



# Skeleton-based shape analysis of protein models<sup>☆</sup>



Zhong Li<sup>a,\*</sup>, Shengwei Qin<sup>a</sup>, Zeyun Yu<sup>b</sup>, Yao Jin<sup>a</sup>

<sup>a</sup> College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China

<sup>b</sup> Department of Computer Science, University of Wisconsin, Milwaukee 53211, USA

## ARTICLE INFO

### Article history:

Accepted 30 June 2014

Available online 12 July 2014

### Keywords:

Protein shapes  
Shape descriptor  
Similarity comparison  
Skeleton  
Local diameter

## ABSTRACT

In order to compare the similarity between two protein models, a shape analysis algorithm based on skeleton extraction is presented in this paper. It firstly extracts the skeleton of a given protein surface by an improved Multi-resolution Reeb Graph (MRG) method. A number of points on the model surface are then collected to compute the local diameter (LD) according to the skeleton. Finally the LD frequency is calculated to build up the line chart, which is employed to analyze the shape similarity between protein models. Experimental results show that the similarity comparison using the proposed shape descriptor is more accurate especially for protein models with large deformations.

© 2014 Elsevier Inc. All rights reserved.

## 1. Introduction

Proteins are vital in cells and play a special role in many biological activities. The surface shape of a protein defines a geometric and biochemical domain where the protein interacts with other proteins or its environment, and thus is an important characteristic of the protein. The similarity comparison between protein shapes had recently become an important means of protein analysis, which can reveal the structural and functional relationship of the associated proteins. With the development of computer graphics, the shape similarity comparison between protein molecules has been applied in many fields, such as computer aided molecular design, drug discovery, protein structure retrieval and so on [1,2].

Many researchers have proposed different methods on shape similarity comparison for surface models and some of them are specifically for protein shape-based analysis. These methods can roughly be classified into three categories [3]. The first one is the similarity comparison based on the outline. For example, Osada et al. [4] provided a shape distribution method based on the statistical histogram to measure the whole model shape. Horn [5] applied the extended Gaussian image mapping for similarity comparison of convex polyhedron models. Michael et al. [6] proposed a functional analysis method according to the spherical harmonic expansion

to compare the protein binding pocket and ligand. Kazhdan et al. [7] put forward an algorithm of analyzing geometric structures by using all kinds of frequent characteristics of three-dimensional shapes to retrieve these models.

The second category of methods on shape comparison is based on shape projections. Min et al. [8] proposed a method using a 2D sketch interface for a 3D model search engine. It mainly does a projection transformation in different directions on three-dimensional models, and gets a series of two-dimensional projection images for model retrieval. However, it has some limitations. For example, it can only describe the feature of brightness distributions and cannot reflect topology characteristics.

The third category of methods on shape comparison is based on geometric features of three dimensional models. Reuter et al. [9] proposed a spectral method using the Laplace–Beltrami operator, which possesses an isometry-invariant global geometric property. Other researchers use topological structures to compare the model shapes. For example, Hilaga et al. [10] provided a shape comparison method by using the multi-resolution Reeb graph (MRG). Though the MRG method describes the topological characteristics, it does not cover full geometric information (for example holes) and may get the mismatching for deformed models [10]. Therefore, for the protein models with holes or with the deformed shape, directly using this method may not obtain satisfactory similarity results. Some researchers used an efficient computation of a simplified medial axis for shape analysis [11,12]. The problem of the medial axis is that the computational complexity is large and sensitive to the holes on 3D models. Fang et al. [1] applied the local-diameter (LD) descriptor to compare the similarity of different flexible proteins. The LD method can improve the efficiency and quality of

<sup>☆</sup> This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LY14A010032 and Scientific Research Foundation of Ministry of Education of China under Grant No. [2009]1590.

\* Corresponding author. Tel.: +86 571 86843224; fax: +86 571 86843224.

E-mail addresses: [lizhong@zstu.edu.cn](mailto:lizhong@zstu.edu.cn), [lizhongzju@hotmail.com](mailto:lizhongzju@hotmail.com) (Z. Li).

similarity analysis, but there are still some problems. For example, it may not be precise for the deformed protein models with large topological changes.

In this paper we propose a new protein shape analysis method. It adopts the MRG algorithm based on collision detection and plane segmentation to obtain the skeleton for approximating the protein shape. The skeleton can describe the whole topological characteristics and optimize the details well. Also, it is insensitive to model noise and large topological changes caused by protein deformations. Based on the skeleton, we calculate the local-diameter of some sample points on the model surface to conduct similarity comparison. The experiment demonstrates that the shape analysis is invariant for rotations and translations of the protein models, and is also robust to shape deformations.

## 2. Materials and methods

### 2.1. Skeleton extraction for protein molecule models

There are many methods to extract the linear skeleton of a three-dimensional model [10,13]. Here we propose an improved MRG (Multi-resolution Reeb Graph) algorithm based on the collision detection and plane segmentation for protein molecule models.

#### 2.1.1. MRG algorithm

The basic idea of the MRG algorithm [10] is to find a good function  $\mu$ , for example, the height function, to extract the skeleton by the Reeb graph. The function  $\mu$  can be defined for each vertex  $v$  as follows

$$\mu = \sum g(v, b_i) \cdot \text{area}(b_i) \quad (1)$$

where  $\{b_i\}$  is the fundamental point set sampled on the surface of the model,  $g(v, b_i)$  is the geodesic distance between a vertex  $v$  and the fundamental points  $b_i$ , and  $\text{area}(b_i)$  is the local surface area related on each point  $b_i$ .

Then we can simplify the function  $\mu$  as

$$\mu = \sum g(v, b_i) \cdot C \quad (2)$$

where  $C$  is the average area defined as  $C = S/N$ , where  $S$  is the surface area of the model and  $N$  is the number of fundamental points. In practice,  $C$  can be chosen as a constant. In our experiments we set it as 1.

After the function  $\mu$  of each vertex is calculated from (2), we then normalize the value of the function  $\mu$  and divide the values of the function  $\mu$  into  $k$  ranges. The point set in each range is then constructed. We finally replace each point set by a joint point according to the barycentric coordinates and automatically connect these joint points based on the connecting relationship to extract the skeleton. Fig. 1 is the skeleton results for two protein molecular models by using the above process.

We find the traditional MRG algorithm is easy to implement, but for some protein molecular models especially with holes, the skeleton results sometimes cannot effectively preserve the topology and shape feature. For example, the skeleton may not be located in the center of the model and the skeleton may beyond the body of the model (see Fig. 1). In the next subsection we aim to improve the MRG algorithm to solve these problems.

#### 2.1.2. Our improved MRG algorithm for the skeleton extraction

In order to extract more accurate skeletons for protein molecular models, we provide an improved MRG algorithm based on the collision detection and plane segmentation. The main algorithm steps are as follows.

Step (1) We extract the origin skeleton  $R$  from model  $M$  based on the traditional MRG algorithm, save the joint points and record the numbers of joint points as  $N$ .

Step (2) We check whether each skeleton segment composed of two neighboring joint points  $(g_i, g_{i+1})$  (where  $i < N$ ) is beyond the surface of the model. If it stays in the model, go to step (4); otherwise, go to step (3).

Step (3) If the skeleton segment is beyond the surface of the model, we calculate the intersection points, record them and count the number  $S$  of these intersections. Then we do the following operations according to the number of intersections.

- (i) If  $S = 1$ , there are three cases. Case 1: one endpoint of segment  $(g_i, g_{i+1})$  is out of the object and another one is in the object. We compute the midpoint between the intersection point and

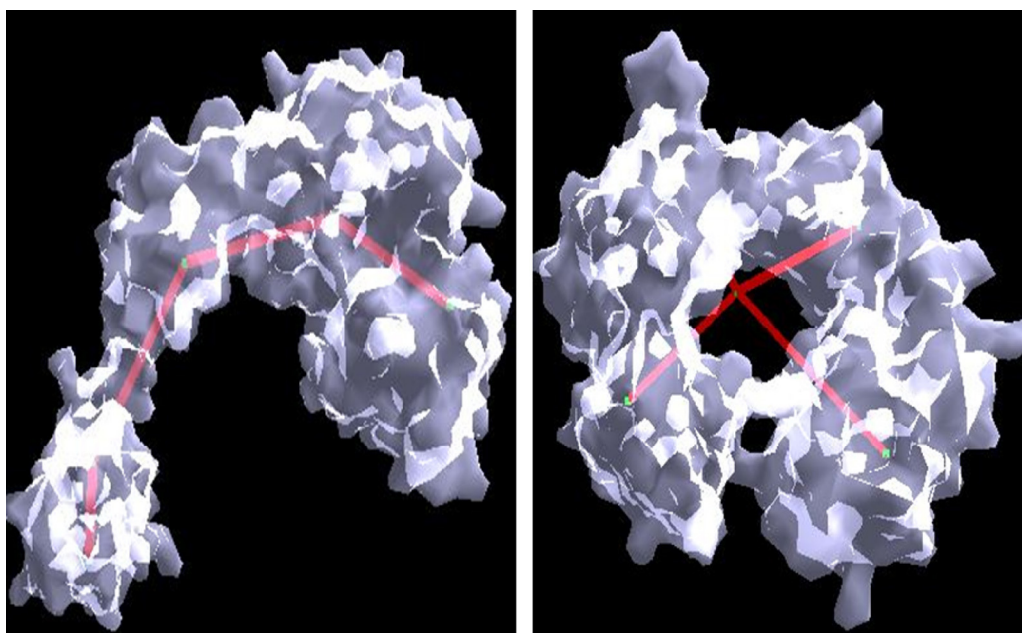


Fig. 1. Skeleton extraction based on the traditional MRG algorithm.

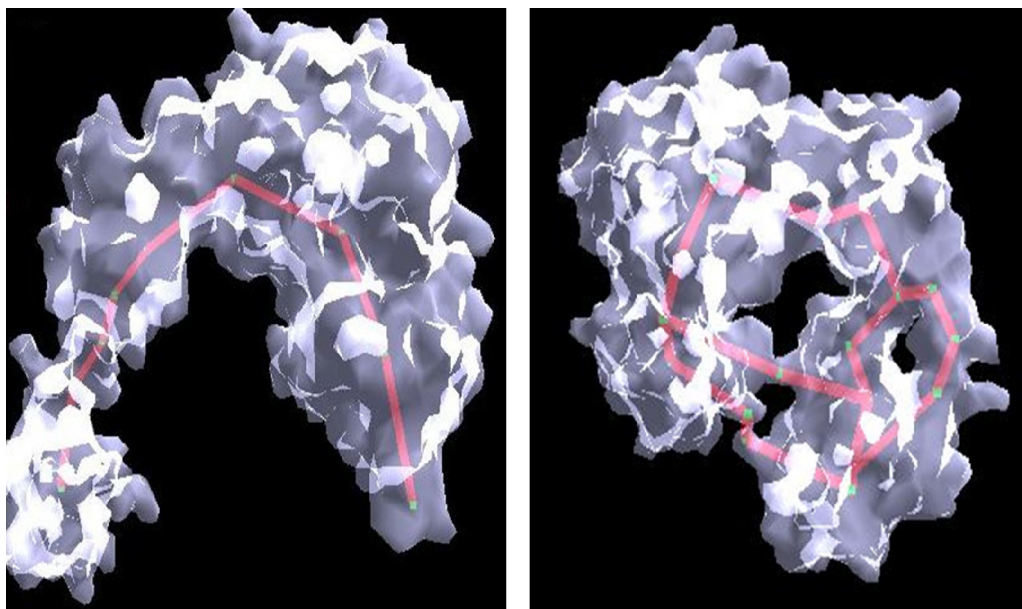


Fig. 2. Skeleton extraction by our improved MRG algorithm.

- one inner endpoint as the new joint point, update this skeleton segment by the segment composed of the intersection and inner endpoint, then go to step (2). Case 2: two endpoints are in the model and the skeleton segment is tangent to the boundary of the object, go to step (4). Case 3: two endpoints are out of the model and the skeleton segment is tangent to the boundary of the object with the point  $T$ , we replace  $g_i, g_{i+1}$  by point  $T$  and update the skeleton segment related in  $g_i, g_{i+1}$ .
- (ii) If  $S = 2$ , we compute the midpoint  $s_m$  between two intersection points. Then, we cast a ray from the midpoint  $s_m$  (The direction of shortest distance from this point to the surface of the model is chosen as the direction of the ray) and find two intersection points  $g_1$  and  $g_2$  with the model  $M$ . We compute the midpoint  $g_{new}$  between points  $g_1$  and  $g_2$ , insert it between two joint points  $g_i$  and  $g_{i+1}$  and replace the skeleton segment  $(g_i, g_{i+1})$  by  $(g_i, g_{new})$  and  $(g_{new}, g_{i+1})$ . Then go to step (2).
- (iii) If  $S \geq 3$ , we first determine the sub skeleton composed of two neighboring intersection points. For each new sub skeleton, we use (ii) to obtain the new joint point. Then we connect the original endpoints of the skeleton  $(g_i, g_{i+1})$  and new joint points sequentially to obtain new sub skeletons. For each sub skeleton, go to step (2).

Step (4) We calculate the shortest distance  $D_j$  ( $\neq 0$ ) from the surface point  $D_{min}$  on the model  $M$  to each skeleton segment. We cast a ray from the point  $D_{min}$  (the direction of the shortest distance from this point to the skeleton is chosen as the direction of the ray) and obtain another shortest intersection point  $C$  on the model surface  $M$ . Then, we compute the distance  $L$  from the intersection point  $C$  to the skeleton segment. If  $|L - D_j| \leq d$  (we set it as 0.001), this step is finished. Otherwise, we choose the midpoint  $g_{new}$  of  $D_{min}$  and  $C$  as a joint point and repeat step (4) until  $|L - D_j| \leq d$ . If  $D_j = 0$ , namely, the skeleton is tangent to the boundary of the object with point  $C$ , we construct the tangent plane which passes through the point  $C$  and its normal direction is vertical to the skeleton segment direction, we search the shortest intersection point as  $D_{minp}$  between the model surface  $M$  and tangent plane. We set the midpoint of points  $D_{minp}$  and  $C$  as the new joint point, insert it between endpoints of the skeleton segment to form two skeleton segments.

Fig. 2 is our skeleton extraction results for the same protein molecular models shown in Fig. 1. Because we add Step (2) and

(3) to keep the skeletal segments inside the molecular model with holes and add Step (4) to let the joint point of the skeleton be in the center as much as possible, we see that the results using our improved MRG algorithm are better than those using the original MRG algorithm. It provides a guarantee for the similarity comparison of protein molecule shape analysis. In the following Section 3, we will provide the detailed shape similarity comparison by using our improved skeleton and original method.

## 2.2. Shape analysis of protein molecule using new local diameter (LD)

In this section, we propose the similarity comparison method of protein molecule models based on a new local diameter (LD). Firstly, we use the improved MRG algorithm to extract the approximate skeleton of a protein molecule. Secondly, we apply the skeleton to compute the new LD. The main algorithm steps are first summarized below, followed by more algorithmic details.

- Step (1) We evenly extract a certain number of points on the protein surface (we normally collect 260–300 sample points).
- Step (2) We calculate the shortest distance from each sample point to the skeleton line and set the twice of the shortest distance as its LD.
- Step (3) We draw a line chart (histogram) constructed by dividing the LD values of all the sample points into 128 equal intervals and counting the number of sample points that fall into each interval.
- Step (4) We use the LD histograms to compare the similarity between different protein molecule models.

### 2.2.1. Sample point selection

The sample points are evenly selected by using the Dijkstra's algorithm. Dijkstra's algorithm is a graph search algorithm which is used to solve the single-source shortest path problem for a graph with non-negative edge path costs and produce a shortest path tree [14]. Firstly, we select an untreated vertex as the sample point each time and set a certain distance length. Secondly, we mark the vertices whose shortest path from the sample point to the vertex around the sample point is less than the distance length as the treated point around the sample point. We continue the operation

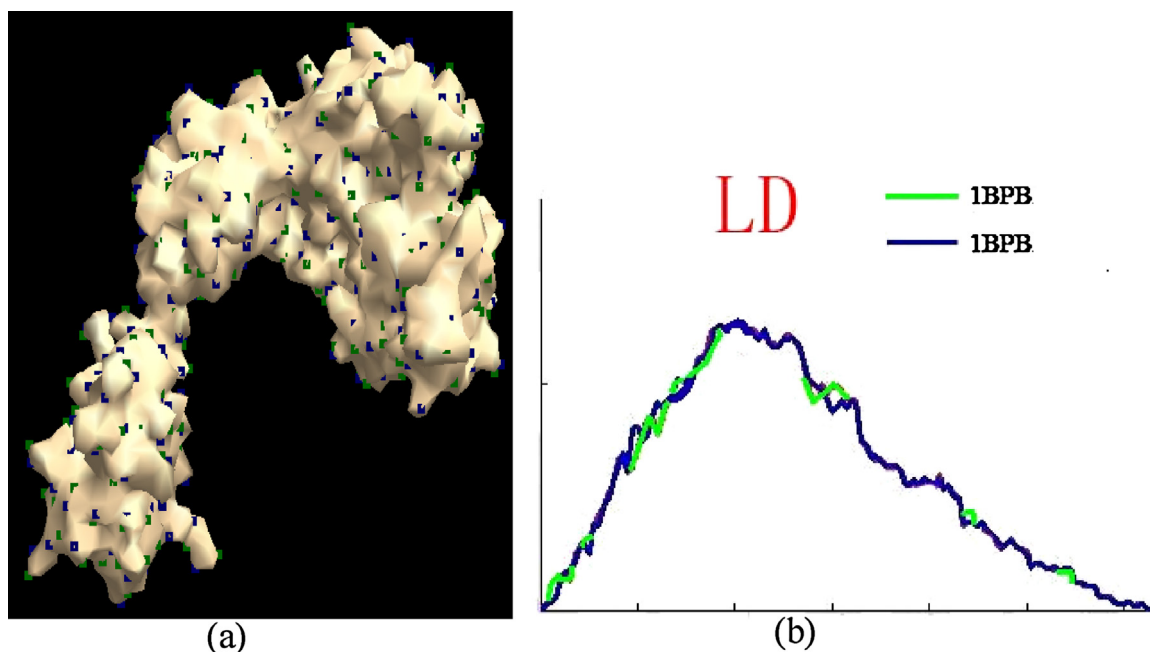


Fig. 3. Sample point selection from 1BPB model and corresponding LD curve.

until every vertex is marked as either a sample point or a treated point.

We stochastically choose the sample points on the protein model and find the variance of the random sample points nearly does not influence the similarity analysis. We give the LD curve comparison by choosing different sample points in blue color and green color, as shown in Fig. 3(b). We find their LD curves have no distinct differences. For the given sample points, we also run the program code for several times and find the results of their LD curves are almost same.

#### 2.2.2. Local diameter (LD)

The local diameter at a surface point is normally defined as the doubled distance from the surface point to the nearest object medial axis [1,15]. Because our approximate skeleton is in the center of the model and can effectively describe the model shape, we use the skeleton as the approximation of the medial axis to compute the LD. Note that the LD is invariant to rigid-body transformations and certain types of deformations [16], hence our skeleton-based LD computation is well suited for similarity comparison in case of protein shape deformations.

As shown in Fig. 4, the idea of the LD computation is that we calculate the distance from the sample points to the nearest skeleton, and then the doubled distance is used as the LD. We sort the distance of LD from the maximum to minimum value, divide it into equal intervals, and count the number of sample points in each range to construct the shape vector of each protein model for similarity comparison of different models [1,15,17].

#### 2.3. Similarity comparison

We do the similarity measurement for the protein models by using the shape vectors. There are two kinds of common similarity measurements: distance similarity and Similarity function [18–20]. We list some popular methods as follows.

##### 2.3.1. Distance similarity

The distance measure between two protein molecules is computed by the vector distance. Here are several common distances for computing sequence  $x = (x_1, x_2, \dots, x_n)^T$  and  $y = (y_1, y_2, \dots, y_n)^T$ .

##### (1) Minkowsky distance

$$d(x, y) = \left( \sum |x_i - y_i|^p \right)^{1/p} \quad (i = 1, 2, \dots, n)$$

##### (2) Euclidean distance

$$d(x, y) = \left( \sum |x_i - y_i|^2 \right)^{1/2} \quad (i = 1, 2, \dots, n)$$

##### (3) Manhattan distance

$$d(x, y) = \left( \sum |x_i - y_i| \right) \quad (i = 1, 2, \dots, n)$$

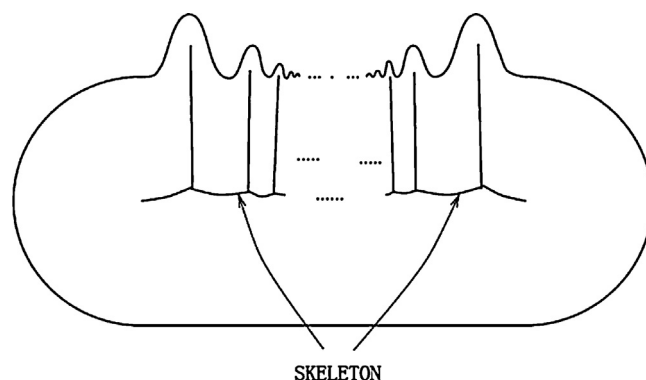
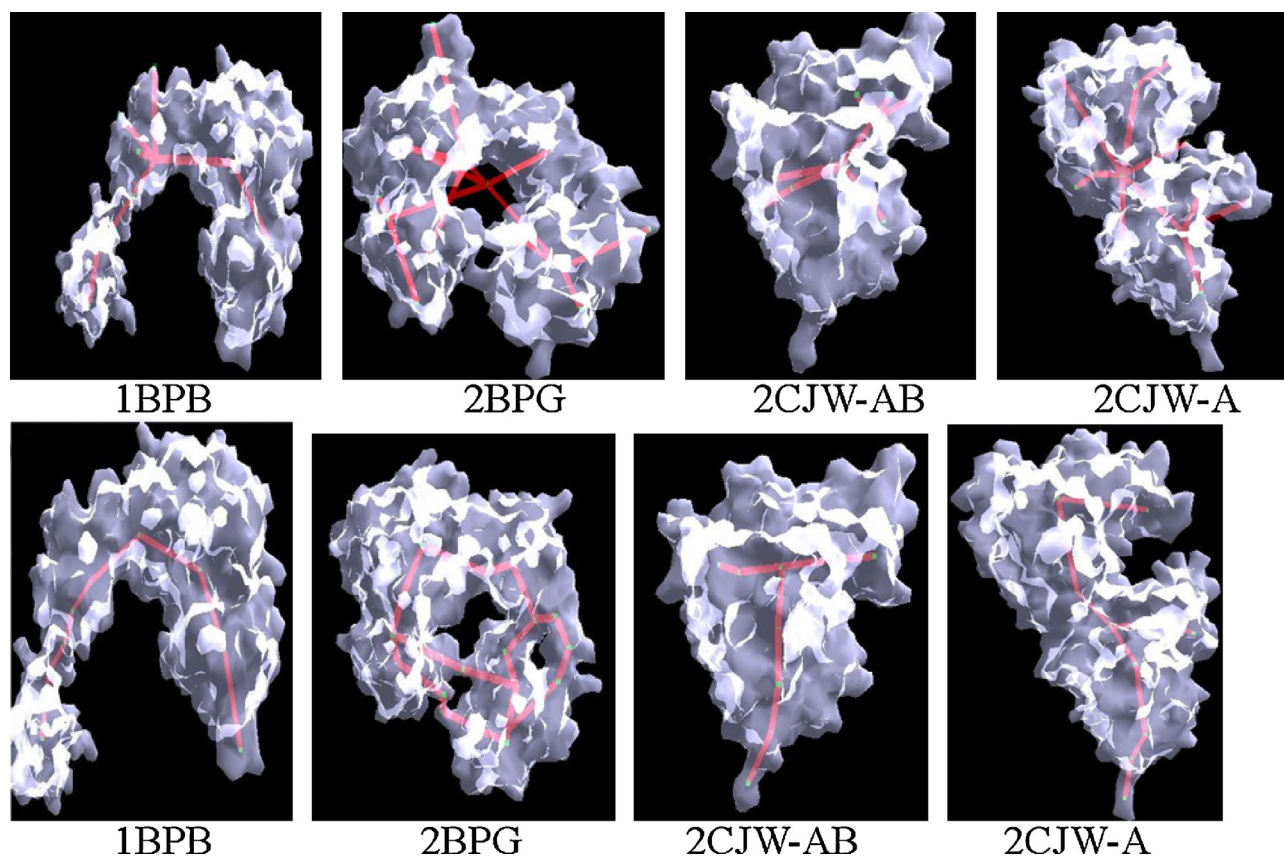


Fig. 4. The distance of the model surface point to the skeleton.





**Fig. 5.** Comparison between the original MRG algorithm and our improved method (the first line is the results of origin MRG algorithm and the second line is our results).

#### (4) Relative error distance

$$d(x, y) = \left( \sum [(x_i - y_i)/y_i]^2 \right)^{1/2} \quad (i = 1, 2, \dots, n).$$

#### 2.3.2. Similarity function

The similarity function is applied more widely than the distance measure for the shape analysis of protein molecule models [20]. Common methods are

##### (1) Cosine of angle

$$\text{Cos}(x, y) = \frac{\sum x_i \times y_i}{[(\sum x_i^2) \times (\sum y_i^2)]^{1/2}} \quad (i = 1, 2, \dots, n).$$

##### (2) Correlation coefficient

$$\text{Corr}(x, y) = \frac{\sum (x_i - \bar{x}) \times (y_i - \bar{y})}{[\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2]^{1/2}}$$

where  $\bar{x} = 1/n \cdot \sum x_i$ ,  $\bar{y} = 1/n \cdot \sum y_i$  ( $i = 1, 2, \dots, n$ ).

##### (3) Generalized Jaccard coefficient

$$\text{EJ}(x, y) = \frac{\sum x_i \times y_i}{[(\sum x_i^2) + (\sum y_i^2) - \sum x_i \times y_i]} \quad (i = 1, 2, \dots, n).$$

##### (4) Biggest dissimilarity coefficient

It first defines the relative error value as  $\lambda_i = |(x_i - y_i)/y_i|$ . Then it sorts the relative error values from the maximum to minimum and takes the first  $k$  values out of these values to build a set  $\xi = \{\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_k\}$ . Finally, it does the weighted mean operation for the set  $\xi$  and obtains the biggest dissimilarity coefficient by

$$\text{Ali}_{\text{biggest}} = \frac{\sum (k - j + 1) \xi_j}{\sum j}, \quad j = 1, 2, \dots, k$$

The value range of the biggest dissimilarity coefficient is in  $[0, \infty)$ . When the degree of similarity is high,  $\text{Ali}_{\text{biggest}}$  is close to 0. Otherwise, it is large.

### 3. The results and discussion

The algorithm presented in the paper is implemented on a Intel(R) Core(TM) i5-3210M CPU @2.5Ghz computer with 8 GB RAM running Windows 7. The proteins for the experiments have been chosen from the protein data bank (<http://www.rcsb.org/pdb>). The environment of experiment is based on VS2010 and OpenGL. We firstly use the improved MRG algorithm based on the collision detection and plane segmentation to extract the shape skeletons. Fig. 5 shows the comparison of our skeleton method and the original MRG algorithm for different protein models. We randomly choose one point A on the protein model, and compute the shortest distance  $L_1$  from this point to the skeleton point B, then extend the line AB and compute another point C between the line AB and the protein surface. We calculate the distance  $L_2$  of segment BC and analyze the result of  $|L_1 - L_2|$ . If  $|L_1 - L_2|$  is smaller, then the corresponding skeleton is more accurate. We find our method can obtain better approximate skeletons than original skeleton method for these protein models.

**Table 1**

Similarity test for similar group and dissimilar group by using different distance measure methods.

Models (group)	A	B	C	D	E	F	G	H
First group	0.894	13.097	0.053	2.012	0.999	0.988	0.997	0.090
Second group	0.460	5.100	0.019	6.199	0.999	0.997	0.993	0.469

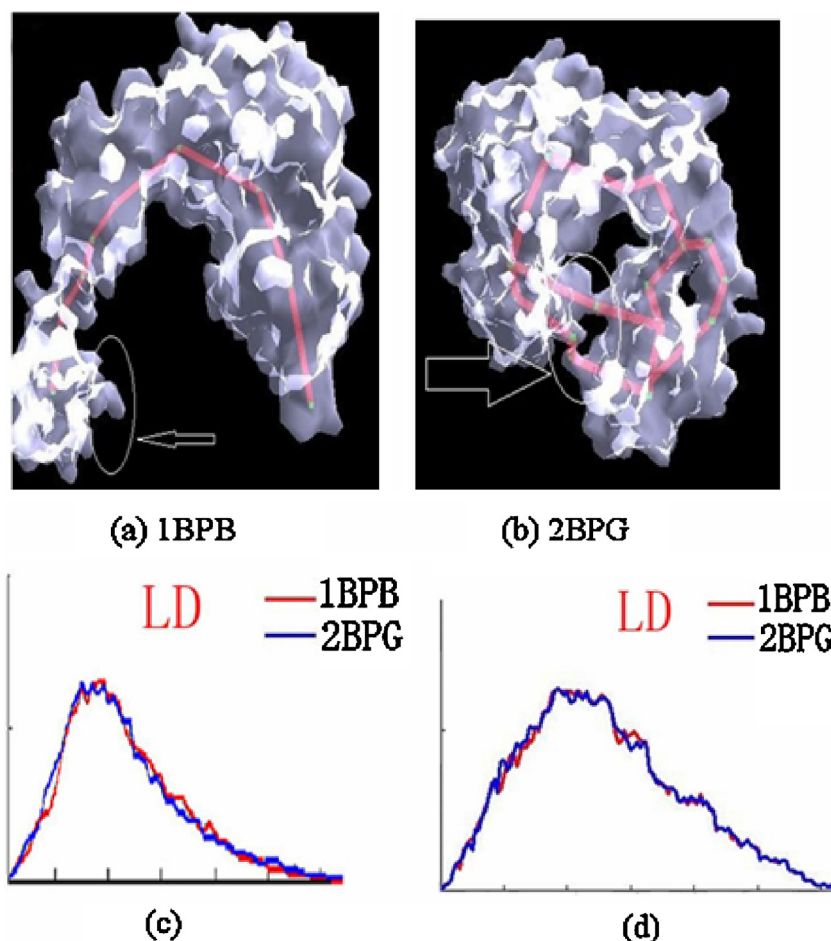
Notes that the first group contains similar models (model 1BPB and model 2BPG). The second group contains dissimilar models (model 2CJW-A and model 2CJW-AB). A is Euclidean distance, B is Manhattan distance, C is Canberra distance, D is Relative error distance, E is Cosine of Angle, F is Generalized Jaccard coefficient, G is Correlation coefficient and H is Biggest dissimilarity coefficient.

We set a group with two similar models (1BPB and 2BPG) and another group with two dissimilar models (2CJW-A and 2CJW-AB) from R [1] in Table 1. We use different distance measure methods based on our skeleton extraction to compute the similarity magnitude between two similar models (First group) and two dissimilar models (Second group) respectively. By analyzing these data for the first similar group and the second dissimilar group, we can verify the effect of different distance measures. We find that using the relative error distance (D) and the biggest dissimilarity coefficient (H) can get relatively accurate similarity results.

We then use the relative error distance to analyze the shape similarity between the original MRG algorithm and our improved MRG method. The results are shown in Table 2. For the second row that protein 2BPG is compared to other protein models, as we know 2BPG and 1BPB are similar models from R [1]. If we use original skeleton extraction method A, the result is that 2BPG and 2CJW-AB are similar (because the value 3.244 is larger than 2.236). But if we use our skeleton extraction method B, we can obtain the same result as R [1] (because 2.602 is smaller than 4.003). For the 2BPG model, its overall shape is close to the 2CJW-AB model,


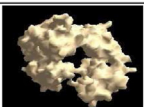
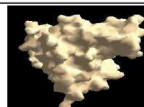
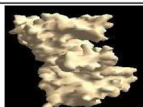
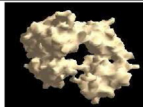











but it contains three holes in the body of the model. Because the method A using the original MRG cannot guarantee the skeleton outside three holes, 2BPG and 2CJW-AB are treated as similar models. However, our method B keeps the skeleton in the center of the model and outside the holes, so it can obtain satisfying similarity results.

Fig. 6 shows two protein models with similar structures but one is deformed from the other. The arrows indicate the deformation area between two proteins. Fig. 6(c) and (d) is the statistical results based on the LD measurement. Here each histogram is constructed by dividing the LD from the maximum to minimum measurement into 128 equal ranges and counting the number of observations that fall into each range. Fig. 6(c) is constructed using the approximate LD method [1] and Fig. 6(d) is the result of our method. For two match curves which reflect the LD distribution result in Fig. 6(c) and (d), we compute the accumulated error between two curves to show the difference by two skeleton extractions. The accumulated error in Fig. 6(c) is 1.083 which is larger than 0.379 in Fig. 6(d). This means that our method can be used for similarity analysis of deformed protein molecule models.



**Fig. 6.** Model (b) is the deformed result from model (a). (c) and (d) Show the statistical results by the method in [1] and the proposed method respectively.

**Table 2**  
Comparison by the relative error distance between the original MRG algorithm and our skeleton algorithm.

Protein Models	Experimental results of using the relative error distance for compared protein models			
 (1BPB) Method A Method B	 (2BPG) 2.522 2.012	 (2CJW-AB) 2.937 2.519	 (2CJW-A) 5.103 6.935	
 (2BPG) Method A Method B	 (1BPG) <u>3.244</u> <u>2.602</u>	 (2CJW-AB) <u>2.236</u> <u>4.003</u>	 (2CJW-A) 27.539 11.012	
 (2CJW-AB) Method A Method B	 (1BPB) 4.565 3.984	 (2BPG) 2.595 2.978	 (2CJW-A) 6.780 6.199	
 (2CJW-A) Method A Method B	 (1BPB) 5.083 3.154	 (2BPG) 3.125 4.224	 (2CJW-AB) 4.154 2.313	

Notes that method A uses original MRG method and method B uses our improved skeleton algorithm. Because the relative error distance does not satisfy the symmetry property of the distance, the experimental results of different sequences of the same model group are different.

Two protein models (1WRP and 3WRP) are shown in Fig. 7(a) and (b), where 1WRP is deformed from 3WRP. We do the similarity analysis of two proteins by using our local diameter based on the improved MRG algorithm. Fig. 7(c) and (d) are the skeleton result of two protein models. This figure supports that for the protein model and the deformed model, although two extracted skeletons are not close, the method based the local diameter (LD) combining with our improved skeleton and the relative error distance can be used to analyze the shape similarity. The value computed by the relative error is 0.118 which is close to 0, so we think our method is applicable for the similarity analysis for protein models with large deformations while the similarity measurement in [1] cannot obtain satisfactory measurement result.

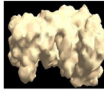
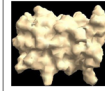
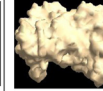
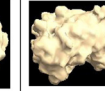
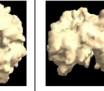
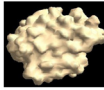
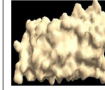

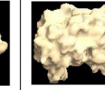
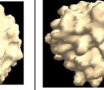
We consider the Skolnick’s dataset for validation of the proposed protein structure comparison algorithms. We randomly choose 10 proteins from the Skolnick’s dataset which is shown in Table 3, and

compute the relative error distances by our method. The similar protein models corresponding to 10 proteins are shown in Fig. 8. We can see that these similarity results of 10 proteins are in accord with Pelta et al’s method [21].

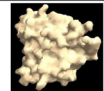
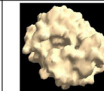
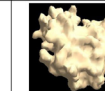
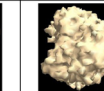
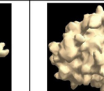
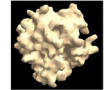
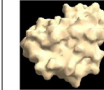
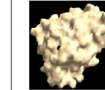
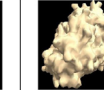
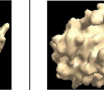
And we also do the protein similarity analysis from different dataset, for example, Chew–Kedem dataset [30]. We randomly choose 10 proteins from the Chew–Kedem dataset which are shown in Table 4, and compute the relative error distances by our method. The similar protein models corresponding to 10 proteins are shown in Fig. 9. We find these similarity results of 10 proteins are in accord to the results in [30].

We provide the protein similarity comparison between our method and some conventional algorithms. We compute the RMSD [22] values for different methods in Table 5. As we know the RMSD of two similar protein models should be close to 0 and otherwise the RMSD should be some large. Although the RMSD value of our

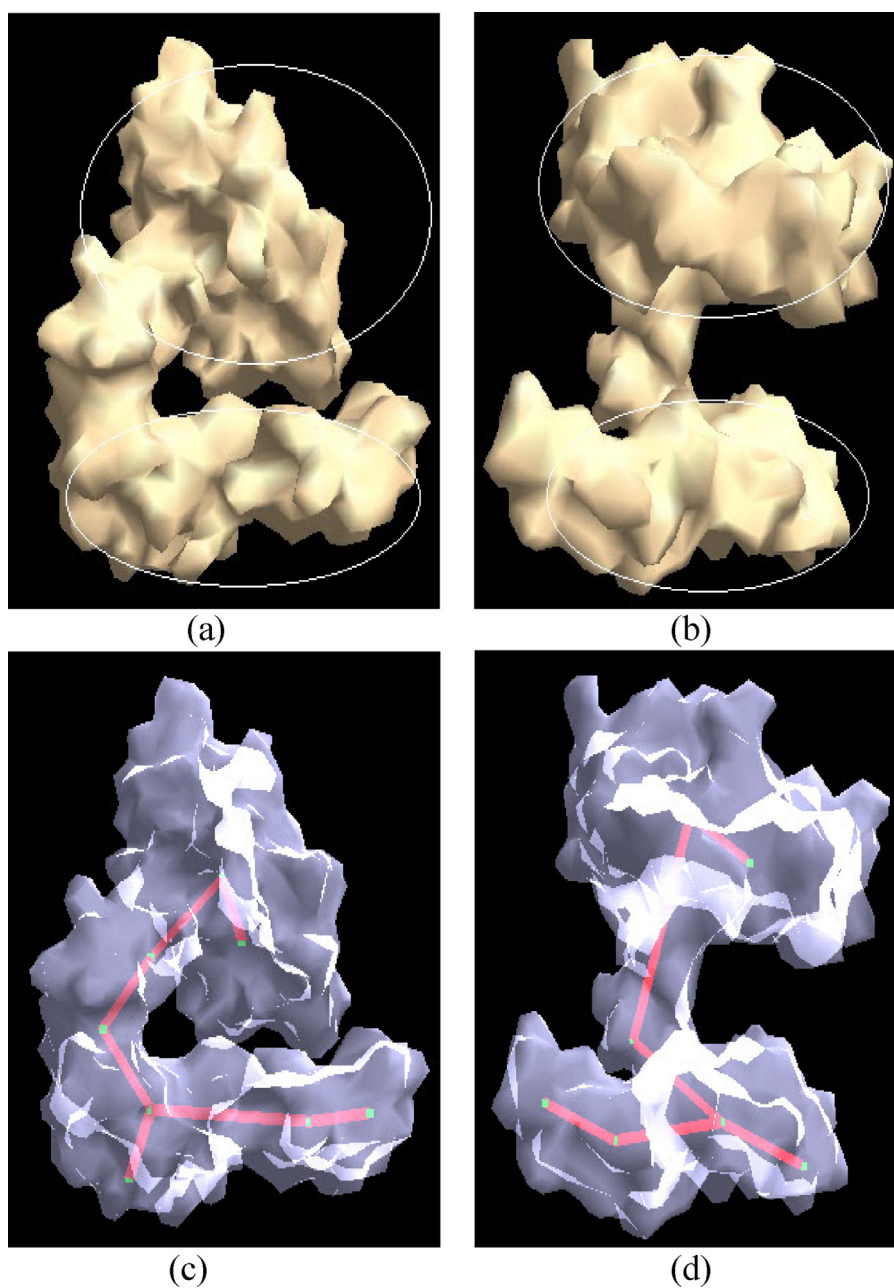
**Table 3**  
10 protein models randomly chosen from the Skolnick dataset.

				
1AW2	2B3I	3YPI	1HTI	1TRE
				
1B00	1RCD	1DBW	1BTM	1NAT

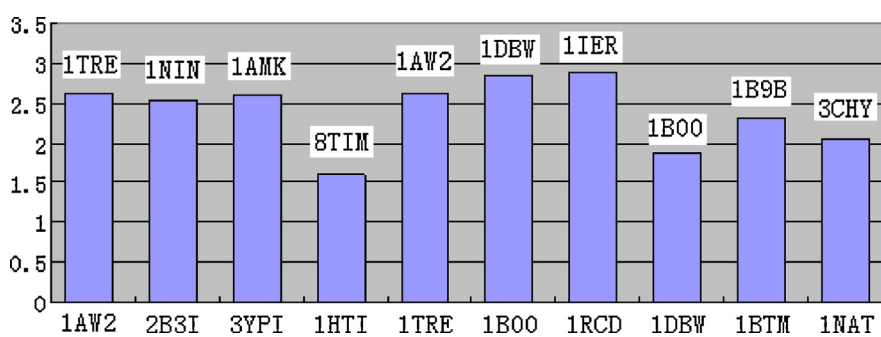
**Table 4**  
10 protein models randomly chosen from the Chew–Kedem dataset.

				
1HLM	2LHB	5MBN	1CHR	5P2I
				
1HLB	1MBA	1LH2	2MNR	1GNP





**Fig. 7.** Model (a) is protein molecule 1WRP and model (b) is protein molecule 3WRP. (c) Shows the skeleton of the protein 1WRP and (d) is the skeleton of the protein 3WRP based on the improved MRG algorithm.



**Fig. 8.** Protein similarity comparison by the relative error distance for randomly choosing 10 proteins from the Skolnick dataset.



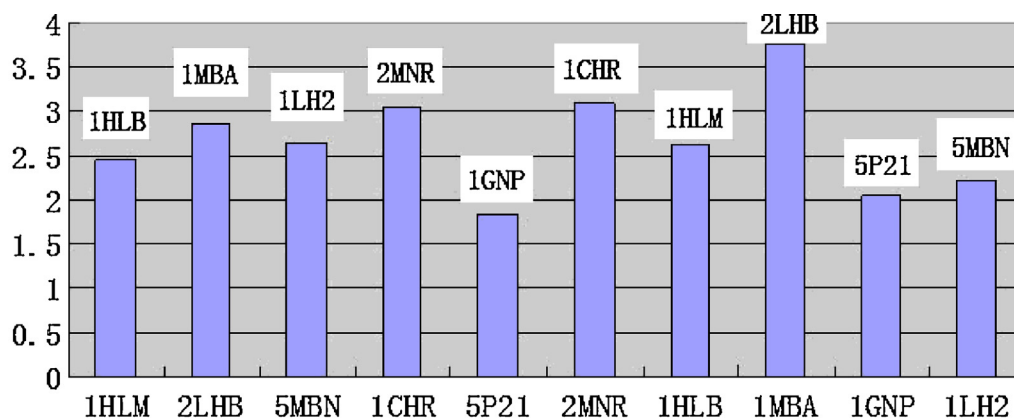


Fig. 9. Protein similarity comparison by the relative error distance for randomly choosing 10 proteins from the Chew–Kedem dataset.

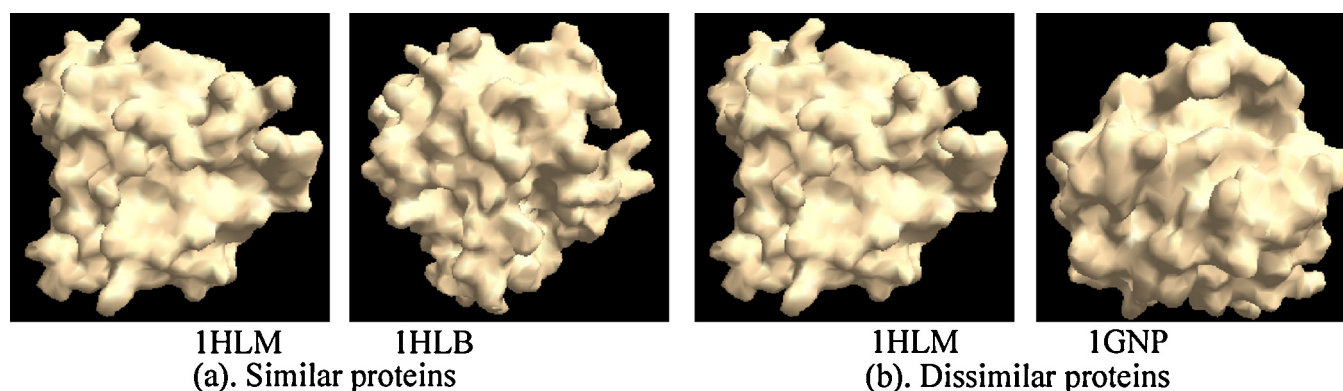


Fig. 10. Comparison of two similar proteins and two dissimilar proteins.

Table 5

Similarity comparing with different methods according to the RMSD (Root-Mean-Square Deviation) value for two group models.

Models (group)	A	B	C	D	E	F	G	H
First group	2.95	10.15	2.72	10.34	2.38	1.76	5.10	1.80
Second group	8.54	4.2	4.41	19.43	5.03	3.36	6.45	2.89

Notes that the first group contains similar models (model 1BPB and 2BPG). The second group contains dissimilar models (model 2CJW-A and 1BPB). A is our method, B is Dali-light method [23], C is TM-align method [24], D is TM-Score method [25], E is CE method [26], F is FATCAT method [27], G is Superpose method [28], H is iPBA method [29].

method for the first group (similar models) is not the smallest one and for the second group (dissimilar models) is not the largest one respectively, the ratio of RMSD value of the second group to that of the first group is largest, which can also effectively do the similarity analysis for the protein models.

Finally, we give the comparison between our method and alpha-carbon traces. In Fig. 10(a), 1HLM and 1HLB models are regarded as similar models in R [30]. We use the online server TM-Align (alpha-carbon traces) to compare the similarity between two proteins and the value of RMSD is 2.75. When using our method, we get the result value of RMSD is 2.70. In Fig. 10(b), 1HLM and 1GNP models are regarded as dissimilar models. We use the online server TM-Align (alpha-carbon traces) to compare the similarity between two proteins and the value of RMSD is 5.35. When using our method, the result value of RMSD is 9.52. Through the comparison, we find our method can improve the shape analysis result for similar or dissimilar models.

#### 4. Conclusion

A novel shape analysis for protein molecule models based on shape skeletons is presented in this paper. It firstly extracts the skeleton by an improved MRG method using collision detection

and plane segmentation. Then it uses the skeleton to compute a new local diameter. Finally it applies the vector distance or the distance function to measure the similarity for protein models. The experimental results show that this method can obtain consistent shape similarity results with alpha-carbon traces for most molecule models. For some protein models with holes and the large deformations, our method can achieve better similarity analysis than other methods.

Our current work has some limitations too. The skeleton extraction takes the majority of running time during the shape analysis, mainly because of the geodesic distance computation and the skeleton optimization. In our method, it is also related to the stochastic process. How to improve our method with the multiple, independent calculation and speed up the skeleton extraction proceed will be our main future work. And how to combine our method with the secondary structure for the protein structure-based retrieval in large protein data banks will also be one of our future researches.

#### References

- [1] Y. Fang, Y. Liu, K. Ramani, Three dimensional shape comparison of flexible proteins using the local-diameter descriptor, *BMC Struct. Biol.* 9 (29) (2009) 1–15.

- [2] S. Lee, B. Li, D. La, et al., Fast protein tertiary structure retrieval based on global surface shape similarity, *Proteins Struct. Funct. Bioinform.* 72 (4) (2008) 1259–1273.
- [3] Z. Lian, A. Godil, B. Bustos, et al., A comparison of methods for non-rigid 3D shape retrieval, *Pattern Recognit.* 46 (1) (2013) 449–461.
- [4] R. Osada, T. Funkhouser, B. Chazelle, et al., Shape distributions, *ACM Trans. Graph.* 21 (4) (2002) 807–832.
- [5] B. Horn, Extended Gaussian image, *Proc. IEEE* 72 (12) (1984) 1671–1676.
- [6] K. Michael, F. Thomas, R. Szymon, Rotation invariant spherical harmonic representation of 3D shape descriptions, in: *Proceedings of the Euro-graphics/ACM SIGGRAPH Symposium on Geometry Processing*, Aachen, Germany, 2003, pp. 156–164.
- [7] M. Kazhdan, B. Chai, D. Dobkin, A reflective symmetry descriptor for 3D models, *Algorithmica* 38 (1) (2004) 22–26.
- [8] P. Min, J. Chen, T. Funkhouser, A 2D sketch interface for a 3D model search engine, in: *Computer Graphics, Proceedings, Annual Conference Series*, ACM SIGGRAPH, 2002, Technical Sketch, Texas, USA, 2002, pp. 22–35.
- [9] M. Reuter, F. Wolter, N. Peinecke, Laplace-spectra as fingerprints for shape matching, in: *Proceedings of the 2005 ACM symposium on Solid and physical modeling*, Cambridge, ACM, New York, 2005, pp. 101–106.
- [10] M. Hilaga, Y. Shinagawa, T. Komura, et al., Topology matching for fully automatic similarity estimation of 3D shapes, in: *Computer Graphics, Proceedings Annual Conference Series*, ACM SIGGRAPH, Los Angeles, CA, 2001, pp. 203–212.
- [11] M. Foskey, M. Lin, D. Manocha, Efficient computation of a simplified medial axis, in: *Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications*, Seattle, WA, USA, 2003, pp. 96–107.
- [12] T. Culver, J. Keyser, D. Manocha, Accurate computation of the medial axis of a polyhedron, in: *Proceedings of Symposium Solid Modeling*, Ann Arbor, MI, United States, 1999, pp. 179–190.
- [13] J. Tierny, J. Vandeboer, M. Daoudi, 3D mesh skeleton extraction using topological and geometrical analyses, in: *Proceedings of the 14th Pacific Conference on Computer Graphics and Applications*, vol. 85, Taipei, 2006, p. 94.
- [14] T. Cormen, C. Leiserson, R. Rivest, et al., *Introduction to Algorithms*, second ed., MIT Press and McGraw-Hill, 2001, pp. 595–601.
- [15] R. Gal, A. Shamir, D. Cohen-Or, Pose-oblivious shape signature, *IEEE Trans. Vis. Comput. Graph.* 13 (2) (2007) 261–271.
- [16] L. Shapira, A. Shamir, D. Cohen-Or, Consistent mesh partitioning and skeletonisation using the shape diameter function, *Vis. Comput.* 24 (4) (2008) 249–259.
- [17] H. Choi, S. Choi, H. Moon, Mathematical theory of medial axis transform, *Pac. J. Math.* 181 (1) (1997) 57–88.
- [18] J. Rodegers, W. Nicewander, Thirteen ways to look at the correlation coefficient, *Am. Statist.* 42 (1) (1988) 59–66.
- [19] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, third ed., Elsevier, Singapore, 2012, pp. 44–52.
- [20] S. Choi, S. Cha, C. Tappert, A survey of binary similarity and distance measures, *Syst. Cybern. Inform.* 8 (1) (2010) 43–48.
- [21] D.A. Pelta, J. Gonzalez, M. Vega, A simple and fast heuristic for protein structure comparison, *BMC Bioinform.* 9 (2008) 161.
- [22] M. Betancourt, J. Skolnick, Universal similarity measure for comparing protein structures, *Biopolymers* 59 (5) (2001) 305–309.
- [23] L. Holm, S. Kaariainen, P. Rosenstrom, et al., Searching protein structure databases with DaliLite v.3, *Bioinformatics* 24 (23) (2008) 2780–2781.
- [24] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic Acids Res.* 33 (7) (2005) 2302–2309.
- [25] Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality, *Proteins* 57 (4) (2004) 702–710.
- [26] I. Shindyalov, P. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng.* 11 (9) (1998) 739–747.
- [27] Y. Ye, A. Godzik, FATCAT: a web server for flexible structure comparison and structure similarity searching, *Nucleic Acids Res.* 32 (2004) W582–W585.
- [28] R. Maiti, G. Domselaar, H. Zhang, et al., SuperPose: a simple server for sophisticated structural superposition, *Nucleic Acids Res.* 32 (2004) W590–W594.
- [29] J. Gelly, A. Joseph, N. Srinivasan, et al., iPBA: a tool for protein structure comparison using sequence alignment strategies, *Nucleic Acids Res.* 39 (2011) W18–W23.
- [30] N. Krasnogor, D.A. Pelta, Measuring the similarity of protein structures by means of the universal similarity metric, *BMC Bioinform.* 20 (7) (2004) 1015–1021.