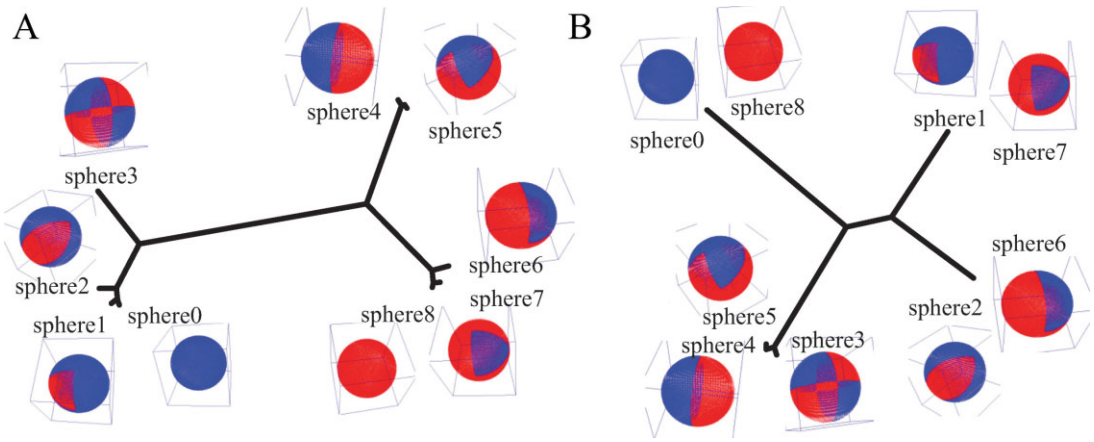

Figure 1

*Analysis of coloring patterns on spheres using complete linkage clustering. (**A**) the clustering results when the 3DZD of the areas of positive values and the negative values are separately computed (thus 242 invariants are used); (**B**) the result using the original 3DZD (i.e., 121 invariants are used).*

where Z_{Ai} and Z_{Bi} is the i-th invariant of 3DZD of protein A and B, respectively. N is set to 121 for comparing surface shape and 242 for comparing surface electrostatics of the two proteins. The correlation coefficient based distance (CC), is defined as

$$\mathrm{CC} = 1 - r(Z_A, Z_B), \tag{12}$$

where $\{r | -1 \leq r \leq 1\}$ is the correlation coefficient of two 3DZDs. CC is 0 when two 3DZDs correlate perfectly, that is, when they have the correlation coefficient of 1. Both EUC and CC range from 0 to 2.

The similarity indices by Hodkins and Carbo were calculated using the "similar" program in the APBS package. To make the range consistent with EUC and CC, we modified the distance as follows:

$$\mathrm{CL2} = 1 - SI_{\mathrm{Carbo}}(A, B) \tag{13}$$

$$\mathrm{HL2} = 1 - SI_{\mathrm{Hodgkins}}(A, B) \tag{14}$$

HL2 and CL2 are modified L2 inner products of voxel values which equals to 0 when identical, 1 when unrelated, and 2 when opposite.

Calculation speed

Comparison of two 3DZDs is much faster than HL2 and CL2 indices, as the latter considers the whole grid while the former only evaluates coefficients. Pairwise comparison of 3DZD takes about 0.05 sec while HL2 or CL2 takes 50 sec on an average when the grid size is set to 193.[3] Preprocessing consists of running APBS and computing the 3DZD. Calculating electrostatics for a protein using APBS takes ~1.5 min, and computing a single 3DZD from an APBS output file takes under a 1 min.

RESULTS

Clustering of color patterns on spheres

Similarity assessments combining both surface shape and property provide a meaningful comparison of proteins. Previous works have shown that 3DZDs can be used for identifying global molecular surface shape similarity.[19,27] Here to examine the effectiveness of 3DZD in distinguishing different surface properties, we cluster nine unit spheres with different color patterns.

The surface of a unit sphere is equally partitioned into eight sections and voxels of each section are assigned a value of either +1.0 (blue) or −1.0 (red). The nine spheres, S0–S8 are distributed according to the complete linkage clustering method using CC [Eq. (12)] of 3DZDs [Fig. 1(A)]. These spheres are primarily clustered by the number of positive voxels relative to the number of negative voxels, that is, S0 to S8 are arranged in the decreasing order of blue sections on the spheres. However, an interesting cluster is observed with S3, S4, and S5. The total area of blue sections of the three spheres is the same. However, S4 is more similar to S5 in that both S4 and S5 are partitioned into two areas while S3 is partitioned into four. In contrast, naïve application of the original 3DZD[18] only recognizes the contrast of patterns of the positive and the negative value but not the value itself [Fig. 1(B)]. Thus it is not able to distinguish S0 from S8 or S1 from S7. Our approach is able to obtain

the clustering of the nine spheres by combining two 3DZDs separately computed for blue and red sections.

Distance distribution of random protein pairs and the globin family

Next, in order to obtain an idea of the distance distribution of proteins using the 3DZD, we have performed a comparison study on 184 representative proteins and 43 globins. The CC and the EUC [Eq. (11)] of the surface shape, the electrostatic potential, and the hydrophobicity are shown in Figure 2. The globin family is used because they are known to have a conserved fold with a wide variation in sequence[28] and function[29,30]; thus high similarity in surface shapes but some diversity in electrostatic potential and hydrophobicity were expected. The root mean square deviation (RMSD) of pairs of the 43 globins ranges from 0.57 to 3.66 Å computed by the CE program and the sequence identity from 6.2 to 69.0%.

The CC of surface shape of the representative proteins ranges from 0.008 to 0.723 with the peak value of 0.095 [Fig. 2(A)]. The majority of the protein pairs have a small distance because they are globular and compact.

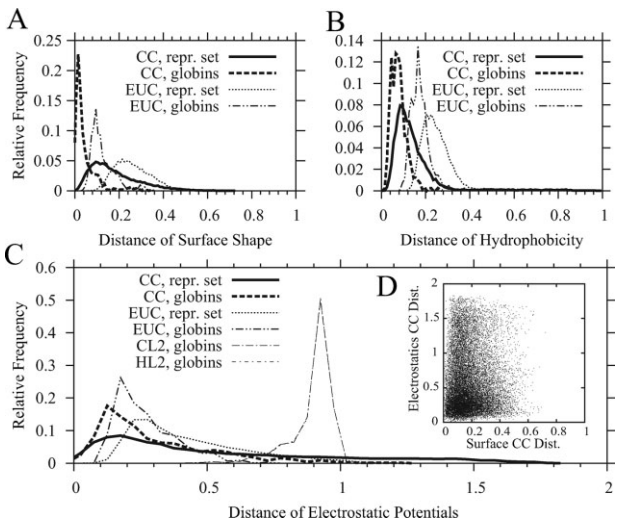

Figure 2

*Distribution of the 3DZD distances using representative (repr.) protein set and the globins. (**A**) the distance of protein surface shape. The minimum (min) and maximum (max) value of the CC of the repr. set and the globin set is (0.008, 0.723) and (0.000, 0.354), respectively. The min and max value of the EUC of the representative set and the globin set is (0.059, 0.546) and (0.001, 0.366), respectively. (**B**) the distances of surface hydrophobicity. The (min, max) value of the CC of the repr. set, the globin set, the EUC of the repr. set, and the globin set is (0.0188, 0.993), (0.0267, 0.274), (0.107, 0.760), and (0.099, 0.384), respectively. (**C**) the distance of the surface electrostatic potentials. The (min, max) value of the CC of the repr. set, the globin set, the EUC of the repr. set, and the globin set is (0.011, 1.829), (0.043, 1.275), (0.090, 1.241), and (0.151, 0.787), respectively. The distribution of the CL2 and HL2 are also shown. HL2 and CL2 happened to have the almost identical distribution. The (min, max) value of the CL2 (HL2) of globins is (0.473, 1.019). (**D**) Correlation of the distance of surface shape and electrostatic potential of proteins in the representative set.*

The hydrophobicity has a narrower distribution [Fig. 2(B)]. This is partially due to the residue-based assignment of hydrophobicity values using the Kyte-Doolittle scale,[31] which does not produce much diversity in patterns compared to that of electrostatic potential. The CC of the electrostatic potentials [Fig. 2(C)] does not have a strong correlation to that of the surface shape [Fig. 2(D)]. This result is advantageous for comparing proteins because a combination of shape and electrostatics can provide a hierarchical classification of protein surfaces.

Compared to the representative set of proteins, the surface shape of globins is significantly more similar to each other [Fig. 2(A)]. In contrast, the electrostatic potentials of globins show less variability as compared to that of the representative set but distributed over a wide range. This is suggestive of the diversity of functions within the globin family [Fig. 2(C)]. The distance distributions of CL2 and HL2 [Eqs. (13) and (14)] are also shown in Figure 2(C). Note that both CL2 and HL2 rely on superimposition, which in the case of the representative set could not be performed given the poor structural similarity. The CL2 and HL2 have a skewed distribution near 1.0, which means that there is very low similarity between the compared globin proteins, implying that these two metrics are sensitive to the difference in the electrostatic potentials of the globin proteins. Structural superimposition of globins for computing CL2 and HL2 is computed by PyMol in the data shown in Figure 2(C). We also used CE for the superimposition, which yielded essentially the same distribution as Figure 2(C). Therefore, CL2 and HL2 may be suitable to compare electrostatics of very closely related proteins, but difficult to provide meaningful similarity metrics for more general use.

Using the representative protein set, we have examined how changing the order of 3DZDs affects to calculated distances of the surface electrostatic potentials. Figure 3 shows histograms of the CC distance of the electrostatic potentials of pairs of proteins in the representative set using different orders. Significantly different histograms are obtained when 3DZDs with a smaller order is used (e.g., the order of less than 10) but the histograms almost converge when a higher order is used. Therefore, it will not be necessary to use further higher order in describing the surface electrostatic potentials.

Figure 4 shows examples of the surface electrostatics and 3DZD of several globin proteins. First, the pair of globins that exhibit the largest electrostatic potential difference according to 3DZD are shown [Fig. 4(A)]. 1h97A and 1hbg are monomeric hemoglobins from different organisms which are known to display extremely different oxygen affinity because of the different disposition of amino acids in their heme binding pockets.[32,33] This apparent difference in the surface electrostatics [Fig. 4(A), top panel] is captured by the 3DZD (the middle panel). The higher value in the first 121 invariants of 1hbg indicates dominance of the positive electrostatic

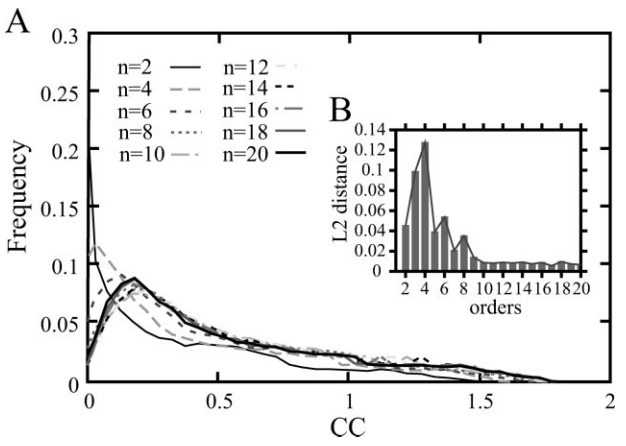

Figure 3

*Histograms of the CC of the electrostatic potentials of the representative protein set using different number of orders. (**A**) histograms of 10 orders are used, ranging from 2 to 20. (**B**) the difference of pairs of histograms of adjacent orders, that is, 1 with 2, 2 with 3, 19 with 20 are shown using the L2 distance. The L2 distance of two histograms is the average difference of frequencies at each bin.*

potential on the surface. Likewise, 1hbg has more hydrophobic regions than 1h97A, which is conveyed by the higher values in the first 121 invariants in the third panel of Figure 4(A).

Figure 4(B) further shows examples of globin proteins of diverse functions. 1gdi is a leghemoglobin from *Lupinus luteus*, that functions as a monomer in root nodules to regulate oxygen by the nitrogen-fixing aerobic bacteria.[34] 1myt is a monomeric myoglobin from yellowfin tuna. 1oj6A is human brain neuroglobin, whose role is to sustain oxygen supply at highly oxygen-demanding and metabolically active cells like neurons.[35] Besides, it has a larger binding cavity that displays a hexa-coordinated heme. It has a more negative surface electrostatic potential, which produces a large 3DZD distance value for 1myt and 1ux8A. 1ux8A belongs to the group II truncated oxygen-avid hemoglobin from *Bacillus subtilis*. Truncated hemoglobin is about 20 residues shorter than the full length globins.[36] This structural difference is reflected in the large RMSD values obtained for the other three globins. The four globins shown here are all monomeric, but have a distant evolutionary relationship (i.e., the sequence identity between them is low), a varied range of affinity to oxygen, and different functions and also are located in different environments. These differences coincide with the relatively large distance of surface electrostatic potentials measured by 3DZD.

Thermophilic and mesophilic proteins

Thermophilic proteins have gained substantially higher thermal stabilities as compared to their mesophilic orthologs, and the underlying principle for the stability has been a subject of intensive discussion for the past few years.[37–39] Among the different properties studied, electrostatic contributions, especially surface electrostatics, has been identified as one of the major stabilization factors.[37,39] Therefore, a method for robust and quantitative comparison of surface electrostatics will lead to a better understanding of thermostability of proteins. Here, as a demonstration that 3DZD can cluster proteins into functionally relevant groups, surface electrostatics of a total of 14 thermophilic and mesophilic proteins from three families are compared: the dihydrofolate reductase (DIR) family, the glutamate dehydrogenase (GDH) family, and the TATA box binding protein (TBP) family.

The three DIRs [Fig. 5(A)] are interesting examples where the similarity based on surface electrostatics is not inferred from either sequence or structure. 1dyjA

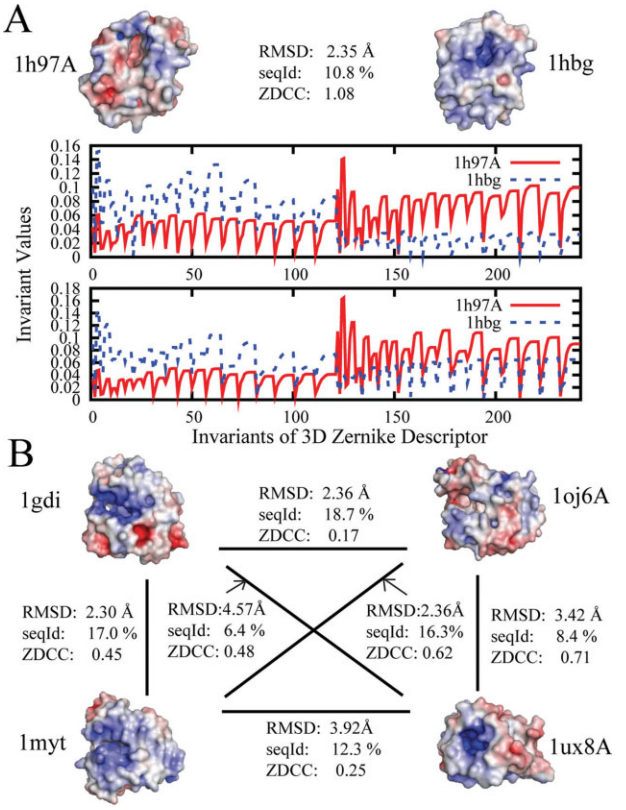

Figure 4

*Distances based on the electrostatic potentials of globin proteins. (**A**) the pair of globin proteins with the largest CC of the electrostatic potentials. Monomeric hemoglobin of Paramphistomum epiclitum (1h97A) and Glycera dibranchiata (1hbg). Positive and negative electrostatic potentials on the protein surface are represented in blue and red, respectively. The RMSD (Å) of the main-chain conformation; the sequence identity(%) (SeqId); and the CC of the surface electrostatic potentials by 3DZD of the two proteins are shown. The 3DZD invariants of the electrostatic potentials and the hydrophobicity of the two proteins are shown in the middle and the bottom panel, respectively. (**B**) four globin proteins of diverged evolutionary distance, leghemoglobin from Lupinus luteus (1gdi), human brain neuroglobin (1oj6A), myoglobin from yellowfin tuna (1myt), and group II truncated hemoglobin from Bacillus subtilis (1ux8A) are shown with the mutual RMSD, seqId, and the 3DZD CC of the electrostatic potentials.*

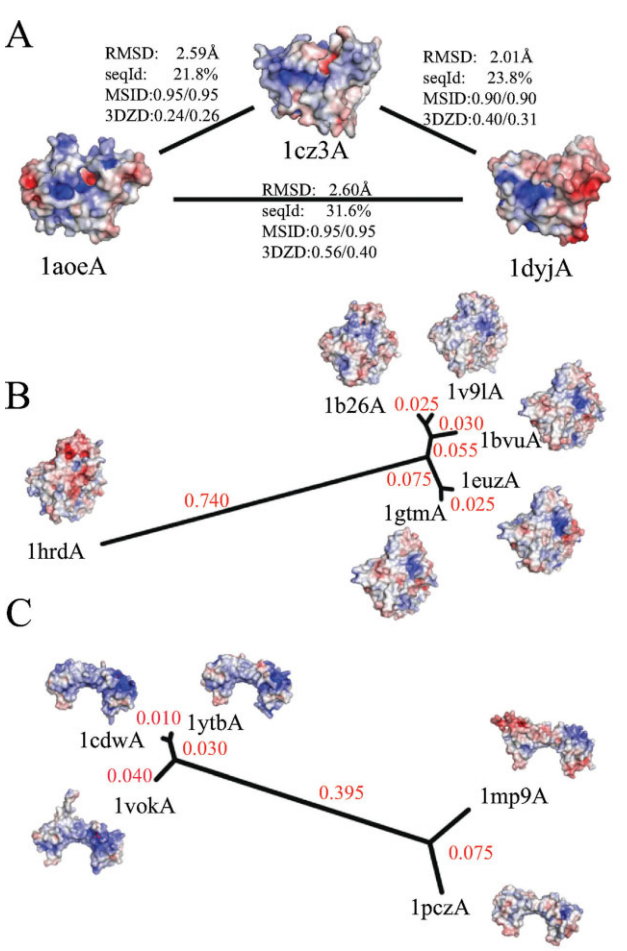

Figure 5

*Comparison of surface electrostatic potentials of thermophilic and mesophilic proteins. (**A**) RMSD, SeqId, modified similarity index based distances (MSID: HL2/CL2), and 3DZD distances (3DZD: CC/EUC) of the surface electrostatic potentials of three DIR family proteins. 1cz3A from a thermophilic organism, Thermotoga maritima; 1aoeA and 1dyjA from mesophilic organisms, Candida albicans and Escherichia coli, respectively. (**B**) 3DZD CC distances of the surface electrostatic potentials of proteins GDH family. 1hrdA is a mesophilic protein from Clostridium symbiosum, and the rest are thermophilic proteins: 1b26A (Thermotoga maritima), 1v9lA (Pyrobaculum islandicum), 1bvuA (Thermococcus litoralis), 1euzA (Thermococcus profundus), and 1gtmA (Pyrococcus furiosus). Completely linkage clustering is used. The distance is indicated in red on branches. (**C**) Proteins of TBP family: two thermophilic proteins, 1mp9A (Sulfolobus acidocaldarius) and 1pczA (Pyrococcus woesei) with three mesophilic proteins, 1ytbA (Saccharomyces cerevisiae), 1cdwA (human), and 1vokA (Arabidopsis thaliana).*

has a larger region with negative electrostatic potentials (colored in red), which is represented by the larger 3DZD CC distance to 1aoeA and 1cz3A. This characteristic surface electrostatics of 1dyjA is not obvious by sequence or structure similarity, as the largest sequence identity is observed between 1dyjA and 1aoeA and the smallest RMSD is observed between 1dyjA and 1cz3A. The CL2 and HL2 again fail to provide meaningful values.

Figure 5(B) shows that the 3DZD of the surface electrostatic potentials clearly discriminates a mesophilic homolog of the GDH family (1hrdA) from the other five thermophilic proteins. All these six proteins exhibit significant sequence similarity ranging from 87% to 33%. Despite the high sequence similarity, 1hrdA has a distinct surface electrostatics compared with the thermophilic proteins with the average CC of 0.714. The thermophlic proteins are very similar in terms of the surface electrostatic potentials, having an average CC of 0.108. The classification of thermophilic and mesophilic homologs of the TBP family is shown in Figure 5(C). The three mesophilic proteins, 1ytbA, 1cdwA, and 1vokA are well clustered with significantly small CC distance. The sequence identity for the two thermophilic proteins 1mp9A and 1pczA is 44.8% while that for 1mp9A and 1ytbA (mesophilic) is 45.0%. One can see that sequence identity alone is unable to distinguish between the two types. Similarly, clustering using HL2 for GDH [Fig. 6(A)] and TBP [Fig. 6(B)] does not show clear separation between the thermophilic and mesophilic proteins. Figure 6(A) indicates that some thermophilic protein pairs (e.g., 1euzA and 1v9lA) are as distant to each other as against 1hrdA. Figure 6(B) shows that the two thermophilic proteins, 1mp9A and 1pczA are very distinct; indeed more distant than between 1mp9A and 1vokA. 3DZD on the other hand provides a clear delineation of the protein families.

Local active sites of TIM barrel proteins

The TIM β/α barrel are one of the most prevalent folds adopted by a variety of enzymes.[21] Active sites of TIM barrel enzymes, which are usually located at the cleft with cluster of loops of the barrel, show wide ranging behavior in terms of electrostatics. This is also reflected in the nature of binding of the ligands.[40–42] As a demonstration that 3DZD can effectively compare local surface electrostatics, we classified ligand binding sites of 19 TIM barrel fold enzymes[21] of different families. Figure 7 shows the 19 active sites clustered into three groups, two of which have negative electrostatic potentials while the other has a dominant positive potential.

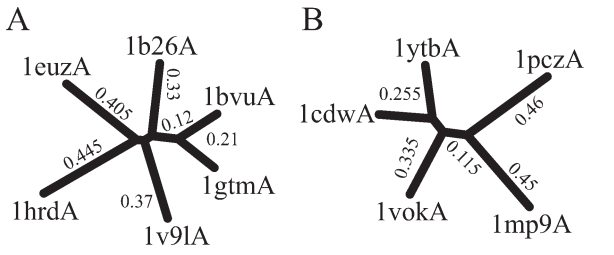

Figure 6

*Complete linkage clustering of surface electrostatic potentials of GDHs and TBPs using HL2 measure. The HL2 distance is shown on branches. (**A**) GDH family; (**B**) TBP family.*

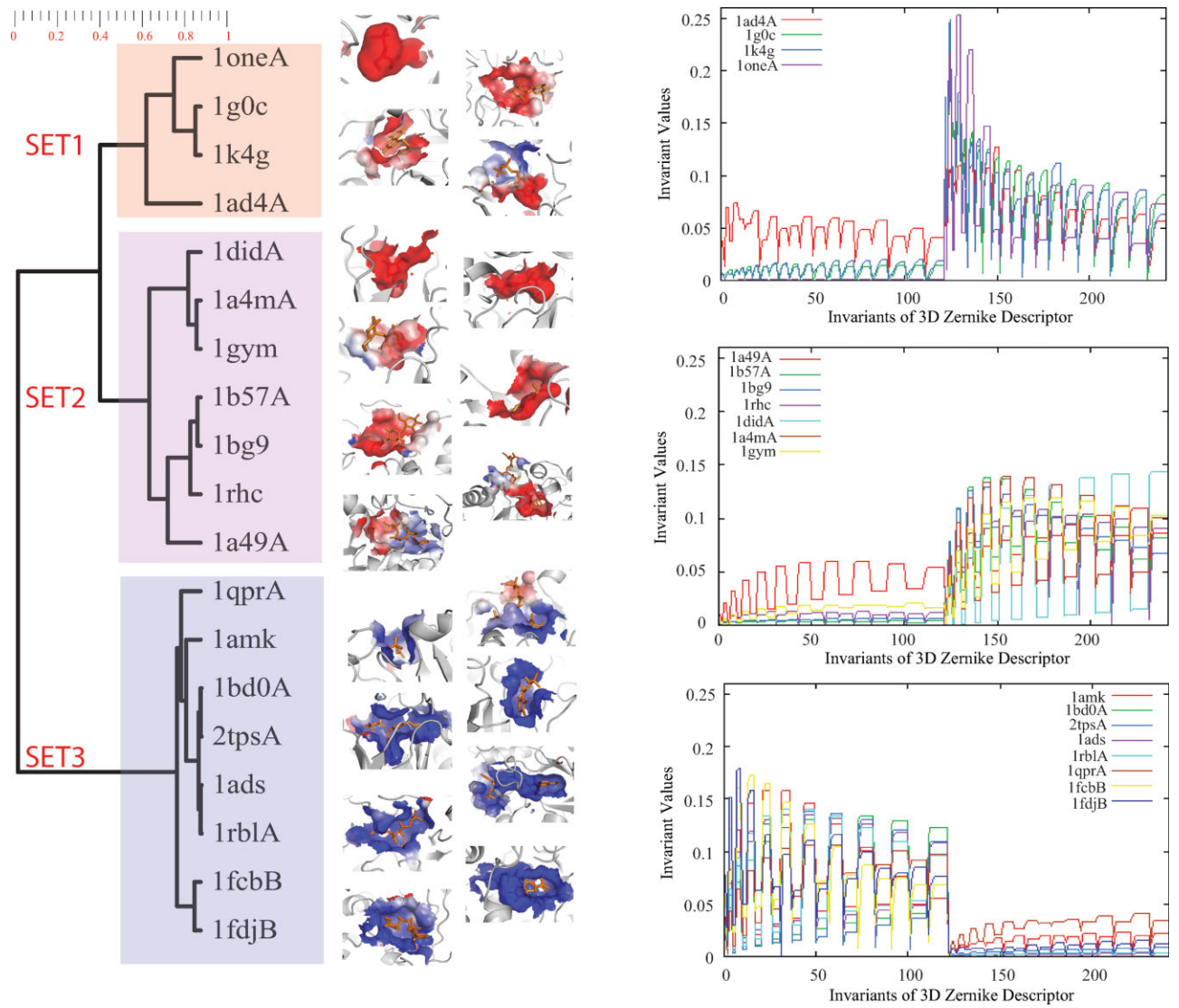

Figure 7

Binding site electrostatic potential of 19 TIM barrel structures: The tree structure shows complete linkage clustering using 3DZD CC of electrostatic potentials on the ligand binding interface. Ligand binding site is computed as a surface region within 3.5 Å of the binding ligand. Two clusters are formed using a CC threshold of 1.22: Set1 + Set2 and Set3. Three clusters are formed using a CC threshold of 0.61: Set1, Set2, and Set3. The three plots on the right show 3DZD invariants for each of the sets.

All of the bound ligands in group 3 have one or more phosphate groups, which complements binding sites with positive electrostatic potentials. In contrast, enzymes in the groups 1 and 2 bind ligands with positively charged groups (e.g., amino pteridine and purine) or metal ions (e.g., Mg^{2+}, Mn^{2+}, and Zn^{2+}) in the binding pockets. Despite groups 1 and 2 having negative potentials, strong peaks in the first few invariants of the group 1 differentiate it from group 2. These peaks correspond to common sphere-like pockets among binding sites in the group 1. The group 1 clustering is again consistent with the surface shape based 3DZD of the 19 binding sites.

CONCLUSIONS

We have introduced 3DZD for fast quantitative comparison of physicochemical properties defined on protein surfaces. Using 3DZD, similarities based on properties such as the electrostatics and hydrophobicity can be quantified in a way that matches our intuition. 3DZD performs better than CL2 and HL2 in its ability to pro-

vide meaningful distances with minimum computational effort. 3DZDs can compare not only global surfaces but also local regions, e.g., binding sites of proteins, even in the absence of sequence or the tertiary structure similarity. Application of 3DZD can be further extended for comparison of the other properties, such as residue conservation. There is an urgent demand for structure-based protein function characterization, due to the structural genomics projects,[43,44] which solve protein tertiary structures of unknown function in an increasing pace.[45] There are earlier works for structure-based function prediction,[46,47] but most of the algorithms are not fast enough to realize a real-time structure search. We believe that the method introduced here opens up a way to develop new methods for quick real-time protein surface based function assignment,[48] which are similarly fast as routinely used global and local sequence motif based annotation.[49,50]

ACKNOWLEDGMENTS

The authors thank Vishwesh Venkatraman for invaluable assistance in the preparation of this paper.

REFERENCES

1. 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.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
3. Mizuguchi K, Go N. Seeking significance in three-dimensional protein structure comparisons. Curr Opin Struct Biol 1995;5:377–382.
4. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
5. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
6. Hodgkin EE, Richards WG. A semiempirical method for calculating molecular similarity. J Chem Soc-Chem Commun 1986;1342–1344.
7. Carbo R, Leyda L, Arnau M. An electron density measure of the similarity between two compounds. Int J Quantum Chem 1980;17:1185–1189.
8. Hodgkin EE, Richards WG. Molecular similarity based on electrostatic potential and electric-field. Int J Quantum Chem 1987;14: 105–110.
9. Nikolova N, Jaworska J. Approaches to measure chemical similarity—a review. QSAR Comb Sci 2004;22:1006–1026.
10. Blomberg N, Gabdoulline RR, Nilges M, Wade RC. Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. Proteins 1999;37:379–387.
11. De RF, Gabdoulline RR, Menziani MC, Wade RC. Blue copper proteins: a comparative analysis of their molecular interaction properties. Protein Sci 2000;9:1439–1454.
12. De RF, Gabdoulline RR, Menziani MC, De Benedetti PG, Wade RC. Electrostatic analysis and Brownian dynamics simulation of the association of plastocyanin and cytochrome f. Biophys J 2001; 81:3090–3104.
13. Schleinkofer K, Wiedemann U, Otte L, Wang T, Krause G, Oschkinat H, Wade RC. Comparative structural and energetic analysis of WW domain-peptide interactions. J Mol Biol 2004;344:865–881.
14. Winn PJ, Religa TL, Battey JN, Banerjee A, Wade RC. Determinants of functionality in the ubiquitin conjugating enzyme family. Structure 2004;12:1563–1574.
15. Sasin JM, Godzik A, Bujnicki JM. SURF'S UP!—protein classification by surface comparisons. J Biosci 2007;32:97–100.
16. Pawlowski K, Godzik A. Surface map comparison: studying function diversity of homologous proteins. J Mol Biol 2001;309:793–806.
17. Zhang X, Bajaj C, Kwon B, Dolinsky T, Nielsen J, Baker N. Application of new multiresolution methods for the comparison of biomolecular electrostatic properties in the absence of structural similarity. Multiscale Model Simul 2006;5:1196–1213.
18. Novotni M, Klein R. 3D Zernike descriptors for content based shape retrieval. ACM Symposium on solid and physical modeling, In Proceedings of the 8th ACM Symposium on Solid Modeling and Applications, Seattle, Wash, USA, 2003. pp 216–225.
19. 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, online ahead of print; DOI: 10.1002/prot.22030.
20. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
21. Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–765.
22. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
23. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 2001;98:10037–10041.
24. Kinoshita K, Nakamura H. eF-site and PDBjViewer: database and viewer for protein functional sites. Bioinformatics 2004;20:1329–1330.
25. Canterakis N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. In: Proceedings of 11th Scandinavian Conference on Image Analysis, Kangerlussuaq, Greenland, 1999. pp 85–93.
26. Dym H, McKean H. Fourier series and integrals. New York: Academic Press; 1972.
27. 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 2008; 26:1035–1045.
28. Kapp OH, Moens L, Vanfleteren J, Trotman CN, Suzuki T, Vinogradov SN. Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume. Protein Sci 1995;4:2179–2190.
29. Aronson HE, Royer WE, Jr, Hendrickson WA. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. Protein Sci 1994;3:1706–1711.
30. Lecomte JT, Vuletich DA, Lesk AM. Structural divergence and distant relationships in proteins: evolution of the globins. Curr Opin Struct Biol 2005;15:290–301.
31. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
32. Tarricone C, Galizzi A, Coda A, Ascenzi P, Bolognesi M. Unusual structure of the oxygen-binding site in the dimeric bacterial hemoglobin from *Vitreoscilla* sp. Structure 1997;5:497–507.
33. Pesce A, Dewilde S, Kiger L, Milani M, Ascenzi P, Marden MC, Van Hauwaert ML, Vanfleteren J, Moens L, Bolognesi M. Very high resolution structure of a trematode hemoglobin displaying a TyrB10-TyrE7 heme distal residue pair and high oxygen affinity. J Mol Biol 2001;309:1153–1164.
34. Harutyunyan EH, Safonova TN, Kuranova IP, Popov AN, Teplyakov AV, Obmolova GV, Valnshtein BK, Dodson GG, Wilson JC. The binding of carbon monoxide and nitric oxide to leghaemoglobin

in comparison with other haemoglobins. J Mol Biol 1996;264: 152–161.
35. Hankeln T, Ebner B, Fuchs C, Gerlach F, Haberkamp M, Laufs TL, Roesner A, Schmidt M, Weich B, Wystub S, Saaler-Reinhardt S, Reuss S, Bolognesi M, De SD, Marden MC, Kiger L, Moens L, Dewilde S, Nevo E, Avivi A, Weber RE, Fago A, Burmester T. Neuroglobin and cytoglobin in search of their role in the vertebrate globin family. J Inorg Biochem 2005;99:110–119.
36. Lecomte JT, Vuletich DA, Lesk AM. Structural divergence and distant relationships in proteins: evolution of the globins. Curr Opin Struct Biol 2005;15:290–301.
37. Torrez M, Schultehenrich M, Livesay DR. Conferring thermostability to mesophilic proteins through optimized electrostatic surfaces. Biophys J 2003;85:2845–2853.
38. Kinjo AR, Nishikawa K. Comparison of energy components of proteins from thermophilic and mesophilic organisms. Eur Biophys J 2001;30:378–384.
39. Xiao L, Honig B. Electrostatic contributions to the stability of hyperthermophilic proteins. J Mol Biol 1999;289:1435–1444.
40. Spassov VZ, Karshikoff AD, Ladenstein R. Optimization of the electrostatic interactions in proteins of different functional and folding type. Protein Sci 1994;3:1556–1569.
41. Madura JD, McCammon JA. Brownian dynamics simulation of diffusional encounters between triose phosphate isomerase and D-glyceraldehyde phosphate. J Phys Chem 1989;93:7285–7287.
42. Raychaudhuri S, Younas F, Karplus PA, Faerman CH, Ripoll DR. Backbone makes a significant contribution to the electrostatics of α/β-barrel proteins. Protein Sci 1997;6:1849–1857.
43. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287.
44. Bourne PE, Allerston CK, Krebs W, Li W, Shindyalov IN, Godzik A, Friedberg I, Liu T, Wild D, Hwang S, Ghahramani Z, Chen L, Westbrook J. The status of structural genomics defined through the analysis of current targets and structures. Pac Symp Biocomput 2004;375–386.
45. Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The protein data bank and structural genomics. Nucleic Acids Res 2003;31:489–491.
46. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 2005;15:275–284.
47. Kinoshita K, Nakamura H. Protein informatics towards function identification. Curr Opin Struct Biol 2003;13:396–400.
48. 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.
49. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 2006;15:1550–1556.
50. Hawkins T, Kihara D. Function prediction of uncharacterized proteins. J Bioinform Comput Biol 2007;5:1–30.