Evaluating Protein Similarity from Coarse Structures

Yong Wang, Ling-Yun Wu, Ji-Hong Zhang, Zhong-Wei Zhan, Xiang-Sun Zhang, and Luonan Chen

Abstract—To unscramble the relationship between protein function and protein structure, it is essential to assess the protein similarity from different aspects. Although many methods have been proposed for protein structure alignment or comparison, alternative similarity measures are still strongly demanded due to the requirement of fast screening and query in large-scale structure databases. In this paper, we first formulate a novel representation of a protein structure, i.e., Feature Sequence of Surface (FSS). Then, a new score scheme is developed to measure the similarity between two representations. To verify the proposed method, numerical experiments are conducted in four different protein data sets. We also classify SARS coronavirus to verify the effectiveness of the new method. Furthermore, preliminary results of fast classification of the whole CATH v2.5.1 database based on the new macrostructure similarity are given as a pilot study. We demonstrate that the proposed approach to measure the similarities between protein structures is simple to implement, computationally efficient, and surprisingly fast. In addition, the method itself provides a new and quantitative tool to view a protein structure.

Index Terms—Protein structure, structure comparison, protein surface.

1 Introduction

Elucidation of protein functions is one of the key tasks in molecular biology. It is well known that how a protein functions or interacts with other molecules is closely related to its structure information. On the other hand, the fact that structures deposited in PDB increase by about 30-50 entries every week also calls for the development of efficient data mining techniques that extract the useful information effectively and quickly.

A direct and important method to meet this challenge is protein structure comparison. There are several reasons for using structure comparison [1], [2]. First, it comes from the need for managing and organizing the great amount of structural data. Existing secondary databases such as SCOP, CATH, and FSSP provide clustering and classification for protein structures. However, the inconvenience for the combination of a good deal of human expertise requires fully automatic methods. Also, the need to achieve fast or efficient search and query in large structure databases provides another motivation. Second, it is important to detect a distant evolutionary relation by structure resemblance from the evolutionary viewpoint. For instance, it can be applied to infer the function of a new protein by comparing its structure to the known ones to find distant evolutionary neighbors. Third, many existing structure comparison algorithms are used to find the maximal common substructure, i.e., a motif or a domain, which is closely related to biological functions. Finally, structure comparison method is also a useful assessment tool for the protein structure prediction.

To compare protein structures, a number of different automatic methods have already been proposed as indicated in the comprehensive reviews [1], [2], [3], [4], [5]. Generally, these works can be roughly classified into two categories. Early research works mainly focus on finding the optimal rigid-body superposition of two structures such that the root mean square deviation (RMSD) between the aligned C_α atoms is minimized. These methods are simply named as a structure alignment dedicated to get the best correspondence between two proteins. The common characteristics of these methods are the element-based representation of a structure such as atoms, residues, and secondary structure elements (SSEs) and the RMSD scoring scheme to measure the similarity [6], [7], [8], [9]. The fundamental difficulties encountered come from three aspects. First, RMSD is a useful measure of the similarity but only effective between nearly identical structures [10]. It lacks good mathematical properties as a distance and also introduces many undetermined parameters. In addition, it is known that the RMSD depends not only on conformational differences but also on the dimensions of the structures [11], [12]. Second, to identify the correspondence between the elements of two proteins requires time-consuming computation [9]. Third, proteins do possess internal degrees of freedom, which violate the basic assumptions in those alignment methods, where a protein

- *Y. Wang, L.-Y. Wu, and X.-S. Zhang are with the Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, No. 55 Zhongguancun East Road, Beijing, 100080, China. E-mail: ywang@amss.ac.cn, lywu@amt.ac.cn, zxs@amt.ac.cn.*
- *J.-H. Zhang is with the School of International Business, Beijing Foreign Studies University, Beijing 100081, China. E-mail: zhangjihong@bfsu.edu.cn.*
- *Z.-W. Zhan is with the China Aerospace Engineering Consultation Center, China. E-mail: wallen2006@163.com.*
- *L. Chen is with the Department of Electronics, Information, and Communication Engineering, Osaka Sangyo University, Nakagaito 3-1-1, Daito, Osaka 574-8530, Japan. E-mail: chen@eic.osaka-sandai.ac.jp.*

Manuscript received 22 Mar. 2006; revised 11 June 2006; accepted 20 Aug. 2007; published online 30 Aug. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0086-0306.
Digital Object Identifier no. 10.1109/TCBB.2007.70250.

TABLE 1
List of Recent Alignment-Free Methods of Measuring Protein Structure Similarity

Method	Representation of structure	Ref
PRIDE	The protein structure is described by a set of distributions of $C_\alpha(i) - C_\alpha(i+n)$ distances, where i is the residue number in the protein chain and n is an integer ranging from 3 to 30.	[15], [16]
MolCom	The automated MolCom method incorporates an octtree strategy to partition and examine molecular properties in three-dimensional space at multiple levels of analysis.	[17]
CMO	A 3D fold of a protein structure is concisely represented by a contact map which is an undirected graph.	[18]
GSM	Through computing the topological invariants of the polygonal curve connecting the C_α atoms. Each domain is assigned a 30-dimensional vector.	[19], [20]
USM	Every protein is denoted by its contact map.	[21]
Topology	A protein is mapped to a three-dimensional array making use of a Delaunay-based topological mapping which is the representative of the global structural topology.	[22]
MATRAS	A combination of local structure and solvent accessibility, residue-residue distance and secondary structure elements (SSE).	[2], [23]

is treated as a rigid body [1]. As a result, it has been indicted that there is no perfect method to solve this problem [13].

A number of novel methods to measure the similarity between protein three-dimensional (3D) structures without superposing them or aligning their equivalent residues have recently been proposed (see Table 1). In contrast to the direct comparison of two structures, these methods compare features between two proteins. In particular, they all accord with such a framework that the relevant features are extracted and represented in structure descriptions, and the equivalence is obtained by the specific score scheme [3], [14]. A representative of these methods is the GSM, which classifies 20,937 protein domains into multiple levels and achieves 96 percent agreement with the CATH. The recent progress shows that a protein can be represented from many aspects, and more and more abstract mathematics tools are involved in this field. Furthermore, as pointed in paper [5], alternative similarity measures and fast methodologies are still strongly demanded because different aspects of protein 3D structures may be relevant to various biological problems.

In this paper, a novel screening approach for protein structure comparison is proposed. There are two main contributions, i.e., the representation of a protein structure and the score scheme. First, a protein structure is represented by a coarse structure, which is approximated by supporting planes of the convex hull of its backbone. Then, an FSS (feature sequence of surface) is defined and easily computed with respect to each plane. Hence, every 3D protein structure is represented by a feature sequence and, further, a feature distribution after normalization. Second, a similarity score based on information theory called the function of degree of disagreement (FDOD function) [24] is applied to achieve the comparison of two feature distributions. It provides a new measure of protein similarity and has many good mathematical properties. Combining the new representation of protein structures and the score scheme, a pairwise comparison algorithm is designed, and it can be easily generalized to the problem of multiple-structure comparison [25], which is explained in detail in this paper.

The remainder of the paper is organized as follows: In Section 2, the definition of the FSS representation and descriptions of the FDOD comparison method are presented. Then, the numerical results are given in unsupervised cases on several existing protein data sets and in supervised cases on a pilot classification of whole CATH database. Finally, a detailed discussion and conclusion about the new macrostructure similarity and further development of screening database search method are given.

2 Methods

2.1 FSS Representation of a Protein Structure

Different protein shape representations are used to mine a huge amount of protein 3D structure data. The choice of a proper representation depends on the biological task at hand. For example, a protein can be viewed as a set of its individual atoms, a folded 3D curve of amino acids, or a much coarser assemble of SSEs. Most of existing comparison methods often view a protein as a sequence of C_α atoms described by the positions of their centers, which is called the backbone of the protein. The backbone representation of a protein is adopted in this paper as the start point.

Our motivation to develop a new protein representation comes from the successful applications of contour analysis in mechanics, aerography, and iatrology, where the contour map is used to visually explain the information that the data provide. The basic idea of this technology is to project the 3D data set onto a fixed two-dimensional (2D) plane, and the information in the third dimension is characterized by isopleth. Through this projection, the data can be interpreted in a 2D view, and the patterns contained in the data can be identified in a relatively easy way. Also, a three-dimensional object can be easily constructed (such as machinist processes hardware according to component drawing) through its contour maps without loss of the information. In this paper, we use the contour maps to analyze the protein backbone by projecting its 3D data onto 2D planes. For example, the backbone of protein 1acj is showed by its contour maps on three planes in its reference frame, respectively, in Fig. 1.

To describe a protein's global shape with a set of contour maps and keep the representation independent of the rotation transformation, the reference frame must depend on the protein's inherent structure. Protein surfaces satisfy this requirement and ensure that a set of contour maps by projecting a protein to its surfaces will keep invariant when the protein is rotated. Also, protein surfaces provide a special view from outside to look into the protein structure. Furthermore, since protein function is closely related with its surface, it is expected that structural analysis of protein surface can identify function determinants that are independent of sequences or secondary structures. It is also a powerful tool to highlight cases of possible convergent or divergent evolution [26], [27]. However, on the other hand, the protein surface is a complex concept and hard to be expressed exactly. In our protein model, the surface is approximated by a set of facets of a protein's convex hull with respect to its backbone [28] (see Fig. 2 and the detailed

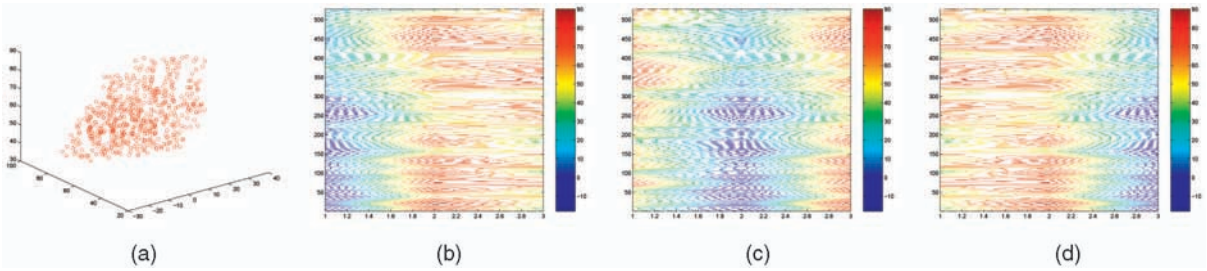

Fig. 1. The C_α atoms of protein 1acj in the reference frame and the contour maps of protein 1acj corresponding to three planes. (a) The C_α atoms of 1acj in reference frame. (b) The contour map of XY plane. (c) The contour map of XZ plane. (d) The contour map of YZ plane.

mathematical expression and computing algorithm for convex hull are presented in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70250). Then, the facet-containing supporting planes of the protein's convex hull are taken as a set of reference frames for the contour maps of that protein.

The problem rising in the protein representation is how to express and compare the contour maps. Supposing that a protein's backbone is projected to one of its supporting planes, as shown in Fig. 3, then, the structure information is represented by the projected points located on the supporting plane and the distance between every C_α atom and its projected point. To simplify the expression, we only consider the distance sequence corresponding to the C_α atom sequence. The distance sequence of the backbones to the supporting plane possess abundant information, but here, only the average value of the distance sequence is extracted as the main feature related to that supporting plane.

With the feature defined on every supporting plane of the protein, or every polygonal facet of the convex hull, a protein can be encoded as an FSS, which is a novel protein structure representation. It should be noted that the FSS so far cannot be used directly to compare two proteins for two reasons. First, two different proteins have different lengths of FSS; second, each protein can have exponentially many FSSs because the sequence of supporting planes can be put in different facet order configuration. To overcome these difficulties, it is convenient to convert the sequence to a distribution of the distance value. Then, a feature distribution is extracted based on the FSS, and the protein structure comparison problem is expressed as a problem to calculate the distance between two feature distributions. As shown in Fig. 4, the procedure of mapping a protein to the corresponding feature distribution can be summarized in three steps.

Now, we formally state the FSS representation of a protein structure. Based on the coordinates of C_α atoms, assume that the structure of a protein is completely determined by its amino acid set:

$$\boldsymbol{X} = \{\boldsymbol{x}^k\} = \left\{\left(x_1^k, x_2^k, x_3^k\right), k = 1, \ldots, N\right\}, \tag{1}$$

where (x_1^k, x_2^k, x_3^k) is the coordinate of the kth C_α atom, and N is the total number of C_α atom in the protein. Let $conv(\boldsymbol{X})$ be the convex hull of the set $\boldsymbol{X}$ (see the supplementary material for a detailed definition and computation algorithm of the convex hull, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70250).

Let $\{F_h\}$ be the set of facets of $conv(\boldsymbol{X})$ and $\{P_h\}$ be the corresponding set of planes, or supporting planes. For a

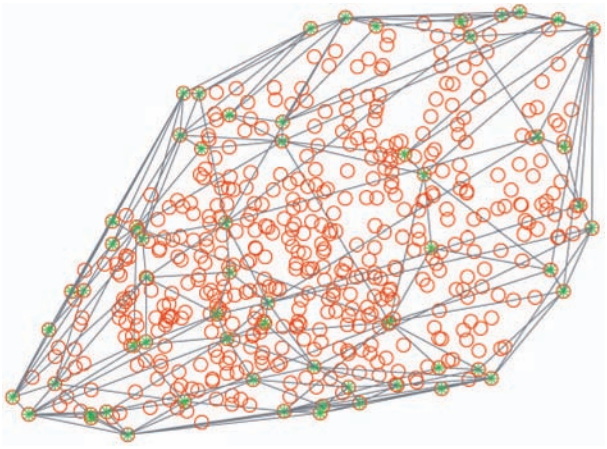

Fig. 2. The convex hull representation of protein 1acj. The circles denote the C_α atoms inside the hull, and the stars are the C_α atoms located on the surface of the convex hull. The protein surface is approximated by a set of polygonal facets, i.e., the convex hull surface.

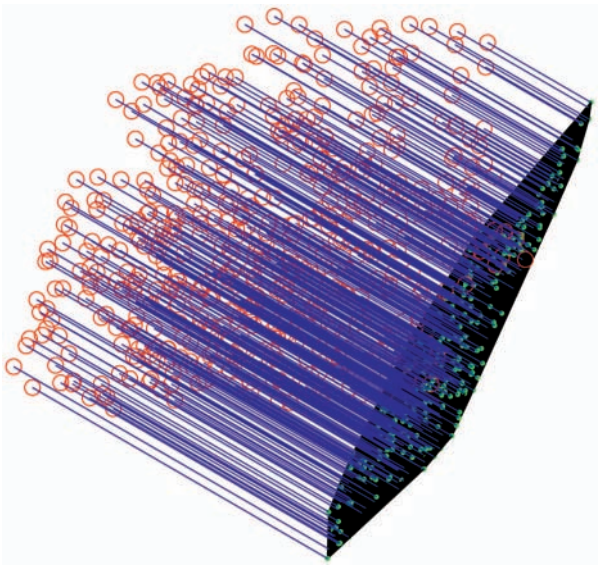

Fig. 3. The sketch map of projecting a protein's backbone on one of its supporting planes. A red circle in the map denotes the C_α atom, and a green star represents the projected point on the supporting plane. The length of the line between them (distance between the red circle and the green star) and the location of projected points record the backbone's structure information.

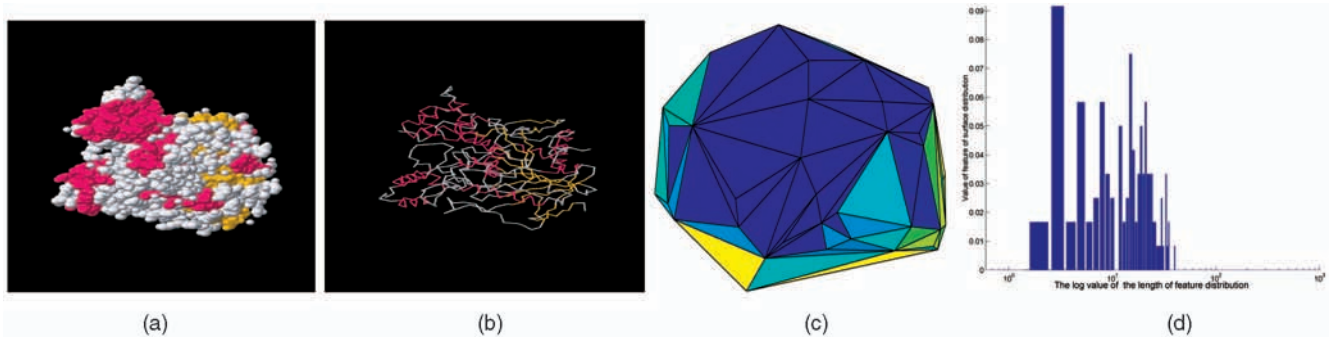

Fig. 4. The sketch map of the procedure of how a protein is mapped to a feature distribution. The representative protein is the PDB entry 1acj. (a) The protein graph by Rasmol 2.7.1.1. (b) The backbone with only C_α atoms. Through computation, every plane of the convex hull of the protein is shown in (c). Finally, the feature corresponding to every plane is calculated, and the feature distribution of the protein is shown in (d). (a) The space-fill view of 1acj. (b) The backbone view of 1acj. (c) The convex hull of 1acj. (d) The feature distribution of 1acj.

given supporting plane $P_h : \boldsymbol{a}_h^T \boldsymbol{x} = b_h, h = 1, \cdots, H$, where H is the number of facets, the projected point of the C_α atom $\boldsymbol{x}^k$ on the plane is written as $\boldsymbol{y}^k$. The distance between them can be calculated by $\|\boldsymbol{x}^k - \boldsymbol{y}^k\|$. With these preparations, the feature of P_h can be extracted as follows:

A feature W^h corresponding to a plane P_h of the convex hull is defined by the following points-to-plane projection problem:

$$W^h = \sum_{k=1}^{N} \frac{1}{N} \|\boldsymbol{x}^k - \boldsymbol{y}^k\|,$$
$$\boldsymbol{y}^k : \begin{cases} \min & \|\boldsymbol{x}^k - \boldsymbol{y}^k\|, \\ \text{s.t.} & \boldsymbol{a}_h^T \boldsymbol{y}^k = b_h, \\ & \|\boldsymbol{a}_h\| = 1. \end{cases} \tag{2}$$

Here, we provide geometric explanations to clearly demonstrate the concept of the new feature. In fact, the calculated $\|\mathbf{x}^k - \mathbf{y}^k\|$ is geometrically the vertical distance of $\mathbf{x}^k$ to the given plane, and the objective function of the optimization problem (2) is the average of vertical distances for all the C_α atoms in the protein. On the other hand, the feature can also be interpreted as a kind of average potential energy. Given a fixed supporting plane, the distance between a C_α atom and its projected point can be viewed as the potential energy of that atom, provided that all the residues represented by their C_α atoms are assumed as basic elements of the structure. Thus, the objective function of the optimization formulation (2) is the average potential energy for all the projected C_α atoms.

Suppose that a protein A has H convex hull facets, from which we extract H features $\{W_i^A, i \in \{1, \cdots, H\}\}$ by model (2). Then, the protein is denoted by its FSS $\boldsymbol{W}^A \in \mathbb{R}^H$:

$$\boldsymbol{W}^A = (W_1^A, \ldots, W_H^A).$$

By the following normalization for $\boldsymbol{W}^A$, we have

$$\boldsymbol{w}^A = \boldsymbol{W}^A / E,$$

we obtain $\boldsymbol{w}^A = (w_1^A, w_2^A, \cdots, w_H^A)$, where

$$E = \max\{W_1^A, \cdots, W_H^A\}.$$

Now, define a casting mapping $f : [0,1] \longrightarrow \mathbb{R}^S$, $S \in N$ as

$$f(x) = \boldsymbol{e}_i^T, x \in \left[\frac{i-1}{S}, \frac{i}{S}\right], i \in \{1, \ldots, S\},$$

where $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_S$ are the normal base vectors in $\mathbb{R}^S$, and $\boldsymbol{e}_i^T$ is the transpose of $\boldsymbol{e}_i$. The casting mapping f can be expressed in a precise way as

$$f(x) = \begin{cases} \boldsymbol{e}_1^T, & x \in \left[0, \frac{1}{S}\right], \\ \cdots \\ \boldsymbol{e}_i^T, & x \in \left[\frac{i-1}{S}, \frac{i}{S}\right], \\ \cdots \\ \boldsymbol{e}_S^T, & x \in \left[\frac{S-1}{S}, 1\right]. \end{cases}$$

Generally, S should be a larger integer. In the computation, we take $S = 10{,}000$. Extensive numerical tests show that different choice of parameter S has little effect on the results.

The feature of surface distribution $\boldsymbol{P}^A = (p_1^A, \cdots, p_S^A)$ of a protein is defined by the following formula:

$$\boldsymbol{P}^A = \frac{\sum_{i=1}^{H} f(w_i^A)}{H},$$

where f is the casting mapping, and w_i^A is the ith element of the normalized feature sequence $\boldsymbol{w}^A$ defined above. Obviously, $\sum_{i=1}^{S} p_i^A = 1$.

2.2 FDOD Scoring Scheme

FDOD is a new measure of information discrepancy [25]. It has been successfully used to measure the discrepancy between DNA sequences and amino acid sequences from different species in the study of phylogeny and prediction of protein structural classes. This measure has a close connection with Shannon entropy and has many good mathematical characteristics such as symmetry, boundedness, triangle inequality, absolute continuity, symmetric recursiveness, monotonicity, effectiveness in singular case, and convexity. Also, this measure is applicable to the multiple sequence comparison [25]. In this paper, we aim to develop a method that is able to apply to the problems of both protein pairwise and multiple structure comparisons.

Given a set of n distributions of m elements, we have

$$\boldsymbol{U}_1 = (p_{11}, \ldots, p_{m1}),$$
$$\cdots$$
$$\boldsymbol{U}_n = (p_{1n}, \ldots, p_{mn}),$$

where $\sum_{i=1}^{m} p_{ik} = 1$, $k = 1, \cdots, n$.

The FDOD measures are defined as

$$F(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n)=\sum_{k=1}^{n}\sum_{i=1}^{m}p_{ik}\log\frac{p_{ik}}{\sum_{k=1}^{n}p_{ik}/n}, \tag{3}$$

$$F_k(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n)=\sum_{i=1}^{m}p_{ik}\log\frac{p_{ik}}{\sum_{k=1}^{n}p_{ik}/n}, \tag{4}$$

where $0\log 0=0$ and $0\log(0/0)=0$ are defined. $F(U_1,\cdots,U_n)$ denotes a measure of discrepancy among n distributions, while $F_k(U_1,\cdots,U_n)$ denotes a measurement of discrepancy between the kth distribution and all other distributions in the group.

2.3 Classification Scheme and Data Preparation

As indicated in this paper, the FDOD score scheme is proposed for the comparison between one FSS distribution and a group of FSS distributions. Then, we adopt the new similarity measure in the framework of machine learning. The classification procedure can be described as follows:

- **Step 1: Feature extraction.** For each protein in the training data set and the testing data set, the FSSs are extracted from their PDB files.
- **Step 2: Training.** According to known classification, we assign proteins in the training data set into several different groups such that the proteins in the same group have same class label. Supposing that the training data set T is divided in n groups or subsets, then T is the union of these subsets as

$$T=T_1\cup T_2\cdots\cup T_n.$$

- **Step 3: Recognition.** For each query protein q in the testing data set, the discrepancy of q and group T_i is computed from (4) and is denoted by $d_q^{T_i}$. Similarly, we have $d_q^{T_1},\cdots,d_q^{T_n}$. Accordingly, the query protein q is assigned to group T_q when

$$d_q^{T_q}=\min\left\{d_q^{T_1},\ldots,d_q^{T_n}\right\},$$

  where $T_q\in\{T_1,\cdots,T_n\}$.
- **Step 4: Performance analysis.** According to the nature of the research work, we define the sensitivity and specificity of the group k as widely used in classification performance measure by

$$\text{Sensitivity}=\frac{\text{TP}_k}{\text{TP}_k+\text{FN}_k},$$

  where TP_k (True Positive) is the number of proteins in the kth group that have been classified correctly. FN_k is the number of proteins in the kth group that have been classified into other groups:

$$\text{Specificity}=\frac{\text{TN}_k}{\text{TN}_k+\text{FP}_k},$$

  where TN_k (True Negative) is the number of proteins in the kth group that have been classified wrongly. FP_k is the number of proteins wrongly classified in the kth group.

Also, the overall correct rate of classification is defined as

$$\text{Overall correct rate}=\frac{\sum_{k=1}^{n}\text{TP}_k}{\sum_{k=1}^{n}(\text{TP}_k+\text{FN}_k)}.$$

The Matthew correlation coefficient (MCC) uses all four numbers TP_k, TN_k, FP_k, and FN_k and can provide a more balanced evaluation of the prediction than the percentages:

$$\text{MCC}_k=\frac{\text{TP}_k\times\text{TN}_k-\text{FP}_k\times\text{FN}_k}{\sqrt{(\text{TP}_k+\text{FN}_k)(\text{TP}_k+\text{FP}_k)(\text{TN}_k+\text{FP}_k)(\text{TN}_k+\text{FN}_k)}}. \tag{5}$$

CATH is a novel hierarchical classification of protein domain structures, and clusters proteins at four major levels, i.e., Class(C), Architecture(A), Topology(T), and Homologous superfamily (H). The CATH classifies the proteins on the domain level using manual and automatic methods, and every domain is assigned to a unique CATH ID and four explicit classification numbers to denote its Class(C), Architecture(A), Topology(T), and Homologous superfamily(H) belongings. Therefore, we choose the CATH database as our data set to examine, duplicate, and extend by the classification method based on the proposed new similarity measure.

To start, we obtained the version of CATH v2.5.1 released January 2004 and extracted domains according to the CATH v2.5.1 definitions from PDB protein structures. The inconsistency between PDB and CATH caused by entry update was removed. There are total 47,427 domains in the testing data set for the experiment, whereas the training data set was selected according the design of the experiments (see details in Section 3).

3 Results

The purpose of this paper is to develop a fast screening comparison method for protein structures, in particular, for coarse structures. The implementation procedure is simply as 1) first representing protein structures by FSSs and 2) then comparing the FSSs by FDOD scheme. Based on the new similarity score, the experiments and analysis listed in this section can be roughly cataloged into unsupervised and supervised cases. Partial results are also listed and explained in details in http://zhangroup.aporc.org/bioinfo/strucomp/ for concision.

3.1 Protein Structure Classification in Unsupervised Cases

3.1.1 Benchmarks on Existing Data Sets

Four data sets were picked from different articles as the initial examples to assess the ability of the new approach to measure the similarity of proteins from their macrostructures.

Leluk-Konieczny-Roterman data set. The Leluk-Konieczny-Roterman data set is a small data set first employed in Leluk et al. [29] and then used by USM [21] to test the different similarity measures. There are six proteins that belong to the same Alpha and Beta class, the same Serpins fold and the same Serpins family and superfamily in the

SCOP classification. The difference appears at the Protein and Species level. Our approach surprisingly properly clusters 1att, 1azx, and 2antL into the Antithrombin, 7apiA, and 2achI into Antitrypsin, Alpha-1, and 1ovaA into Ovalbumin.

David data set. The David data set was introduced in the Bostick and Vaisman [22] to test a new topological method for measuring protein structure similarity. There are 10 proteins: 1mli, 1ris, 2acy, 1a79A, 1avqA, 1a6m, 2hbg, 1b8dA, 1bu2A, and 1aisB, which belong to three different classes in the SCOP classification. The clustering result shows that only two proteins, i.e., 1a79A and 1avqA, are misplaced.

Chew-Kedem data set. This data set was used in Chew and Kedem [30] to assess the quality of a newly proposed method to measure consensus shapes, then recently employed in USM [21]. These are 36 medium size proteins in five different families: globins (1eca, 5mbn, 1hlb, 1hlm, 1babA, 1babB, 1ithA, 1mba, 2hbg, 2lhb, 3sdhA, 1ash, 1flp, 1myt, 1lh2, 2vhbA, and 2vhb), Alpha-Beta (1aa9, 1gnp, 6q21, 1ct9, 1qra, and 5p21), tim-barrels (6xia, 2mnr, 1chr, and 4enl), all Beta (1cd8, 1ci5, 1qa9, 1cdb, 1neu, 1qfo, and 1hnf), and Alpha (1cnp and 1jhg). Protein 2vhb was repeated two times (as 2vhb and 2vhbA) in order to check whether the approach can detect that the two are identical and induce a cluster where both appear together. There are only three (out of 36) proteins that seem to be misplaced, i.e., 1jhg, 1ct9, and 1cnp. Surprisingly, again, the result is the same as the clustering result in [21], which uses the USM method. In [21], the reason why the two proteins are not classified properly in their contact map overlapping view is explained in details.

Skolnick data set. This protein data set was first suggested by Jeffery Skolnick, containing 40 large proteins that are used in various recent papers [31]. The edition used in this paper is the 39 selections from it by Krasnogor and Pelta [21]. In Fig. 5, it is easy to see that the classification of this data set is almost perfect except the entry 1awzA. The tree in Fig. 5 is constructed by a fully automatic procedure based on the proposed new similarity measure between two proteins. The cluster is computed by the Neighbor-joining and UPGMA methods provided in Phylogeny Inference Package (PHYLIP) package and shown by software TreeView.

3.1.2 Exploring the Structure Classification of SARS Coronavirus

The epidemic of severe acute respiratory syndrome (SARS) is an atypical highly contagious pneumonia. Yang et al. [32] summarized that SARS coronavirus had affected 32 countries in the period from February to June 2003. In total, $\approx$8,500 people were infected, and $>$900 died from the disease. After the SARS, coronavirus is named publicly by the World Health Organization and member laboratories as the "SARS virus," a number of research groups worldwide began to undertake the identification of the causative agent. The important results appeared in [33] and [34], which show that from genome organization SARS coronavirus is similar to that of other coronaviruses. While phylogenetic analysis and sequence comparisons show that SARS-CoV is not closely related to any of the previously characterized coronaviruses.

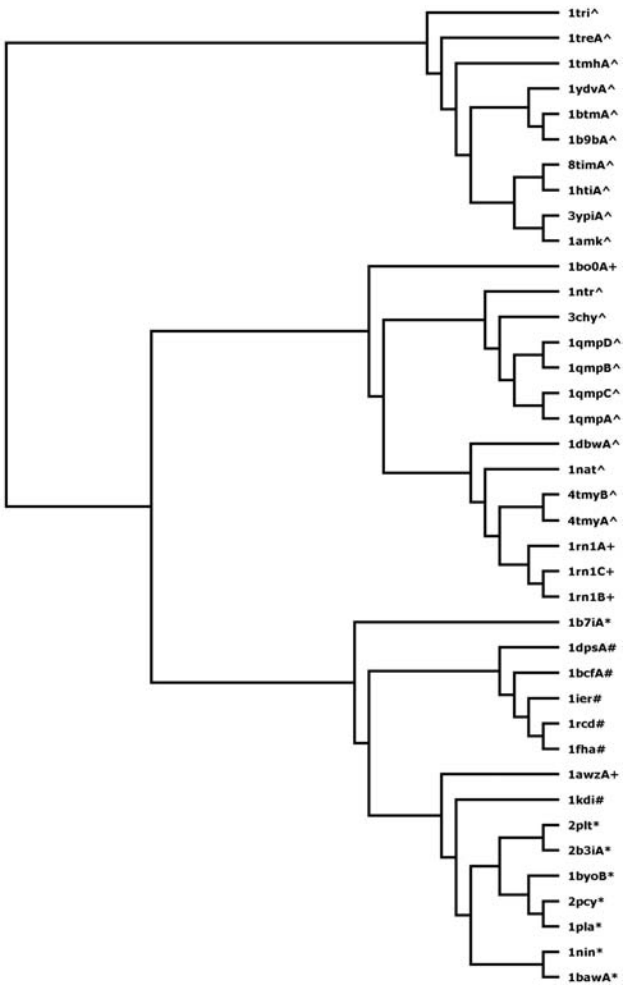

Fig. 5. Clustering result of the proteins from Skolnick data set according to the new similarity measure. The character following the PDB ID denotes the different class. + is Alpha and Beta proteins (a + b), * is all Alpha class, ^ denotes the Alpha and Beta proteins (a/b), and # is the all Beta proteins.

Since the protein structure information is more conservative than a protein sequence and related more closely with the protein function, many laboratories in the world conducted experiments to determine the 3D structure of SARS main protein. In [35], a predicted structure of SARS coronavirus protease was constructed through homology model and was deposited in PDB as entry 1p9t. Later, the crystal structures of the SARS-CoV main protease at different PH values and in complex with a specific inhibitor were reported in [32]. In this section, we applied the proposed comparison technique to structurally classify SARS coronavirus. To our best knowledge, this is the first time to study the classification of SARS main protein from structure comparison viewpoint.

To collect the structure data of coronaviruses, we searched the whole Protein Data Bank with the keyword "coronavirus." The query found 14 structures in the current PDB release. There are nine entries of them including the theoretical entry 1p9t relating to the Hydrolase function, as listed in Table 2. To be noted that, all the proteins except 1p9t have multiple chains. We constructed a SARS coronaviruses data set by extracting every chain from its PDB data file and named it as PDB ID plus its chain ID,

TABLE 2
Nine Entries of SARS Coronavirus Related to the Hydrolase Function

PDB ID	Description
1lvo	Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra-helical domain
1p9s	Coronavirus main proteinase (3Clpro) structure: basis for design of anti-sars drugs, 3C-Like protease of the human corona virus 229E
1p9u	Coronavirus main proteinase (3Clpro) Structure: basis for design of anti-sars drugs, Complex of 3C-Like protease and inhibitor
1q2w	X-Ray crystal structure of the SARS coronavirus main protease, Coronavirus main proteinase (3Cl-pro, putative coronavirus nsp2)
1p9t	Predicted structure of SARS coronavirus protease through the theoretical model
1uj1	Crystal structure of SARS coronavirus main proteinase (3Cl-pro), SARS-CoV Mpro, apoenzyme at pH 6.0
1uk3	Crystal structure of SARS coronavirus main proteinase (3Cl-pro), SARS-CoV Mpro, apoenzyme at pH 7.6
1uk2	Crystal structure of SARS coronavirus main proteinase (3Cl-pro), SARS-CoV Mpro, apoenzyme at pH 8.0
1uk4	Crystal structure of SARS coronavirus main proteinase (3Cl-pro), SARS-CoV Mpro in complex with the hexapeptidyl CMK inhibitor

respectively. Therefore, the data set is composed of the following 25 entries: 1lvoA, 1lvoB, 1lvoC, 1lvoD, 1lvoE, 1lvoF, 1p9sA, 1p9sB, 1p9uA, 1p9uB, 1p9uC, 1p9uD, 1p9uE, 1p9uF, 1q2wA, 1q2wB, 1p9t, 1uj1A, 1uj1B, 1uk2A, 1uk2B, 1uk3A, 1uk3B, 1uk4A, and 1uk4B.

The phylogenetic tree in Fig. 6 was constructed based on the proposed new similarity measure between two proteins using the Neighbor-joining and UPGMA methods provided in Phylogeny Inference Package (PHYLIP) package.

From the figure, we have two basic conclusions as follows:

1. The figure shows that SARS-CoV is not closely related to any of the previously characterized coronaviruses. Clearly, both the crystal structure and predicted structure of SARS coronaviruses belong to one group. Such result indicates that SARS is a novel coronavirus, which agrees with reports in [33] and [34].
2. When exploring deep inside the SARS coronavirus group in the Phylogenetic tree, we found that the 1P9T is much different from the other crystal structures. This result is supported by Yang et al. [32], which reported that the crystallographic data reveal the limitations of the theoretical model [35]. The RMSD of the C_α atoms is 3.8 Å between the crystal structure and the theoretical model (PDB ID code 1p9t).

3.2 Protein Structure Classification in Supervised Cases

After showing the basic ability of identifying difference of protein structures by their coarse structures, we test the ability of fast screening search and apply the new similarity measure to automatic classification of a protein structure. There are two motivations to do so. One is that the new FSS representation is simple but loses structure information. Hence, it is necessary to objectively assess the accuracy of the similarity measure and study at which level the new

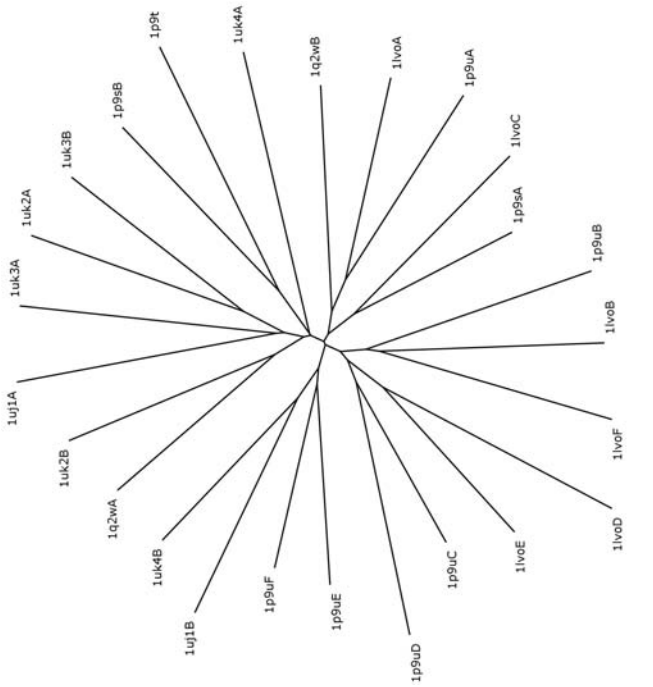

Fig. 6. Phylogenetic analysis from a pairwise structure similarity of coronavirus proteins using the Neighbor-joining and UPGMA methods.

representation can depict the protein structure. The second one lies in that there are biological databases, which can make the lines between functionally different proteins clear. By incorporating the expert judgement in the existing classification scheme as a training set, we can perform the protein structure classification based on the new representation and score scheme in the machine learning framework.

3.2.1 Preliminary Classification Results

With the classification algorithm and the experiment of whole CATH data set (see Section 2), preliminary results were obtained based on the different experiment schemes. First, an examination of the classification was conducted on the Class level in which proteins are assigned into four groups: "Mainly Alpha," "Mainly Beta," "Mixed Alpha-Beta," and "Few secondary structures." The training data were selected as the representative protein of every group, i.e., every group only contains one protein as teacher data. In fact, the problem is simplified so that every protein of CATH is compared with four template proteins and assigned to the group represented by the closest template protein. The classification results and the representative protein information are summarized in Table 3. The table shows that the new similarity measure has a good discriminating ability to the mixed Alpha-Beta class because the sensitivity reaches 90.49 percent, and the overall correct rate of classification is 45.59 percent.

Another experiment was conducted also on the Class level to avoid the affection of the choice of the representative protein and to extract the information provided by training set. The template proteins in the training set were randomly selected from the whole CATH database, and every training group contains 80 proteins as teacher data. From the results collected in Table 4, we can see that the sensitivity of the Mixed Alpha-Beta class drops to 86.81 percent, but the sensitivity of mainly Alpha class increases considerably.

TABLE 3
Classification Results of CATH v2.5.1 on the Class Level with a Representative Training Set

Class	Mainly Alpha	Mainly Beta	Mixed Alpha-Beta	Few Secondary Structures
Representative	1cuk03	1pdc00	1rthA1	1bg503
Sensitivity	6.61%	2.26%	90.49%	26.07%
Specificity	96.49%	97.12%	32.57%	85.47%
Correlation coefficient	0.063	0.017	0.279	0.045

TABLE 4
Classification Results of CATH v2.5.1 on the Class Level with a Randomly Selected Training Set

Class	Mainly Alpha	Mainly Beta	Mixed Alpha-Beta	Few Secondary Structures
Sensitivity	33.13%	2.78%	86.81%	18.56%
Specificity	76.46%	97.42%	45.13	98.58
Correlation coefficient	0.089	0.023	0.348	0.181

Also, overall correct rate raises to 49.42 percent, indicating that the accuracy is improved with the increase of training data. Clearly, the whole CATH classification needs more CPU time when the number of the template proteins increases. Besides, the Mainly Alpha and Mainly Beta proteins tend to be wrongly predicted into class Mixed Alpha-Beta. The similar conclusions were drawn in [36].

In the architecture level, the CATH v2.5.1 is classified to 37 different groups. The third experiment was conducted on this level as a pilot study. Similar to the first experiment, only one template protein was selected for every architecture group (the representative protein of that architecture). Classification results recorded in Table 5 show that many groups have high sensitivity and specificity.

The classification experiments in the topology level and homologous family level were not performed because the simple classification rule cannot handle cases with over several hundred groups. On the other hand, there are many sophistic tools to perform this job in an automatic way based on protein sequence similarity.

This pilot study for the new similarity measure on the existing protein structure classification shows that the global shape distribution from the FSS representation of a protein structure has discriminating ability, although it is only applied to the comparison of protein structures from the macrostructure viewpoint. Also, the combination of fast similarity comparison method with simple classification scheme can perform the screening search well for large-scale database. Since the purpose of this paper is to develop a fast screening comparison method of protein structures only from their coarse structures, we aim to remove the drawback of clustering methods by a classification scheme with supervision learning. Therefore, although the whole CATH structure database was tested, we only took a very small number of data as the training data set in contrast to the whole database.

Preliminary classification results were reported in this paper. Since only one protein in each class was selected as the training data set, it is not surprising that the sensitivity and correlation coefficient are not satisfactory for some classes. Specifically, the whole CATH structure database (47,027 items) are used as test data set, but there are only a very small number of training data. According to the classification scheme, there is only one template structure in each class as training data set. In addition, the unbalance structure of the data set also causes low sensitivity. For example, at class level, there are 9,976, 13,972, 22,547, and 932 items in the Mainly alpha, Mainly Beta, Mixed Alpha-Beta, and Few secondary structures, respectively. Furthermore, since the number of classes increases rapidly at the architecture level, it also increases the difficulty for correct classification. However, these results are useful and informative to analyze the ability of the proposed method because we aim to develop a fast screening comparison method of protein structures for large database searching. By combining with other computationally intense but more accurate methods [9], we can further explore and analyze the query results.

TABLE 5
Classification Results of CATH v2.5.1 on the Architecture Level with a Representative Training Set

Class/Architecture	Representative	Sensitivity	Specificity	Correlation coefficient
Mainly Alpha Alpha solenoid	1pprM1	66.67%	97.85%	0.049
Mainly Beta Clam	3bcl00	100%	98.35%	0.035
Mainly Beta Trefoil	1hcd00	50%	98.14%	0.123
Mainly Beta Orthogonal Prism	1jpc00	71.43%	99.57%	0.182
Mainly Beta 3 Solenoid	1lxa01	50%	94.06%	0.116
Mainly Beta Complex	1oen02	50%	98.26%	0.072
Mixed Alpha-Beta Super Roll	1bp101	50%	99.79%	0.099
Mixed Alpha-Beta Alpha-Beta prism	1ejdA1	81.25%	97.02%	0.119
Mixed Alpha-Beta 5-stranded Propeller	4jdwA0	91.67%	95.31%	0.065

Except the computational efficiency, the accuracy and reliability of the new similarity criterion can be improved further. There are two ways, the first one is to design the learning system carefully so as to explore more information of data, especially in multiclass cases. The other way is to mine the data from the coarse structures.

4 Discussion and Conclusion

One of the advantages of the new FSS representation is the effect of data compression demanded by fast screening of structure comparison. It is easy to see that two similar proteins have similar feature sequences of surfaces. However, two proteins having similar FSSs are not necessarily similar in their structures. Hence, the FSS is an approximate expression of characteristics for a protein structure. For example, for a protein with N amino acid residues, there are approximately a total of N^3 planes, but most of them cannot be supporting planes. If a protein is folded in a cubic with edge length N, then the total number of planes constituted by the atoms on the surface is bounded by $O(N^2)$. In fact, only a few number ($\ll N$) of supporting planes pass through each atom. Therefore, we expect that the total number of supporting planes in the convex hull is $O(N)$ with a small constant c, i.e., cN. For example, for the protein 1acj in Fig. 4, there are a total of 4,095 atoms with 528 C_α atoms, but when

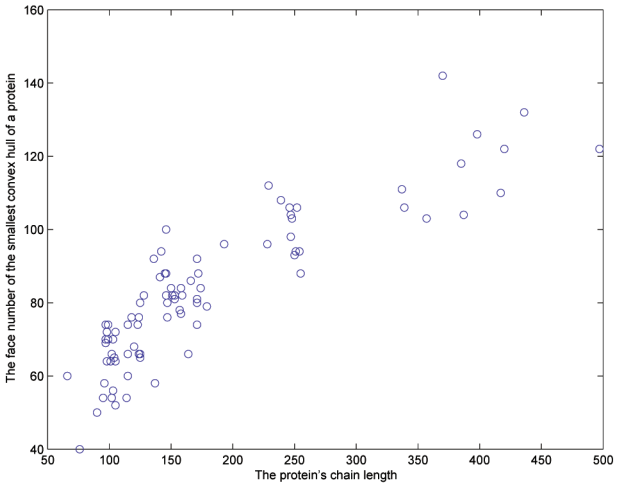

Fig. 7. Relationship between the protein length and the number of faces of its convex hull. A total 94 proteins in four existing data sets listed in this paper are used.

considering the spherical expression, there are only 120 supporting planes. Hence, the protein data are reduced to a feature sequence with length 120. Also, we give the relationship between the protein length and the number of faces of the convex hull in Fig. 7, where evidently, the new representation reduces the data input for comparison and such a tendency is clearer when the protein size is bigger.

Describing the protein's surface by its convex hull not only is simple but also can infer the functional relationships of proteins to some extent, as shown by the experiments. Comparing with the geometric model to describe the concave surface regions in the form of pockets and voids [37], [38], the new approach extracts the information of concave surface regions in a protein structure in a global way, i.e., by exploiting the difference of the projection of all atoms to every supporting planes. Hence, the ability of the new approach to discover the similarity relationships of protein structures comes from the global surface similarity.

The proposed new representation of protein structure is conservative. First, the FSS is independent of the translation and rotation transformation of a protein, which is viewed as a 3D rigid body. Second, the FSS is stable under the interior flexibility of a protein. To test such ability, we remove some individual C_α atoms from a known protein structure, but the FSS representation is a little bit affected. The reason lies in that the atoms are separated into two parts by the smallest convex hull. The smaller part of C_α atoms is located on the surface of the hull and is more important than the interior atoms. Therefore, despite of loss of some atom's information or poorer resolution than 3.0 Å, the new representation can tolerate it and give reasonable discrimination.

The feature distributions provide a measure of global shape of a protein, which only explores the macrostructure relative to a surface and neglects the interior flexibility of other atoms. Although certain information including the sequence of C_α atoms may be lost, the proposed approach really simplifies the representation and computation. It also provides some insight about the macrostructure including the distributions of mass and compact or looser structure when using as a screening method of protein structure comparison. As indicated in [39], the improvement in efficiency is offset by loss of discriminatory power. Also, from the viewpoint of machine learning, the new shape distribution can be regarded as the feature extracted from a protein structure, which is worth exploring further if the feature grasps the main structure characteristics of proteins.

Compared with sequence similarity-based method, protein structure comparison method can reveal the remote homology because a structure is believed more conservative than the corresponding sequence. However, generally, similarity between structures is difficult to be quantified and requires intense computation. Our method in this paper aims to evaluate the similarity between protein structures in an efficient way. Since the structure information is used to compare proteins, we are able to find remote similarity compared to sequence similarity-based methods. On the other hand, compared with other structure-based methods, we not only provide a new way to represent a protein structure but also develop a quantitative measure to efficiently and accurately compare the proteins.

Measuring protein structure similarity is one of the core topics of today's bioinformatics research. Many algorithms trying to quantify the similarity between protein structure pair have been proposed because of the increasingly accumulated protein structures in databases. We developed a novel method in this paper, which provides insight from the viewpoint of the coarse structures of proteins.

In this paper, we describe the details of the new method from the aspects of protein representation and score scheme, respectively. Numerical experiments were conducted for four existing different protein data sets and also for classification of SARS coronavirus to verify the effectiveness of FDOD score scheme. Furthermore, preliminary results of fast classification of the whole CATH v2.5.1 database based on the new macrostructure similarity were given as a pilot study. We showed that such a measure of protein similarity is able to provide certain insight into the protein 3D structure and capture some factors for assessing protein structure similarity in a fast and automatic way. The work to analyze where the structure similarity originates from is in progress, and the key lies in what is the accurate representative of a protein global structure topology. In our method, the protein representation is a one-dimensional array called FSS, which is a rough depiction of the macrostructure of a protein. Also, other effort is to mine deep information from the convex hull representation of a protein structure. The features such as the normal vector and interception also can be combined to our method to evaluate the similarity between proteins.

4.1 Availability

The results not listed in this paper and the data sets used are presented in http://zhangroup.aporc.org/bioinfo/strucomp/. All the programs and materials used in this paper are available on request from the authors.

4.2 List of Abbreviations Used

- FSS is the feature sequence of surface,
- PDB is protein data bank,
- FDOD is the function of degree of disagreement,

- RMSD is the root mean square deviation,
- SSE is the secondary structure element, and
- SARS is severe acute respiratory syndrome.

ACKNOWLEDGMENTS

This work is partly supported by Grant 5039052006CB from the Ministry of Science and Technology, China, and the National Natural Science Foundation of China under Grants 10801131, 10631070, and 60873205. This work is also partly supported by the JSPS-NSFC Collaboration Project.

REFERENCES

[1] C. Guerra and S. Istrail, *Mathematical Methods for Protein Structure Analysis and Design.* Springer, 2003.
[2] T. Kawabata and K. Nishikawa, "Protein Structure Comparison Using the Markov Transition Model of Evolution," *Proteins: Structure, Function and Genetics,* vol. 41, pp. 108-122, 2000.
[3] I. Eidhammer, I. Jonassen, and W.R. Taylor, "Structure Comparison and Structure Patterns," *J. Computational Biology,* vol. 7, no. 7, pp. 685-716, 2000.
[4] L. Holm and C. Sander, "Mapping the Protein Universe," *Science,* vol. 273, no. 2, pp. 595-602, Aug. 1996.
[5] O. Carugo and S. Pongor, "Recent Progress in Protein 3D Structure Comparison," *Current Protein and Peptide Science,* vol. 3, no. 4, pp. 441-449, Aug. 2002.
[6] L. Chen, T. Zhou, and Y. Tang, "Protein Structure Alignment by Deterministic Annealing," *Bioinformatics,* vol. 21, pp. 51-62, 2005.
[7] T. Zhou, L. Chen, Y. Tang, and X.-S. Zhang, "Aligning Multiple Protein Structures by Deterministic Annealing," *J. Bioinformatics and Computational Biology,* vol. 3, no. 4, pp. 837-860, 2005.
[8] L. Chen, L.-Y. Wu, R. Wang, Y. Wang, S. Zhang, and X.-S. Zhang, "Comparison of Protein Structures by Multi-Objective Optimization," *Genome Informatics,* vol. 16, no. 2, 2005.
[9] L. Chen, L.-Y. Wu, Y. Wang, S. Zhang, and X.-S. Zhang, "Revealing Divergent Evolution, Identifying Circular Permutations and Detecting Active-Sites by Protein Structure Comparison," *BMC Structural Biology,* vol. 6, no. 1, p. 18, 2006, http://www.biomedcentral.com/1472-6807/6/18.
[10] F.E. Cohen and M.J.E. Sternberg, "On the Prediction of Protein Structure: The Significance of the Root-Mean-Square Deviation," *J. Molecular Biology,* vol. 138, pp. 321-333, 1980.
[11] O. Carugo, "How Root-Mean-Square Distance (R.M.S.D.) Values Depend on the Resolution of Protein Structures That Are Compared," *J. Applied Crystallography,* vol. 36, no. 1, pp. 125-129, Feb. 2003.
[12] O. Carugo and S. Pongor, "A Normalized Root-Mean-Square Distance for Comparing Protein Three-Dimensional Structures," *Protein Science,* vol. 10, pp. 1470-1473, 2001.
[13] A. Godzik, "The Structural Alignment between Two Proteins: Is There a Unique Answer?" *Protein Science,* vol. 5, pp. 1325-1338, 1996.
[14] Y. Wang, L.-Y. Wu, and X.-S. Zhang, "Supervised Classification of Protein Structures Based on Convex Hull Representation," *Int'l J. Bioinformatics Research and Applications,* vol. 3, no. 2, 2007.
[15] O. Carugo and S. Pongor, "Protein Fold Similarity Estimated by a Probabilistic Approach Based on C(alpha)-C(alpha) Distance Comparison," *J. Molecular Biology,* vol. 315, no. 4, pp. 878-898, Jan. 2002.
[16] K. Vlahovicek, O. Carugo, and S. Pongor, "The PRIDE Server for Protein Three-Dimensional Similarity," *J. Applied Crystallography,* vol. 35, pp. 648-649, 2002.
[17] S.D. O'Hearn, A.J. Kusalik, and J.F. Angel, "Molcom: A Method to Compare Protein Molecules Based on 3D Structural and Chemical Similarity," *Protein Eng.,* vol. 16, no. 3, pp. 169-178, 2003.
[18] A. Caprara, R. Carr, and S. Istrail, "1001 Optimal PDB Structure Alignments: Integer Programming Methods for Finding the Maximum Contact Map Overlap," *J. Computational Biology,* vol. 11, no. 1, pp. 27-52, 2004.
[19] P. Rogen and B. Fain, "Automatic Classification of Protein Structure by Using Gauss Integrals," *Proc. Nat'l Academy of Sciences USA (PNAS '03),* vol. 100, no. 1, pp. 119-124, Jan. 2003.
[20] P. Rogen and H. Bohr, "A New Family of Global Protein Shape Descriptors," *Math. Biosciences,* vol. 182, pp. 167-181, 2003.
[21] N. Krasnogor and D.A. Pelta, "Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric," *Bioinformatics,* vol. 20, no. 7, 2004.
[22] D. Bostick and I.I. Vaisman, "A New Topological Method to Measure Protein Structure Similarity," *Biochemical and Biophysical Research Comm.,* vol. 304, pp. 320-325, 2003.
[23] T. Kawabata, "MATRAS: A Program for Protein 3D Structure Comparison," *Nucleic Acids Research,* vol. 31, no. 13, pp. 3367-3369, 2003.
[24] W. Fang, "The Characterization of a Measure of Information Discrepancy," *Information Sciences,* vol. 125, nos. 1-4, pp. 207-232, 2000.
[25] W. Fang, F.S. Roberts, and Z. Ma, "A Measure of Discrepancy of Multiple Sequences," *Information Sciences,* vol. 137, pp. 75-102, 2001.
[26] A. Via, F. Ferre, B. Brannetti, and M. Helmer-Citterich, "Protein Surface Similarities: A Survey of Methods to Describe and Compare Protein Surfaces," *Cellular and Molecular Life Sciences,* vol. 57, pp. 1970-1977, 2000.
[27] M.L. Connolly, "Molecular Surfaces: A Review," http://www.netsci. org/Science/Compchem/feature14.html, 1996.
[28] X.-S. Zhang, Z.-W. Zhan, Y. Wang, and L.-Y. Wu, "An Attempt to Explore the Similarity of Two Proteins by Their Surface Shapes," *Operations Research and Its Applications,* vol. 5, pp. 276-284, World Publishing Corp., 2005.
[29] J. Leluk, L. Konieczny, and I. Roterman, "Search for Structural Similarity in Proteins," *Bioinformatics,* vol. 19, no. 1, pp. 117-124, 2003.
[30] L.P. Chew and K. Kedem, "Finding the Consensus Shape for a Protein Family," *Proc. 18th ACM Symp. Computational Geometry (SoCG),* 2002.
[31] A. Caprara and G. Lancia, "Structural Alignment of Large Size Proteins via Lagrangian Relaxation," *Proc. Sixth Ann. Int'l Conf. Computational Biology (RECOMB '02),* pp. 100-108, 2002.
[32] H. Yang et al., "The Crystal Structures of Severe Acute Respiratory Syndrome Virus Main Protease and Its Complex with an Inhibitor," *Proc. Nat'l Academy of Sciences USA (PNAS '03),* vol. 100, no. 23, pp. 13190-13195, 2003, http://www.pnas.org/cgi/content/abstract/100/23/13190.
[33] M.A. Marra et al., "The Genome Sequence of the SARS-Associated Coronavirus," *Science,* vol. 300, no. 5624, pp. 1399-1404, 2003, http://www.sciencemag.org/cgi/content/abstract/300/5624/1399.
[34] P.A. Rota et al., "Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome," *Science,* vol. 300, no. 5624, pp. 1394-1399, 2003, http://www.sciencemag.org/cgi/content/abstract/300/5624/1394.
[35] K. Anand, J. Ziebuhr, P. Wadhwani, J.R. Mesters, and R. Hilgenfeld, "Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs," *Science,* vol. 300, no. 5626, pp. 1763-1767, 2003, http://www.sciencemag.org/cgi/content/abstract/300/5626/1763.
[36] L. Jin, W. Fang, and H. Tang, "Predicting Protein Structure Class by a New Method of Information Theory," *J. Computational Biology and Chemistry,* vol. 27, no. 3, 2003.
[37] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design," *Protein Science,* vol. 7, pp. 1884-1897, 1998.
[38] T.A. Binkowski, L. Adamian, and J. Liang, "Inferring Functional Relationship of Proteins from Local Sequence and Spatial Surface Patterns," *J. Molecular Biology,* vol. 332, pp. 505-526, 2003.
[39] M.B. Swindells, C.A. Orengo, D.T. Jones, E.G. Hutchinson, and J.M. Thornton, "Contemporary Approaches to Protein Structure Classification," *Bioessays,* vol. 20, no. 11, pp. 884-891, 1998.

Yong Wang received the bachelor's degree in mathematics and physics from Inner Mongolia University in 1999, the master's degree in operations research and control theory from the Dalian University of Technology in 2002, and the PhD degree in operations research and control theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, in 2005. He is an assistant professor at the Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences. His current interests include mathematical modeling and algorithm analysis in bioinformatics.

Ling-Yun Wu received the PhD degree in operations research and control theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He is an associate professor at the Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He was with the Department of Industrial Engineering and Engineering Management, Hong Kong University of Science and Technology. His current interests are in the bioinformatics and systems biology.

Ji-Hong Zhang received the PhD degree in operations research and control theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He is a full professor in the School of International Business, Beijing Foreign Studies University. He was with the School of Economics and Management, Tsinghua University. His current interests are in the bioinformatics and supply chain management.

Zhong-Wei Zhan received the BS degree from the College of Mathematics, Peking University, and the PhD degree from the Academy of Mathematics and Systems Science, Chinese Academy of Science. He is a project manager at the China Aerospace Engineering Consultation Center. His current interests are in the bioinformatics and system engineering.

Xiang-Sun Zhang received the degree from the Department of Applied Mathematics, Chinese University of Science and Technology, in 1965. He is a full research professor in the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He has extensive experience in operations research, including optimization theory and application, artificial neural networks, and management information system (MIS) theory and application. His current interests are in the application of operations research methods in bioinformatics. He is the honorary president of the Operations Research Society of China.

Luonan Chen received the bachelor's degree from the Department of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 1984, and the master's and PhD degrees from the Department of Electrical and Communication Engineering, Tohoku University, Sendai, Japan, in 1988 and 1991, respectively. He is a professor in the Department of Electrical Engineering and Electronics, Osaka Sangyo University. He is also with ERATO Aihara Complexity Modeling Project of JST, the Institute of Industrial Science, the University of Tokyo, Japan, and the Institute of Systems Biology, Shanghai University, China. His current interests are in systems biology and bioinformatics. He is a senior member of the IEEE and a member of the IEE Japan.