BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 7 2014, pages 949–955
doi:10.1093/bioinformatics/btt659

Structural bioinformatics
Advance Access publication November 29, 2013

Fingerprinting protein structures effectively and efficiently

Xuefeng Cui[1], Shuai Cheng Li[2,*], Lin He[1] and Ming Li[1,*]

[1]David R. Cheriton School of Computer Science, University of Waterloo, Ontario, Canada and [2]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: One common task in structural biology is to assess the similarities and differences among protein structures. A variety of structure alignment algorithms and programs has been designed and implemented for this purpose. A major drawback with existing structure alignment programs is that they require a large amount of computational time, rendering them infeasible for pairwise alignments on large collections of structures. To overcome this drawback, a fragment alphabet learned from known structures has been introduced. The method, however, considers local similarity only, and therefore occasionally assigns high scores to structures that are similar only in local fragments.

Method: We propose a novel approach that eliminates false positives, through the comparison of both local and remote similarity, with little compromise in speed. Two kinds of contact libraries (ContactLib) are introduced to fingerprint protein structures effectively and efficiently. Each contact group of the contact library consists of one local or two remote fragments and is represented by a concise vector. These vectors are then indexed and used to calculate a new combined hit-rate score to identify similar protein structures effectively and efficiently.

Results: We tested our method on the high-quality protein structure subset of SCOP30 containing 3297 protein structures. For each protein structure of the subset, we retrieved its neighbor protein structures from the rest of the subset. The best area under the Receiver-Operating Characteristic curve, archived by ContactLib, is as high as 0.960. This is a significant improvement compared with 0.747, the best result achieved by FragBag. We also demonstrated that incorporating remote contact information is critical to consistently retrieve accurate neighbor protein structures for all-β query protein structures.

Availability and implementation: https://cs.uwaterloo.ca/∼xfcui/contactlib/.

Contact: shuaicli@cityu.edu.hk or mli@uwaterloo.ca

Received on April 6, 2013; revised on October 16, 2013; accepted on November 9, 2013

1 INTRODUCTION

Although assessing the similarities and differences between protein structures is a common practice in structural biology, efficiently performing this comparison is critical in some applications. For example, once a new protein structure is determined, researchers often need to infer its function or evolution by studying proteins of similar structures. A few databases, such as SCOP (Chandonia *et al.*, 2004; Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997), maintain hierarchical classifications of known protein structures. The need to obtain structures similar to the new protein from these databases motivates the *neighbor protein structure retrieval problem*: *given a query protein structure and a database of protein structures, retrieve all the structures in the database that are similar to the query structure.*

One intuitive solution to the neighbor protein structure retrieval problem is to align the query protein structure and every protein structure of the database using a pairwise protein structure alignment tool. One successful approach of pairwise protein structure alignment is to represent protein structures as 3D coordinates and to find the optimal residue mapping and orientation (rotation and translation) together, as stralign (Akutsu, 1996), Combinatorial Extension (CE) (Shindyalov and Bourne, 1998), Local-Global Alignment (LGA) (Zemla, 2003), secondary-structure matching (SSM) (Krissinel and Henrick, 2004), TMalign (Zhang and Skolnick, 2005) and SPalign (Yang *et al.*, 2012) proposed. An orientation-free approach is possible by encoding each structure as a 2D matrix of residue–residue interaction distances; comparison between two structures can be performed by an 'alignment' of their respective matrices as proposed in DALI (Holm and Sander, 1993). One drawback of adopting these approaches is inefficiency, especially when the protein structure database is large. Either using 3D coordinates or using distance matrices, solving the pairwise protein structure alignment problem is NP-hard (Kolodny and Linial, 2004). Thus, all the aforementioned pairwise protein structure alignment tools adopt heuristic approaches without global optimality guarantee. Unfortunately, such heuristic approaches are still time-consuming (Aung and Tan, 2007; Kolodny *et al.*, 2005).

The concern for efficiency has prompted the use of 1D protein structure profiles, which often perform well. In particular, the current state-of-the-art method called FragBag (Budowski-Tal *et al.*, 2010) has been shown experimentally to be fast and accurate on average. Specifically, FragBag represents a protein structure as a profile that contains counts of structure fragments in a fragment alphabet learned from known structures. Then, neighbor protein structures can be retrieved by comparing the profiles efficiently. One drawback, however, is that although FragBag is capable of delivering good average accuracy, its accuracy is sometimes significantly worse than the average accuracy, as shown in Section 3.1. This accuracy drop occurs when two structures are similar in many local fragments but differ significantly in their overall structure; because FragBag compares only local contacts, it fails to identify the large non-local discrepancy in these structures.

*To whom correspondence should be addressed.

In this article, we present ContactLib, a complete contact group library defined in Section 2, which is to be used as fingerprints of protein structures. FragBag (Budowski-Tal *et al.*, 2010) and local feature frequency profile (LFFP) (Choi *et al.*, 2004) are two promising tools that are closely related. Our ContactLib is different from FragBag and LFFP in following ways: (i) FragBag and LFFP are developed on general structure fragments, whereas ContactLib introduces both local and remote contact groups eliminating potentially weak contact groups; (ii) FragBag and LFFP use 3D coordinates or 2D distance matrices, whereas ContactLib introduces 1D distance vectors that can be efficiently indexed; (iii) FragBag and LFFP require a predefined word alphabet, whereas ContactLib avoids such word alphabet and introduces some freedom of specifying similarity thresholds at runtime; and (iv) FragBag and LFFP use word frequency profiles and distance functions from the text information retrieval problem, whereas ContactLib introduces a combined hit-rate scoring function for the neighbor protein structure retrieval problem. Because the word alphabet of LFFP is not publicly available, we focus on comparing ContactLib and FragBag in this article.

As an initial study, we built two ContactLibs: ContactLib-9L, which models local contacts, and ContactLib-3R, which models remote contacts. Using one or both of ContactLib-9L and ContactLib-3R, we tested our method on the high-quality protein structure subset of SCOP30 (Chandonia *et al.*, 2004; Murzin *et al.*, 1995) containing 3297 protein structures. For each protein structure, we retrieved its neighbor protein structures from the rest.

According to the Receiver-Operating Characteristic (ROC) curve analysis (Fawcett, 2006) in Section 3.1, the best area under the ROC curve (AUROC), archived by ContactLib, is as high as 0.960. This is a significant improvement compared with 0.747, the best result achieved by FragBag (Budowski-Tal *et al*., 2010). Specifically, when ContactLib-3R is used, 75% of the AUROCs is >0.936 and the lowest AUROC is 0.504. When ContactLib-9L is used, 75% of the AUROCs are >0.823 and 3% of the AUROCs are <0.5. However, when FragBag is used, 75% of the AUROCs are >0.657 and 10% of the AUROCs are <0.5. Therefore, the worst-case AUROC is significantly improved by using ContactLib, and ContactLib-3R is even able to guarantee an AUROC higher than a random method, which has an AUROC equals to 0.5.

We also demonstrated that incorporating remote contact information is critical to consistently retrieve accurate neighbor protein structures for all-β query protein structures, which is more challenging than that for all-α query protein structures, in Section 3.2. Moreover, we demonstrated that if two contact groups have similar structures with low root mean square deviation (RMSD) values, they tend to have similar distance vectors, which are used to index contact groups, in Section 3.3. Finally, we discussed several future extensions and applications of ContactLib in Section 4.

2 CONTACTLIB NEIGHBOR PROTEIN STRUCTURE RETRIEVAL

In this section, we first define a *contact group*. Then, we build a comprehensive library of contact groups as fingerprints of all existing protein structures and we call such a contact group library *ContactLib*. We also propose an indexing technique for ContactLib, which may be applied to neighbor contact group retrieval. Finally, we introduce a *combined hit-rate score* to retrieve neighbor protein structures.

A *contact group* in this article refers to a small collection of residues that may have a high density of contacts among the residues. As two residues in contact should not be far apart, we require that all residues are within a sphere. The position of each residue here is represented by its C_α atom. A *local contact group* models contact within a protein structure fragment and a *remote contact group* could involve two or more structure fragments. Owing to chemical and physical constraints within limited sphere space, it is rare for a contact group to contain a large number of fragments. For conciseness, we require a remote contact group to involve exactly two fragments. Hence, we define a contact group as follows:

DEFINITION 1 (contact group): *a contact group is a set of residues, represented by the respective C_α atoms, of either a single fragment with l_1 residues, called a local contact group, or a pair of fragments with l_2 residues, called a remote contact group, such that all the C_α atoms are located within a sphere of radius r.*

Here, we set $l_1 = 9$ and $l_2 = 3$ as we find that they are sufficient to accurately model a local and a remote contact group, respectively. The fragment length of nine has also been used and shown to be the optimal fragment length to model protein structure fragments (Maadooliat *et al.*, 2012; Simons *et al.*, 1997). Moreover, the radius of the sphere is set to be $r = 16$ Å, so that it is large enough to capture most contacts. Then, we define a *ContactLib* as follows:

DEFINITION 2 (ContactLib): *a ContactLib is a contact group library containing local and/or remote contact groups in all protein structures of the search protein structure database.*

We use the contact groups to fingerprint protein structures. To create an efficient and effective index of the ContactLib, we devise a strategy to represent a contact group by a low-dimensional vector. Before defining such a representation, we examine the number of dimensions or the degree of freedom of a contact group; that is, we want to know how many values are necessary to reconstruct a contact group.

We determine the dimension of a contact group as follows: a protein structure can be represented by bond angles, bond lengths and dihedral angles (Cui *et al.*, 2013; Rice and Brünger, 1994). The fixed bond length and bond angle structures have been widely used in modeling protein structures (Canutescu *et al.*, 2003; Güntert and Wüthrich, 1991; Li *et al.*, 2008; Simons *et al.*, 1997). The peptide dihedral angles (i.e the Ω angles) are also rounded to either 0 or π. Because <2% of the Ω dihedral angles have a value closer to 0, it is treated as a rare case (Engh and Huber, 2006). Hence, it is acceptable to use $\Omega = \pi$ as a good approximation, which results in the distance between two adjacent C_α atoms to be 3.8 Å. If we connect any two adjacent C_α atoms by such a pseudo bond, the number of dihedral angles in this pseudo molecule of a local contact group is $l_1 - 3$, and the number of bond angles in it is $l_1 - 2$. Thereafter, the dimension required to represent a local contact group is $2l_1 - 5 = 13$.

Similarly, the number of dihedral angles in the pseudo molecule of a remote contact group is $2l_2 - 3$, the number of bond angles in it is $2l_2 - 2$ and the number of bond lengths between non-adjacent C_α atoms in it is 1. Thereafter, the dimension required to represent a remote contact group is $(2l_2 - 3) + (2l_2 - 2) + 1 = 4l_2 - 4 = 8$. The number of dimensions is proportional to the number of residues in the contact group.

Given the desired number of dimensions, we create distance vectors to represent contact groups. Denote $D(a, b)$ as the distance between two points a and b. Given a local contact group of a single protein structure fragment $\{P_1, P_2, ..., P_{l_1}\}$, the distance vector is defined as

$$V_1 = \{D(P_i, P_{i+g}) | 1 \leq i, i+g \leq l_1, g = 2^k, k \geq 1\}$$

For a remote contact group of a protein structure fragment pair $\{P^1_1, P^1_2, ..., P^1_{l_2}, P^2_1, P^2_2, ..., P^2_{l_2}\}$, we define the distance vector as follows:

$$V_2 = \left\{ \begin{array}{l} D(P^1_i, P^1_{i+2}) | 1 \leq i, i+2 \leq l_2 \\ D(P^2_i, P^2_{i+2}) | 1 \leq i, i+2 \leq l_2 \\ D(P^1_i, P^2_{i+1}) | 1 \leq i < l_2 \\ D(P^2_i, P^1_{i+1}) | 1 \leq i < l_2 \\ D(P^1_{l_2}, P^2_1) \\ D(P^2_{l_2}, P^1_1) \end{array} \right\}$$

Here, V_1 and V_2 have 13 dimensions and 8 dimensions, respectively. In addition, our definition of V_1 and V_2 covers different types of distances (as shown in Fig. 1a and b). One critical feature of V_1 and V_2 is that if two contact groups have similar structures with low RMSD, they should have similar pairwise distances (Holm and Sander, 1993) and hence similar V_1 or V_2, as described in Section 3.3.

The number of similar contact groups shared by two proteins can be used as an indicator of their structure similarity. Here, we introduce an index to efficiently find all contact groups that are similar to a query contact group in ContactLib by using a 13-by-256 table of bit vectors for a local ContactLib and an 8-by-256 table of bit vectors for a remote ContactLib. Here, each row of the table represents a dimension of the distance vector. For each dimension of the distance vector, the value space is discretized into 256 bins, and each column represents a bin. Each element of the table is a bit vector indicating if a contact group belongs to the associated bin on the associated dimension for all contact groups of the ContactLib. Then, these tables can be effectively used to retrieve the set of contact groups in a particular bin along a given dimension. Contact groups in m consecutive bins along a particular dimension (or column) can be calculated by bitwise OR operations, and then contact groups in m consecutive bins along all dimensions (or rows) can be calculated by bitwise AND operations. Here, we carefully choose a parameter m, such that contact groups similar to the query contact group are within m bins from the query bins along each dimension.

To compare two structures, we introduce a *combined hit-rate score* to rank and select protein structures in the search database. We observed that, for a pair of similar protein structures, most of the contact groups for one structure tend to have similar contact groups from the other structure. Conversely, for a pair of dissimilar protein structures, the opposite scenario was observed.

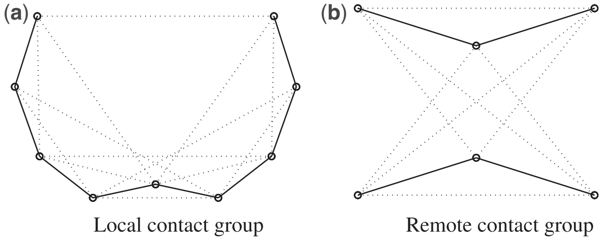


Fig. 1. Captured distances of local and remote contact groups: each circle represents a C_α atom, each solid line represents a pseudo bond between two adjacent C_α atoms (captured implicitly in our distance vector) and each dashed line represents a distance captured by our distance vector

These observations suggest a combined hit-rate score for a pair of protein structures, as the geometric mean of the similar contact group hit-rates of the two protein structures:

$$S = \sqrt{\frac{h_1}{n_1}\frac{h_2}{n_2}}$$

where h_1 is the number of hit contact groups for the first protein structure that have similar contact groups from the second protein structure, h_2 is the number of hit contact groups for the second protein structure that have similar contact groups from the first protein structure, n_1 is the number of contact groups for the first protein structure and n_2 is the number of contact groups for the second protein structure.

In summary, we find all pairs of neighbor contact groups between the query protein structure and the search database using our indexes of ContactLib, and then we calculate the combined hit-rate score to rank and select protein structures in the search database. Let p be the number of contact groups in a query, q be the number of contact groups in the database and N be the number of structures in the database. Recall that m is the number of consecutive bins that defines similarity on a dimension of the distance vector, and l is the dimension of the distance vector. For each query contact group, $O(m)$ bitwise OR operations and $O(l)$ bitwise AND operations are performed, and each bitwise OR or AND operation takes $O(q)$ time. Thus, the runtime complexity to find all similar contact group pairs between the query protein structure and the search database is $O(pq(m + l))$, and the combined hit-rate scores can be calculated simultaneously. Moreover, the runtime complexity to rank structures according to the combined hit-rate scores is $O(N \log N)$. Therefore, the running time for our neighbor protein structure retrieval method is $O(pqm + N \log N)$. Here, the indexes can be prebuilt, and the runtime complexity is not included.

3 RESULT

For performance analysis of our neighbor protein structure retrieval program, we used the high-quality protein structure subset of SCOP30 1.75B (Chandonia *et al.*, 2004; Murzin *et al.*, 1995) that has a minimum Summary PDB ASTRAL Check Index (SPACI) of 0.5. In this article, we simply refer this dataset as SCOP30. Then, we built the local contact group library, *ContactLib-9L*, and the remote contact group library, *ContactLib-3R*, of SCOP30. For each protein structure of

SCOP30, we retrieved its neighbor protein structures from the rest of SCOP30. For reference, there are 3297 protein structures in SCOP30, 375 299 local contact groups in ContactLib-9L and 6 309 469 remote contact groups in ContactLib-3R.

To find neighbor protein structures of each query protein structure, we used SCOP (Murzin *et al.*, 1995) and the best alignment found by six popular protein structure alignment tools: DALI (Holm and Sander, 1993), CE (Shindyalov and Bourne, 1998), LGA (Zemla, 2003), SSM (Krissinel and Henrick, 2004), TMalign (Zhang and Skolnick, 2005) and SPalign (Yang *et al.*, 2012). Specifically, we considered two protein structures as neighbors if and only if both protein structures are from the same SCOP superfamily and the best pairwise structure alignment has a structure alignment score (SAS) (Kolodny *et al.*, 2005) below 2.0 Å. Such neighbor protein structures tend to have globally similar structures and functional features, but are not necessary to have similar sequences. Because different SCOP levels and SAS thresholds conduct similar conclusions, we focus on the aforementioned neighbor protein structure definition in this article. For the best alignments with SAS below 2.0 Å, 50% are contributed by SPalign, 31% are contributed by LGA and 16% are contributed by SSM.

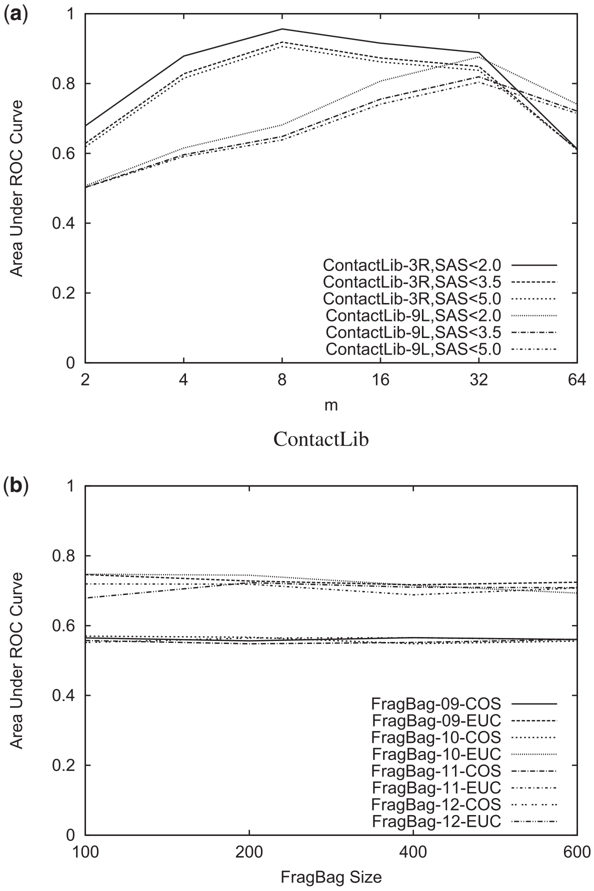
The accuracy of neighbor protein structure retrieval is evaluated by the AUROC, which has been used in many research areas (Fawcett, 2006), including the protein structure alignment area (Budowski-Tal *et al.*, 2010; Kolodny *et al.*, 2005). For instance, an AUROC of 0.9 means that a neighbor protein structure should be scored higher than a non-neighbor protein structure with a probability of 0.9, and a random method has an AUROC equals to 0.5. When the query protein structure does not have any neighbor protein structures in SCOP30, the AUROC is not defined. Thus, such cases are eliminated in our analysis.

3.1 General ROC curve analysis

In this experiment, we demonstrate that ContactLib significantly outperforms FragBag for the neighbor protein structure retrieval problem in terms of AUROC. For ContactLib-9L and ContactLib-3R, we tested $m \in \{2, 4, 8, 16, 32, 64\}$ (recall that m is the number of neighboring bins we should use around the query bin along each dimension). For FragBag, we tested the bag-of-words datasets of lengths between 9 and 12 and of all sizes from the FragBag Web site (Budowski-Tal *et al.*, 2010).

The AUROC of our combined hit-rate score, using ContactLib-9L and ContactLib-3R, are shown in Figure 2a. We see that the best accuracy of ContactLib-9L is achieved when $m = 32$, where the average AUROC is 0.876. Moreover, the best accuracy of ContactLib-3R is achieved when $m = 8$, where the average AUROC is 0.956. Thus, the best result for ContactLib-3R is 9% more accurate on average than that for ContactLib-9L. This indicates that remote contacts carry critical information that is not carried by local contacts and are capable of identifying neighbor protein structures more accurately.

The AUROC of the neighbor protein structure retrieval that defined on different SAS thresholds are also shown in Figure 2a. Specifically, when the SAS threshold of 3.5 Å is used, the best average AUROCs of ContactLib-3R and ContactLib-9L are 0.918 and 0.819, respectively; when the SAS threshold of 5.0 Å is used, the best average AUROCs of ContactLib-3R and ContactLib-9L are 0.906 and 0.804. Moreover, the AUROC of the neighbor protein structure retrieval that defined on different SCOP levels are used in our experiment but not shown here. Although the results are slightly different, our neighbor protein structure retrieval method, with either a local or a remote contact group library, is always capable of delivering high accuracies with high AUROCs.



ContactLib

FragBag

Fig. 2. ROCcurve analysis: (**a**) the highest average AUROC is 0.876 when the ContactLib-9L with $m = 32$ is used; the highest average AUROC is 0.956 when the ContactLib-3R with $m = 8$ is used; (**b**) the highest average AUROC is 0.747 when the FragBag with an Euclidean distance function, a fragment length of 10 and a bag size of 100 is used

We also combined ContactLib-9L and ContactLib-3R to retrieve neighbor protein structures from SCOP30 (Chandonia *et al.*, 2004; Murzin *et al.*, 1995). This is done by linearly combining the score for ContactLib-9L with $m = 32$ and the score for ContactLib-3R with $m = 8$. When a weight of 1 : 16 is used between ContactLib-9L and ContactLib-3R, the average AUROC is improved slightly to the highest value of 0.960. Thus, ContactLib-3R contributes more than ContactLib-9L to deliver more accurate results.

For comparison, we tested bag-of-words for FragBag (Budowski-Tal *et al.*, 2010) with different fragment lengths and bag sizes as shown in Figure 2b. Different experiment settings, such as eliminating the query protein structures that do not have any neighbor protein structures in SCOP30, lead to a few new observations. First, the Euclidean distance function performs significantly more accurately than the cosine distance function. Moreover, the choice of FragBag, with different fragment lengths or different sizes, has no significant impact on the accuracy obtained. According to our result, the optimal FragBag is the one with a Euclidean distance function, a fragment length of 10 and a bag size of 100, that has an average AUROC of 0.747.

By comparing Figure 2a and b, we find that our ContactLib outperforms FragBag (Budowski-Tal *et al.*, 2010) in terms of AUROC. This is further supported by looking at the AUROC distributions of ContactLib and FragBag in Figure 3. Specifically, when ContactLib-3R is used, 75% of the AUROCs are >0.936 and the lowest AUROC is 0.504. When ContactLib-9L is used, 75% of the AUROCs are >0.823 and 3% of the AUROCs are <0.5. However, when FragBag is used, 75% of the AUROCs are >0.657 and 10% of the AUROCs are <0.5. Recall that a random method has an AUROC equals to 0.5. Although FragBag is capable of delivering good average accuracy, the worst case may not be acceptable for many accuracy sensitive applications. In our experiment, the worst-case AUROC is significantly improved by using ContactLib, and ContactLib-3R is even able to guarantee an AUROC, which is higher than a random method.

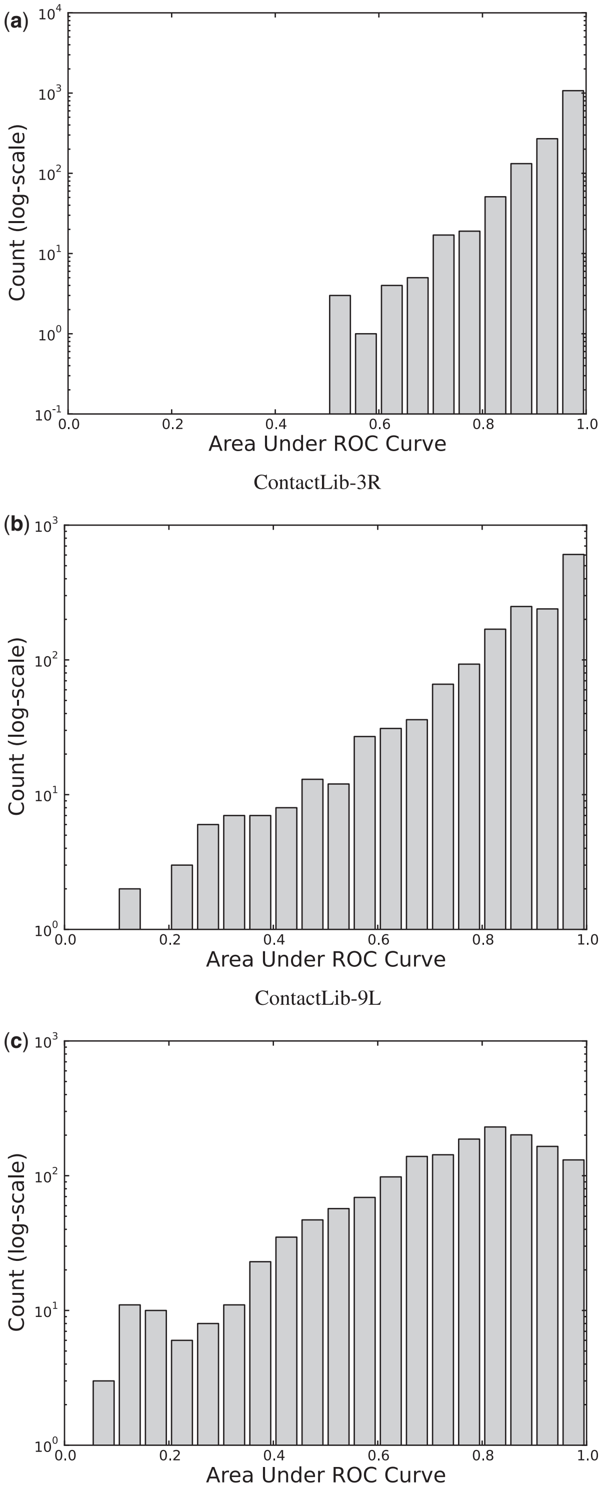
Therefore, the best accuracy is archived when ContactLib-3R with $m = 8$ is used. If only the top three ranked protein structures according to our combined hit-rate score are considered, there is a probability of 58% that we found at least one neighbor protein structure. The probability is increased to 73% when only the top 10 are considered. The excellent result suggests that ContactLib-3R can be used as a highly accurate and efficient filter to remove most unrelated protein structures while keeping many neighbor protein structures.

3.2 ROC curve analysis of all-α and all-β proteins

To understand the influence of secondary structure to the neighbor protein structure retrieval problem, we studied the AUROC of those all-α and all-β query protein structures in the previous section. From the 1574 query protein structures in the previous section, there are 157 all-α protein structures and 313 all-β protein structures.

The AUROCs of our neighbor protein structure retrieval with ContactLib-3R and ContactLib-9L for $m \in \{2, 4, 8, 16, 32, 64\}$ are shown in Figure 4. Comparing the AUROCs of all-α and those of all-β query protein structures, the AUROCs of all-α query protein structures tend to be higher. Comparing the AUROCs of ContactLib-9L and ContactLib-3R, the impact on the type of query protein structures is significantly smaller when ContactLib-3R is used. This is because our remote contact groups are also capable of modeling hydrogen bonds in α-helices. However, local contact groups are incapable of modeling hydrogen bonds in β-strands.

Therefore, the neighbor protein structure retrieval problem for all-β query protein structures is more challenging than that for all-α query protein structures, and incorporating remote contact information is critical to produce accurate results consistently for all-α and all-β query protein structures.



ContactLib-3R

ContactLib-9L

FragBag

Fig. 3. AUROC distributions (the AUROC of a random method equals to 0.5): (**a**) when ContactLib-3R is used, 75% of the AUROCs are >0.936, and the lowest AUROC is 0.504; (**b**) when ContactLib-9L is used, 75% of the AUROCs are >0.823, and 3% of the AUROCs are <0.5; and (**c**) when FragBag is used, 75% of the AUROCs are >0.657, and 10% of the AUROCs are <0.5

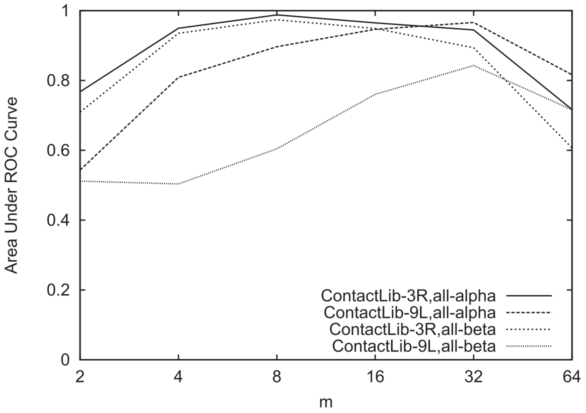


Fig. 4. ROC curve analysis of all-α and all-β query protein structures: the AUROCs of all-α query protein structures tend to be higher than those of all-β query protein structures; the impact on the type of query protein structures is significantly smaller when ContactLib-3R is used than when ContactLib-9L is used

3.3 Correlation analysis of distance functions

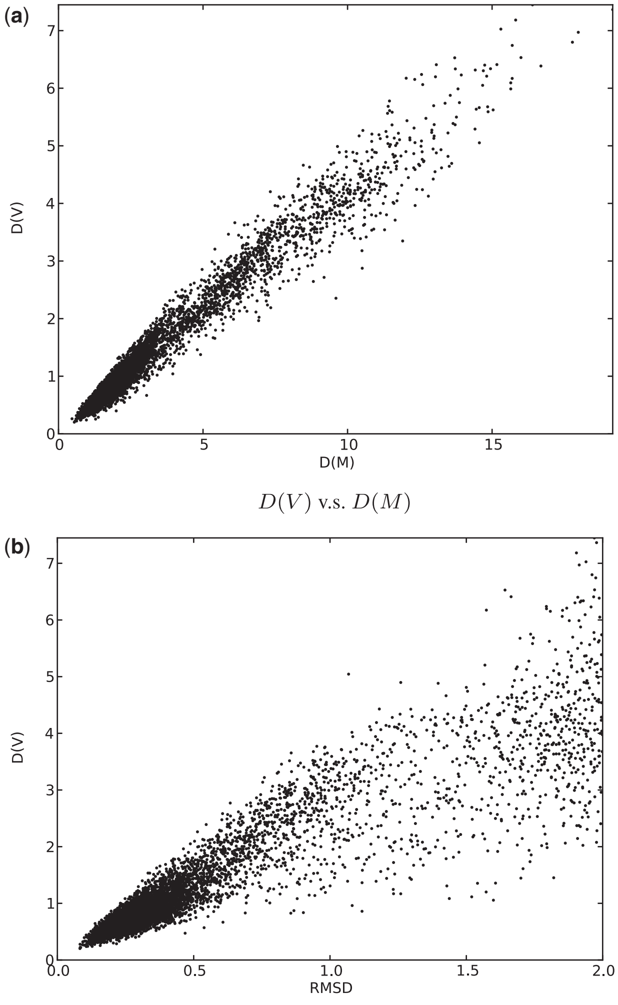
In this experiment, we demonstrated that if two contact groups have similar structures of low RMSD values, they tend to have similar distance vectors (defined in Section 2). This was done by studying the correlations among RMSD, the Euclidean distance between distance matrices $D(M)$ and the Euclidean distance between distance vectors $D(V)$. The data shown in Figure 5 were collected from similar local contact groups, with RMSD <2.0 Å, from 100 random pairs of proteins, such that each pair of proteins belonged to the same SCOP domain.

From Figure 5a, we find a strong correlation between $D(V)$ and $D(M)$ for local contact groups with RMSD <2.0 Å. This is also true for remote contact groups. Specifically, the correlation coefficients are 0.98 and 0.96 between $D(V)$ and $D(M)$ of local and remote contact groups, respectively. Therefore, our distance vector is as good as the distance matrix, which is used by the popular and successful pairwise protein structure alignment tool, DALI (Holm and Sander, 1993), to capture similar contact groups.

Although both of RMSD and pairwise distance matrix have been shown to be capable of capturing similarities between protein structures, they are not required to have strong correlations. This is also supported by our results. From Figure 5b, we find that small values of $D(V)$ suggest small values of RMSD among local contact groups. Specifically, the correlation coefficient is 0.92 between $D(V)$ and RMSD between local contact groups. However, neither $D(V)$ nor $D(M)$ has such strong correlations to RMSD between remote contact groups, and the correlation coefficients are ~0.6.

3.4 Running time

Our protein neighbor retrieval program is not only accurate, but it is also fast. On a computer with an Intel(R) Core(TM) i7 2.80 GHz CPU, both of our program with ContactLib-9L and FragBag (Budowski-Tal *et al.*, 2010) finishes in less than a second. With ContactLib-3R, our program finishes in 5.1 s on average on the same computer. Here, we assumed that our program was in server mode, such that both ContactLib-9L and ContactLib-3R were loaded in memory only once. For reference, note that the ContactLib-9L binary file has a size of 43 MB and the ContactLib-3R binary file has a size of 494 MB. Therefore, when retrieving neighbor protein structures, ContactLib is capable of delivering a significantly faster running time than pairwise protein structure alignment tools are, with little compromise in accuracy.



$D(V)$ v.s. $D(M)$

$D(V)$ v.s. RMSD

Fig. 5. Correlation analysis among RMSD, $D(M)$ and $D(V)$ of local contact groups, where RMSD is <2.0 Å, D is the Euclidean distance function, M is the distance matrix used by DALI and V is our distance vector: (**a**) the correlation coefficient is 0.98 between $D(V)$ and $D(M)$; (**b**) the correlation coefficient is 0.92 between $D(V)$ and RMSD

4 DISCUSSION

In conclusion, we have shown that ContactLib is an effective and efficient neighbor protein structure retrieval method. Most importantly, ContactLib was able to maintain a consistent level of accuracy in our tests. The key to consistently retrieve accurate neighbor protein structures for all-β query protein structures is incorporating remote contact information in ContactLib. This is

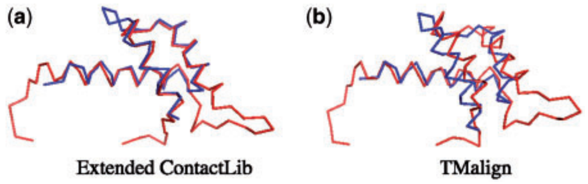


Fig. 6. Pairwise protein structure alignment between SCOP proteins d2cufa1 (red) and d3k2aa_ (blue): (**a**) our alignment with a TM score of 0.81797, an RMSD of 1.03 and an alignment length of 50; (**b**) TMalign alignment with a TM score of 0.52190, an RMSD of 2.49 and an alignment length of 49

unmatched by existing neighbor protein structure retrieval method, FragBag (Budowski-Tal *et al.*, 2010).

However, our method can be improved in several ways. One possibility is to discover and study new types of contact groups. We will look for new definitions of distance vectors representing remote contact groups based on free energies (Zhou and Zhou, 2002). Another promising extension of ContactLib would be to use contact groups to improve pairwise protein structure alignment tools, such as TMalign (Zhang and Skolnick, 2005). One such case is provided in Figure 6, which shows two alignments between SCOP proteins d2cufa1 and d3k2aa_, drawn by The PyMOL molecular graphics system (Schrödinger, 2010). The alignment shown in Figure 6b was found by TMalign, and the alignment shown in Figure 6a was found by our neighbor protein structure retrieval method with two extra steps. First, we clustered the rotation and transition matrices that yield the RMSD of similar contact groups between SCOP proteins d2cufa1 and d3k2aa_, found by our neighbor protein structure retrieval method. Second, we used the representative rotation and transition matrix of the largest cluster to generate the alignment.

We will also look for new applications for ContactLib. One promising application for ContactLib is the 'structural BLAST' approach of PrePPI (Dey *et al.*, 2013), whose performance depends mainly on the accuracy and on the speed of its neighbor protein structure retrieval. Moreover, ContactLib is also capable of finding neighbor protein structures if the query protein structure is only partially known in the process of protein structure prediction (Zhang *et al.*, 2010) or determination (Wüthrich, 1990). Then, ContactLib may use the incomplete $C_\alpha - C_\alpha$ pairwise distance matrix to find template candidates to enable it to predict or to determine the query protein structure.

Funding: GRF/ECS Grant No. 9041805 [CityU 124512] of Hong Kong (7002731), an NSERC Grant (OGP0046506), the Canada Research Chair program, an NSERC Collaborative Grant, OCRiT, the Premier's Discovery Award, the Killam Prize and SHARCNET.

Conflict of Interest: none declared.

REFERENCES

Akutsu,T. (1996) Protein structure alignment using dynamic programing and iterative improvement. *IEICE Trans. Inf. Syst.*, **79**, 1629–1636.

Aung,Z. and Tan,K.-L. (2007) Rapid retrieval of protein structures from databases. *Drug Discov. Today*, **12**, 732–739.

Budowski-Tal,I. *et al.* (2010) FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc. Natl Acad. Sci. USA*, **107**, 3481–3486.

Canutescu,A.A. *et al.* (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.

Chandonia,J.-M. *et al.* (2004) The astral compendium in 2004. *Nucleic Acids Res.*, **32** (**Suppl. 1**), D189–D192.

Choi,I.-G. *et al.* (2004) Local feature frequency profile: a method to measure structural similarity in proteins. *Proc. Natl Acad. Sci. USA*, **101**, 3797–3802.

Cui,X. *et al.* (2013) Protein structure idealization: how accurately is it possible to model protein structures with dihedral angles? *Algorithms Mol. Biol.*, **8**, 5.

Dey,F. *et al.* (2013) Toward a structural blast: using structural relationships to infer function. *Protein Sci.*, **22**, 359–366.

Engh,R.A. and Huber,R. (2006) Chapter 18.3: Structure quality and target parameters. In: *International Tables for Crystallography*. Vol. F, Wiley Online Library, pp. 382–416.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**, 861–874.

Güntert,P. and Wüthrich,K. (1991) Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *J. Biomol. NMR*, **1**, 447–456.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Kolodny,R. and Linial,N. (2004) Approximate protein structural alignment in polynomial time. *Proc. Natl Acad. Sci. USA*, **101**, 12201–12206.

Kolodny,R. *et al.* (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.

Li,S.C. *et al.* (2008) Fragment-HMM: a new approach to protein structure prediction. *Protein Sci.*, **17**, 1925–1934.

Maadooliat,M. *et al.* (2012) Assessing protein conformational sampling methods based on Bivariate lag-distributions of backbone angles. *Brief. Bioinform.*, **14**, 724–736.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A. *et al.* (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Rice,L.M. and Brünger,A.T. (1994) Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins*, **19**, 277–290.

Schrödinger,L.L.C. (2010) The PyMOL Molecular Graphics System, version 1.3r1.

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

Simons,K.T. *et al.* (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.

Wüthrich,K. (1990) Protein structure determination in solution by NMR spectroscopy. *J. Biol. Chem.*, **265**, 22059–22062.

Yang,Y. *et al.* (2012) A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic acid binding prediction. *Proteins*, **80**, 2080–2088.

Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.

Zhang,J. *et al.* (2010) Mufold: a new solution for protein 3D structure prediction. *Proteins*, **78**, 1137–1152.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.

Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.