

RESEARCH ARTICLE

Open Access



# Three-dimensional protein model similarity analysis based on salient shape index

Bo Yao, Zhong Li\*, Meng Ding and Minhong Chen

## Abstract

**Background:** Proteins play a special role in bioinformatics. The surface shape of a protein, which is an important characteristic of the protein, defines a geometric and biochemical domain where the protein interacts with other proteins. The similarity analysis among protein models has become an important topic of protein analysis, by which it can reveal the structure and the function of proteins.

**Results:** In this paper, a new protein similarity analysis method based on three-dimensional protein models is proposed. It constructs a feature matrix descriptor for each protein model combined by calculating the shape index (SI) and the related salient geometric feature (SGF), and then analyzes the protein model similarity by using this feature matrix and the extended grey relation analysis.

**Conclusions:** We compare our method to the Multi-resolution Reeb Graph (MRG) skeleton method, the L1-medial skeleton method and the local-diameter descriptor method. Experimental results show that our protein similarity analysis method is accurate and reliable while keeping the high computational efficiency.

**Keywords:** Protein model, Shape index, Salient geometric feature, Shape analysis

## Background

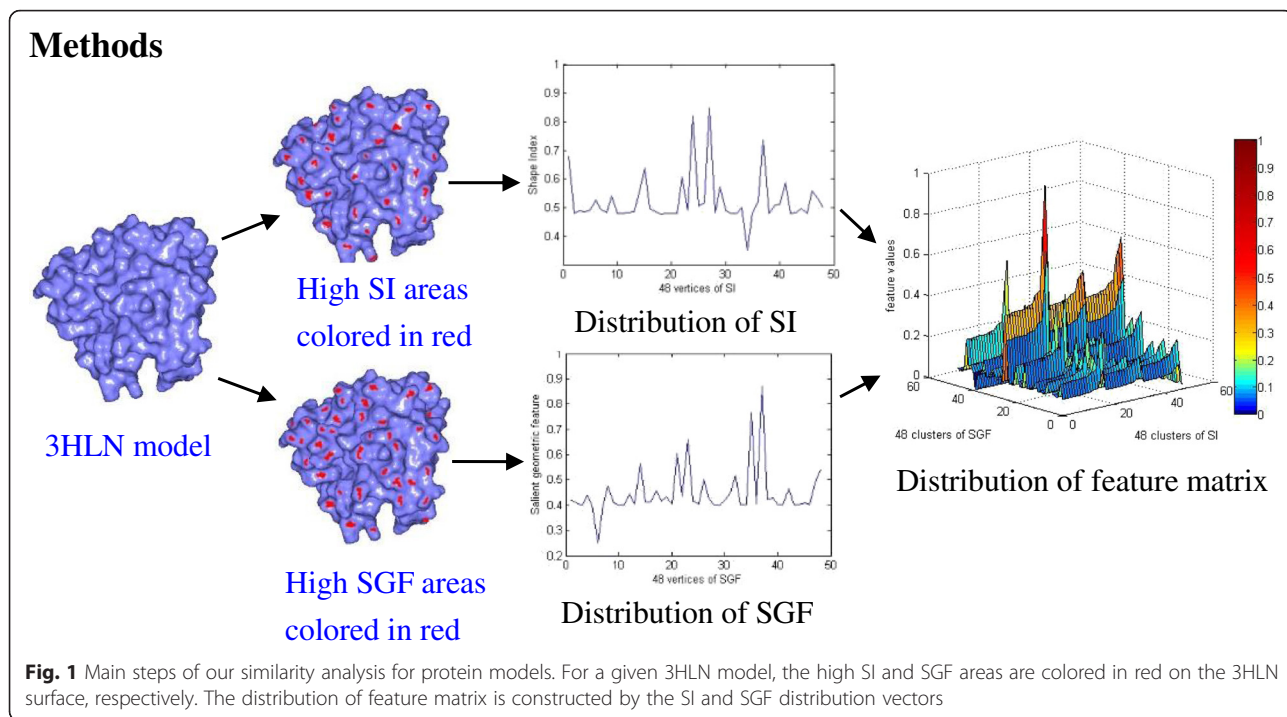
Protein similarity analysis is an important topic in bioinformatics. With it, we can help understand the structure and the function of proteins. The protein shape analysis plays an important role in medical research, computer aided molecular design, protein structure retrieval and prediction, among others. However, the analysis is highly challenging due to the complexity of a protein's three-dimensional surface shape, which can deform significantly enough to change the topological structure during molecular interactions [1].

Many researchers have contributed to similarity analysis methods for comparing protein shapes. Via et al. [2] gave a survey on the current knowledge of the protein surface similarity. The similarity analysis method based on comparison of shape feature is a common approach. Sael et al. [3] proposed a popular protein surface similarity method based on a 3D Zernike descriptor. Compactness and rotational invariance of this descriptor enable fast comparison suitable for protein database searches. However, in order to capture the high resolution

of the protein surface similarity, computation increases as the number of terms in its series expansion increases. And it is not applicable for the protein models with complex topology such as holes. Osada et al. [4] provided a shape distribution method based on the statistical histogram that measures the vertex distribution of the whole model surface, from which it forms a shape feature distribution histogram, and finally obtains a three-dimensional model's geometric similarity measure by comparing two similar distances. Horn et al. [5] proposed an algorithm based on an extended Gaussian image, in which it maps each grid of the model surface to a unit sphere, thus obtains an extended Gaussian ball vector. Ohbuchi et al. [6] presented a statistical histogram algorithm in which the three-dimensional model vertices are sampled and then a three-dimensional coordinate axis histogram is used to generate three statistics about the model's geometric features. Vranic et al. [7] introduced a functional analysis method that assesses the three-dimensional model similarity using the modulus of a spherical harmonic analysis coefficient.

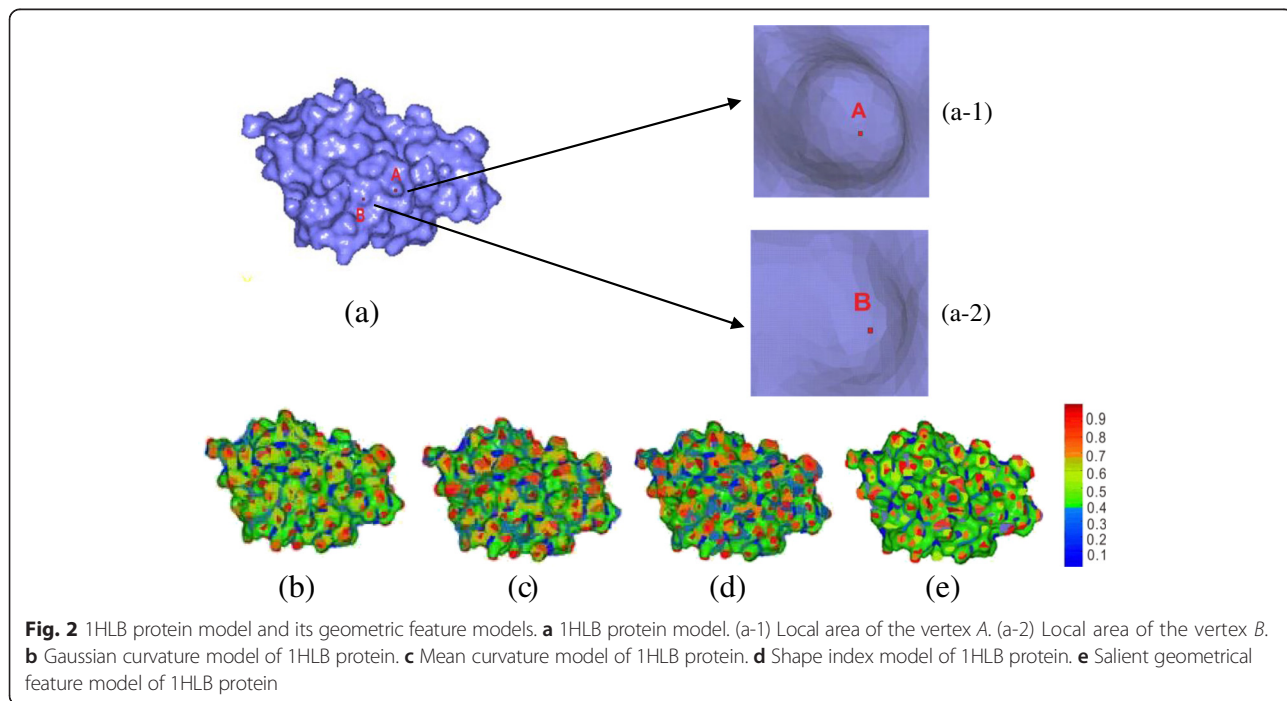
Other shape similarity methods based on topology are also widely studied. For example, Hilaga et al. [8] proposed a multi-resolution Reeb graph (MRG) method. It uses the

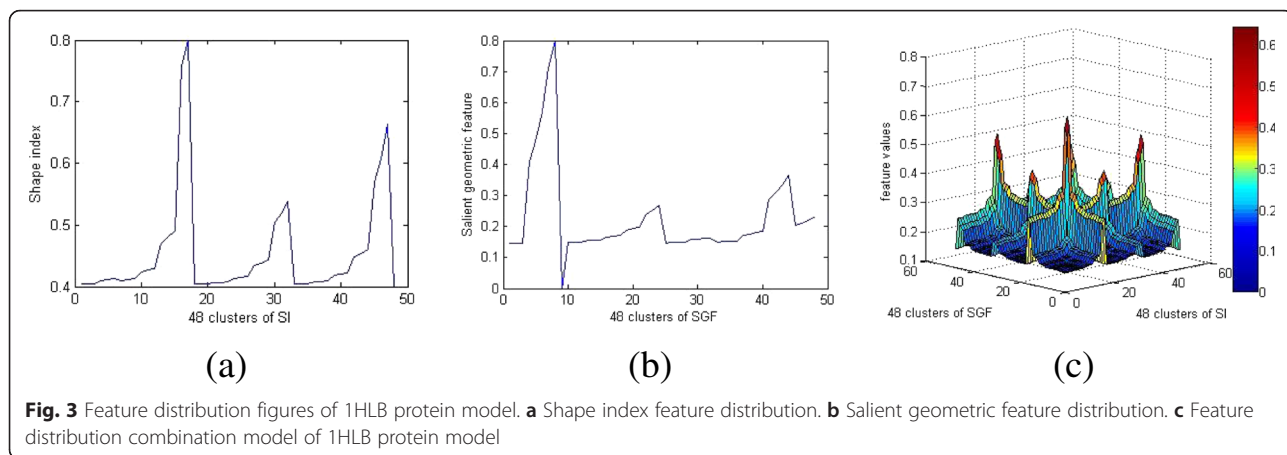
\* Correspondence: lizhong@zstu.edu.cn  
Departments of Mathematical Sciences, Zhejiang Sci-Tech University,  
Hangzhou 310018, China



model surface's geodesic distance as a Morse function to draw the multi-resolution Reeb graph of a three-dimensional model. Bronstein et al. [9] provided a method based on heat kernel signatures (HKS). It draws analogies with feature-based image representations to construct shape descriptors, which are invariant to a wide class of transformations on one hand and are discriminative on

the other hand. Forked et al. [10] proposed a method based on the simplified medial axis, which is parameterized by a separation angle. The angle is formed by the vectors connecting a point on the medial axis to the closest points on the boundary. Du et al. [11] proposed a method based on the skeleton graph. It first calculates the skeleton node of a three-dimensional model and then constructs

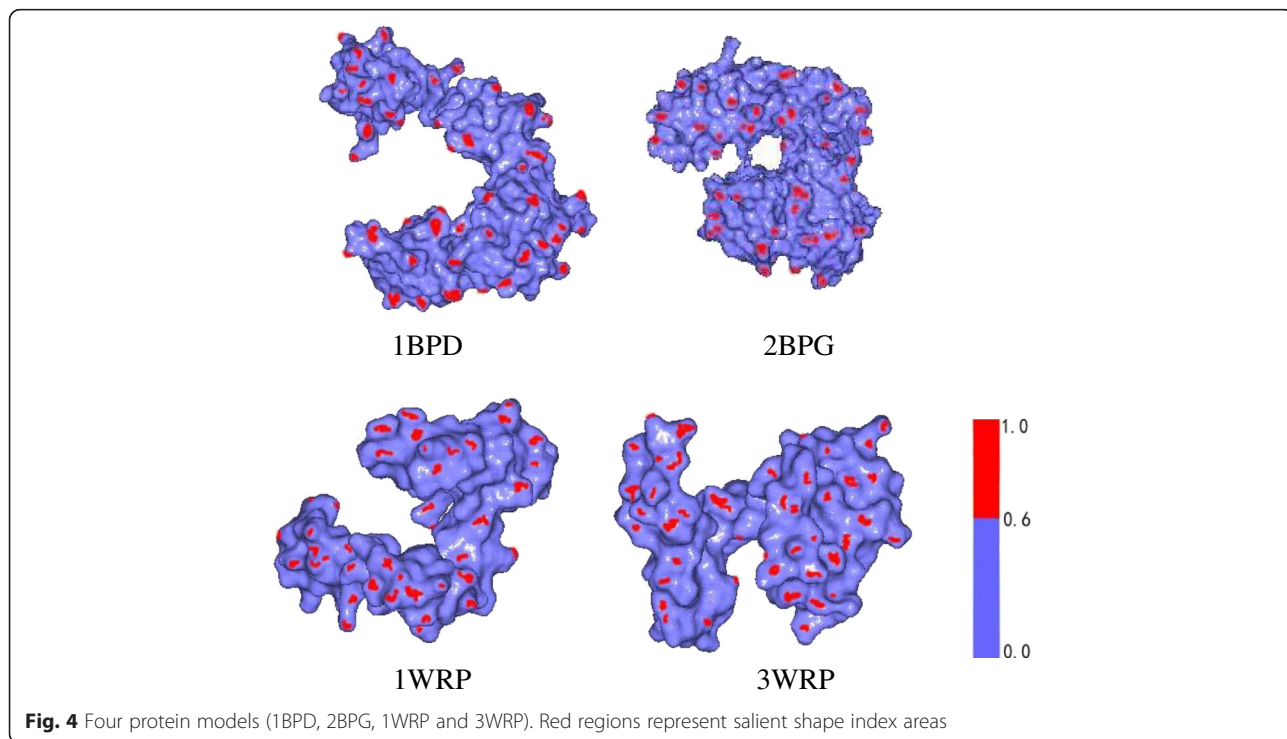




the corresponding skeleton graph between nodes. Both of these methods are computationally expensive, are more sensitive to holes in the three-dimensional models, and are lack of robustness to noise. Morris et al. [12] obtained a similarity comparison of three-dimensional protein models by spherical harmonic expansion. Fang et al. [13] proposed a shape comparison method based on the local diameter (LD). Qin et al. [14] introduced an improved MRG skeleton algorithm. In this process, sample points are used to build the local diameter (LD) for a model similarity comparison, but it is a computationally expensive approach. Li et al. [15] presented a method based on an improved L1-medial protein skeleton, but they only apply it to the CPK protein model. Hence, their method is

not generally applicable to all three-dimensional protein models.

Motivated by the salient theory of a three-dimensional model proposed by Hoffman and Singh [16], and by the shape index (SI) concept by Bradford et al. [17], we propose a new shape comparison method for three-dimensional protein models. We first compute the shape index which reflects the protein surface’s geometric feature including the concave and convex properties. Then we construct the salient geometric feature (SGF) through the region-related shape index information. The shape index and the salient geometric feature are then combined to form the feature matrix of each protein model. We finally use the extended grey relation analysis to



**Table 1** Comparison by the grey relation distance between Qin et al’s algorithm [15] and our algorithm

| Similarity values by different algorithms |               |        |               |
|---|---------------|--------|---------------|
| 1BPD                                      | 2BPG          | 1WRP   | 3WRP          |
| Algorithm in [15]                         | <b>0.9713</b> | 0.5625 | 0.5816        |
| Our algorithm                             | <b>0.9856</b> | 0.3853 | 0.3605        |
| 2BPG                                      | 1BPD          | 1WRP   | 3WRP          |
| Algorithm in [15]                         | <b>0.9713</b> | 0.6497 | 0.6823        |
| Our algorithm                             | <b>0.9856</b> | 0.3844 | 0.3596        |
| 1WRP                                      | 1BPD          | 2BPG   | 3WRP          |
| Algorithm in [15]                         | 0.5625        | 0.6497 | <b>0.9597</b> |
| Our algorithm                             | 0.3853        | 0.3844 | <b>0.9723</b> |
| 3WRP                                      | 1BPD          | 2BPG   | 1WRP          |
| Algorithm in [15]                         | 0.5816        | 0.6823 | <b>0.9597</b> |
| Our algorithm                             | 0.3605        | 0.3596 | <b>0.9723</b> |

Bold numbers mean the similarity values measured by different methods for similar proteins

analyze the feature matrix and obtain the final shape similarity results of protein models.

**Methods**

A three-dimensional protein model can be represented by the form with the triangular mesh. We first estimate the curvature of each vertex on a protein model surface, and calculate the shape index (SI) and the salient geometric feature (SGF) based on the shape index of each vertex. Then, we construct the protein model’s feature matrix through the shape index and the salient shape index. Finally, we do the similarity analysis for protein models by the matrix-based grey relation analysis. The main process of our algorithm is shown in Fig. 1.

**Shape index (SI)**

The concept of shape index was proposed in [17]. It is a curvature-related parameter that describes the protein surface’s concave and convex properties. As we know, surface curvature controls the surface orientation and provides information about its degree of concavity or convexity. Thus, the shape index is thought to play an important role in determining the stability of the protein molecules in the process of molecular recognition and structure prediction. The shape index of a protein model can help us study the atomic-level geometry of the interacting versus non-interacting regions of a protein, and therefore help us understand protein interaction mechanisms.

Here, we focus on using the shape index to represent the shape characteristics of a protein surface. The shape index (SI) of a protein model is defined as

$$SI = -\frac{2}{\pi} \arctan \frac{k_1 + k_2}{k_1 - k_2},$$

where  $k_1$  and  $k_2$  denote the maximum and the minimum principal curvatures, respectively. From the above formula, we know SI is between -1 and 1. When the shape index is close to 1, it indicates the convex shape of the given vertex on the protein surface. On the contrary, when the shape index is close to -1, it indicates the concave shape of the given vertex on the surface. When  $k_1 = -k_2$ , the shape index is 0.

SI relates to the curvature estimation of each vertex on the model surface. We use Dyn and Hormann’s method [18] to estimate the discrete Gaussian curvature  $k_G$  and the discrete mean curvature  $k_M$ . Then, the maximum principal curvature  $k_1$  and the minimum principal curvature  $k_2$  are obtained by

$$k_1 = k_M + \sqrt{k_M^2 - k_G}, \quad k_2 = k_M - \sqrt{k_M^2 - k_G}.$$

For the 1HLB protein model in Fig. 2a, we give the corresponding Gaussian curvature figure and mean curvature figure as shown in Fig. 2b and c, where the red and the blue areas represent large and small curvature regions, respectively. We also show the corresponding shape index figure in Fig. 2d, where the red and the blue areas represent convex and concave regions, respectively.

**Salient Geometric Feature (SGF)**

Salient geometric feature is built on the theory of saliency of visual parts proposed by Hoffman and Singh [16]. They regarded that the saliency of a part depends on two factors: its size relative to the whole object, and the number of curvature changes and their strength. This concept has been applied to three-dimensional mesh model matching [19]. It constructed a salient feature formula based on geometric information, which can detect some areas that the topology and numerical calculations may not be similar, but they are considered to be substantially similar. Here, we focus on using the shape index to construct the salient shape index of a protein surface. It similarly includes the local area size and its shape index variance by

**Table 2** Running time comparison between Qin et al’s [15] algorithm and our algorithm (The time unit is ms)

| Model | 1BPD and 2BPG     |               | 1WRP and 3WRP     |               |
|-------|-------------------|---------------|-------------------|---------------|
| Time  | Algorithm in [15] | Our algorithm | Algorithm in [15] | Our algorithm |
|       | 2153              | 112           | 5716              | 216           |



$$SGF = \sum_{i \in F} w_1 Area(i)SI(i)^3 + w_2 N(SI)Var(SI),$$

where  $F$  is a cluster consisting of each vertex  $i$ ,  $w_1$  and  $w_2$  are the weights, we set them as 0.5.  $Area(i)$  is the area of the patch associated with vertex  $i$  relative to a cluster size,  $N(SI)$  is the number of local minimum(s) or maximum(s) shape index in the cluster,  $Var(SI)$  is the shape index variance in the cluster,  $SI(i)$  is the shape index associated with vertex  $i$ .

For the 1HLB protein, we give its salient geometric feature model in Fig. 2e. The regions with the red color represent the more salient parts, and the regions with the blue color are the less salient parts. And we also use the 1HLB model as the example to address the difference scales of SI and SGF on the protein surface. For vertex A in Fig. 2(a-1), its SI value is 0.5368 and its SGF value is 0.6425. For vertex B in Fig. 2(a-2), its SI value is 0.5279 and its SGF value is 0.2981. We find their SI values are close which are hard to reflect the difference of local feature. Whereas, their SGF values have a big difference because SGF value is related to the local geometric region. When the local geometric region varies saliently, the SGF value is high. So from this model, we conclude that point A has a salient geometric feature since it has a high SGF value.

### Feature descriptor structure

The shape index and the salient geometric feature of all vertices on a three-dimensional model constitute an  $n$ -dimensional vector (where  $n$  is the number of model vertices), respectively. Because the number of vertices on each model surface is not the same for different proteins, these vectors cannot be directly compared and analyzed. In our approach, we cluster all feature values into the same group number through a clustering algorithm [20]. For the number of clusters representing different features in  $K$ -means clustering, the high value of  $K$  will improve the accuracy of shape analysis, but it also increases the computation of shape comparison. The low value of  $K$  does not need the high running time of computation, but it can not guarantee the accuracy of shape analysis. Here, we set it as  $K = 48$ . Then, we calculate the mean for each group  $t_i$  and obtain a feature described vector of shape indexes  $T = (t_1, t_2, \dots, t_K)$ . For the shape index feature clustering of a protein model, we randomly select  $K$  data points from the database of  $n$  values as the initial cluster centers for use with the  $K$ -means clustering algorithm. We perform clustering until the change in cluster centers reaches a convergence condition. From this, we obtain the final  $K$  data point clusters.

Similarly, the salient geometric features of a protein model can be represented as a vector  $P = (p_1, p_2, \dots, p_K)$ . The shape index feature and salient geometrical feature of the 1HLB protein model are shown in Fig. 3a and b,

**Table 3** Grey relation distance comparison by different similarity measure methods

| Similarity measure methods                           | 1BPD and 2BPG | 1WRP and 3WRP |
|--|---------------|---------------|
| Feature vectors of SI ( $T$ )                        | 0.9346        | 0.9517        |
| Feature vectors of SGF ( $P$ )                       | 0.9419        | 0.9486        |
| Combined feature vectors of SI and SGF ( $(T+P)/2$ ) | 0.9357        | 0.9541        |
| Matrix-based feature descriptor ( $Q^TQ$ )           | 0.9856        | 0.9723        |

where the horizontal axis represents 48 representative groups obtained by clustering, and the ordinate axis represents the features. We notice that there is no correspondence between clusters in Fig. 3a and b, because each cluster is determined by the randomly selected initial vertices on the protein surface.

In order to better reflect the shape feature of a three-dimensional protein model, the method based on the feature matrix expression has become a popular method for the model shape analysis [21]. Here, we apply above two vectors to construct a matrix which contains rich feature information as a feature descriptor. We denote  $Q_{K \times 2} = [T; P]^T$  and use  $Q^TQ$  to represent a  $K \times K$  ( $K = 48$ ) feature matrix, then do the similarity analysis for the protein models with this feature descriptor. We give the feature matrix figure of 1HLB protein model in Fig. 3c.

### Similarity measurement

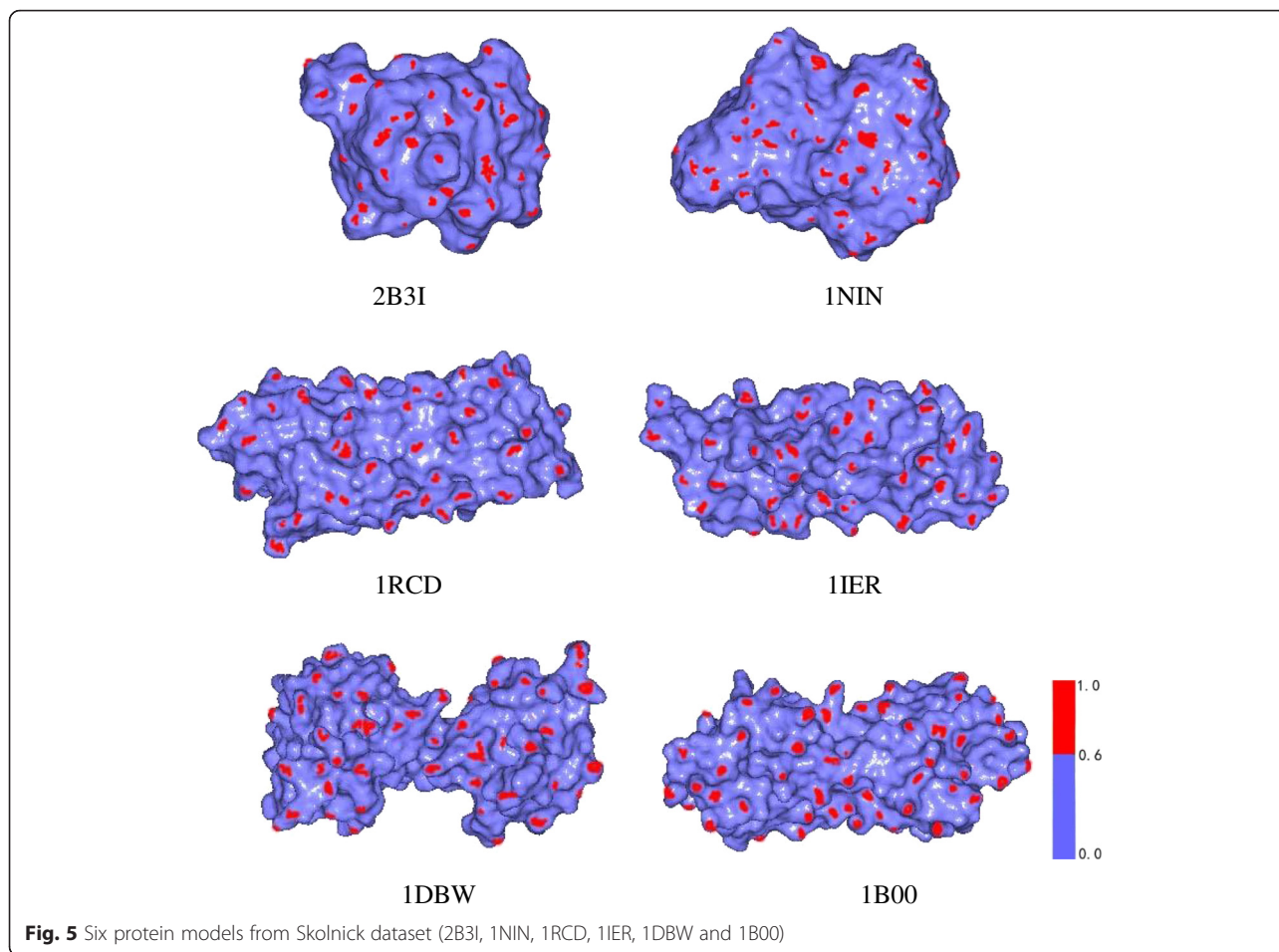
For the shape analysis of a protein sequence or its surface model, common methods use distance measurements such as Euclidean distance, Manhattan distance, angle cosine method and correlation coefficient method, etc [22]. One problem using these methods is that the measure value is normally not guaranteed to lie in the standard interval  $[0, 1]$ . If we use the normalization to transform the values into  $[0,1]$ , it relates to the maximum and the minimum measure values of all protein models and this transformation will influence the accuracy of the shape analysis for protein models.

For our similarity analysis of three-dimensional protein models, because we construct a matrix-based feature descriptor, the previous vector-based method is not directly

**Table 4** Similarity results for six protein models from Fig. 5

| Models | 2B3I          | 1NIN          | 1RCD          | 1IER          | 1DBW          | 1B00          |
|--------|---------------|---------------|---------------|---------------|---------------|---------------|
| 2B3I   | 1.0000        | <b>0.9500</b> | 0.8812        | 0.4460        | 0.9350        | 0.9244        |
| 1NIN   | <b>0.9500</b> | 1.0000        | 0.9269        | 0.4918        | 0.9241        | 0.9176        |
| 1RCD   | 0.8812        | 0.9269        | 1.0000        | <b>0.9630</b> | 0.9109        | 0.7513        |
| 1IER   | 0.4460        | 0.4918        | <b>0.9630</b> | 1.0000        | 0.4749        | 0.4785        |
| 1DBW   | 0.9350        | 0.9241        | 0.9109        | 0.4749        | 1.0000        | <b>0.9930</b> |
| 1B00   | 0.9244        | 0.9176        | 0.7513        | 0.4785        | <b>0.9930</b> | 1.0000        |

Bold numbers mean the similarity values measured by different methods for similar proteins



applicable for our measurement. At the same time, we hope to advocate the use of a scalar value of similarity directly between 0 and 1, where higher values represent greater similarity between two protein models. Here we popularize the vector-based grey relation analysis [23] to the matrix-based grey relation analysis, which also keeps the value in [0,1] and other properties of the grey relation analysis. Then, we apply it to measure the similarity of three-dimensional protein models.

Suppose that  $X$  and  $Y$  are matrices with the same  $m$  rows and  $n$  columns

$$X = [x(i, j)]_{m \times n}, \quad Y = [y(i, j)]_{m \times n},$$

$$i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n.$$

For the  $k^{\text{th}}$  row, we produce the image of zero starting point of matrix  $X$  and  $Y$

$$x'(k, q) = x(k, q) - x(k, 1), \quad y'(k, q) = y(k, q) - y(k, 1),$$

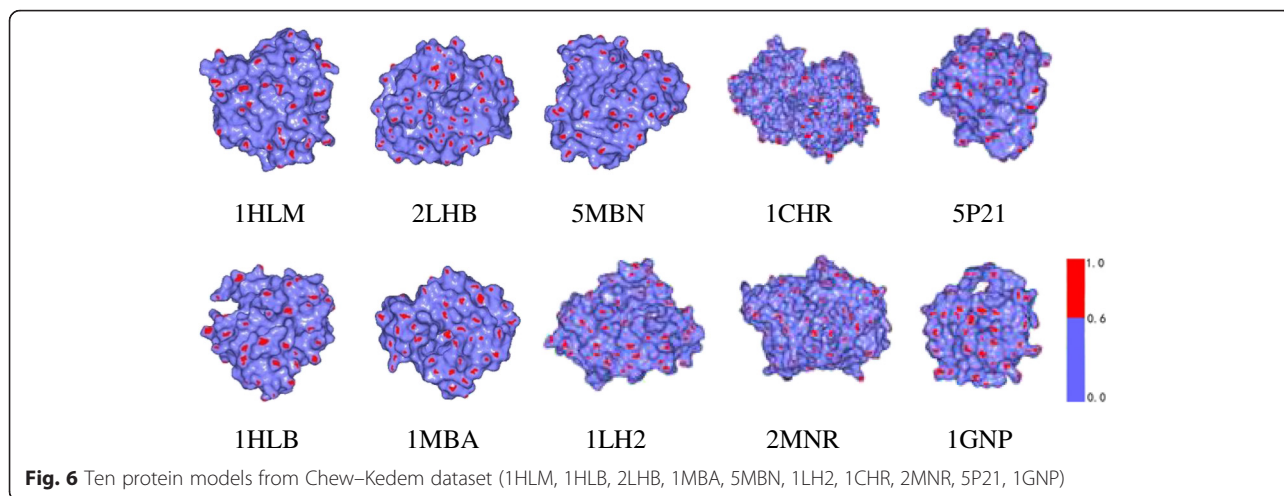
$$q = 1, 2, \dots, n.$$

Then, we compute the grey relation degree of matrix  $X$  and  $Y$  for each row

**Table 5** Running time comparison between the skeleton extraction algorithm [14] and our algorithm

| Model | 2B3I |     | 1NIN  |     | 1RCD |     | 1IER  |     | 1DBW  |     | 1B00  |     |
|-------|------|-----|-------|-----|------|-----|-------|-----|-------|-----|-------|-----|
|       | A    | B   | A     | B   | A    | B   | A     | B   | A     | B   | A     | B   |
| 2B3I  | 2480 | 104 | 2481  | 104 | 2465 | 105 | 2481  | 105 | 2480  | 104 | 2464  | 105 |
| 1NIN  | 7472 | 215 | 8472  | 202 | 7753 | 208 | 6131  | 210 | 11809 | 198 | 10733 | 205 |
| 1B00  | 7457 | 208 | 10733 | 205 | 7738 | 207 | 10743 | 200 | 10732 | 194 | 10749 | 202 |
| 1RCD  | 7488 | 210 | 7753  | 208 | 7737 | 212 | 7759  | 198 | 7753  | 207 | 7738  | 207 |
| 1DBW  | 7456 | 204 | 11809 | 198 | 7753 | 207 | 6130  | 194 | 11825 | 195 | 10732 | 194 |
| 1IER  | 7472 | 199 | 6131  | 210 | 7753 | 198 | 6147  | 194 | 6130  | 194 | 10732 | 200 |

Notes that A is the running time of algorithm in [14], B is the running time of our algorithm. The time unit is ms



$$\varepsilon_{ij}^m(k) = \frac{1 + |s(k)| + |t(k)|}{1 + |s(k)| + |t(k)| + |s(k) - t(k)|},$$

(where)

$$s(k) = \sum_{q=1}^n x'(k, q), \quad t(k) = \sum_{q=1}^n y'(k, q),$$

$$k = 1, 2, \dots, n.$$

Similarly, we get the grey relation degree  $\varepsilon_{ij}^n(k)$  of matrix  $X$  and  $Y$  for the  $k^{\text{th}}$  column. Finally, we obtain the grey relation degree of matrix  $X$  and  $Y$  by

$$\eta = \frac{1}{2} \left( \frac{1}{m} \sum_{k=1}^m \varepsilon_{ij}^m(k) + \frac{1}{n} \sum_{k=1}^n \varepsilon_{ij}^n(k) \right).$$

From the above calculation process, it is easily known that the grey relation degree is between 0 and 1, and the degree indicates the high similarity of two models when it is close to 1.

### Results

The algorithm presented in this paper is implemented on a Intel(R) Core(TM) i3-3110 M CPU @2.5 Ghz desktop computer with 4GB RAM running MS Windows 7. The software environment of the experiment is based on Mathworks' MATLAB R2010a.

We first chose four protein models from the Protein Data Bank [24], which are shown in Fig. 4. We already know that the 1BPD and 2BPG models are similar and the 1WRP and 3WRP models are similar [15]. Table 1 shows the results of comparing our algorithm with Qin et al's algorithm [15] which is based on the improved L1-medial skeleton extraction. The similarity measurement values of two methods are both between 0 and 1, and the more similar two protein models, the more close to 1 their values. Our analysis method obtains a reasonable similarity comparison result because our value is closer to 1 by comparing bold data in Table 1. In Table 2, we compared the execution time of two algorithm's implementations, which shows that our method runs faster than Qin et al's algorithm [15]. We also compare our matrix-based feature descriptor to the vector-based method

**Table 6** Similarity results for ten protein models from Fig. 6

| Models | 1HLM  | 1HLB  | 2LHB  | 1MBA  | 5MBN  | 1LH2  | 1CHR  | 2MNR  | 5P21  | 1GNP  |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1HLM   | 1.000 | 0.973 | 0.562 | 0.676 | 0.437 | 0.752 | 0.653 | 0.549 | 0.789 | 0.752 |
| 1HLB   | 0.973 | 1.000 | 0.608 | 0.763 | 0.574 | 0.647 | 0.564 | 0.653 | 0.742 | 0.698 |
| 2LHB   | 0.562 | 0.608 | 1.000 | 0.969 | 0.579 | 0.695 | 0.745 | 0.659 | 0.732 | 0.634 |
| 1MBA   | 0.676 | 0.763 | 0.969 | 1.000 | 0.614 | 0.713 | 0.697 | 0.705 | 0.720 | 0.744 |
| 5MBN   | 0.437 | 0.574 | 0.579 | 0.614 | 1.000 | 0.987 | 0.353 | 0.438 | 0.493 | 0.464 |
| 1LH2   | 0.752 | 0.647 | 0.695 | 0.713 | 0.987 | 1.000 | 0.434 | 0.468 | 0.464 | 0.413 |
| 1CHR   | 0.653 | 0.564 | 0.745 | 0.697 | 0.353 | 0.434 | 1.000 | 0.993 | 0.563 | 0.615 |
| 2MNR   | 0.549 | 0.653 | 0.659 | 0.705 | 0.438 | 0.468 | 0.993 | 1.000 | 0.595 | 0.685 |
| 5P21   | 0.789 | 0.742 | 0.732 | 0.720 | 0.493 | 0.464 | 0.563 | 0.595 | 1.000 | 0.981 |
| 1GNP   | 0.752 | 0.698 | 0.634 | 0.744 | 0.464 | 0.413 | 0.615 | 0.685 | 0.981 | 1.000 |

**Table 7** Running time comparison including searching the dataset between the skeleton extraction algorithm [14] and our algorithm

| ModelsMethods | 1HLM          | 2LHB          | 5MBN          | 1CHR          | 5P21          |
|---------------|---------------|---------------|---------------|---------------|---------------|
| A             | 20mins 28 s   | 25mins 05 s   | 22mins 49 s   | 18mins 58 s   | 21mins 36 s   |
| B             | 5mins<br>43 s | 7mins<br>19 s | 6mins<br>57 s | 5mins<br>34 s | 6mins<br>26 s |

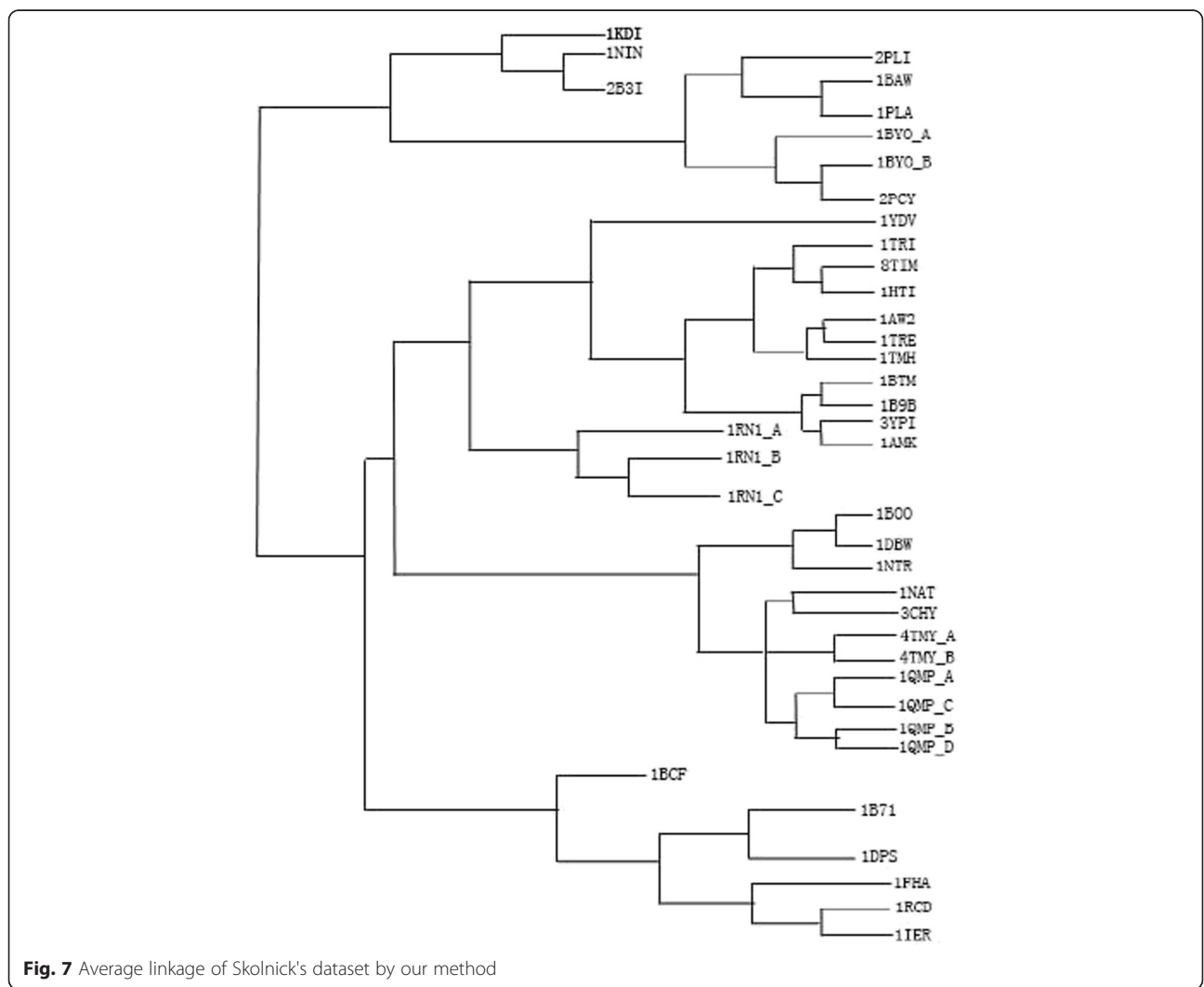
Note that A is the algorithm in [14] and B is our algorithm

directly by SI, SGF, and the simply combined feature vector method ((SI + SGF)/2). The results are shown in Table 3. For two pairs of similar proteins, we find our similarity result is more close to 1.

Then, we chose three groups of protein models from the Skolnick dataset [25], which are shown in Fig. 5. We already know that the 2B3I and 1NIN proteins are similar because they are in the same clustering [14,25]. For other four protein models, the 1RCD and 1IER models are similar, and the 1DBW and 1B00 models are similar [14,15]. We used our method to compare similarities of

these three pairs of proteins and find that they are in accordance with the results of [14,15,25]. Table 4 indicates that two models with bold underlined data are similar (The similarity of a model with itself is always 1.000). We also compared the execution time of Li et al's method [14] based on the improved skeleton extraction by Reeb Graph and our method in Table 5. We find our method is obviously faster than the method in [14] because it does not need to conduct the skeleton extraction.

Next, we chose 10 protein models from the Chew-Kedem dataset [26], which are shown in Fig. 6. We have



**Fig. 7** Average linkage of Skolnick's dataset by our method



known that the 1HLM and 1HLB proteins are similar because they are in the globin family, the 5P21 and 1GNP proteins are similar because they are the alpha-beta family [26]. We computed the matrix-based grey relation distance by using our method. The similarity results of protein models corresponding to 10 proteins are shown in Table 6. We find our similarity results are in agreement with the results in [26].

We also demonstrate a total running time including searching the most similar protein model for 5 protein models in the Chew-Kedem dataset. In Table 7, we find our method has a fast searching speed for obtaining the similar protein model.

To increase the robustness of our method, we added another testing dataset as Skolnick’s dataset from R[25] for the experiment, which includes 40 proteins models. We use our method to construct the average linkage of Skolnick’s dataset in Fig. 7 and find that our result is almost consistent with the result in R[25]. For example, the 3YPI and 1AMK proteins are in the same cluster, the 1NAT and 3CHY proteins are in the same cluster. These results are in accord with the current evolutionary research [14,25].

Finally, we compared two proteins (1BAR and 1RRO) that have a similar shape surface but have completely different secondary structure elements [13] (Fig. 8). The algorithm [13] based on the shape analysis by the local diameter construction resulted in a similarity of 0.9956, which is close to 1. It falsely reflects the similarity

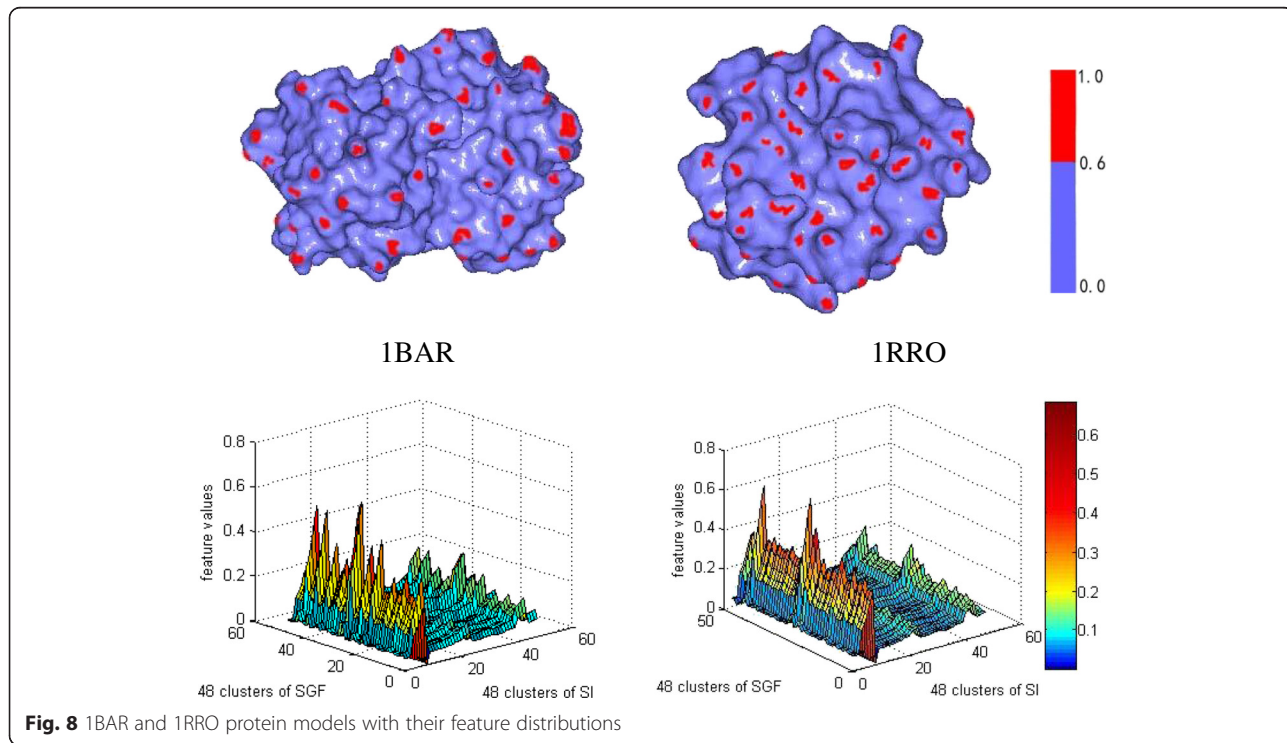
properties of the two proteins with different secondary structures. Our method produced a similarity value of 0.7546 which is comparatively some smaller than 1. It can infer the non-similarity for two protein models although they have similar shape surface. We conclude that our approach, in this specific case, improves the similarity analysis for non-homologous protein models with similar shapes.

**Discussion**

Our method is based on the surface analysis and the advantage is that the running time is fast because it does not need to conduct the skeleton extraction. The disadvantage is that it can not be applied for other protein models such as protein CPK models. The advantage of the skeleton based method is that it can be applied for both the protein surface (triangular mesh model) and the protein CPK model (point cloud representation), the disadvantage is the skeleton extraction requires a time-consuming process.

**Conclusion and future work**

In this paper, we propose a three-dimensional protein model’s similarity analysis algorithm based on salient shape index. We first calculate the shape index (SI) and salient geometric feature (SGF) of the protein models. And then we construct the matrix-based feature descriptor by SI and SGF information. Finally, we compare the similarity of protein models by the matrix-based grey



**Fig. 8** 1BAR and 1RRO protein models with their feature distributions

relation degree. Experimental results show the effectiveness of our protein similarity analysis method.

Currently, we only consider the shape index (convex and concave properties of the protein surface) and the salient geometric feature to analyze the similarity of the protein models. We do not take account of the physical properties of the protein molecules. In fact, these properties such as pH, polar and non-polar, hydrophilic, also affect the structure and the function of the protein molecules. How to combine these factors for the protein shape similarity analysis will be our future research. For the clustering in our similarity analysis, we find the cluster size is normally not equal and the clustering is sometimes dominated by several big clusters. The size of the clusters might be highly relevant in describing the global shape of the protein model. This also gives us an interesting work for our future research.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

Developed the method: ZL and BY. Conceived and designed the experiments: ZL, BY, MD and MC. Analyzed the data: BY and ZL. Wrote the first draft of the manuscript: BY and ZL. Contributed to the writing of the manuscript: ZL, MD and MC. Agree with the manuscript results and conclusions: ZL, BY, MD and MC. Jointly developed the structure and arguments for the paper: BY, ZL, MD and MC. Made critical revisions and approved final version: ZL, BY, MD and MC. All the authors reviewed and approved of the final manuscript.

#### Acknowledgements

We thank Prof. Stefan Heller in the school of Medicine, Stanford University, USA, for helpful discussion and suggestion. This research was supported by Scientific Research Foundation of Ministry of Education of China under Grant No. [2009]1590, Zhejiang Provincial Natural Science Foundation of China under Grant No. LY14A010032, Zhejiang Province Key Science and Technology Innovation Team Project (2013TD18) and Project of 521 Excellent Talent of Zhejiang Sci-Tech University.

Received: 14 September 2015 Accepted: 9 March 2016

Published online: 18 March 2016

#### References

- Xu SC, Li Z, Zhang SP, et al. Primary structure similarity analysis of proteins sequences by a new graphical representation. *SAR QSAR Environ Res*. 2014;25(10):791–803.
- Via A, Ferre F, Brannetti B, et al. Protein surface similarities: a survey of methods to describe and compare protein surfaces. *CMLS Cell Mol Life Sci*. 2000;57:1970–7.
- Sael L, La D, Li B, et al. Rapid comparison of properties on protein surface. *Proteins Struct Funct Bioinforma*. 2008;73:1–10.
- Osada R, Funkhouser T, Chazelle B, et al. Matching 3D models with shape distributions. Geneva: *Proceeding of Shape Modeling International*; 2001. p. 07–11.
- Horn BK. Extended Gaussian image. *Proc IEEE*. 1984;1671–1686.
- Ohbuchi R, Nakazawa M, Takei T. Retrieving 3D shapes based on their appearance. Berkeley, California, USA: *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*; 2003.
- Vranic DV, Saupe D. 3D model retrieval with spherical harmonics and moments. London: *The 23rd DAGM Symposium on Pattern Recognition*, SpringerVerlag; 2001. p. 392–7.
- Hilaga M, Shinagawa Y, Komura T, et al. Topology matching for fully automatic similarity estimation of 3D shapes. Los Angeles, California, USA: *Computer Graphics, Proceedings of Annual Conference Series, ACM SIGGRAPH*; 2001. p. 203–12.
- Bronstein MM, Kokkinos I. Scale-invariant heat kernel signatures for non-rigid shape recognition. *CVPR*. 2010;1704–1711.
- Foskey M, Lin MC, Manocha D. Efficient computation of a simplified medial axis. Seattle Washington, USA: *Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications*; 2003. p. 96–107.
- Du HX, Qin H. Medial axis extraction and shape manipulation of solid objects using parabolic PDEs. Genoa: *ACM Symposium on Solid Modeling and Applications*; 2004. p. 25–34.
- Morris RJ, Najmanovich RJ, Kahraman A, et al. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*. 2005;21:2347–55.
- Fang Y, Liu YS, Ramani K. Three dimensional shape comparison of flexible proteins using the local-diameter descriptor. *BMC Struct Biol*. 2009;29(9):01–15.
- Li Z, Qin SW, Yu ZY, et al. Skeleton-based shape analysis of protein models. *J Mol Graph Model*. 2014;53:72–81.
- Qin SW, Li Z, Jin Y, et al. Shape similarity comparison of CPK models based on improved L1-medial skeleton. *SAR QSAR Environ Res*. 2014;25(9):747–59.
- Hoffman D, Singh MD. Saliency of visual parts. *Dep Cogn Sci*. 1997;63(1):29–78.
- Bradford R, Westhead R. Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*. 2008;1487–1494.
- Dyn N, Hormann K, Kim S-J, Levin D. Optimizing 3d triangulations using discrete curvature analysis. In: Lyche T, Schumaker LL. (eds.) *Mathematical Methods in CAGD*, Oslo 2000, pp. 135–146 (2001).
- Gal R, Cohen-Or D. Salient Geometric Features for Partial Shape Matching and Similarity. *ACM Trans Graph*. 2006;25(1):134–8.
- Kanungo T, Mount DM, Netanyahu NS, et al. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans Pattern Anal Mach Intell*. 2002;24(7):881–92.
- Hu R, Fan L, Liu L. Co-segmentation of 3D shapes via subspace clustering. *Eurographics Symp Geom Process*. 2012;31(5):1703–13.
- Choi S, Tappert S. A survey of binary similarity and distance measures. *Cybern Inform*. 2010;8(1):43–8.
- Zhang S. Matrix absolute grey relational degree of B-Mode and its significance. *J Grey Syst*. 2012;9:135–41.
- Protein Data Bank [<http://www.rcsb.org/pdb/home/home.do>].
- Pelta DA, Gonzalez JR, Vega MM. A simple and fast heuristic for protein structure comparison. *BMC Bioinforma*. 2008;9(1):156–61.
- Krasnogor N, Pelta DA. Measuring the similarity of protein structure by means of the universal similarity metric. *BMC Bioinforma*. 2004;20(7):1015–21.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

