Protein structure similarity based on multi-view images generated from 3D molecular visualization

Chendra Hadi Suryanto, Shukun Jiang, Kazuhiro Fukui
Graduate School of Systems and Information Engineering, University of Tsukuba
{*chendra,kyou*}*@cvlab.cs.tsukuba.ac.jp, kfukui@cs.tsukuba.ac.jp*

Abstract

Comparing the structures of proteins is one of the most challenging problems in structural biology. Root Mean Square Distance (RMSD) has become a standard measurement to calculate the similarity between two protein structures. However, to get the best result one has to align and superpose the two protein structures, which raises issues related to finding the best alignment technique. In this paper, we propose a new approach to protein structure comparison using canonical angles between two subspaces generated from multiple views of the protein structure visualization. The main advantage of our approach is that no protein alignment is required. Moreover, since we also consider the various visualization types of the 3D protein structures (backbone, ribbons, and rockets), our protein descriptors contain more elaborate structures and characteristics of the protein, which possibly cannot be represented by only a single visualization geometry. The validity of our proposed method is shown by experiments on classifications of four classes of protein in which our approach exhibited better performance than the two well-known methods of combinatorial extension alignment and the Gauss integral tuning.

1. Introduction

In structural biology, finding the similarity between protein structures is a fundamental problem, especially to understand protein function and evolution as applied to various purposes, such as designing new drugs and investigating the evolution of organisms. To support the study of protein structure, SCOP [1][2], a database of comprehensive structural classification of proteins, has been proposed. The SCOP database is constructed by manual visual inspection and utilization of various automatic tools in which protein structures were obtained from X-ray crystallography and NMR spectroscopy.

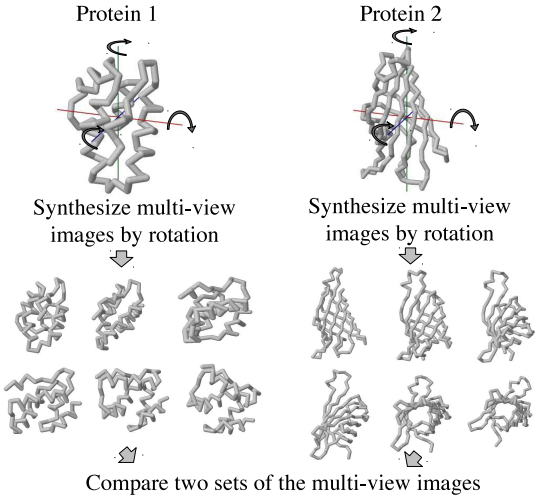

Figure 1. Overview of the proposed method for comparing the similarity of two proteins.

The Protein Data Bank (PDB) [3] provides a massive dataset of such 3D protein structures. Nevertheless, currently more than half of the known protein structures in the PDB are yet to be classified in the SCOP database, which indicates a need for an automatic classification framework with higher accuracy to reduce or omit the necessity of manual inspection in comparing the protein structures.

Most protein structure comparisons are based on information, such as amino acid positions, about the molecular structure of proteins. In such a method, the similarity of two proteins is computed in two steps. The first step is an alignment search to find the largest number of aligned amino acid positions within the pairs. Some common structure alignment methods are the Distance Alignment Matrix Method (DALI) [4] and Combinatorial Extension (CE) [5]. The second step is

to superpose the aligned rigid body structure and compute the similarity based on the spatial distance between them. However, this conventional method has a disadvantage in that it highly depends on the backbone length of the protein structures, which makes it sensitive to local error due to non-optimal alignment. To address such problems of finding a global protein measure, a compact representation of 3D protein structure as a 31-dimensional feature vector called Gauss Integral Tuning (GIT) [6] has been proposed, in which the similarity is defined simply by the Euclidean distance between the two GIT vectors.

Based on the idea that a biochemist can identify the similarity between a pair of proteins by observing the 3-D structure of the proteins from different viewpoints, we propose an approach to compute the similarity of protein structures using multi-views of the 3D shapes of the protein structures captured from different angles. Figure 1 shows the basic idea of our approach. In our approach, the multiple viewpoints of the images are synthesized by random rotation of the protein model. After the multi view images are collected, feature vectors are extracted and the subspace is generated by applying PCA. The similarity between protein structures is defined by the cosine square of the canonical angles between the two subspaces. The use of canonical angles as the similarity is based on the Mutual Subspace Method (MSM) [7], which is one of the widely used 3D object recognition methods based on multi-view of images.

The most important contribution of our approach is to provide the similarity between proteins with high extendability while avoiding the performance loss due to non-optimal alignment. To view the characteristics of a target protein in various ways, we can obtain various visualizations of the 3D protein structures (backbone, ribbons, rockets, etc.) using 3D molecular graphics software. For example, the backbone visualization shows the backbone of a protein by connecting the alpha carbon atoms. The ribbons visualization represents the backbone using a wide flat band along the adjacent alpha carbon. The rockets visualization shows the protein conformation in which the cylinders represent helices and arrows represent the strands of a sheet. This implies that we can use these various visualizations of the target proteins to define the similarity between the proteins. In our framework, this idea can be realized by combining different protein visualizations as a feature vector which later can be used for computing the similarity. We show the validity of our proposed method by comparing the classification results of four classes of protein based on the SCOP database using our proposed method, with the commonly used CE alignment method [5] and GIT descriptor [6].

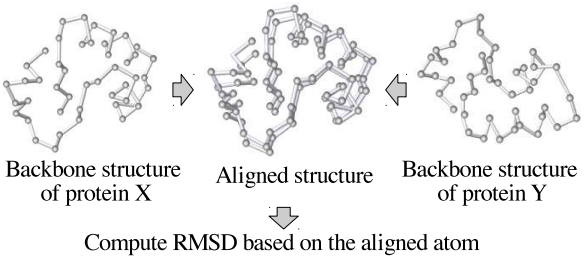

Figure 2. Conventional method to compute the similarity of protein structures.

This paper is organized as follows. First, we briefly review the conventional method of RMSD. Next, we describe the proposed method in section 3. The experimental conditions and results are shown in section 4. Finally, section 5 concludes.

2. Conventional Similarity Using RMSD

In the conventional method, after a pair of protein structures is aligned by using a particular alignment method such as DALI[4] or CE[5], their similarity is computed by the root mean square distance (RMSD), as illustrated by Figure 2. RMSD is defined as follows:

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\delta_i^2}, \tag{1}$$

where δ_i is the distance between each of the ith pair of alpha carbon atoms and n is the number of atom pairs.

3. Proposed Similarity Method

3.1 Overview

Since proteins have a complex structure, relying on a single representation may be inadequate. Unlike conventional methods that rely solely on alpha carbon coordinates of the backbone structure of the protein, our approach considers several protein visualizations generated using molecular graphics software packages such as Jmol [8]. Some protein visualizations are shown in Figure 3. By considering the multiple types of visualizations of the protein structure such as backbones, ribbons, and rockets, the geometry of the protein structures can be described more elaborately as they can complement each other.

Figure 3. Three types of protein visualization.

3.2 Similarity Calculation with MSM

In the learning phase of MSM, we take a number of f dimensional features belonging to class c $(= 1, ..., C)$, where C is the number of the class, and apply PCA (without centering) to generate an N-dimensional reference subspace $\mathcal{P}_c$. In the test phase, the M dimensional input subspace $\mathcal{Q}$ is generated by applying PCA to the input patterns. The similarity between the input subspace $\mathcal{Q}$ and reference subspace $\mathcal{P}_c$ is defined by the canonical angle θ_i between them.

M canonical angles $(0 \leq \theta_1 \leq \ldots \leq \theta_M \leq \frac{\pi}{2})$ between M-dimensional subspace and N-dimensional subspace $(M \leq N)$ are defined. The i-th canonical angle θ_i is defined as follows [9]:

$$\cos\theta_i = \max_{\boldsymbol{u}_i \in \mathcal{Q}} \max_{\boldsymbol{v}_i \in \mathcal{P}_c} \boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{v}_i \tag{2}$$

$s.t.\ \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_i = \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_i = 1, \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j = \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_j = 0, i \neq j.$

In practice, $\cos\theta_i$ is obtained by computing the singular value of the matrix $\mathbf{X}$, where $\mathbf{X} = \mathbf{A}^{\mathrm{T}}\mathbf{B}$. $\mathbf{A} = [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_M]$, $\mathbf{B} = [\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_N]$. $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$ are the orthogonal basis vectors of the subspace $\mathcal{Q}$ and $\mathcal{P}_c$ respectively. The similarity between two subspaces is finally calculated as follows:

$$Sim = \frac{1}{M}\sum_{i}^{M} \cos^2\theta_i, \tag{3}$$

where $0 \leq Sim \leq 1$. Higher Sim values indicate that two proteins are highly similar, while lower values of Sim indicate that the two proteins are less similar.

3.3 Process Flow

Figure 4 shows the flow of the proposed similarity. The steps are as follows:

Step 1: Normalization of the center coordinate of the protein's chain atoms based on the inertia tensor of molecular physics.

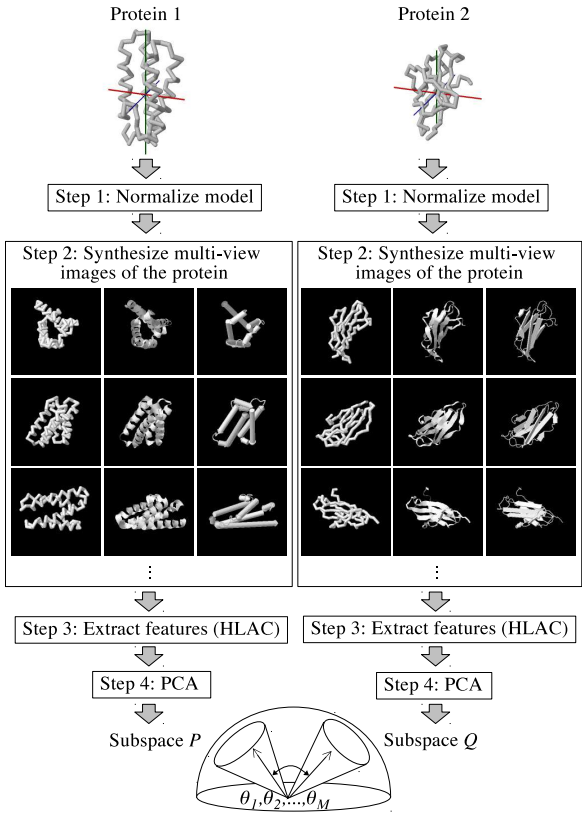

Figure 4. Flow of the proposed method.

Step 2: The protein model is rotated randomly around its three principal viewing axes and multi-view of images from several types of visualization (backbone, ribbons, and rockets) are synthesized.

Step 3: Feature extraction using HLAC (Higher-Order Local Autocorrelation) [10]. We use a 35-dimensional HLAC feature that is position invariant, because we want to compare only the shape of the protein structure regardless of positions. Let the image be denoted by I. The Nth order of the autocorrelation function with N displacements $a_1, ..., a_N$ is defined as follows:

$$x(a_1, .., a_N) = \sum I(r)I(r + a_1)...I(r + a_N), \tag{4}$$

where r is the image coordinate vector. The order N is limited to the second order $(N \in \{0, 1, 2\})$. $a_{ix}, a_{iy} \in \{\pm\Delta r, 0\}$. Duplicate configurations of $r, r+a_1, ..., r+a_N$ are removed so that the final local mask patterns are reduced to 35. Since we use three types of visualization of the protein, the combined HLAC features produce a 105-dimensional HLAC feature vector.

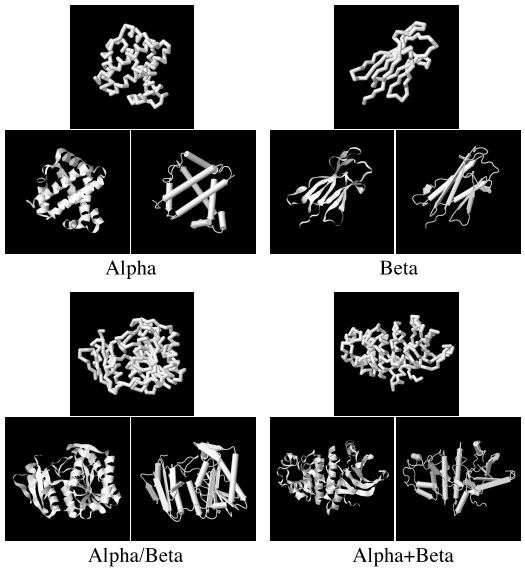

Figure 5. Examples of each protein class.

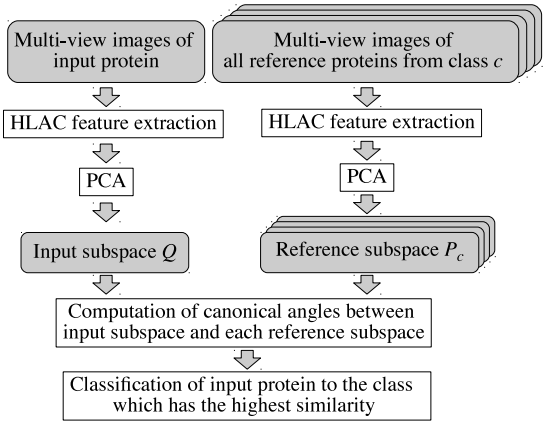

Figure 6. Flow of the protein classification.

Step 4: Subspace generation by applying PCA and the similarity computation based on MSM.

4. Experiment

To evaluate the effectiveness of the proposed method, we did experiments on classification of four classes of protein: Alpha (α), Beta (β), Alpha/Beta (α/β), and Alpha+Beta (α+β) as shown in Figure 5, based on the SCOP database [2]. Alpha proteins contain alpha helices. Beta proteins contain beta sheets. Alpha/Beta proteins contain alpha-beta motifs (mainly parallel beta sheets). In Alpha+Beta class, we used α+β proteins which has segregated alpha and beta regions (mainly anti-parallel beta sheets) protein and also the multi-domain proteins (alpha and beta).

4.1 Dataset and Experimental Conditions

We collected 80 proteins randomly from the RCSB PDB [3] and cropped the protein chain accordingly by referring to the class label from the SCOP using the Matlab Bioinformatics toolbox. Jmol [8] was used to synthesize the protein and acquire 9000 images at 32×32 pixels for each protein (3000 images $\times 3$ visualizations type: backbone, ribbons, and rockets). The visualization color was set to white (gray-scaled) with black background. We adopt the leave-one-out cross validation method so that one protein is used as test data and the rest were used to generate four reference subspaces. The classification process flow is shown in Figure 6. The dimensions of the reference subspaces were varied from 1 to 60. The dimension of the input subspace was varied from 1 to 20. The experimental results are based on the best result from the various parameters.

4.2 Results and Discussions

Table 1 shows the classification results of the proposed method. As a comparison, we also did other experiments based on CE alignment and GIT descriptor.

The experiment with CE alignment was conducted as follows. First, we calculated the RMSD between the test protein and the other proteins using the jCE tool [11]. The protein chains were cropped beforehand using the Bio3d package [12]. Next, we computed the average of the RMSD of the test protein and the training data from each class and categorized the test protein to the class with the smallest average RMSD. Table 2 shows the classification results of the CE alignment.

In the GIT experiment, we employed the tool in [6] to extract GIT features. First, we computed the Euclidean distance of the test and the training data. Then, we categorized the test protein to the class with the smallest average distance. However, since GIT does not consider protein backbones with more than 3 missing alpha carbons, we only used 70 of the 80 proteins. The experimental results of GIT are shown in Table 3.

The experimental results demonstrate that our approach using multi-view and visualizations of the protein achieves better performance than that of both CE and GIT. The average classification rate of our proposed method was 83.75%, while the CE and GIT only achieved average classification rates of 71.25% and

Table 1. Classification results of MSM.

Class	α	β	α/β	$\alpha+\beta$	Correct Rate
α	17	0	1	2	85%
β	0	17	1	2	85%
α/β	0	1	19	0	95%
$\alpha+\beta$	2	0	4	14	70%
Average rate					**83.75%**

Table 2. Classification results of CE.

Class	α	β	α/β	$\alpha+\beta$	Correct Rate
α	16	1	3	0	80%
β	1	18	0	1	90%
α/β	0	1	17	2	85%
$\alpha+\beta$	5	8	1	6	30%
Average rate					71.25%

Table 3. Classification results of GIT.

Class	α	β	α/β	$\alpha+\beta$	Correct Rate
α	18	0	0	0	100%
β	0	16	0	1	94.12%
α/β	6	0	13	0	68.42%
$\alpha+\beta$	3	5	0	8	50%
Average rate					78.57%

78.57% respectively. Despite the good performance of the conventional methods in classifying α and β proteins, CE and GIT have difficulty in classifying complicated protein with characteristics like alpha helices and beta sheets that co-exist separately in different parts of the structure. In contrast, by taking advantage of multiple visualizations, the proposed method can classify such complicated structures more accurately.

5. Conclusion

In this paper we proposed a new approach to compare 3D protein structure by using multi-views of the synthesized 3D protein structure images. Our proposed method has the benefit that no protein alignment is required. In addition, we can have more heterogeneous descriptors by combining different visualizations of the structures that complement one another in representing the complex structure. The effectiveness of our proposed method of using canonical angles as the similarity metric is shown by experimental results in which we achieved better performance than the conventional methods in classifying four classes of 3D protein structure based on SCOP.

Since this is still an early work, we only used small datasets. In the future, we will collect large amounts of protein data from the PDB and consider different feature extraction methods for the protein images. We will also consider the extensions of the MSM [13][14] to further improve the performance of our method.

References

[1] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Molecular Biology*, 247:536–540, 1998.

[2] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the scop database: new developments. *Nucleic Acids Research*, 36:D419–D425, 2007.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[4] L.Holm and C.Sander. Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20:478–480, 1995.

[5] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, 11:739–747, 1998.

[6] P. Røgen. Evaluating protein structure descriptors and tuning gauss integral based descriptors. *Nucleic Acids Research*, 17:1523–1538, 2005.

[7] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *Int. Conf. on Face and Gesture Recognition*, pages 318–323, 1998.

[8] A. Herráez. Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255–261, 2006.

[9] F. Chatelin. *Eigenvalues of matrices*. John Wiley & Sons, Chichester, 1993.

[10] N. Otsu and T. Kurita. A new scheme for practical flexible and intelligent vision systems. *Proceeding of IAPR Workshop on CV*, pages 431–435, 1988.

[11] A. Prlić, S. Bliven, P. W. Rose, W. F. Bluhm, C. Bizon, A. Godzik, and P. E. Bourne. Precalculated protein structure alignments at the RCSB PDB website. *Bioinformatics*, 26:2983–2985, 2010.

[12] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. D. Caves. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22:2695–2696, 2006.

[13] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. *11th International Symposium of Robotics Research (ISRR'03)*, pages 192–201, 2005.

[14] K. Fukui, B. Stenger, and O. Yamaguchi. A framework for 3D object recognition using the kernel constrained mutual subspace method. *ACCV06*, 3851:315–324, 2006.