Interdiscip Sci Comput Life Sci (2010) 2: 308–319
DOI: 10.1007/s12539-010-0019-8

Protein Structure Alignment Based on Internal Coordinates

Yue-Feng SHEN[1,2], Bo LI[1,2], Zhi-Ping LIU[1*]
[1](Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China)
[2](Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

Received 6 October 2008 / Revised 5 January 2010 / Accepted 6 January 2010

Abstract: Protein structure alignment provides an effective way to detect and analyze fold mechanism, evolutionary history and biological function of proteins. In this work, we introduced a novel method named SABIC for protein structure alignment based on the internal coordinates (i.e. bond lengths, bond angles and torsion angles) of structure representation. SABIC provides multi-alignments as the output, from which various aspects of structural similarities between proteins can be identified. The experimental results on benchmark datasets show that SABIC performs better than other popular algorithms, such as DALI, CE and SSM. Using a new defined mQ-score of alignment, SABIC performs consistently better in detecting structural classifications of proteins. In addition, we detected the extreme value distribution form statistics of mQ-score, and then the statistical significance P-value of alignment can be obtained simultaneously. The presented SABIC algorithm has been implemented in C++ and the software is available (http://www.aporc.org/doc/wiki/SABIC).

Key words: protein structure alignment, internal coordinate, difference matrix, structural bioinformatics.

1 Introduction

The number of proteins with known structures in PDB has exceeded 50,000, and it will grow larger in the near future (Berman *et al.*, 2000). These known structures provide us with rich information about the folding mechanism, evolutionary history and biological function of proteins. A common fold pattern might be shared by many proteins and it is widely believed that proteins with significant structure similarity are more likely to have similar functions or origins (Chothia and Lesk, 1986; Kinch and Grishin, 2002). To investigate such similarities in protein structures and functions, protein structure alignment methods are developed (Koehl, 2001). Structure alignment plays important roles in classifying protein folding space, detecting evolutionary relationship, annotating protein function, identifying structural motifs and designing drug targets in bioengineering (Eidhammer *et al.*, 2000; Sierk and Pearson, 2004; Thornton *et al.*, 2000; Stark *et al.*, 2003; Whisstock and Lesk, 2003). Many structure alignment algorithms have been developed, e.g. SSAP (Taylor and Orengo, 1989), DALI (Holm and Sander, 1993), CE (Shindyalov and Bourne, 1998), and SSM (Krissinel and Henrick, 2004). Most of them represent a protein structure by a distance matrix, which is often considered to be highly redundant (Godzik, 1996; Stark *et al.*, 2003; Miao *et al.*, 2008). Kolodny *et al.* (2005) presented a comprehensive evaluation of several well-known structure alignment methods. They highlighted the limitations of these existing methods and expected further development of better ones. Protein structure alignment still needs improving, especially in detailed analysis of structural similarities and functional implications (Thornton *et al.*, 2000; Watson *et al.*, 2005; Sippl, 2008).

For most existing methods, there are two major barriers to identify the optimal alignment between two proteins. One is the measurement of structure similarity, and the other is the algorithms themselves. Generally, two protein structures are considered to be similar if the alignment between them has long length of aligned residues and small RMSD. However, it is always possible to enlarge the length of alignment at the expense of RMSD because they are the tradeoff factors during alignment process (Chen *et al.*, 2006). To completely describe the quality of each alignment, various similarity scores have been designed, such as Z-score (Shindyalov and Bourne, 1998), Q-score (Krissinel and Henrick, 2004), TM-score (Zhang and Skolnick, 2005), and the score of geometric measures (Kolodny *et al.*, 2005). Although these scores have considered the balance between the alignment length and RMSD, no generally accepted standards have been proposed (Sippl, 2008). As for the algorithms, finding the optimal alignment between two proteins is an NP-hard

*Corresponding author.
E-mail: zpliu@amss.ac.cn

problem (Goldman *et al.*, 1999; Chen *et al.*, 2005) and none of the proposed algorithms gives an exact solution in a polynomial time. Due to the two barriers, only one alignment produced by these previous methods may not be enough for describing the similarities between two proteins (Godzik, 1996). Hence, generating multi-alignments does not only give various aspects of structure similarities between them, but also produces candidate alignments for choosing the best one from a biological perspective (Feng and Sippl, 1996; Oldfield, 2007).

In this work, we proposed a novel structure alignment algorithm named SABIC based on internal coordinates (ICs: bond lengths, bond angles, torsion angles), which represent the natural connectivity of protein chemical structures. Compared to traditional coordinates, internal coordinates can reduce the redundancy in the distance matrices of protein structures and put emphasis on the local backbone structures. Due to ICs, multi-alignments can easily be produced by SABIC. These different alignments of two proteins can lead to the detection of repeat structures or various similarities between them. The algorithm performs better than DALI, CE and SSM on Fisher's and Novotny's benchmark datasets, respectively. Moreover, we defined a new similarity mQ-score, a variant of Q-score (Krissinel *et al.*, 2004) combining more alignment information, to assess the quality of alignments. Using mQ-score, SABIC performs much better than CE in detecting fold classifications and evolutionary relationships among protein domain structures.

2 Results

We propose a novel approach to detect the multi-alignments between two protein structures based on their internal coordinates. The effectiveness of our algorithm has been tested and compared with other methods. We firstly give the multi-alignment results found by our method for some protein pairs in literature. Then, we choose the relatively optimal one among these multi-alignments and compare it with those found by DALI, CE and SSM on Fischer's and Novotny's benchmark datasets. Finally, we show the ability of SABIC to detect fold classifications of protein domains.

2.1 Multi-alignments between pairs of protein structures

Due to the hardness of the structure alignment problem, no existing methods can give an optimum solution within an acceptable time (Godzik, 1996; Holm and Sander, 1996; Jung and Lee, 2000; Kolodny and Linial, 2004). Existing algorithms often solely give one a self-defined optimal alignment. However, it is doubted that this alignment makes sense in biology (Feng and Sippl, 1996; Oldfield, 2007). Therefore, SABIC gives multi-alignments, which possibly give more information about the similarities between two protein structures and provide various alternatives for alignment. Specifically, we can recognize flexible alignments and structural repeats between proteins from these multi-alignments.

Many proteins have multiple domains (Murzin *et al.*, 1995). For example, 1EHD:A has two domains (i.e., d1ehda1 and d1ehda2) and 1ULK:A has three domains (i.e., d1ulka1, d1ulka2 and d1ulka3). All the five domains are in the same domain family (g.3.1.1) in SCOP (Murzin *et al.*, 1995). There were 9 alignments produced by SABIC (Table 1). We found that three of them were redundant. Specifically, alignments 5, 7 were the same, 6 and 9 were also the same. We regard two alignments are the same one if the atom-pairs are identical and the positions of atom-pairs in proteins are the same. Alignment 8 was similar to alignment 4 because both of them were aligned based on the domain pair of d1ehda2 and d1ulka2, and they overlapped nearly as one. Hence, there were total six multi-alignments left between the two proteins. The results are shown in Fig. 1, in which we find that each alignment is based on a pair of domains. For this protein pair, DaliLite (Holm and Park, 2000) gives exactly 6 alignments, which are nearly the same to those found by our algorithm. The results provide evidence for the effectiveness of our method.

The loops connecting different domains in proteins are usually flexible (Shatsky *et al.*, 2002). Chen and Crippen (2005) gave such an example, where protein 8FAB:A contained two domains d8faba1 and d8faba2, and protein 1DCL:B also contained two domains d1dclb1 and d1dclb2. The domains d8faba1 and d1dclb1 belonged to the same SCOP family (b.1.1.1), while the domains d8faba2 and d1dclb2 belonged to family (b.1.1.2). As rigid bodies, these two proteins were aligned based on the larger domain pair, i.e., d8faba1 and d1dclb1. Using flexible alignment, they could be aligned very well at two domain pairs only if a twist was allowed. For the multi-alignments produced by SABIC, the alignments based on both domains were yielded (Fig. 2). Fig. 2(a) shows the alignment based on the domains d8faba2 and d1dclb2, and Fig. 2(b) shows the alignment based on the domains d8faba1 and d1dclb1. From the two alignments, it is easy to see that if a twist is allowed between domains d1dclb1 and d1dclb2, then a good alignment is generated between 8FAB:A and 1DCL:B which can cover both domain pairs (Fig. 2(c)). From this viewpoint, multi-alignments in SABIC do propose an alternative to the end of flexible alignment. In contrast, DaliLite produced rather bad alignments for this pair of proteins (Supplementary materials).

Moreover, our algorithm can detect the structural repeats between proteins. To align 1ABW:A and 1ABW:B, SABIC gave 30 ranked multi-alignments which are listed in Table 2. Among these alignments,

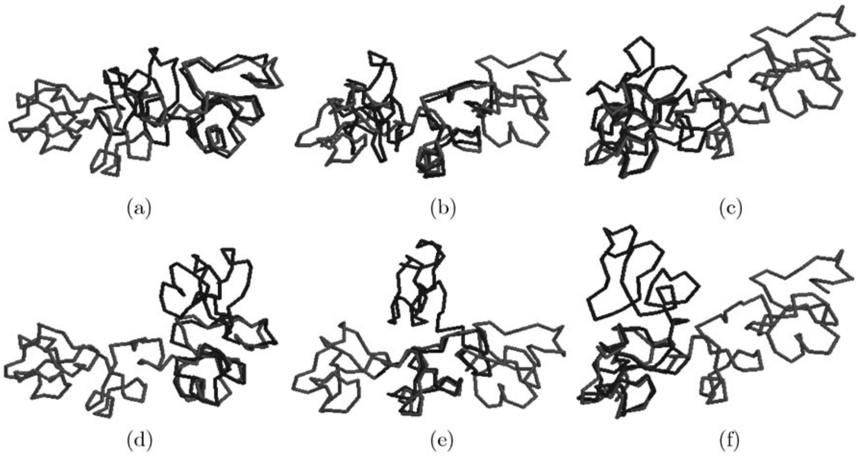


Fig. 1 Six alignments between proteins 1EHD:A and 1ULK:A produced by SABIC. (a) is aligned based on the domain pair d1ehda1 and d1ulka1, (b) is based on the pair d1ehda1 and d1ulka2, (c) is based on the pair d1ehda1 and d1ulka3, (d) is based on the pair d1ehda2 and d1ulka1, (e) is based on the pair d1ehda2 and d1ulka2 and (f) is based on the pair d1ehda2 and d1ulka3

Table 1 Multi-alignments of 1EHD:A(88) and 1ULK:A(126) produced by SABIC

Id	AlignedLength	RMSD	Gaps	SequenceIdentity	Q-score	mQ-score	P-value
1	56	2.3270	9	20	0.1770	0.1990	4.70E-03
2	59	2.6540	11	22	0.1760	0.2101	3.96E-03
3	42	1.4660	1	16	0.1280	0.1380	1.44E-02
4	45	2.0500	4	17	0.1240	0.1399	1.39E-02
5	42	2.0490	3	14	0.1080	0.1202	2.16E-02
6	46	1.9120	3	18	0.1360	0.1524	1.07E-02
7	42	2.0490	3	14	0.1080	0.1202	2.16E-02
8	44	1.9320	3	17	0.1230	0.1381	1.44E-02
9	46	1.9120	3	18	0.1360	0.1524	1.07E-02

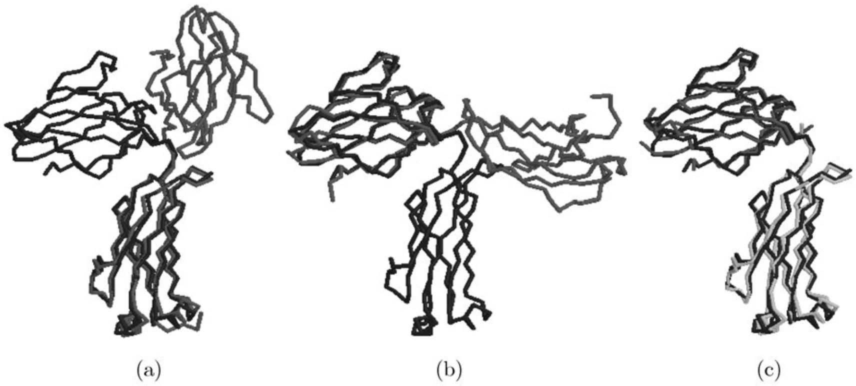


Fig. 2 The alignments between protein 8FAB:A and 1DCL:B produced by SABIC. (a) shows the alignment between 8FAB:A and 1DCL:B, which is based on domain pair d8faba2 and d1dclb2. (b) shows the alignment based on domain pair d8faba1 and d1dclb1. (c) shows the alignment if a twist is allowed in 1DCL:B, where the black part is the domain d1dclb1 and the gray is d1dclb2

alignments 1, 4 and 21 were the same, and alignments 2, 3 and 17 were also the same. But alignments 1 and 2 were different because the positions of the atom-pairs were different in the two alignments. This is illustrated in Fig. 3. The difference between them is also exhibited by their corresponding RMSDs. The sequence alignments based on structure alignments 1 and 2 are presented in the Supplementary materials. The two subfigures clearly indicate that 1ABW:A contains two domains, which are from residue 1 to 141 and from 143

Table 2 Multi-alignments of 1ABW:A(283) and 1ABW:B(146) produced by SABIC

Id	AlignedLength	RMSD	Gaps	SequenceIdentity	Q-score	mQ-score	P-value
1	139	1.2668	4	61	0.3969	0.4217	3.98E-04
2	139	1.4207	5	59	0.3820	0.4107	4.35E-04
3	139	1.4207	5	59	0.3820	0.4107	4.35E-04
4	139	1.2668	4	61	0.3969	0.4217	3.98E-04
5	39	3.2717	6	2	0.0168	0.0157	8.01E-01
6	48	2.9351	10	7	0.0285	0.0283	4.90E-01
7	51	3.0092	13	9	0.0314	0.0317	4.21E-01
8	39	3.3027	6	2	0.0166	0.0155	8.05E-01
9	47	2.8518	11	9	0.0281	0.0286	4.82E-01
10	34	2.8308	8	1	0.0148	0.0125	8.78E-01
11	49	2.9646	12	8	0.0294	0.0292	4.69E-01
12	20	1.2280	1	2	0.0083	0.0081	9.55E-01
13	34	3.0229	10	6	0.0139	0.0138	8.48E-01
14	21	1.5669	1	2	0.0084	0.0083	9.53E-01
15	30	3.1269	5	2	0.0104	0.0098	9.30E-01
16	30	3.1474	5	2	0.0104	0.0098	9.30E-01
17	139	1.4207	5	59	0.3820	0.4107	4.35E-04
18	29	3.0142	5	2	0.0101	0.0094	9.36E-01
19	30	2.8135	7	2	0.0116	0.0102	9.22E-01
20	46	3.1798	19	5	0.0241	0.0208	6.69E-01
21	139	1.2668	4	61	0.3968	0.4216	3.98E-04
22	29	2.9786	6	2	0.0102	0.0093	9.38E-01
23	30	2.8089	7	2	0.0116	0.0102	9.22E-01
24	64	3.6180	26	7	0.0404	0.0364	3.43E-01
25	32	3.1035	5	2	0.0120	0.0113	9.02E-01
26	42	3.0786	14	3	0.0208	0.0175	7.54E-01
27	50	3.3111	16	6	0.0273	0.0255	5.52E-01
28	32	3.1261	6	2	0.0119	0.0109	9.09E-01
29	57	3.9224	27	5	0.0290	0.0249	5.67E-01
30	48	3.1516	16	6	0.0265	0.0244	5.79E-01

to 283, respectively. The specific structure features underlying the protein complex have been identified. This result is highly consistent with the classification of the protein in SCOP (Murzin *et al.*, 1995).

The multi-alignments of protein pairs can also provide valuable information of related structures. Various alignments are generated easily from the novel concept of the difference matrix in SABIC. Representation of structure difference from the internal coordinates has the advantage of detecting the multi-alignments in proteins pairs (see Methods and discussions). From the above three instances, we demonstrate that SABIC can successfully detect flexible alignments, structural repeats or multiple domains between proteins by producing multi-alignments.

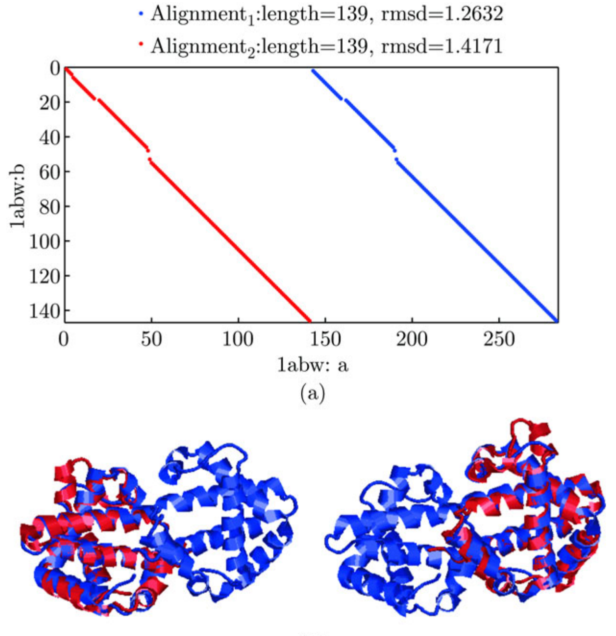
2.2 Results on benchmark datasets

To show the effectiveness of SABIC in aligning large-scale proteins, we implemented the tests in several well-known benchmark datasets. We simply regarded the longest alignment in the given multi-alignments as the optimal one. If there were several such alignments, we would choose the one with relatively smaller RMSD. Generally, an alignment was said to be "better" than the other if it had both longer length of aligned residues and lower RMSD, and to be "worse" if it had shorter length and higher RMSD. Otherwise, the two alignments were said to be "level". In this part, we compared SABIC with three popular methods, DALI, CE and SSM, on the ten difficult pairs (Chen and Crippen, 2005), Fischer's (Fischer *et al.*, 1996) and Novotny's (Novotny *et al.*, 2004) benchmark datasets. Bhattacharya *et al.* (2007) has recently proposed an alignment method by sequence nbhd and structure nbhd. They have also compared the alignment results with DALI, CE and SSM. Thus we also presented the results by their algorithm.

Table 3 gives the alignment results on the ten difficult standard protein pairs, where the comparisons between

Table 3 Comparisons between SABIC and SSM on 10 difficult pairs of proteins

Protein pairs	Length SABIC	Length SSM	RMSD SABIC	RMSD SSM	Q-score SABIC	Q-score SSM	mQ-score SABIC	P-value SABIC
1FXI:A-1UBQ	63	60	2.57	2.87	0.31	0.26	0.30	1.31E-03
1TEN-3HHR:B	87	73	1.83	2.1	0.32	0.21	0.32	9.83E-04
3HLA:B-2RHE	80	78	2.85	3.08	0.30	0.26	0.27	1.87E-03
2AZA:A-1PAZ	82	79	2.23	2.41	0.28	0.24	0.28	1.57E-03
1CEW:I-1MOL:A	82	79	2.25	2.12	0.42	0.41	0.44	3.33E-04
1CID-2RHE	96	89	2.42	2.33	0.28	0.25	0.27	1.72E-03
1CRL-1EDE	202	188	2.79	3.81	0.13	0.08	0.12	2.32E-02
2SIM-1NSB:A	287	271	2.84	2.86	0.29	0.26	0.28	1.53E-03
1BGE:B-2GMF:A	100	44	2.97	2.49	0.26	0.06	0.26	2.05E-03
1TIE-4FGF	115	114	2.67	2.85	0.36	0.33	0.34	8.25E-04



(a)

(b)

Fig. 3 Two alignments of 1ABW:A and 1ABW:B produced by SABIC. (a) shows dotted matrix representations of two alignments. (b) shows the two alignments in ribbon form. The left is alignment 1 and the right is alignment 2

SABIC and SSM are simultaneously shown. From the comparisons, we find that all the alignments given by SABIC are longer than those given by SSM and the corresponding RMSD measurements are smaller in 7 pairs. Furthermore, the Q-scores of our alignments are consistently larger than that of SSM. From above standards, SABIC is identified to be "better" alignments than SSM. In addition, the mQ-scores and P-values of these alignments produced by SABIC are listed. We find that all the ten alignments are statistically significant. The corresponding comparisons with CE and DALI also show the advantages of SABIC. The results can be found in the Supplementary materials.

The comparison results on other benchmarks between SABIC and DALI, CE and SSM are given in Tables 4, 5 and 6, respectively. In each table, we also list the corresponding results from Bhattacharya *et al.* (2007). On Fischer's dataset, except for two worse alignments, SABIC produced 31, 20 and 28 alignments better than that produced by DALI, CE and SSM, respectively. From the comparisons, SABIC performs much better than these known methods. Furthermore, the advancement of our method is greater than both sequence nbhd and structure nbhd when it is compared to DALI, CE and SSM in the benchmark. On Novotny's dataset, our method does not give any worse alignments compared to DALI. There are only two worse alignments in CATH (Orengo *et al.*, 1997) classes 1.10.40 and 2.30.110 compared to that of CE. Except for the two alignments, our method gives either "better" or "level" alignments compared to that of CE. Compared to SSM, only one worse alignment in class 2.100.10 is found by our method. For the total 153 tested protein pairs, SABIC performs better than the three other methods respectively. As for the methods of sequence nbhd and structure nbhd, our method also shows obvious advantages. Specifically, for the class 3.70.10, both sequence nbhd and structure nbhd cannot give any good alignments as compared to that of CE and SSM, while our method performs still well. These detailed results are listed in the Supplementary materials.

2.3 Testing the ability to classify protein structures

SCOP is often taken as a gold standard database for protein structure classification (Murzin *et al.*, 1995; Gerstein and Levitt, 1998). It consists of thousands of documented structure similarities based purely on visual inspection. We test our automatic method of structural comparison against the known similarities in SCOP. We randomly choose two data sets (Set A con-

Table 4 Comparison of results obtained by SABIC and DALI on Fischer's and Novotny's datasets. Bhattacharya's results (Bhattacharya *et al.*, 2007) are also listed for references

Date set	Our method Better/Worse/ Level	Sequence nbhd Better/Worse/ Level	Structure nbhd Better/Worse/ Level
Fischer's	31/2/35	4/4/60	5/2/61
Novotny's	44/0/109	27/6/120	19/0/134

Table 5 Comparison of results obtained by SABIC and CE on Fischer's and Novotny's datasets. Bhattacharya's results (Bhattacharya *et al.*, 2007) are also listed for references

Date set	Our method Better/Worse/ Level	Sequence nbhd Better/Worse/ Level	Structure nbhd Better/Worse/ Level
Fischer's	20/2/46	2/1/65	2/0/66
Novotny's	30/2/121	13/1/139	9/1/143

Table 6 Comparison of results obtained by SABIC and SSM on Fischer's and Novotny's datasets. Bhattacharya's results (Bhattacharya *et al.*, 2007) are also listed for references

Date set	Our method Better/Worse/ Level	Sequence nbhd Better/Worse/ Level	Structure nbhd Better/Worse/ Level
Fischer's	28/2/38	13/10/45	23/5/40
Novotny's	61/1/91	28/13/112	29/7/117

sists of proteins from all alpha class and all beta class, and set B consists of proteins from the alpha and beta class) from SCOP. The detailed lists of the domains in A and B are given in the Supplementary materials. SABIC is investigated to recognize the protein structure classification on the above two datasets.

ROC curves have been widely used to evaluate the performances of different structure alignment methods (Chen and Crippen, 2005; Kolodny *et al.*, 2005). We define the protein pairs as true positives when the identification is consistent with the classification in SCOP at homologous family level. ROC curves were generated by calculating specificity and sensitivity at different cutoffs to determine the classifications (Chen and Crippen, 2005). In ROC plot, specificity and sensitivity are defined as follows; $sensitivity = \frac{true\ positives}{true\ positives + false\ negatives}$ and $specificity = \frac{true\ positives}{true\ negatives + false\ positives}$. The results are shown in Fig. 4, where (a) and (b) show the ROC curves of SABIC and CE on sets A and B, respectively. On both datasets, the areas under the ROC curves of SABIC using Q-score or mQ-score are always larger than those of CE. From the statistical perspective, SABIC performs better than CE in classifying protein structures. Also, it shows the robustness and generalization of SABIC for the testing results on two different datasets. Moreover, as expected, mQ-score performs better than Q-score for SABIC. The reason is that Q-score does not take into account gaps and sequence identity while the mQ-score does.

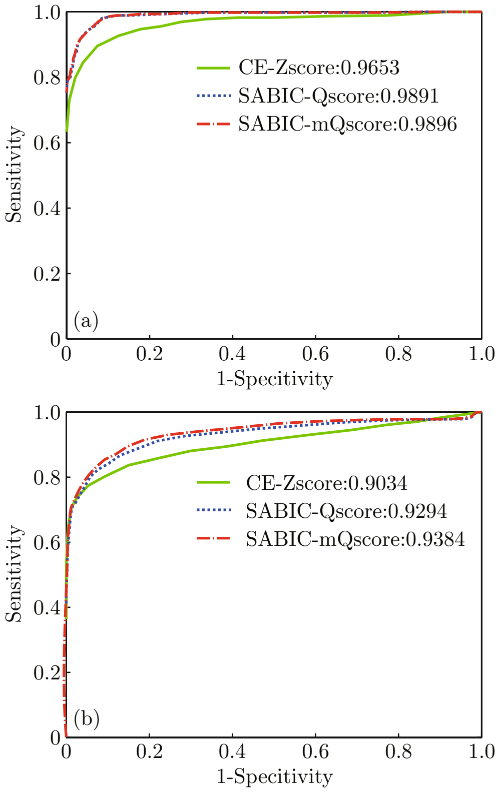


Fig. 4 Receiver operating characteristic (ROC) curves for SABIC and CE on sets A (a) and B (b). The value behind each method is the area under its ROC curve. The average time spent by CE in comparing 19306 pairs in set B is 17.4s, while the average time by our method is 42.3s. From this viewpoint, the time of our method is nearly 2.5 times that of CE on average due to the multi-alignments

3 Discussions and conclusion

In this work, a novel structure alignment method based on the internal coordinates of protein structures has been introduced. SABIC provides a novel difference matrix, a new alignment score and its statistical significance between protein structures. We have shown that SABIC performs more competitively than other popular methods.

3.1 New representation of structural difference

In SABIC, we used the internal coordinates to represent the proteins to be compared. Internal coordinates, invariant under rigid motions, can be easily transformed from the traditional Cartesian coordinates. The new geometric coordinate system provides an alternative representation of protein structure with concerning local structure patterns of the primary chain and peptide plane. It is a concise representation compared to the distance matrix representation, which is redundant in describing the relationship among the atoms (Miao *et al.*, 2008). Using internal coordinates, one structure can easily be represented by vectors of ICs (see Methods) and the structural difference between two proteins can be described by a difference matrix. The difference matrix is a novel representation of the difference between two structures, which is very different from the distance matrix for representing a protein structure. From the difference matrix, the perfectly aligned fragments can be found easily by filtering out those impossible equivalents (i.e. the equivalents with large distance or in a short continuous fragment). However, for the traditional distance matrix representing a protein structure, each item in the matrix implies the distance between atoms in the same protein. If we want to find the similar sub-matrices between the two distance matrices, we need to move a distance matrix to the top of the other and then compare them (Holm and Sander, 1993). Difference matrix is in the spirit of structure difference rather than structure representation. It is a technique to detect the structure difference with a matrix formulation in the novel coordinate system. Hence, our new representation of structure and the novel difference matrix exclude much computational redundancy.

We also defined a new normalization score for assessing the structure alignment with considering the gaps and sequence identities. mQ-score contains much more valuable information about the alignment and provides more details for the compared proteins. The statistical significance P-value measurement allows to compare protein structures in large-scale datasets.

3.2 Similarities between pairs of protein structures

So far, many scores have been proposed for assessing the similarity between a pair of protein structures. These scores are different from each other in balancing the values of alignment length, RMSD, gaps and sequence identities, etc. Although there is not an unified standard for measuring the structure similarity among proteins (Sippl, 2008), there is consensus that long aligned length, low RMSD, few gaps and many sequence identities indicate good alignments (Eidhammer *et al.*, 2000). Fig. 5 shows the alignment performance on large datasets. The RMSDs of the 19306 alignments produced by SABIC and CE are presented in (a) and (b), respectively. We can find that the RMSD distribution produced by SABIC algorithm follows an EVD model. It has an upper bound 4 whereas RMSD produced by CE follows a normal distribution, and has a much higher upper bound. For most of the pairs, SABIC produces lower RMSD than CE. Fig. 5(c) also shows that most alignments found by CE contain more gaps than those found by SABIC. This indicates the good quality of alignments found by SABIC. mQ-score, a novel similarity score taking alignment length, RMSD, gaps and sequence identities into account in an integrative manner, performs better than Q-score on the testing datasets.

3.3 Seed set

Our algorithm performs an extension alignment in the internal coordinate system. Difference matrix can be partitioned into several sub-matrices, which represent the core regions of the proteins to be aligned. Multiple seeds would produce multi-alignments. The difference matrix triggers and powers the generation of multiple candidate alignments. Moreover, the seeds during

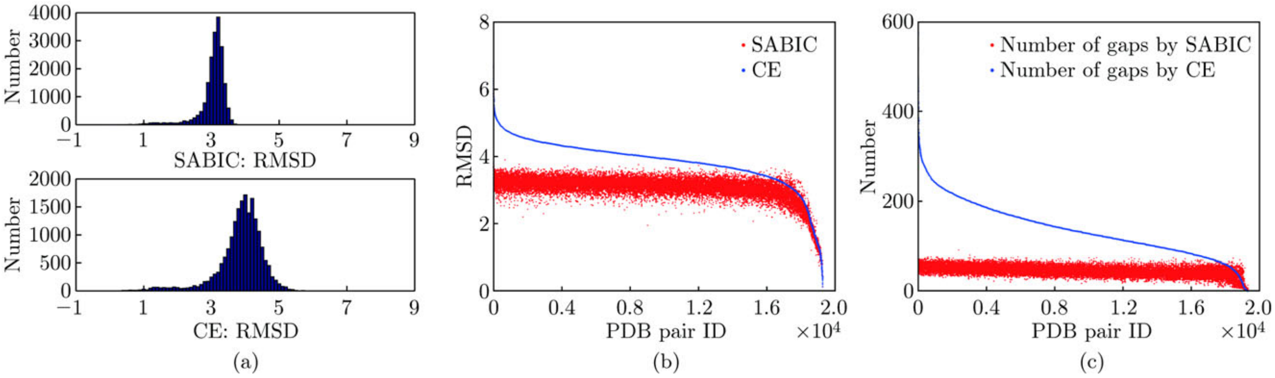


Fig. 5 Comparison between SABIC and CE. (a) shows the histograms of RMSDs of the 19306 alignments given by SABIC and CE. (b) shows the RMSDs produced by SABIC (thick) and CE (thin), where the pairs are sorted by a descending order of RMSDs produced by CE. (c) shows the gaps produced by SABIC (thick) and CE (thin), where the pairs are sorted by a descending order of the number of gaps produced by CE

the alignment provide valuable structural consistency of protein local structures. The attributes can be developed for detecting recurring structural motifs (Chen *et al.*, 2006). SABIC produces a final alignment by first superposing two structures based on a seed and then iteratively tunes the superposition until the final one is achieved. The procedures are similar to CAALIGN (Oldfield, 2007), however, the length of seed of our algorithm is self-adaptive and it is not limited by secondary structures or open frame windows with fixed length. As an extreme case, if two structures are the same, there would be only one seed, which is rightly the structure itself. Such techniques can reduce the number of possible seeds dramatically. The new coordinate representation and the scheme of alignment adjustment can easily detect the multi-alignments between the proteins to be compared, which provide more information for detailed analysis of structural relationships.

3.4 Increasing the efficiency of SABIC

SABIC produces ranked multi-alignments as the output, which redounds to comprehensively recognize various similarities between pair of proteins. The internal coordinates and the features of extending the alignment provide a scheme to generate multi-alignments. These various alignments will provide valuable information for detecting structure repeats or other interesting features. Certainly, it is more computationally expensive than those methods producing only one alignment. Due to the independence in producing an alignment from each seed, this process can be implemented in a parallel manner. Hence, parallelization of the algorithm can increase the efficiency significantly. Another way of increasing the efficiency is to limit the number of output alignments, which is to produce the alignments by selecting the first several seeds and omitting the left ones. Since it is heuristic, some accuracy will be sacrificed. In this work, the seeds are sorted by a descending order of seed length. It is not always the case that the longest seed would produce the best alignment. If we sort the seeds in such a proper order that the best alignment can always be provided from the first or several top seeds, the efficiency of our algorithm would also be increased. It is believed that the sequence information of the seeds would be helpful in such a sorting. These are our research directions in the future.

3.5 Conclusion

We proposed a novel strategy for structure alignment based on the differences of protein internal coordinates. Compared to other popular methods, the concise representation and comprehensive measurement of structure similarity provide more sensitive and detailed protein alignments. Multi-alignments as the output of the algorithm can be used for detecting various similarities between protein structures, such as flexible alignments and structural repeats.

4 Methods

4.1 Internal Coordinates

In the protein structure alignment, the structure is usually represented by the Cartesian coordinate of the C_α atom of each residue along the backbone. Protein A with m residues is described as $A = \{a_1, a_2, \cdots, a_m\}$, where $a_i \in R^3$, $1 \leq \mathrm{i} \leq \mathrm{m}$. Each point a_i represents the position of atom C_α in the ith residue. Similarly, protein B with n residues is represented as $B = \{b_1, b_2, \cdots, b_n\}$, where $b_j \in R^3$, $1 \leq \mathrm{j} \leq \mathrm{n}$. Alternatively, a protein structure can be described by its internal coordinates, i.e., bond lengths, bond angles and torsion angles. Protein A can be represented as $\mathrm{A} = \{\hat{\mathrm{a}}_1, \hat{\mathrm{a}}_2, \cdots, \hat{\mathrm{a}}_{\mathrm{m}-1}\}$, where $\hat{\mathrm{a}}_\mathrm{i} = [bond_{i,i+1}, angle_i, torsion_{i-1,i}]$ (Fig. 6(a)). The $bond_{i,i+1}$ is the bond length between atoms a_i and a_{i+1}, the $angle_i$ is the bond angle from the vectors $\overrightarrow{\mathrm{a_{i-1}a_i}}$ to $\overrightarrow{\mathrm{a_i a_{i+1}}}$, and the $torsion_{i-1,i}$ is the dihedral angel in the clockwise direction along vector $\overrightarrow{\mathrm{a_{i-1}a_i}}$ from the planes $a_{i-2}a_{i-1}a_i$ to $a_{i-1}a_i a_{i+1}$. For completeness, we specified that $\hat{\mathrm{a}}_1 = [bond_{1,2}, 0, 0]$ and $\hat{\mathrm{a}}_2 = [bond_{2,3}, angle_2, 0]$. Similarly, protein B can also be represented as $\mathrm{B} = \{\hat{\mathrm{b}}_1, \hat{\mathrm{b}}_2, \cdots, \hat{\mathrm{b}}_{\mathrm{n}-1}\}$. Internal coordinate representation has a merit of being invariant under rigid motion. Moreover, the representation puts emphasis on the local structure relationship among the atoms. We introduced a difference matrix $\mathrm{D}_{(\mathrm{m}-1)\times(\mathrm{n}-1)}$ to calculate the structure difference between two proteins from $\mathrm{A} = \{\hat{\mathrm{a}}_1, \hat{\mathrm{a}}_2, \cdots, \hat{\mathrm{a}}_{\mathrm{m}-1}$ and $\mathrm{B} = \{\hat{\mathrm{b}}_1, \hat{\mathrm{b}}_2, \cdots, \hat{\mathrm{b}}_{\mathrm{n}-1}$. Specifically, $\mathrm{D(i,j)} = \sqrt{(\mathrm{x}_1 - \mathrm{y}_1)^2 + (\mathrm{x}_2 - \mathrm{y}_2)^2 + (\mathrm{x}_3 - \mathrm{y}_3)^2}$, where $\mathrm{x}_1 = \hat{\mathrm{a}}_{\mathrm{i},1} \times \cos(\hat{\mathrm{a}}_{\mathrm{i},2})$, $\mathrm{x}_2 = \hat{\mathrm{a}}_{\mathrm{i},1} \times \sin(\hat{\mathrm{a}}_{\mathrm{i},2}) \times \sin(\hat{\mathrm{a}}_{\mathrm{i},3})$, $\mathrm{x}_3 = \hat{\mathrm{a}}_{\mathrm{i},1} \times \sin(\hat{\mathrm{a}}_{\mathrm{i},2}) \times \cos(\hat{\mathrm{a}}_{\mathrm{i},3})$ and $\mathrm{y}_1 = \hat{\mathrm{b}}_{\mathrm{j},1} \times \cos(\hat{\mathrm{b}}_{\mathrm{j},2})$, $\mathrm{y}_2 = \hat{\mathrm{b}}_{\mathrm{j},1} \times \sin(\hat{\mathrm{b}}_{j,2}) \times \sin(\hat{\mathrm{b}}_{\mathrm{j},3})$, $\mathrm{y}_3 = \hat{\mathrm{b}}_{\mathrm{j},1} \times \sin(\hat{\mathrm{b}}_{\mathrm{j},2}) \times \cos(\hat{\mathrm{b}}_{\mathrm{j},3})$ (Fig. 6(b)). This matrix is more precise and intuitive than that proposed by Ye *et al.* (2004). The difference matrix represents the structure relationship between proteins. The multi-alignments can then be naturally produced in the coordinate system.

4.2 SABIC algorithm

It is believed that an optimal structure alignment of two proteins contains a sub-alignment that is aligned almost perfectly (Bhattacharya *et al.*, 2007). We defined a perfectly aligned fragment as the alignment with a certain length and no gaps. We started with identification of all the perfectly aligned fragments (named seeds), which possibly contribute to the final optimal alignment. Then we superposed the substructures based on each seed to deduce a new alignment. After several iterations, a final alignment was found. Such kinds of operations were also used by many other algorithms, such as STRUCTAL (Gerstein and Levitt, 1998), but the details were different from one to another. Obviously, the final alignments might be differ-

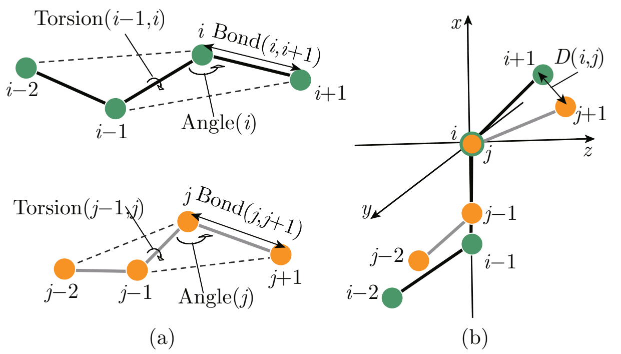


Fig. 6 Internal coordinates and the distance between them. (a) shows the internal coordinates (bond length, bond angle and torsion angle) of two atoms i and j. (b) The distance between the internal coordinates of atom i and j is defined by the Euclidean distance between atoms i+1 and j+1 under the conditions that atoms i and j are located on the origin, atoms i-1 and j-1 are on the negative axis of X, and both atoms i-2 and j-2 are on the X-Y plane with positive Y

ent upon different seeds. All the alignments yielded by SABIC provided candidates for the final optimal alignments and sources for analyzing the similarities between proteins in detail. Fig. 7 shows the flow chart of the SABIC algorithm.

In our method, we found the seeds in the defined difference matrix $D_{(m-1)\times(n-1)}$. Given two thresholds, the difference matrix D could be split into several sub-matrices. The first threshold θ_1 was to rule out those pairs (i, j) with $D(i, j) > \theta_1$, which meant that we ignored the pairs with weak similarity. Because large $D(i, j)$ indicates atoms i and j cannot be an equivalent in a perfectly aligned fragment, i.e. a seed. The second threshold θ_2 ruled out those pairs in continuous fragments with lengths less than θ_2, which meant we considered only the long fragments since short fragments with small distances were not significant in statistics. After the filtering, for each row or column, if there was no pair left, matrix D would be split at that row or column. Thus, we got several sub-matrices of D. In each sub-matrix, the longest fragments with no gaps were chosen as the seeds, from which the final alignments were generated. In the paper, we set $\theta_1 = 0.2$ and $\theta_2 = 5$ individually. Fig. 8 shows the difference matrix of proteins 1EHD:A and 1ULK:A before and after the filtering, respectively. In this case, the difference ma-

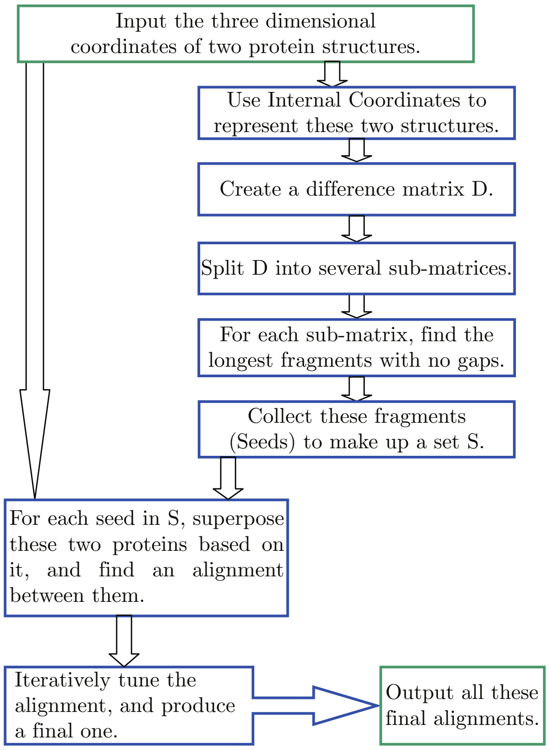

Input the three dimensional coordinates of two protein structures.
Use Internal Coordinates to represent these two structures.
Create a difference matrix D.
Split D into several sub-matrices.
For each sub-matrix, find the longest fragments with no gaps.
Collect these fragments (Seeds) to make up a set S.
For each seed in S, superpose these two proteins based on it, and find an alignment between them.
Iteratively tune the alignment, and produce a final one.
Output all these final alignments.

Fig. 7 An overview of SABIC

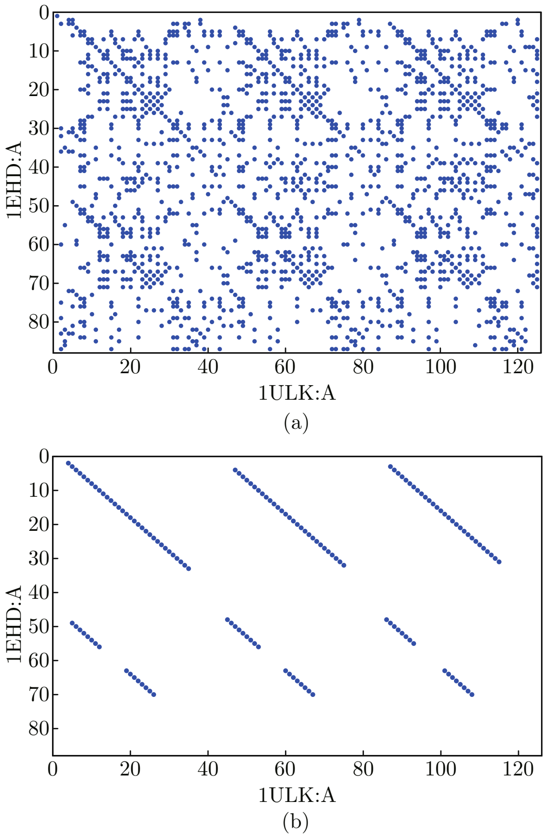


Fig. 8 The difference matrix and the seeds. (a) shows the difference matrix between protein 1EHD:A and 1ULK:A in internal coordinates representation. The dot indicates that the corresponding distance is less than 0.2, and the blank means that the distance is larger than 0.2. (b) shows the difference matrix after filtering out the fragments with length less than 5

trix was divided into 9 sub-matrices. There was only one continuous fragment in each sub-matrix, but it was not always like this case. When there was more than one continuous fragment in a sub-matrix, we chose the longest ones as seeds in the corresponding sub-matrix.

All seeds from every sub-matrix formed a set S. For every seed in S, we superposed the structures of protein A and B based on it. In the processing of superposition, we fixed protein A and then rotated and translated protein B. After the superposition, another threshold θ_3 was needed to produce a new structure alignment. Suppose $A = \{a_1, a_2, \cdots, a_m\}$ and $B = \{b_1, b_2, \cdots, b_n\}$ weres the coordinates after the superposition. Then a new distance matrix $\mathrm{M}_{\mathrm{m}\times\mathrm{n}}$ could be calculated by $\mathrm{M(i,j)} = \sqrt{(\mathrm{a_i(1)} - \mathrm{b_j(1)})^2 + (\mathrm{a_i(2)} - \mathrm{b_j(2)})^2 + (\mathrm{a_i(3)} - \mathrm{b_j(3)})^2}$. Using a dynamic programming similar to the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), we computed a minimum-score global alignment of A and B (Ye *et al.*, 2004). The recursion formula is as follow:

$$v(i,j) = \min\{v(i,j-1) + \theta_3, \quad v(i-1,j) + \theta_3, \\ v(i-1,j-1) + M(i,j)\}.$$

Specifically, $v(1,1) = 0$, $v(i,1) = v(i-1,1) + \theta_3$, $v(1,j) = v(1,j-1) + \theta_3$, $i = 2, \cdots, m$, $j = 2, \cdots, n$. In this paper, we set $\theta_3 = 3.5$.

Therefore, a new alignment was produced. Based on this alignment, another new alignment was produced by the above superpose-and-realign process. Such iterations stopped either when the number of iterations exceeded a specified limit or when the motion of protein B was tiny. In this paper, we set the iteration limit of 20. And we constrained the motion of B by $\sqrt{\sum_{\mathrm{i}=1,\cdots,\mathrm{n}} \left\|\mathrm{b_i^k} - \mathrm{b_i^{k-1}}\right\|^2} \leq 0.05$ Å, where k was the number of iterations, $\mathrm{b_i^k}$ was the coordinate of the ith C_α of protein B after k iterations.

Then we got a final alignment from one seed in set S. From different starting seeds in S, we could get different final alignments. Some of the final alignments might be similar or even the same because the optimal alignment often contained several perfectly aligned fragments, which formed different seeds. Obviously, such alignments were optimal or near optimal. Those with significant differences provide some candidates to find alignments that make sense in biology. Sometimes, these results give us more information to learn about the compared structures.

To make a comparison between SABIC and the other existing method, we chose the longest alignment as the object in all the generated multi-alignments. If there were a number of such alignments, we would choose the one with the smallest RMSD. In SABIC, all thresholds have been optimized based on various tests.

4.3 mQ-score for normalization and statistical significance

There are many scores, such as Z-score (Shindyalov and Bourne, 1998), Q-score (Krissinel and Henrick, 2004), and TM-score (Zhang and Skolnick, 2005), which have been proposed to describe quality of an alignment between two proteins. However, none of them is a gold standard. In this work, we designed a modified Q-score (mQ-score) to assess the quality of alignment based on Q-score by incorporating gaps and sequence identity.

$$\mathrm{mQ} = \frac{\mathrm{N_{align}Q}}{\mathrm{N_{align}} + \mathrm{N_{gap}}}(1 - \mathrm{seqIden}) + \frac{\mathrm{N_{align}^2}}{\mathrm{N_1 N_2}}\mathrm{seqIden},$$

where $\mathrm{N_{align}}$ is the length of the alignment, $\mathrm{N_{gap}}$ is the number of gaps, $\mathrm{N_1}$ and $\mathrm{N_2}$ are the lengths of the two proteins, seqIden is the ratio of sequence identity, and Q is Q-score (Krissinel and Henrick, 2004). In this score, we penalized the gaps and encouraged sequence identities. If an alignment has no gaps and zero RMSD, the sequence identity does not affect the final score anymore, which indicates that mQ-score concerns structural similarity comprehensively.

The statistical significance is different from the similarity score. The former focuses on the characteristics of a particular alignment in the population of alignments between any pair of proteins, while the latter just considers the quality of an alignment individually (Karlin and Altschul, 2002; Zhu and Weng, 2005). We considered the statistical significance of the alignments produced by SABIC. It depent on the mQ-score of a particular alignment and the total distribution of mQ-score. We randomly chose 200,000 pairs of proteins from the 40% ID filtered subset of Astral 1.73 (Brenner *et al.*, 2000). The distribution of mQ-scores by SABIC is shown in Fig. 9. Obviously, it could be fitted to an EVD approximately with parameters $k = 0.2741$,

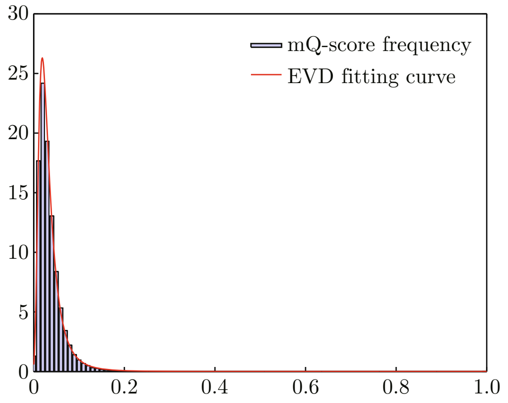


Fig. 9 The distribution of mQ-score obtained by aligning 200,000 pairs of domains randomly selected from the 40% ID filtered subset of Astral 1.73. The distribution can be fit to an Extreme Value Distribution, with parameters $k = 0.2741$, $\lambda = 68.9655$ and $\mu = 0.0222$

$\lambda = 68.9655$ and $\mu = 0.0222$. Hence, we could give the statistical significance such as P-value of an alignment. The formula is as follows:

$$P(s > x) = 1 - exp(-(1 + k\lambda(x - \mu))^{-1/k}),$$

where x is the mQ-score of an alignment, $P(s > x)$ is the corresponding P-value.

Abbreviations

PDB: protein data bank

SABIC: structure alignment base on internal coordinates

IC: internal coordinate

RMSD: root mean square deviation

ROC: receiver operating characteristic

Sequence nbhd: sequence neighborhood

Structure nbhd: structure neighborhood

mQ-score: modified Q-score

EVD: extreme value distribution

Electronic Supplementary Material

Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s12539-010-0019-8 and is accessible for authorized users.

Acknowledgments This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grants No. 10801131 and No. 10771206. Some of the authors are also supported by 973 Project under Grants No. 2004CB318000. We also appreciate the Center of Bioinformatics of Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences.

References

[1] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P. 2000. The protein data bank. Nucleic Acids Res 28, 235–242.

[2] Bhattacharya, S., Bhattacharyya, C., Chandra, N.R. 2007. Comparison of protein structures by growing neighborhood alignments. BMC Bioinformatics 8, 77.

[3] Brenner, S., Koehl, P., Levitt, M. 2000. The ASTRAL compendium for sequence and structure analysis. Nucleic Acids Res 28, 254–256.

[4] Chen, L., Wu, L., Wang, Y., Zhang, S., Zhang, X.S. 2006. Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. BMC Structural Biology 6, 18.

[5] Chen, Y., Crippen, G. 2005. A novel approach to structural alignment using realistic structural and environmental information. Protein Sci 14, 2935–2946.

[6] Chen, L., Zhou, T., Tang, Y. 2005. Protein structure alignment by deterministic annealing. Bioinformatics 21, 51–62.

[7] Chothia, C., Lesk, A. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J 5, 823–826.

[8] Eidhammer, I., Jonassen, I., Taylor, W. 2000. Structure comparison and structure patterns. J Comput Biol 7, 685–716.

[9] Feng, Z.K., Sippl, M. 1996. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1, 123–132.

[10] Fischer, D., Elofsson, A., Rice, D., Eisenberg, D. 1996. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pac Symp Biocomput 1996, 300–318.

[11] Gerstein, M., Levitt, M. 1998. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Sci 7, 445–456.

[12] Godzik, A. 1996. The structural alignment between two proteins: Is there a unique answer? Protein Sci 5, 1325–1338.

[13] Goldman, D., Papadimitriou, C., Istrail, S. 1999. Algorithmic aspects of protein structure similarity. In FOCS'99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science. Washington DC, USA. IEEE Computer Society 1999, 512–522.

[14] Holm, L., Park, J. 2000. DaliLite workbench for protein structure comparison. Bioinformatics 16, 566–567.

[15] Holm, L., Sander, C. 1993. Protein structure comparison by alignment of distance matrices. J Mol Biol 233, 123–138.

[16] Holm, L., Sander, C. 1996. Mapping the protein universe. Science 273, 595–602.

[17] Jung, J., Lee, B. 2000. Protein structure alignment using environmental profiles. Protein Eng 13, 535–543.

[18] Karlin, S., Altschul, S. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87, 2264–2268.

[19] Kinch, L., Grishin, N. 2002. Evolution of protein structures and functions. Curr Opin Struct Biol 12, 400–408.

[20] Koehl, P. 2001. Protein structure similarities. Curr Opin Struct Biol 11, 348–353.

[21] Kolodny, R., Koehl, P., Levitt, M. 2005. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 346, 1173–1188.

[22] Kolodny R, Linial, N. 2004. Approximate protein structure alignment in polynomial time. Proc Natl Acad Sci USA 101, 12201–12206.

[23] Krissinel, E., Henrick, K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst D 60, 2256–2268.

[24] Miao, X., Waddell, P., Valafar, H. 2008. TALI: Local alignment of protein structures using backbone torsion angles. J Bioinform Comput Biol 6, 163–181.

[25] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247, 536–540.

[26] Needleman, S., Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443–453.

[27] Novotny, M., Madsen, D., Kleywegt, G. 2004. Evaluation of protein fold comparison servers. Proteins 54, 260–270.

[28] Oldfield, T. 2007. CAALIGN: A program for pairwise and multiple protein-structure alignment. Acta Crystallogr D Biol Crystallogr 63, 514–525.

[29] Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., Thornton, J. 1997. CATH - A hierarchic classification of protein domain structures. Structure 5, 1093–1108.

[30] Shatsky, M., Nussinov, R., Wolfson, H. 2002. Flexible protein alignment and hinge detection. Proteins 48, 242–256.

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

[32] Sierk, M., Pearson, W. 2004. Sensitivity and selectivity in protein structure comparison. Protein Sci 13, 773–785.

[33] Sippl, M.J. 2008. On distance and similarity in fold space. Bioinformatics 24, 872–873.

[34] Stark, A., Sunyaev, S., Russell, R.B. 2003. A model for statistical significance of local similarity in structure. J Mol Biol 326, 1307–1316.

[35] Taylor, W., Orengo, C. 1989. Protein structure alignment. J Mol Biol 208, 1–22.

[36] Thornton, J., Todd, A., Milburn, D., Borkakoti, N., Orengo, C. 2000. From structure to function: Approaches and limitations. Nat Struct Biol 7, 991–994.

[37] Watson, J.D., Laskowski, R.A., Thornton, J.M. 2005. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15, 275–284.

[38] Whisstock, J., Lesk, A. 2003. Prediction of protein function from protein sequence and structure. Q Rev Biophys 36, 307–340.

[39] Ye, J., Janardan, R., Liu, S. 2004. Pairwise protein structure alignment based on an orientation-independent backbone representation. J Bioinform Comput Biol 2, 699–718.

[40] Yona, G., Kedem, K. 2005. The URMS-RMS hybrid algorithm for fast and sensitive local protein structure alignment. J Comput Bio 12, 12–32.

[41] Zhang, Y., Skolnick, J. 2005. TM-align: A protein structure alignment algorithm based on TM-score. Nucleic Acids Res 33, 2302–2309.

[42] Zhu, J., Weng, Z. 2005. FAST: A novel protein structure alignment algorithm. Proteins 58, 618–627.