PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Geometrical comparison of two protein structures using Wigner-D functions

S. M. Saberi Fathi,[1] Diana T. White,[2] and Jack A. Tuszynski[3]*

[1] Department of Physics, Ferdowsi University of Mashhad, Mashhad, Iran
[2] Department of Oncology, University of Alberta, Edmonton, Alberta, Canada
[3] Department of Physics, University of Alberta, Edmonton, Alberta, Canada

ABSTRACT

In this article, we develop a quantitative comparison method for two arbitrary protein structures. This method uses a root-mean-square deviation characterization and employs a series expansion of the protein's shape function in terms of the Wigner-D functions to define a new criterion, which is called a "similarity value." We further demonstrate that the expansion coefficients for the shape function obtained with the help of the Wigner-D functions correspond to structure factors. Our method addresses the common problem of comparing two proteins with different numbers of atoms. We illustrate it with a worked example.

Proteins 2014; 82:2756–2769.

Key words: protein structure; protein classification; computational methods; Wigner-D functions; root-mean-square deviation; tubulin; Fourier transform; reciprocal space; similarity value; correlation value.

INTRODUCTION

A quantitative comparison of two protein tertiary structures to assess their similarity is a major challenge, but if properly investigated, it can offer answers to important questions in biochemistry and cell biology.[1] In particular, structural similarity between proteins is a very good predictor of their functional similarity. In order to classify proteins according to their structural characteristics, we first have to be able to determine the three-dimensional (3D) structures of the proteins in question, which typically involves X-ray or electron crystallography, or in some cases other techniques such as nuclear magnetic resonance or mass spectroscopy.[2] In the absence of crystallographic structures for a given protein, computational methods may still be used to predict a 3D structure based on sequence similarity with crystallographically resolved protein structures using a technique called homology modeling.[3] Assuming structural information is available, a number of methods have been developed to compare protein structures.[4,5] Some methods are based on numerical techniques such as geometric hashing[6] or spherical harmonic descriptors.[7] A recently reported method uses so-called Zernike descriptors.[8]

Traditionally, protein classifications have been performed manually with the aid of automated tools, and they take into account information available to biologists regarding both the functions and the phylogenetic origins of the proteins investigated. Examples of relevant databases include Structural Classification of Proteins,[9,10] Class, Architecture, Topology, and Homologous superfamily,[11] and Families of Structurally Similar Proteins[12] to name but a few.

In order to match two distinct protein structures, there should exist a one-to-one map between their structural elements, which is called "correspondence." In addition, proper alignment of the structural elements of these proteins should be generated. A common measure that is used for this type of alignment is root-mean-square deviation (RMSD).[13,14] Until now, a complete geometrical comparison of two proteins has rarely been possible mainly because most proteins have different sizes and/or different numbers and types of atoms. Therefore, a complete match between an arbitrary pair of

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Ferdowsi University of Mashhad (to SMSF); Grant number: 2/22306.
*Correspondence to: Jack A. Tuszynski, Department of Physics, 4–181 CCIS, University of Alberta, Edmonton, Canada AB T6G 2E1. E-mail: jackt@ualberta.ca
Received 24 December 2013; Revised 20 May 2014; Accepted 18 June 2014
Published online 5 July 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24640

proteins is a difficult task to accomplish in general. This is why either partial or local similarity tests have frequently been used in the past.[13] An example of using RMSD for partial similarity analysis is the STRUCTAL software.[15] In Results and Discussion section, we discuss in more detail different methods used for protein structure comparisons and compare and contrast them with our method.

In this article, we introduce a fully automated method that enables one to compare protein structures and to perform identification of proteins. To this end, we expand the protein shape function in terms of Wigner-D functions[16] and demonstrate mathematically that the expansion coefficients can be regarded as the structure factors of a protein. We then compare them to assess their similarity by introducing a new parameter referred to here as the "similarity value" (SV). Our method obtains the similarity value in the reciprocal space (in relation to the spatial domain) where two proteins have the same dimension (values of their structure factors). However, it is important to note that these proteins are allowed to have different numbers of atoms in the spatial domain. We demonstrate below that the SV can be used as an alternative measure of structural similarity, as opposed to the RMSD measure. In particular, we show that when comparing similarly sized proteins, either RMSD or SV may be the preferred measure. However, when comparing different-size structures, we show that using the similarity value (SV) is strongly preferred as it permits a quantitative comparison between any protein structures independently of their sizes.

BASIC MATHEMATICAL IDEA

The Wigner-D functions describe the rotation on a sphere in 4D space (4-sphere), and they are analogous to the well-known spherical harmonic functions, which are commonly used to describe the rotation on a sphere in 3D space (3-sphere).[16] A rigid body can be projected on a 4-sphere; thus, its shape function can be expanded using the Wigner-D functions. Proteins are not typically thought of as rigid bodies due to their weak bonds, but instead they undergo sizeable thermal fluctuations at finite temperature and conformational changes due to ligand binding. However, the different conformations of a protein, which are explored over time, can be quantitatively characterized using shape functions in time series representations.

We start by expanding a hypothetical protein shape function, f, in terms of Wigner-D functions as

$$f(\alpha,\beta,\gamma)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\sum_{n=-l}^{l}C_{lmn}D^{l}_{mn}(\alpha,\beta,\gamma) \qquad (1)$$

where the C_{lmn} factors are the coefficients of the series expansion, D^{l}_{mn} is a Wigner-D function and the parameters l, m, and n satisfy: $l \geq 0$, $|m| \leq l$, and $|n| \leq l$. The Wigner-D function is defined by[17]

$$D^{l}_{mn}(\alpha,\beta,\gamma)=e^{-im\alpha}e^{-in\gamma}d^{l}_{mn}(\cos\beta) \qquad (2)$$

where

$$d^{l}_{mn}(x)=\epsilon\sqrt{\frac{\left(l-\frac{\nu+\sigma}{2}\right)!\left(l+\frac{\nu+\sigma}{2}\right)!}{\left(l-\frac{\nu-\sigma}{2}\right)!\left(l+\frac{\nu-\sigma}{2}\right)!}}2^{-\frac{\nu+\sigma}{2}}(1-x)^{\frac{\nu}{2}}(1+x)^{\frac{\sigma}{2}}P^{(\nu,\sigma)}_{l-\frac{\nu+\sigma}{2}}(x) \qquad (3)$$

where $\nu=|n-m|$ and $\sigma=|n+m|$, and

$$\varepsilon=\begin{cases}1 & \text{if } n\geq m\\ (-1)^{n-m} & \text{if } n<m\end{cases} \qquad (4)$$

while $P^{m}_{l}(x)$ is the associated Legendre polynomial and is defined as

$$P^{m}_{l}(x)=\frac{(-1)^{m}}{2^{l}l!}\left(1-x^{2}\right)^{\frac{m}{2}}\frac{d^{l+m}}{dx^{l+m}}\left(x^{2}-1\right)^{l}. \qquad (5)$$

The dimension of a Wigner-D function is given by

$$Dim(D)=\sum_{l=0}^{N}(2l+1)^{2}=\frac{1}{3}(N+1)(2N+1)(2N+3). \qquad (6)$$

We can express Eq. (1) in matrix notation simply as $\mathbf{f}=\mathbf{CD}$. Indeed, the discrete Fourier transform on SO(3) can be written in terms of the Wigner-D functions as[17,18]

$$f(\alpha,\beta,\gamma)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\sum_{n=-l}^{l}\hat{f}_{lmn}D^{l}_{mn}(\alpha,\beta,\gamma) \qquad (7)$$

where $\hat{f}$ is the Fourier transform of f. We can express the above relation in matrix form as $\mathbf{f}=\hat{\mathbf{f}}\mathbf{D}$. Thus, the C_{lmn} coefficients can be viewed as Fourier transforms of a given function f. On the other hand, we know from crystallography that the Fourier transform of the shape function of an object is defined as the corresponding structure factor.[19] Thus, the C_{lmn} coefficients describe the structure factors of a given protein with the shape function f (which is obtained from the positions of the atoms of the protein).

Having generated the shape function f, we can obtain the C_{lmn} coefficients of the expansion by

$$C_{lmn}=\frac{(2l+1)}{8\pi^{2}}\int\int\int f(\alpha,\beta,\gamma)D^{l}_{mn}{}^{*}(\alpha,\beta,\gamma)\sin\beta\,d\beta\,d\alpha\,d\gamma \qquad (8)$$

where we use the orthogonality of the Wigner-D function:

Table I
Comparison Between Pairs of Proteins Using Correlation Value (CV) and Similarity Value (SV)

First protein's PDB ID	Second protein's PDB ID	Correlation value (CV)	Similarity value (SV)
1JFF	*1ATN*	*0.3962*	*0.0002*
1JFF	1FSZ	0.9661	0.3453
1JFF	1SA0	0.9361	0.0981
1JFF	1TUB	0.9933	0.4872
1JFF	1JFF-A	0.9537	0.2091
1JFF	1JFF-B	0.9987	0.4967
1JFF-A	1JFF-B	0.9460	0.2271
1JFF-A	*1ATN*	*0.5634*	*0.0010*
1JFF-A	1SA0	0.9955	0.3771
1JFF-A	1TUB	0.9300	0.2425
1JFF-A	1FSZ	0.9795	0.4131
1JFF-B	1SA0	0.9270	0.1073
1JFF-B	*1ATN*	*0.3782*	*0.0002*
1JFF-B	1TUB	0.9969	0.4962
1JFF-B	1FSZ	0.9637	0.3689
1ATN	*1SA0*	*0.5973*	*0.0023*
1ATN	*1FSZ*	*0.5016*	*0.0005*
1ATN	*1TUB*	*0.3639*	*0.0002*
1SA0	1FSZ	0.9806	0.2322
1SA0	1TUB	0.9085	0.1154
1FSZ	1TUB	0.9489	0.3852

1JFF, 1SA0, 1TUB, and 1FSZ are all structures of proteins in the tubulin-FtsZ superfamily. 1ATN is a structure for actin. Comparisons between unrelated protein pairs (tubulin-FtsZ superfamily with actin) are italicized.

$$\int\int\int D^{l}_{m'n'}{}^{*}(\alpha,\beta,\gamma)\,D^{l}_{mn}(\alpha,\beta,\gamma)\sin\beta\,d\beta\,d\alpha\,d\gamma=\frac{8\pi^2}{(2l+1)}\delta_{ll'}\delta_{mm'}\delta_{nn'}. \tag{9}$$

METHOD AND ALGORITHM

In this section, we discuss practical aspects of implementing our method for particular proteins. First, we download each protein's atom positions from the Protein Data Bank (PDB) and convert these positions to obtain the corresponding Euler angles. Then, we define the protein shape function f as follows: if a voxel contains a protein's atom then f is equal to one, otherwise f is taken to be zero. The next steps are to compute the resultant Wigner-D functions up to $l=l_{max}$ and obtain the C_{lmn} matrix elements.

One simple way to measure the similarity between two arbitrarily selected proteins is equivalent to computing the correlation value between the structure factors of the two proteins:

$$\text{Correlation Value}=CV=\frac{\langle abs(\mathbf{C})|abs(\mathbf{C}')\rangle}{\langle abs(\mathbf{C})|abs(\mathbf{C})\rangle\langle abs(\mathbf{C}')|abs(\mathbf{C}')\rangle} \tag{10}$$

where $\langle\cdot|\cdot\rangle$ indicates the inner product and $abs(\cdot)$ indicates the absolute value of a variable. However, the CV measure does not provide proper comparison results for proteins, as is explained below.

Representing a 3D shape by expansion in terms of Wigner-D functions effectively projects this shape on a 3-manifold as a part of the hyper-surface of a 4-sphere. The C_{lmn} matrix elements are the points on the manifold constructed in this manner. The CV computed in Eq. (10) gives a fractional rate of the overlap between the two manifolds.

We illustrate this with a specific example. We have chosen a crystal structure for the tubulin heterodimer with PDB code 1JFF.[20] This PDB has two subunits: 1JFF-A for the α-tubulin monomer and 1JFF-B for the β-tubulin monomer. As shown in Table I, the CV for 1JFF and 1JFF-A is approximately 1. This is because the 1JFF-A manifold is a sub-manifold of 1JFF, and all the points of 1JFF-A are subsumed by 1JFF. A discussion about SV*s* between 1JFF, 1JFF-A, and 1JFF-B, which are obtained in Table I, is given in "Results and Discussion" section.

Instead of using the CV measure, we define a solid measure by applying the RMSD concept to the structure factor distances using the following procedure. A structure factor is a complex number, so we can embed it as a vector in a 2D Euclidean space. Thus, for each protein, we can define a space with the dimension equals two times the number of computed structure factors. For example, for $l_{max}=17$ we will have $Dim=7770$ structure factors, so our space's dimensionality is $2\times7770=15540$-dimension (we represent this space by a 2-column and a 7770-row matrix). Subsequently, we compute the distances between each pair of elements in this matrix. We obtain an $n\times n$ matrix of the distances. In a similar way we obtain another matrix for the second protein. The next step is to compute the parameter

$$D^2=2\sum_{i<j}^{n}\sum_{j=2}^{n}\left(d_{ij}-d'_{ij}\right)^2=2\sum_{i<j}^{n}\sum_{j=2}^{n}\left(d_{ij}^2+d'^2_{ij}-2d_{ij}d'_{ij}\right) \tag{11}$$

where d_{ij} and d'_{ij} are the elements of the distance matrix of each of the two proteins. This is an RMSD relation, except we have eliminated the average coefficient $1/(n(n-1))$. Equation (11) in the vector form is

$$D^2=d^2+d'^2-2\langle\mathbf{d}|\mathbf{d}'\rangle \tag{12}$$

where $d^2=\langle\mathbf{d}|\mathbf{d}\rangle=2\sum_{i<j}^{n}\sum_{j=1}^{n}d_{ij}^2$, d'^2 is defined similarly (these are the square of the vector lengths, that is, the sum of the squares of arrays), and $\langle\mathbf{d}|\mathbf{d}'\rangle$ is the scalar product of the two protein vectors (i.e., the sum of the corresponding array multiplications). This scalar product indicates the correlation between two proteins, because if there is no correlation, then $D^2=d^2+d'^2$, and if we have a maximum correlation (the two proteins are the same), then $D^2=0$. To obtain a direct measure of the similarity

between two proteins, we define SV by rewriting Eq. (12) as follows:

$$\text{Similarity Value}=\text{SV}=\frac{\langle \mathbf{d}|\mathbf{d}'\rangle}{d^2+d'^2}=\frac{1}{2}\left(1-\frac{D^2}{d^2+d'^2}\right). \quad (13)$$

Based on the above discussion, SV will satisfy the following inequality:

$$\text{not correlated} \leftarrow 0 \leq \text{SV} \leq \frac{1}{2} \rightarrow \text{totally correlated}. \quad (14)$$

In other words, when the two proteins are the same, then SV=1/2, and when they are completely different and there is no correlation between them, then SV=0.

The Fourier transformation is a linear transform,[21] and it preserves lengths and the inner product. Thus, the Fourier transform is an isometric mapping.[19,22,23] We have shown earlier in this article that the C_{lmn} coefficients are the Fourier transforms of f. Therefore, SV is a good measure to compare two proteins. In the following box, we summarize the algorithm for computing the similarity between two proteins in several simple steps.

Algorithm

1. Obtain protein data from the PDB website (x, y, z) position coordinates of all atoms.
2. Sort atoms by their distances to the center of mass. It is assumed that all masses are distributed equally for all atoms. Another possibility could be to consider real masses of atoms.
3. Convert Cartesian x, y, z coordinates of all atoms to the corresponding Euler angles, α, β, γ relative to the center of mass of the protein.
4. Define the shape function, $f_i=f(\beta_i, \alpha_i, \gamma_i)=1$, $(i=1, 2, \ldots, N)$, N: the number of atoms $(\beta_i, \alpha_i, \gamma_i)$ corresponding to the (x_i, y_i, z_i) position coordinates of the ith-atom in Euler angles.
5. Compute the results of $D^l_{mn}(\beta_i, \alpha_i, \gamma_i)$ Eqs. (2–4).
6. Compute the structure factor, C_{lmn} from the discrete form of Eq. (8):
   $C_{lmn}=\frac{2l+1}{8\pi^2}\sum_{i=1}^{N}$
   $$f(\beta_i, \alpha_i, \gamma_i) D^l_{mn}{}^*(\beta_i, \alpha_i, \gamma_i) \sin\beta_i \Delta\beta_i \Delta\alpha_i \Delta\gamma_i.$$
   We sum only over occupied atom positions because the shape function is zero when there is no atom in the voxel. Thus, the other terms are zero.
7. Repeat Steps 1–7 for each protein analyzed.
8. Compute SV using Eq. (13) between two proteins selected for comparison.

RESULTS AND DISCUSSION

As an example, Table I lists CV and SV measures for selected pairs of protein structures.

We see that the correlation value, CV, does not give a good comparison between two proteins. This is because it is a criterion to compute the overlap between two manifolds in the reciprocal space. If the two proteins are similar, this criterion gives a good correlation between them because these two proteins have the same structure factors. However, for two different or partially different proteins, the CV is not very accurate.

To check our SV criterion, we have calculated the atomic shape function for the 1JFF-A structure by using the structure factors, C_{lmn}. Figure 1 shows the histograms and plots of f and its reconstructions $f_{\text{reconst.}}=\frac{1}{N}\sum_{l_{max}=1}^{N} f_{l_{max}}$ for $N=9$ and $N=17$ for 1JFF-A. We see that the reconstructed functions, $f_{\text{reconst.}}$, are in good agreement with f, especially when l_{max} increases.

Figure 2 shows that when the surface under a pocket of the structure factors is normalized to one, the structure factor for a given l has the Poisson distribution:

$$\frac{1}{(2l+1)^2}\frac{abs(C_{lmn})}{\sqrt{\sum_{m=-1}^{l}\sum_{n=-1}^{l} abs(C_{lmn})^2}} \rightarrow P_L(\mu)=\frac{\mu^L}{L!}e^{-\mu} \quad (15)$$

where $\mu=Var(|C_{lmn}|)$ and $L=(2l+1)^2$. The Poisson distribution is usually considered to be a continuous distribution. However, here we make it discrete since we need to perform a numerical computation. The maximum probability value for the Poisson distribution occurs when $\mu=L$, and the magnitude of the corresponding peak for probability is then equal to

$$P_L(\mu=L)_{max}=\frac{L^L}{L!}e^{-L}=\frac{1}{\sqrt{\pi\left(2L+\frac{1}{3}\right)}} \quad (16)$$

where we used Stirling's approximation relation, that is, $L!=\sqrt{\pi(2L+1/3)}L^L e^{-L}$. The peaks in Figure 2 are in good agreement with Eq. (11). This is another test to confirm the validity of our method, since it gives the same result as the one obtained in X-ray pattern intensity distributions and in Poisson's distribution for random interactions between radiation and matter.[24,25]

In the following discussion we wish to highlight the differences between our method and other methods. The methods introduced for comparing protein similarities are normally based on the following: the proteome-scale protein structure modeling, score function comparison, obtaining moments or descriptors, or comparing RMSD between residues or chains of two proteins.[7,26–49] Discussing all of these methods is not in the scope of this article. However, we do review some methods, which may appear similar to our method. One of these methods involves spherical polar Fourier shape density functions.[26] This method uses the expansion of the 3D density function in terms of radial and spherical

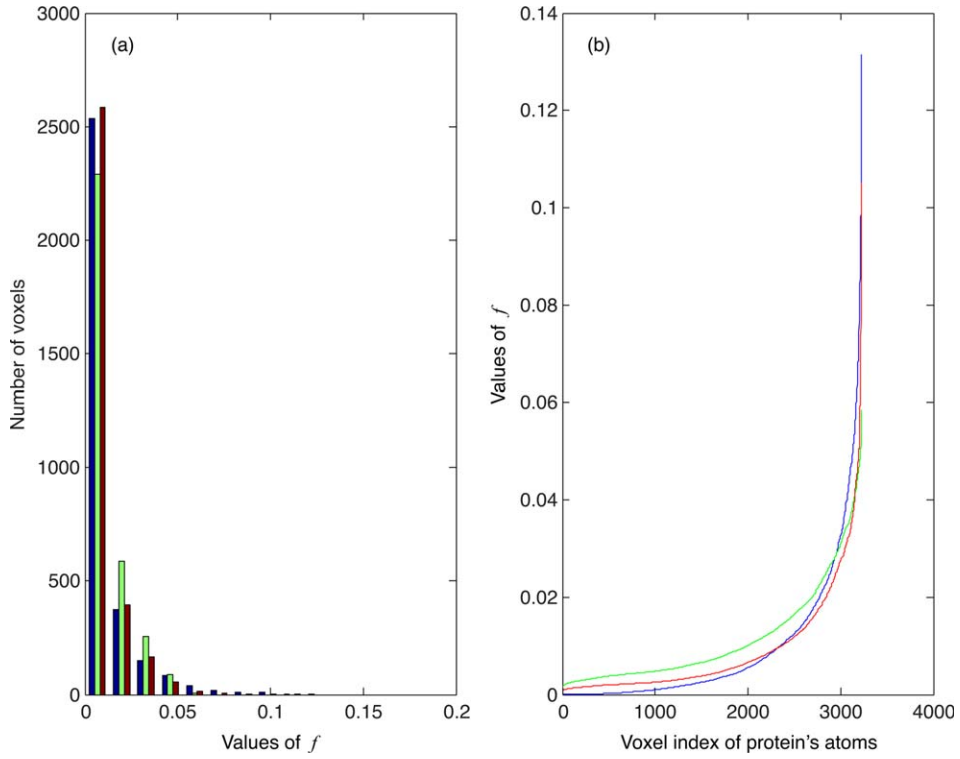

Figure 1

(**a**) Histogram and (**b**) plot of the shape function f (shown in blue) along with its reconstructions $f_{\text{reconst.}}$ for $N=l_{max}=9$ (shown in green) and $N=l_{max}=17$ (shown in red) for 1JFF-A. We see that the reconstructed $f_{\text{reconst.}}$ are in good agreement with f when l_{max} increases.

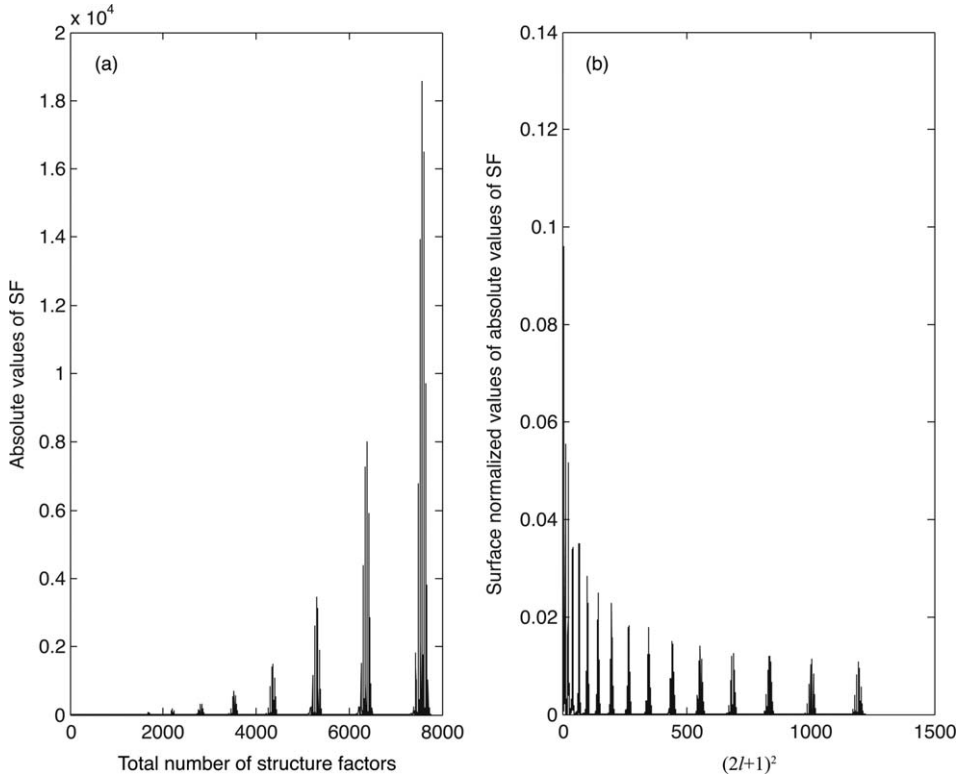

Figure 2

Panel (**a**) shows non-normalized absolute values of structure factors. The abscissa is the total number of structure factors. Panel (**b**) shows the normalized area under the curve to one of the absolute values of the structure factors for each l. The abscissa is the value of $(2l+1)^2$.

Table II
Set of 48 Protein Structures With SV and RMSD From Li *et al.*[1] for Comparison

First protein's PDB ID	Second protein's PDB ID	Similarity value (SV)	RMSD[1]
1A6W	1A6U	0.198	0.34
1MRG	1AHC	0.401	0.43
1RNE	1BBS	0.316	0.61
1RBP	1BRQ	0.108	0.62
1BYB	1BYA	0.499	0.43
1HFC	1CGE	0.399	0.37
3GCH	1CHG	0.070	1.10
1BLH	1DJB	0.497	0.23
1INC	1ESA	0.397	0.21
1GCA	1GCG	0.499	0.32
1HEW	1HEL	0.498	0.21
1IDA	1HSI	0.083	1.07
1DWD	1HXF	0.150	0.27
2IFB	1IFB	0.382	0.37
1IMB	1IME	0.498	0.22
2PK4	1KRN	0.445	0.39
2TMN	1L3F	0.266	0.62
1IVD	1NNA	0.426	1.23
1HYT	1NPC	0.332	0.88
1PDZ	1PDY	0.499	0.66
1PHD	1PHC	0.499	0.17
1PSO	1PSN	0.499	0.33
1SRF	1PTS	0.498	0.26
1ACJ	1QIF	0.497	0.31
1SNC	1STN	0.495	0.70
1STP	1SWB	0.145	0.33
1ULB	1ULA	0.474	0.79
2YPI	1YPI	0.165	1.27
2H4N	2CBA	0.498	0.20
2CTC	2CTB	0.499	0.15
5CNA	2CTV	0.034	0.40
1FBP	2FBP	0.494	1.06
2SIM	2SIL	0.499	0.14
1MTW	2TGA	0.159	0.42
1APU	3APP	0.498	0.40
1QPE	3LCK	0.465	0.28
5P2P	3P2P	0.480	0.42
4PHV	3PHV	0.045	1.23
3PTB	3PTN	0.122	0.26
1BID	3TMS	0.499	0.24
1OKM	4CA2	0.472	0.22
4DFR	5DFR	0.496	0.82
3MTH	6INS	0.381	1.09
6RSA	7RAT	0.440	0.18
1CDO	8ADH	0.403	1.34
7CPA	5CPA	0.132	0.40
1ROB	8RAT	0.469	0.28
1IGJ	1A4J	0.411	0.80

The SVs are computed from structure factors for $l_{max}=7$.

Table III
Set of 86 Protein Structures With SV and RMSD From Li *et al.*[1] for Comparison

First protein's PDB ID	Second protein's PDB ID	Similarity value (SV)	RMSD[1]
1AD4	1AD1	0.499	0.50
1AHX	1AHG	0.499	0.24
1AUR	1AUO	0.499	0.20
1AXZ	1AXY	0.498	0.12
1GN8	1B6T	0.491	0.51
1B9Z	1B90	0.494	0.54
1LRI	1BEO	0.498	1.05
1BUL	1BUE	0.499	0.18
1BYD	1BYA	0.499	0.43
1C3R	1C3P	0.202	0.39
1C5I	1C5H	0.494	0.13
1QJW	1CB2	0.498	0.63
1CTE	1CPJ	0.499	0.29
1SZJ	1CRW	0.499	0.33
1ESW	1CWY	0.498	0.38
1CY7	1CY0	0.155	1.12
1DED	1D7F	0.481	0.26
1P7T	1D8C	0.406	0.66
1DMY	1DMX	0.499	0.19
1DQY	1DQZ	0.052	0.75
1LP6	1DV7	0.471	0.56
1E2S	1E1Z	0.499	0.13
1ESE	1ESC	0.499	0.19
6ALD	1EWD	0.477	0.44
1NLM	1F0K	0.163	1.66
1F4X	1F4W	0.488	0.25
1JBW	1FGS	0.430	1.48
1FR8	1FGX	0.498	0.54
1LD8	1FT1	0.416	0.92
1HVC	1G6L	0.345	0.46
1LSP	1GBS	0.360	0.26
1LC3	1GCU	0.458	0.77
1GJW	1GJU	0.499	0.29
1N75	1GLN	0.422	1.47
1GOY	1GOU	0.476	0.47
1H46	1GPI	0.193	0.15
1GUZ	1GV1	0.383	0.62
1YDD	1HEA	0.491	0.18
1YDA	1HEB	0.498	0.20
1KIC	1HOZ	0.420	0.35
1A80	1HW6	0.466	0.93
1I3A	1I39	0.498	0.40
4AIG	1IAG	0.494	0.26
1JZS	1ILE	0.497	0.69
1JQ3	1INL	0.493	0.35
1JAY	1JAX	0.435	0.60
1UEH	1JP3	0.499	0.67
1JSO	1JSM	0.499	0.10
1JYL	1JYK	0.208	0.94
1JVS	1K5H	0.351	1.16
1K70	1K6W	0.497	1.08
1M6P	1KEO	0.136	1.05
3KIV	1KIV	0.467	0.30
1KMP	1KMO	0.498	0.64
2NGR	1KZ7	0.467	1.61
2MIN	1L5H	0.084	0.55
1LL2	1LL3	0.496	0.37
1LMC	1LMN	0.499	0.10
1EYN	1NAW	0.208	1.02
1BHT	1NK1	0.033	0.58
1PBO	1OBP	0.143	0.38
1OPB	1OPA	0.295	0.68
1I75	1PAM	0.499	0.13

harmonic functions and computes the correlation coefficients between two density function expansion coefficients. The other method uses Zernike descriptors. The Zernike functions are extensions of the spherical harmonic functions. The Zernike descriptors were first used by Novotni and Klein[27] to compare two shapes in shape searching algorithms in computer science. Later they were adapted for protein comparison purposes.[28,29] The 3D Zernike method is a rotational invariant method that finds a descriptor, which represents a given shape. By

Table III
(Continued)

First protein's PDB ID	Second protein's PDB ID	Similarity value (SV)	RMSD[1]
1NME	1PAU	0.499	0.29
1KEV	1PED	0.281	0.81
1PIG	1PIF	0.495	0.32
1PJC	1PJB	0.498	0.61
1KLT	1PJP	0.168	0.97
1QHG	1PJR	0.499	0.23
1CEB	1PKR	0.041	0.58
2PK4	1PMK	0.417	0.71
1BK9	1PSJ	0.494	0.24
1QBB	1QBA	0.497	0.11
1PYY	1QME	0.157	0.59
1OSS	1SGT	0.367	0.27
1SWN	1SWL	0.497	0.31
1LBT	1TCA	0.440	0.24
1WBL	1WBF	0.371	0.39
1YDB	1YDC	0.491	0.12
1H0S	2DHQ	0.499	0.26
1LLO	2HVM	0.498	0.12
43CA	43C9	0.491	0.23
5BIR	4BIR	0.487	0.61
5EUG	4EUG	0.498	0.21
5EAU	5EAS	0.064	0.40
7TAA	6TAA	0.499	0.24

The SVs are computed from structure factors for $l_{max}=7$.

comparing the descriptors, the similarities between any two shapes could be determined. Another method that should be mentioned here is the spherical harmonic method.[7,28] This method expands a shape function in terms of spherical harmonic functions. After some algebraic computations, the spherical harmonic method defines the descriptors and compares them. The above methods use moment or descriptor concepts to compare proteins. Some methods have used RMSD values as a score to compare between two structures. However, because proteins can have different sizes, normally these methods only use partial RMSD calculations. For example, some of these methods have used a difference between the intrastructural residue–residue distances, for example, Dali,[45,46] CE,[47] or between interstructural residue-residue–distances, such as STRUCTAL,[15] SAL,[48] and TM-score.[52] Our method is different from these structural methods for the following reasons:

1. Using the Wigner-D function does not require a definition of a radial function, as is done in the 3D Zernike, spherical harmonic or SPE methods. The three angles in the Wigner-D functions are the Euler angles, and it is well known from classical mechanics that moving through a 3D rigid body is possible by using three Euler angles.
2. We show that the expansion coefficients of the shape function defined by the Wigner-D functions are equivalent to the Fourier transform of the shape function (see more details in Basic Mathematical Idea section). Thus, we introduce the expansion coefficients of a shape function in terms of Wigner-D functions as structure factors.

Table IV
Set of Eight Protein Pairs With High SV and Low RMSD From Li *et al.*[1]

First protein's PDB ID	First protein's total atom #	Second protein's PDB ID	Second protein's total atom #	Similarity value (SV)	RMSD
1a6w	*1774*	*1a6u*	*1736*	*0.4856*	*0.34*
1mrg	2395	1ahc	1933	0.45127	0.43
1rne	2510	1bbs	5018	0.40392	0.61
1byb	3925	1bya	3929	0.49973	0.43
1hfc	1244	1cge	1542	0.47786	0.37
1blh	1999	1djb	2028	0.49806	0.23
1inc	1822	1esa	1822	0.40316	0.21
1gca	2892	1gcg	2893	0.49997	0.32

Methods are in agreement and both are good measures of structure similarity. Side-by-side illustration of protein pair italicized in table is shown in Figure 7(a).

Table V
Set of Eight Protein Pairs With Low SV and High RMSD From Li *et al.*[1]

First protein's PDB ID	First protein's total atom #	Second protein's PDB ID	Second protein's total atom #	Similarity value (SV)	RMSD
1au1	*3510*	*1a2v*	*31080*	*0.0942727*	*15.4917*
1atz	2797	1a2z	6986	0.0767325	5.59876
1aq0	4541	1a5t	3062	0.0899071	12.0075
1apj	1063	1a6f	950	0.168915	4.15105
1aoe	3182	1a7s	1680	0.0977561	11.4711
1aop	3622	1a77	2546	0.0543472	19.6512
1alv	3422	1a99	10720	0.042003	13.7636
1alu	1263	1a9x	44287	0.0271374	12.7272

Methods are in agreement. However, neither is a good measure of structure similarity. Side-by-side illustration of protein pair italicized in table is shown in Figure 7(b).

Table VI
Set of Eight Protein Pairs With High SV and Low RMSD From Li *et al.*[1]

First protein's PDB ID	First protein's total atom #	Second protein's PDB ID	Second protein's total atom #	Similarity value (SV)	RMSD
1aw0	*1063*	*1a0f*	*3224*	*0.425909*	*10.4309*
1auz	1693	1a17	1281	0.45488	12.5662
1aru	2465	1a48	2384	0.422306	14.7692
1aln	2220	1aa6	5454	0.417816	10.0503
1all	2399	1aa7	2980	0.417423	11.2399
1al0	9521	1ab4	3771	0.459777	12.6608
1ajs	6539	1acc	5282	0.396727	14.4862
1air	3271	1aco	5812	0.474606	22.6212

Methods are not in agreement. SV is a better measurement of structure similarity than RMSD. Side-by-side illustration of protein pair italicized in table is shown in Figure 7(c).

Table VII
Set of Seven Protein Pairs With Low SV and Low RMSD From Li *et al.*[1]

First protein's PDB ID	First protein's total atom #	Second protein's PDB ID	Second protein's total atom #	Similarity value (SV)	RMSD
1rbp	*1408*	*1brq*	*1408*	*0.099626*	*0.62*
1pbo	2534	1obp	3158	0.12698	0.38
1oss	1632	1sgt	1620	0.29663	0.27
5bir	1565	4bir	785	0.21514	0.61
2dgd	6984	1h75	620	0.0204437	2.25141
1y6x	669	1n4p	32190	0.025834	2.90993
1tqg	2247	1qbk	8323	0.0651008	1.77265

Methods are not in agreement. RMSD is a better measure of structure similarity than SV. Side-by-side illustration of protein pair italicized in table is shown in Figure 7(d).

3. It is well known that the Fourier transform (consequently an expansion on the Wigner-D function) is a linear transform and hence it preserves isometry.[17,18,22,23] Thus, if we define a RMSD-type criterion, like Similarity Value between structure factors, we show mathematically that the properties obtained in the reciprocal space reflect the same properties in the position space. This means that if in the reciprocal space two proteins are similar, the same result holds true in the position space (provided a method of comparison is defined).
4. The size of a protein analyzed does not affect our comparison. This is because we are able to compare two proteins with the same size in reciprocal space, even though they may be different in position space, and we can choose the dimension of reciprocal space according to a desired level of accuracy. Note that in reality the expansion terms have to go to infinity but similar to other computational calculations, we should choose a cutoff in order to terminate this divergence. The number of expansion coefficients used increases the level of accuracy, but is also computationally costly.
5. We have compared our method results with two other sets, that is, a 48 protein pair data set and an 86 protein pair data set, where both liganded and unliganded proteins are listed, and RMSD values are those reported in Supporting Information of Li *et al.*[1] (These sets are given in http://dragon.bio.purdue.edu/visgrid_suppl). Note that RMSD values are computed in the position space. Tables II and III show computed SV values and RMSD values (taken from Li *et al.*[1]), for the 48 pair data set and the 86 pair data set, respectively. From these tables, we see that, in the case where both protein structures have similar numbers of atoms, both methods are in agreement. That is, a lower value of RMSD corresponds to a higher value of SV (for a similar number of atoms) and vice versa. To highlight the cases where both methods agree, we include Tables IV and V, which summarize a handful of the pairings from Tables II and III. When the number of atoms between two protein structures is different, there is no agreement between the SV and RMSD values. In this case, we see that either SV or RMSD can be useful measures of structural similarity. Again, to highlight the cases where both methods disagree, we include Tables VI and VII, which summarize a handful of the pairings from Tables II and III. We further illustrate the agreement and disagreement between these methods in Figure 7. In some algorithms, using RMSD values in a part of the algorithm, RMSD can be replaced by SV as an alternative parameter. This is because SV is equivalent to RMSD and SV can be computed more precisely than RMSD for

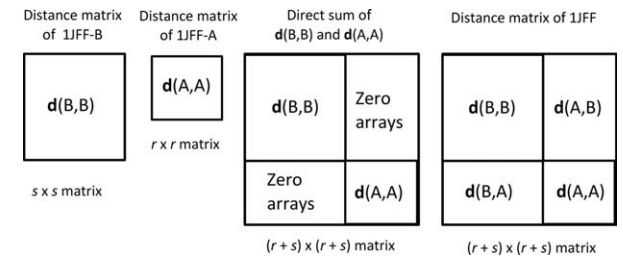

Figure 3
A schematic of the distance matrices **d**(A,A), **d**(B,B), and **d**(AB,AB), and a schematic describing the use of the direct sum.

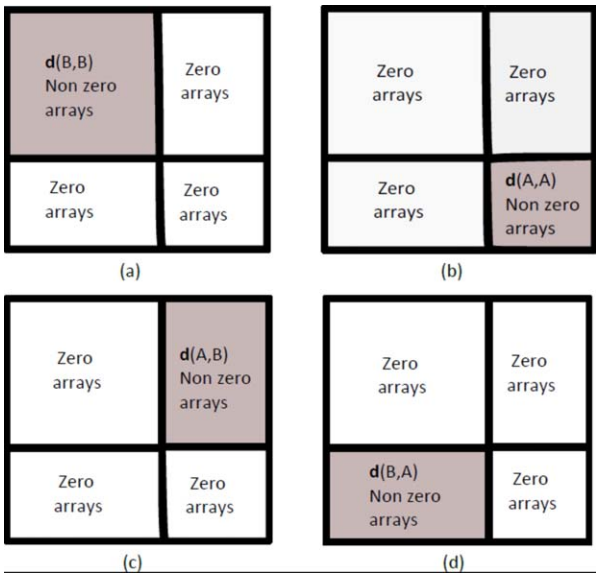

Figure 4
(**a**) Extended $\underline{\mathbf{d}}(B,B)$ matrix. Here we add zeros to change the size of $\mathbf{d}(B,B)$ from $s\times s$ to $(r+s)\times(r+s)$. (**b**) Extended $\underline{\mathbf{d}}(\mathrm{A},\mathrm{A})$ matrix. Here we add zeros to change the size of $\mathbf{d}(\mathrm{A},\mathrm{A})$ from $r\times r$ to $(r+s)\times(r+s)$. (**c**) Extended $\underline{\mathbf{d}}(\mathrm{A},B)$ matrix. Here we add zeros to change the size of $\mathbf{d}(\mathrm{A},B)$ from $r\times s$ to $(r+s)\times(r+s)$. (**d**) Extended $\underline{\mathbf{d}}(B,\mathrm{A})$ matrix. Here we add zeros to change the size of $\mathbf{d}(B,\mathrm{A})$ from $s\times r$ to $(r+s)\times(r+s)$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

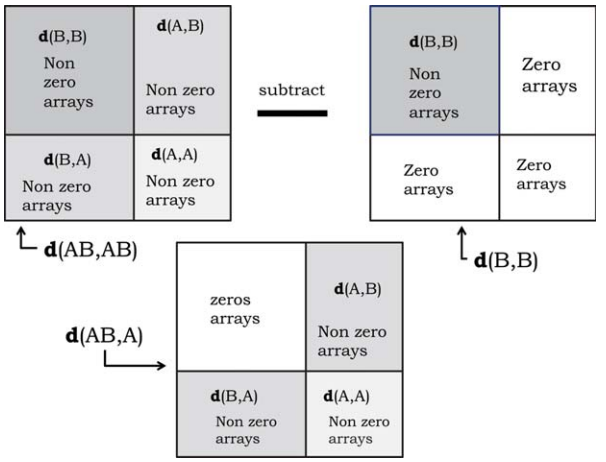

Figure 5

A schematic representation describing how we obtain $\mathbf{d}(\mathrm{AB},\mathrm{B})$. The subtraction of the two top matrices yields the bottom matrix. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

proteins with different sizes. However, it should be kept in mind that the cost of computation for SV is about 33 s for comparing two proteins using a laptop with an Intel Core i7 CPU. An in depth description of how these two methods differ is described in "The difference between SV and RMSD" section.

Below we discuss a specific example by explaining the SVs results for 1JFF and their monomers (1JFF-A and 1JFF-B). As we know, 1JFF-A has $r=3227$-atoms, 1JFF-B has $s=3351$-atoms and 1JFF has the sum of both monomers atoms, that is, $r+s=6578$-atoms. The SVs reported between these macromolecules in Table I are:

1JFF	1JFF-A	SV = 0.2091
1JFF	1JFF-B	SV = 0.4967
1JFF-A	1JFF-B	SV = 0.2271

Here, we discuss these results in more detail. To simplify notation, we represent 1JFF-A with A, 1JFF-B with B, and 1JFF with AB. Their distance matrices are defined by $\mathbf{d}(\mathrm{A},\mathrm{A})$ which has $r\times r$ arrays, $\mathbf{d}(\mathrm{B},\mathrm{B})$ with $s\times s$ arrays, and $\mathbf{d}(\mathrm{AB},\mathrm{AB})$ with $(r+s)\times(r+s)$ arrays. Note that here we define distance matrix $\mathbf{d}$ in position space. We can write $\mathbf{d}(\mathrm{AB},\mathrm{AB})$ as

$$\mathbf{d}(\mathrm{AB},\mathrm{AB})=\mathbf{d}(B,\mathrm{B})\oplus\mathbf{d}(\mathrm{A},\mathrm{A})+\underline{\mathbf{d}}(\mathrm{B},\mathrm{A})+\underline{\mathbf{d}}(\mathrm{A},\mathrm{B}) \quad (17)$$

where $\oplus$ indicates the direct sum between matrices (see Fig. 3). We note that $\mathbf{d}(\mathrm{A},\mathrm{B})$ and $\mathbf{d}(\mathrm{B},\mathrm{A})$ are transpose of each other, and we define $\underline{\mathbf{d}}$(A,B) and $\underline{\mathbf{d}}$(B,A) to be the extended forms of $\mathbf{d}$(A,B) and $\mathbf{d}$(B,A), respectively [matrices with zero arrays added as shown in Figure 4(c,d)]. Let us assume an unknown direct way (in position space and not in reciprocal space), that we can find the RMSD between the above structures. We then compute the following terms

$$\begin{aligned} D^2(\mathrm{A},\mathrm{B})&=\sum\left[\underline{\mathbf{d}}(\mathrm{A},\mathrm{A})-\underline{\mathbf{d}}(\mathrm{B},\mathrm{B})\right].\hat{}2\\ D^2(\mathrm{AB},\mathrm{A})&=\sum\left[\mathbf{d}(\mathrm{AB},\mathrm{AB})-\underline{\mathbf{d}}(\mathrm{A},\mathrm{A})\right].\hat{}2\\ D^2(\mathrm{AB},\mathrm{B})&=\sum\left[\mathbf{d}(\mathrm{AB},\mathrm{AB})-\underline{\mathbf{d}}(\mathrm{B},\mathrm{B})\right].\hat{}2 \end{aligned} \quad (18)$$

where $[\cdot].\hat{}2$ means that all arrays of the matrix in the bracket will be squared, $\sum[\cdot]$ is defined as summation over all matrix arrays in the bracket, $[\cdot]$. The matrices $\underline{\mathbf{d}}$(B,B) and $\underline{\mathbf{d}}$(A,A) are the extended matrices for $\mathbf{d}$(B,B) and $\mathbf{d}$(A,A), respectively [as shown in Figure 4(a,b)]. Now, we expand D^2 between AB and A and B. First, we have

$$\begin{aligned} D^2(\mathrm{AB},B)&=\sum\left[\mathbf{d}(\mathrm{AB},\mathrm{AB})-\underline{\mathbf{d}}(\mathrm{A},\mathrm{A})\right].\hat{}2\\ &=\sum\left[\mathbf{d}(B,\mathrm{B})\oplus\mathbf{d}(\mathrm{A},\mathrm{A})+\underline{\mathbf{d}}(\mathrm{B},\mathrm{A})+\underline{\mathbf{d}}(\mathrm{A},\mathrm{B})-\underline{\mathbf{d}}(\mathrm{A},\mathrm{A})\right].\hat{}2 \end{aligned} \quad (19)$$

Similarly, we find that

$$\begin{aligned} D^2(\mathrm{AB},\mathrm{A})&=\sum\left[\mathbf{d}(\mathrm{AB},\mathrm{AB})-\underline{\mathbf{d}}(\mathrm{B},\mathrm{B})\right].\hat{2}\\ &=\sum\left[\mathbf{d}(B,\mathrm{B})\oplus\mathbf{d}(\mathrm{A},\mathrm{A})+\underline{\mathbf{d}}(\mathrm{B},\mathrm{A})+\underline{\mathbf{d}}(\mathrm{A},\mathrm{B})-\underline{\mathbf{d}}(\mathrm{B},\mathrm{B})\right].\hat{2} \end{aligned} \quad (20)$$

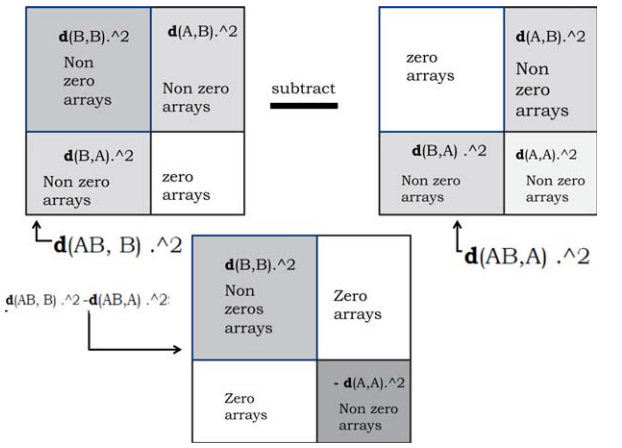

Figure 6

Top: squared arrays of $\mathbf{d}(\mathrm{AB},\mathrm{A})$ and $\mathbf{d}(\mathrm{AB},\mathrm{B})$. ".^2" means that all arrays of the matrix will be squared. The sum over all arrays of these matrices yields D^2 as defined by Eq. (11). Bottom: shows subtraction of the top two matrices. Normally, we should subtract the sum of two top matrices. But, here to show the derivation of our formula in Eq. (21) before summation we subtract two top matrices and we see the result in the bottom. To solve Eq. (21), we have to sum over all arrays of bottom matrix. The minus sign causes the change in shading on $\mathbf{d}$ $(\mathrm{A},\mathrm{A}).\hat{}2$ in the bottom matrix. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We readily see that $D^2(\mathrm{AB},\mathrm{A}) \neq D^2(\mathrm{AB},\mathrm{B})$. This shows that two monomers do not have the same RMSD or SVs.

Now, the following relation can be obtained (a derivation is given by Figures 5 and 6)

$$|D^2(AB,A)-D^2(\mathrm{AB},\mathrm{B})|=|d^2(\mathrm{B},\mathrm{B})-d^2(\mathrm{A},\mathrm{A})| \qquad (21)$$

where $d^2(\mathrm{B},\mathrm{B})=\langle \mathbf{d}(\mathrm{B},\mathrm{B})|\mathbf{d}(\mathrm{B},\mathrm{B})\rangle$ has a similar definition to the one mentioned after Eq. (12). However, this definition does not provide a normalized measure to compare with the SV. Now, we define SS as follows

$$\mathrm{SS}=\left|S(\mathrm{AB},\mathrm{A})-S(\mathrm{AB},\mathrm{B})\right| \cong \left|\frac{d^2(\mathrm{B},\mathrm{B})-d^2(\mathrm{A},\mathrm{A})}{2(d^2(\mathrm{AB},\mathrm{AB})+d^2(\mathrm{A},\mathrm{A}))}\right| \qquad (22)$$

where we have defined $S(\mathrm{AB},\mathrm{A})$ as

$$S(\mathrm{AB},\mathrm{A})=\frac{D^2(\mathrm{AB},\mathrm{A})}{2(d^2(\mathrm{AB},\mathrm{AB})+d^2(\mathrm{A},\mathrm{A}))} \qquad (23)$$

$S(\mathrm{AB},\mathrm{B})$ is defined similarly. Here, to arrive at the right-hand side of Eq. (22) we approximate $d^2(\mathrm{A},\mathrm{A}) \cong d^2(\mathrm{B},\mathrm{B})$ in the dominator. We will show in what follows that this is a very reasonable approximation.

The computation of d^2 for 1JFF, and its two monomers 1JFF-A and 1JFF-B, results in $d^2(\mathrm{AB},\mathrm{AB})=7.1519\times10^{10}$, $d^2(\mathrm{A},\mathrm{A})=8.7874\times10^{9}$ and $d^2(\mathrm{B},\mathrm{B})=9.7127\times10^{9}$ in units of square Angstroms. From these values, one can obtain $d^2(\mathrm{A},\mathrm{B})=d^2(\mathrm{B},\mathrm{A})=2.6509\times10^{10}$. To make more sense of the numerical values obtained, in addition to our real case, we compute SS in Eq. (22) for a totally correlated case ($d^2(\mathrm{AB},\mathrm{AB})=0$) and for an uncorrelated case($d^2(\mathrm{AB},\mathrm{AB})=d^2(\mathrm{B},\mathrm{B})+d^2(\mathrm{A},\mathrm{A})$). Thus, we have

$$\mathrm{SS} \cong \begin{cases} \frac{1}{6}\left(1-\frac{d^2(\mathrm{A},\mathrm{A})}{d^2(\mathrm{B},\mathrm{B})}\right)=0.0159, & \text{A and B are totally different } d^2(\mathrm{A},\mathrm{B})=0 \\ 0, & \text{A and B are totally similar} \\ 0.0058, & \text{In our case} \end{cases} \qquad (24)$$

The above equation shows that $0 \leq \mathrm{SS} \leq 0.0159$. We normalize SS to 1/2 so that we have

$$\mathrm{SS}_{\mathrm{Normalized}}=\frac{1}{2}\left(1-\frac{\mathrm{SS}}{0.0159}\right)=\begin{cases} 0 & \text{A and B are totally different} \\ \frac{1}{2} & \text{A and B are totally similar} \end{cases} \qquad (25)$$

Thus, for our case we have: $\mathrm{SS}_{\mathrm{Normalized}}=0.3176$. Now, we come back to SVs obtained in Table I, where $\mathrm{SV}(\mathrm{AB},\mathrm{A})-\mathrm{SV}(\mathrm{AB},\mathrm{B})=0.4967-0.2091=0.2876$. We see that these two results are close in value. Note that $\mathrm{SS}_{\mathrm{Normalized}}$ is obtained by using an approximation. Here in the structural form, we have shown why the two monomers of 1JFF are different. In particular, we have shown that, if one of these monomers is similar to 1JFF, then the other cannot be similar (and vice versa).

The difference between SV and RMSD

When two structures (A and B) do not have the same size (either different numbers of atoms or residues), the RMSD is an approximate value of structural similarity and, depending on the order of the two structures begin compared, RMSD can calculate two different values. That is, RMSD(A,B) is not equal to RMSD(B,A). This can be verified by using the Matlab batch pdbsuperpose function to compute RMSD. For example, for the two proteins 2nsg (with 1852 atoms) and 1adn (with 731 atoms), the pdbsuperpose function gives two different values. That is, `pdbsuperpose(2nsg,1adn)` gives a value of 7.39249 while `pdbsuperpose(1adn,2nsg)` gives a value of 0.382794. However, the SV values in both cases are the same and equal to a value of 0.352874.

Mathematically, we see this difference in what follows. We find the RMSD between two datasets A and B, RMSD(A,B), within a coordinate system, when two structures have the same size n-point (e.g., the same number of atoms) and centroid at the origin of the coordinate system. Firstly, we find a rotational transformation R of the coordinate vector of B, $\mathbf{r}_\mathrm{B}$ onto the coordinate vectors of A, $\mathbf{r}_\mathrm{A}$. Then, to compute RMSD(A,B) we need to minimize $|\mathbf{r}_\mathrm{A}-R\mathbf{r}_\mathrm{B}R^t|^2$. Expanding this term gives

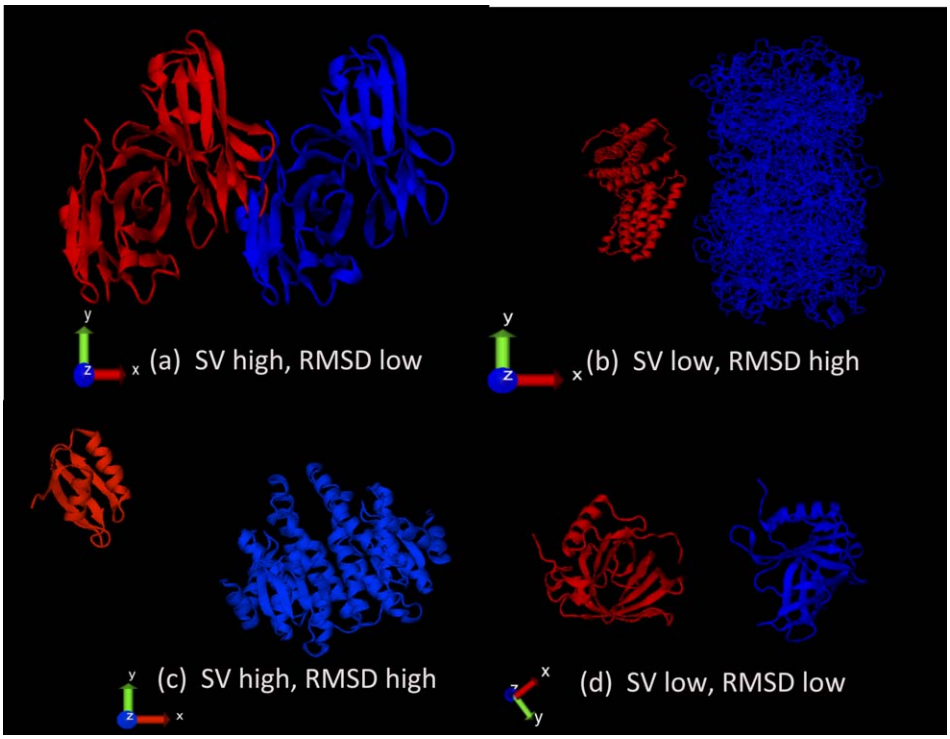

Figure 7

(Top row: Methods agree; **a**) 1a6w in blue and 1a6u in red. Two methods agree that both structures are similar. (**b**) 1au1 in red and 1a2v in blue. Two methods agree that both structures are very different. (Bottom row: Methods disagree; **c**) 1aw0 in red and 1aof in blue. SV is a better predictor of structure similarity because of the large difference in atom numbers. (**d**) 1rbp in red and 1brq in blue. RMSD is a better predictor of structure similarity. Here, both structures have the same number of atoms. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

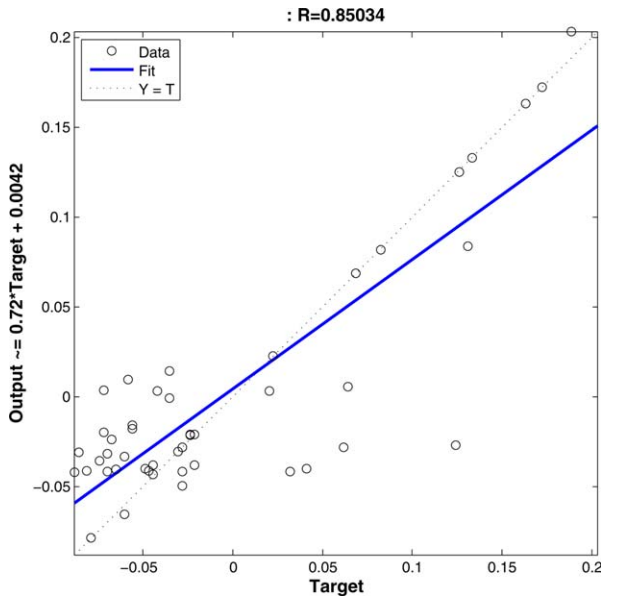

Figure 8

$RMSD_N$ versus $RMSD_N^{fit}$ and correlations for the 48 data set taken from Li *et al.*[1] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

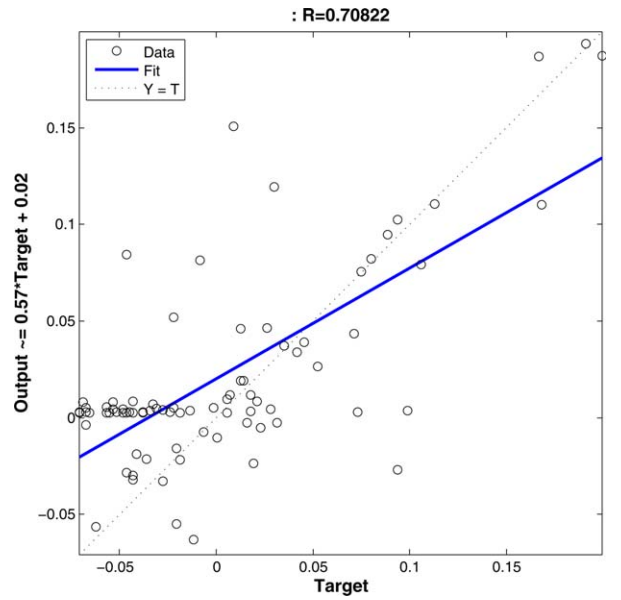

Figure 9

$RMSD_N$ versus $RMSD_N^{fit}$ and correlations for the 86 data set taken from Li *et al.*[1] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

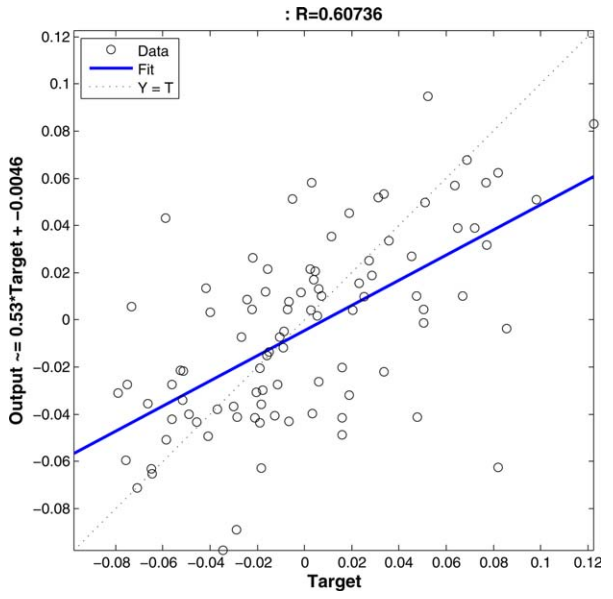

Figure 10

$RMSD_N$ versus $RMSD_N^{fit}$ and correlations for the dataset1, an 89-pair comparison taken from Kolodny *et al.*[4] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$\left(|\mathbf{r}_A|^2+|\mathbf{r}_B|^2-2\mathbf{r}_A\cdot R\mathbf{r}_B R^t\right).$$

The last term in the above expression is positive. Thus, the worst-case superposition occurs when the last term is zero. Thus, to minimize RMSD, we should maximize the last term, $(\mathbf{r}_A\cdot R\mathbf{r}_B R^t)$, in the above expression. The minimization can be competed by computing the correlation matrix

$$C_{ij}=\frac{1}{n}\sum_{k=1}^{n} r_{A,ik}r_{B,jk}\quad (i,j=1,2,3),$$

and replacing $(\mathbf{r}_A\cdot R\mathbf{r}_B R^t)$ by the maximum eigenvalue of **C**, λ_{Max}.[13,50] When the two structures are different sizes in the RMSD(A,B) computation (for example if B has fewer points than A), its void points are replaced by zeroes. This difference is size results in the discrepancy between the RMSD values given above for 2nsg and 1adn.

Another thing to note is that the RMSD has a range of potential values in the interval $[0,\infty)$, thus it has no upper limit. On the other hand, SV does have an upper limit of 1/2. This means that if the RMSD between two structures is very large, for example if RMSD(A,B) = 56 and RMSD(B,A) = 99, then both cases show full dissimilarity between two structures and, in general, the SVs are approximately zero for both these cases. To avoid inaccuracy from this difference of domain size between RMSD values and SV, we use a centroid and then normalize both datasets to unity (as is typically done for wave functions in quantum mechanics). Now we denote the normalized quantities as $RMSD_N$ and SV_N. Then, we define the following measure to compare them.

Taking into account the RMSD concept, we try to reproduce $RMSD_N$ values by using SV_N. We can define a transformation, T of SV_N onto $RMSD_N$ so that $T(SV_N)T^t=RMSD_N^{fit}$. Then, we define the correlation between $RMSD_N^{fit}$ and $RMSD_N$ as a comparative measure. To obtain $RMSD_N^{fit}$ we use the following Matlab batches in the feed-forward neural network fitting:

```
net = newfit(SV_N, RMSD_N,40);
[net,tr] = train(net,SV_N,RMSD_N);
```

$RMSD_N^{fit}$ = sim (net,SV_N;)

Figures 8–11 show $RMSD_N$ versus $RMSD_N^{fit}$ and their correlations, for four datasets: the 48 data set (Table S1 in Supporting Information), the 86 data set (Table S2 in Supporting Information), dataset1, an 89-pair comparison (Table S3 in Supporting Information), and dataset2, a 94-pair comparison (Table S4 in Supporting Information). The sets dataset1 and dataset2 are the same datasets as selected by Kolodny *et al.*[4] However, not all the data is used (since for some of the protein pairs, the Matlab batch cannot compute the RMSD values due to lack of pertinent information). Also, we have taken the PDB files for all datasets from the PDB-site. The 48 set and the 86 set are taken from Li *et al.*[1] The sets dataset1 and dataset2 (non-homologous) have different protein sizes ranging from 40 residues to approximately 1000 residues, and we randomly selected pairs of proteins in

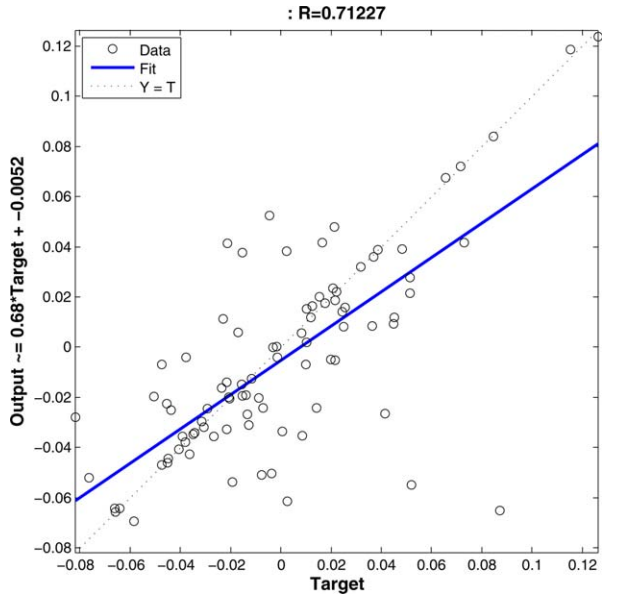

Figure 11

$RMSD_N$ versus $RMSD_N^{fit}$ and correlations for the dataset2, an 94-pair comparison taken from Kolodny *et al.*[4] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

these sets to compute RMSD values and SVs. The RSMD values for the 48 set and the 86 set are the same as those reported by Li *et al.*[1] For dataset1 and dataset2, RSMD values are computed by the following Matlab batch:

```
[d,RMSD]=pdbsuperpose(PDB1,PDB2,'model',
[11],'apply'', all','Display',false');
```

The size of *PDB2* must be less than or equal to the size of *PDB1*.

We do not expect agreement between RMSD values and SV because, unlike the SV, the RMSD value is not sensitive to topology[51] (i.e., the SV changes when the corresponding topology changes). Another reason we do not expect agreement is that RMSD depends on the size of the protein, where small size proteins tend to have a lower RMSD.[41,42,51] SV does not depend on size.

These reasons explained above help to explain why SV can be much different than the corresponding RMSD value. For example, in two liganded 1a6w (1774-atom) and unliganded 1a6u (1736-atom) proteins [illustrated in Fig. 7(a)], their topology is changed but their RMSD value with 1a4j (6626-atom) is the same (2.4 Angstrom) while the SV values are different, namely 0.08 and 0.34, respectively. Thus for the above reasons, we see that the RMSD value between 1a6w and 1a6u is 0.34 Angstrom, but the SV value is 0.19. The two proteins 1RBP and 1BRQ (both have 1408 atoms) are liganded and unliganded [as illustrated in Fig. 7(d)], respectively, but they have a small RMSD value of 0.6 Angstrom. Their geometries are different, so that SV is equal to 0.1.

CONCLUSIONS

This article introduces a new method to compare protein structures; it can be generalized to compare arbitrary shapes defined as a set of 3D coordinates. The novelty of our method lies in expanding the shape function using Wigner-D functions, showing that the expansion coefficients correspond to the structure factors, and using the RMSD measure in the reciprocal space (for the structure factors) to define a similarity value, namely the SV parameter. We show that this measure gives a corresponding similarity in the spatial domain because of the isometric property of the Fourier transform. We have verified our method by obtaining the shape function by using the structure factors and Wigner-D functions (see Fig. 2). The absolute values of the structure factors are the same as the intensities measured by X-ray scattering. We also show that the structure factor distribution is a Poisson distribution; as is well known, the intensity distribution for X-ray scattering is also a Poisson distribution. This result demonstrates the reliability of our method. The numerical results shown in Tables I–VII for SV (as well as the results shown in the tables given in Supporting Information) also confirm the reliability and usefulness of our method.

An important problem for similarity comparison methods is that the number of the protein atoms in an arbitrary pair of proteins is generally not the same. To address this problem, some methods use partial similarity measures between two proteins, like partial RMSD calculations. However, in our method, despite the fact that the number of atoms of the two proteins being compared is different, the number of structure factors is the same in reciprocal space. This is another important advantage of our method.

ACKNOWLEDGMENTS

S.M.S.F. acknowledge grant number 2/22306 from Ferdowsi University of Mashhad. JAT gratefully acknowledges research support received from the National Science and Engineering Research Council (Canada), the Canadian Breast Cancer Foundation, and the Allard Foundation.

REFERENCES

1. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D. Characterization of local geometry of protein surfaces with the visibility criterion. Proteins 2008;71:670–683.
2. Rupp B, Wang J. Predictive models for protein crystallization. Methods (San Diego, Calif.) 2004;34:390–407.
3. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics (Oxford, England) 2006;22:195–201.
4. Kolodny R, Petrey D, Honig B. Protein structure comparison: implications for the nature of "fold space", and structure and function prediction. Curr Opin Struct Biol 2006;16:393–398.
5. Carugo O. Recent progress in measuring structural similarity between proteins. Curr Protein Peptide Sci 2007;8:219–241.
6. Wolfson HJ, Rigoutsos I. Geometric hashing: an overview. IEEE Comput Sci Eng 1997;4:10–21.
7. Funkhouser T, Min P, Kazhdan M, Chen J, Halderman A, Dobkin D, Jacobs D. A search engine for 3D models. ACM Trans Graph 2003;22:83–105.
8. Kihara D, Sael L, Chikhi R, Esquivel-Rodriguez J. Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. Curr Protein Peptide Sci 2011;12:520–530.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
10. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36: D419–425.
11. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure (London, England: 1993) 1997;5:1093–1108.
12. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. A database of protein structure families with common folding motifs. Protein Sci 1992;1:1691–1698.
13. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol 1994;235:625–634.
14. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci 2001; 10:1470–1473.

15. Levitt M, Gerstein M. STRUCTAL. A structural alignment program. Stanford University; 2005. Available from: http://csb.stanford.edu/levitt/Structal.
16. Wigner EP. Gruppentheorie und ihre Anwendungen auf die Quantenmechanik der Atomspektren. Braunschweig: Vieweg Verlag; 1931.
17. Potts D, Prestin J, Vollrath A. A fast algorithm for nonequispaced Fourier transforms on the rotation group. Numer Algorithms 2009; 52:355–384.
18. Hielscher R, Potts D, Prestin J, Schaeben H, Schmalz M. The Radon transform on SO(3): a Fourier slice theorem and numerical inversion. Inverse Problems 2008;24:025011.
19. Lipson H, Taylor CA. Fourier transforms and X-ray diffraction. London: Bell; 1958.
20. Löwe J, Li H, Downing KH, Nogales E. Refined structure of alpha beta-tubulin at 3.5 A resolution. J Mol Biol 2001;313:1045–1057.
21. Arfken JB, Weber HB. Mathematical methods for physicists, 6th ed. Burlington, MA: Elsevier; 2005.
22. Curtis CW. Linear Algebra. Undergraduate texts in mathematics. New York: Springer-Verlag; 1984.
23. Titchmarsh EC. Introduction to theory of the Fourier integrals, 2nd ed. London: Oxford University Press; 1948.
24. Huldt G, Szoke A, Hajdu J. Diffraction imaging of single particles and biomolecules. J Struct Biol 2003;144:219–227.
25. Wilson AJC. The probability distribution of X-ray intensities. Acta Crystallogr 1949;2:318–321.
26. Ritchie DW, Kemp GJ. Protein docking using spherical polar Fourier correlations. Proteins 2000;39:178–194.
27. Novotni M, Klein R. Shape retrieval using 3D Zernike descriptors. Comput Aided Des 2004;36:1047–1062.
28. Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, Kihara D. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins 2008;72:1259–1273.
29. Venkatraman V, Yang YD, Sael L, Kihara D. Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics 2009;10:407.
30. Chikhi R, Sael L, Kihara D. Protein binding ligand prediction using moments-based methods. In: Kihara D, editor. Protein function prediction for omics era. Netherlands: Springer; 2011. pp 145–163.
31. An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 2005;4:752–761.
32. Brunet D, Vrscay ER, Wang Z. On the mathematical properties of the structural similarity index. IEEE Trans Image Process 2012;21: 1488–1499.
33. Heyne S, Costa F, Rose D, Backofen R. GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics 2012;28:i224–i232.
34. Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics (Oxford, England) 2012;28: 2782–2788.
35. Carugo O, Pongor S. Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. J Mol Biol 2002;315:887–898.
36. Huang C-D, Lin C-T, Pal NR. Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification. IEEE Trans Nanobiosci 2003;2:221–232.
37. Rogen P, Fain B. Automatic classification of protein structure by using Gauss integrals. Proc Natl Acad Sci USA 2003;100:119–124.
38. An J, Totrov M, Abagyan R. Comprehensive identification of "druggable" protein ligand binding sites. Genome Inform 2004;15:31–41.
39. Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3D shape descriptors. In: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing. SGP'03. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association; 2003. pp 156–164.
40. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
41. Betancourt MR, Skolnick J. Universal similarity measure for comparing protein structures. Biopolymers 2001;59:305–309.
42. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
43. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci 1998;95: 5913–5920.
44. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng 1994;7: 1059–1068.
45. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
46. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
47. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
48. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003;334:793–802.
49. Mavridis L, Venkatraman V, Ritchie D, Morikawa H, Andonov R, Cornu A, Malod-Dognin N, Nicolas J, Temerinac-Ott M, Reisert M, Burkhardt H, Axenopoulos A, Daras P. SHREC'10 Track: Protein Models; 2010.
50. McLachlan A. Gene duplications in the structural evolution of chymotrypsin. J Mol Biol 1979;128:49–79.
51. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:101186.
52. Pandit S, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics 2008;9:101186.