A Method to Improve Structural Modeling Based on Conserved Domain Clusters

Fa Zhang, *Member, IEEE*, Lin Xu, and Bo Yuan

***Abstract*—Homology modeling requires an accurate alignment between a query sequence and its homologs with known three-dimensional (3D) information. Current structural modeling techniques largely use entire protein chains as templates, which are selected based only on their sequence alignments with the queries. Protein can be largely described as combinations of conserved domains, and already more than two-third of the known protein domains can be found in the Protein Data Bank (PDB). We presented a method to improve structural modeling based on conserved domain clusters. First, we searched and mapped all the InterPro domains in the entire PDB, partitioned and clustered homologous domains into the domain-based template library. For each of the resulting clusters created, a multiple structural alignment was generated based only on the 3D coordinates of all the residues involved. Then we used the structural alignments as anchors to increase the alignment accuracy between a query and its templates, and consequently improve the quality of predicted structure for query protein. We implemented the method on DAWNING 4000A cluster system. The preliminary results show that our domain-based template library and the structure-anchored alignment protocol can be used for the partial prediction for a majority of known protein sequences with better qualities.**

I. Introduction

WITH the completion of the sequencing of the genomes of human and other organisms, attention has now focused on the characterization of 3D structure and function of proteins, the products of genes. Traditionally, protein structures are largely determined experimentally by X-ray crystallography and NMR spectroscopy. Unfortunately, X-ray crystallography is very time-consuming, and NMR spectroscopy is often not accurate and sensitive enough for the structural characterizations of even medium-size proteins. As of to date, there are still only about 9,500 unique structures with less than 95% sequence identity to each other in the PDB [1]. This compares to already more than 2.2 million unique protein sequences in the current UNIPROT database (http://ww.ebi.uniprot.org). As more and more complete genomes have been or are being sequenced, the number of protein sequences will continue to grow exponentially. Obviously, the information gap between sequence and structure is huge.

The newly founded Protein Structure Initiative is aimed at determining representative protein structures for major protein families in a high-throughput mode of operation (http://www.nigms.nih.gov/psi). The idea is that these experimentally determined structures will then be used as templates for the computational modeling of related sequence homologs to produce a structural coverage for a majority of sequenced genes. In some cases, homology modeling and other computational techniques (such as protein threading) might become the only way of obtaining structural information when experimental techniques fail: the proteins are too large for NMR analysis or cannot be crystallized for X-ray diffraction.

Homology modeling uses known protein structures as templates, which is based on two hypotheses:

1. The structure of a protein is uniquely determined by its amino acid sequence. Knowing the sequence should, at least in theory, suffice to obtain its structure [2].
2. During evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structures and distantly related sequences still fold into similar structures [3, 4].

If a protein whose structure is unknown (query) has high sequence similarity (more than 30~40% of sequence identity) to a known structure, homology modeling [5-9] can be used to predict its tertiary structure with a reasonable accuracy. In the homology modeling, the query sequence is first aligned with as many residues as possible to a template, then the backbone of the query is generated based on the sequence alignment, finally all atoms of the query are produced by filling in any gaps and orienting the side chains appropriately [2, 10]. In this regard, the quality of the predicted structure by homology modeling depends on two crucial factors: the template library and the accuracy of the alignment between the query and its templates.

Current modeling techniques still use largely entire proteins or chains as templates, thus dramatically constrain the number and the type of sequences to be modeled. Proteins and their structures can be largely described as combinations of conserved protein domains. Even though the number of

This work was supported in part by the U.S. Department of Commerce under Grant BS123456.

F. Zhang, is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China, and also with the department of Biomedical Information, the Ohio State University, Columbus, OH 43210 USA (e-mail: zf@ ncic.ac.cn).

L. Xu, is with the Institute of Computing Technology, and Graduate Schools, Chinese Academy of Sciences, Beijing 100080, China (e-mail: xulin@ ncic.ac.cn).

B. Yuan, is with the department of Biomedical Information and Pharmacology, the Ohio State University, Columbus, OH 43210 USA (e-mail: yuan.33@osu.edu).

1-4244-0054-6/06/$20.00

unique structures characterized so far remain very limited (<10,000), interesting to note is the fact that already more than two-third of the protein domains (>3,400) in the InterProt database can be found in the PDB. This motivates us to propose an alternative where conserved sequence domains instead of complete protein chains are used as templates for homology modeling. Since more than 85% of all proteins are found to contain at least one or multiple conserved sequence domains, it would be reasonable to imagine that at least partially many protein structures (e.g. > 50%) can be modeled.

It is also important to note that still existing modeling techniques uses sequence alignment to select structural templates. It is generally accepted that structural alignment based only on the three-dimensional coordinates would accurately represent the corresponding residues as well as the boundary and site of any gaps. For this reasons, a number of structure-based alignment tools have been reported. MASS [11] uses secondary structure elements (SSE) to improve sequence alignment; 3Dcoffee [12] and FUGUE [13] combine tertiary information as a more accurate scoring matrix to improve the quality of sequence alignments, recently, the eBLOCKs database [14] enumerates a cascade of conserved blocks anchored by known three-dimensional structures. However, when the sequence similarity is low (e.g. <30%), an accurate alignment between a query protein sequence and its templates remains a major challenge, due to the computational [15] and the biological [16] limitations.

Facing these aforementioned crucial factors in homology modeling, we present a method to improve structural modeling based on conserved domain clusters. We first set out to create a domain-based library aimed at expanding structural coverage to more protein sequences. This was accomplished by partitioning the PDB into domain-based structural clusters, each of which is further consolidated into a multiple structural alignment. These resulting structural alignments are then used as anchors to improve the alignments between query sequences and their templates, and consequently improve the quality of protein structure prodiction. In addition, such conserved structural library will be used for our characterization and validation of protein-protein interactions mediated by many of the conserved domains. Briefly, we used the programs Dali [17] and CE [18] to superimpose all corresponding residues for each of the domains that have been clustered. The resulting structural ensembles are then converted into position-specific profiles as anchors to confine any query-template alignments with the program ClustalW. We showed that at least one-third of the alignments were significantly improved by comparing the sequence- with their structure-anchored alignments. This improvement of using the structure-anchored alignments was further confirmed by the modeling of a benchmark (1,476 conserved domains with known structures) with the program MODELLER [7], which resulted in more accurate structures compared to their originals. In addition, we showed that even just incorporating predicted secondary structure information the accuracy of the query-template alignments could be significantly improved; again pointing to the fact that structural modeling based only on sequence information can be error-prone. Our preliminary results show that our method can firstly consolidate and expand existing structural templates to potentially cover more protein sequences, and secondly improve the quality of the critical query-template alignments.

II. Methods and Materials

We first mapped the InterPro database (ftp://ftp.ebi.ac.uk/pub/databases/interpro) to the PDB by using the accompanying *iprscan* software. All the PDB protein sequences in this project were parsed directly from the structural records reorganized by MSD database (ftp://ftp.ebi.ac.uk/pub/databases/msd version: 20040412,). The consensus sequences for all InterPro domains were obtained from the protein family and superfamily databases include Pfam[19], SCOP[20], SMART[21], TIGRFAM[22], ProDom[23] and PRINT[24]. Then we partitioned all the structural correspondences for all the known protein domains from PDB, and clustered the conserved domains based on InterPro entries. For each of the domain clusters, we superimposed and aligned homologous structures using the Dali or CE program to construct a multiple 3D-structural alignment based only on structure information. The benchmarks consisted of the remotest structures and sequences compared to their corresponding consensus sequences for each of the domain clusters were selected.

A. Construction a Template Library of Structural Clusters for All Conserved Domains

As was stated earlier, current protein modeling techniques use entire protein or whole chain structure as template, which works well only when the majority of the query sequence and its template can be aligned, thus limits the number of sequences to be modeled. Although only about 33,000 individual protein structures have been determined, at least half of the know domains can be found in these structures. Focusing on the template issue in homology modeling, our first goal in this project is to construct a domain template library to increase the likelihood of widely applicable structure templates.

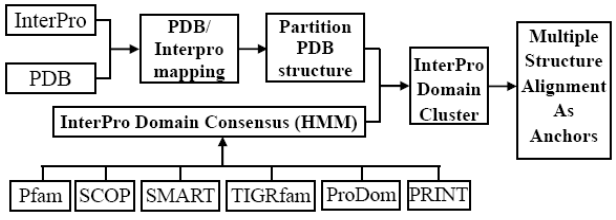
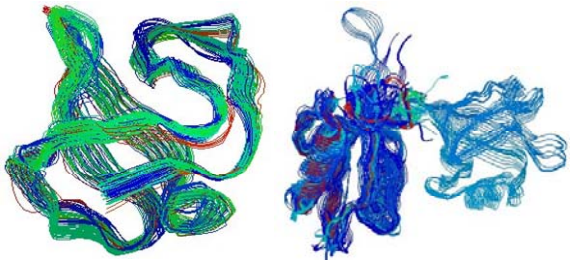


Fig. 1. Flowchart for constructing template library.

The flowchart is shown in the Figure 1. From the protein family and superfamily databases, such as Pfam, SCOP, SMART, TIGRFAM and so forth, we used the HMMER

method [25] to obtain the consensus sequence for each InterPro domain. The InterPro database integrates the information for all known protein families, domains and functional sites. We searched all the InterPro domains in PDB using the program iprscan [26], and mapped the corresponding protein structures in PDB. Since many of the existing protein sequences derived from the PDB do not always correlated precisely with the positions in their corresponding structures, we directly parsed the sequence information from the MSD database, a cleaned-up structural database from the PDB. We observed that there are three possible ways in the mapping of InterPro domains and PDB structures: (1) the entire domain can be found in a protein, (2) the greater part of a domain (> 40% residues) can be found in a protein and (3) only a small part of a domain (< 10 residues or < 40% residues) can be found in a protein. Based on the first two mapping criteria, we partitioned all the structural correspondences from PDB for each InterPro domain, and constructed the primary domain cluster, that is the "partition PDB structure" in Figure 1. For each domain cluster, we compared all the sequences of domains involved with the relevant consensus sequence using the Smith-Waterman algorithm, then chose and refined the domain cluster by removing the structure whose sequence identity to the consensus sequence is less than a predefined threshold (e.g. 30%). Here, we defined the structure (domain), whose sequence has the highest similarity to the consensus sequence, as the reference for that cluster. Finally, we used Dali method to calculate the RMSDs (Root Mean Square Deviation) for the rest structures against the reference, and chose the structures; their RMSD is less than 3Å, to construct the domain cluster. Since all the domains in each of clusters are conserved in both sequence and structure level, we built the domain template library, by adopting the conserved structures as template for the relevant cluster.

B. Structure-Anchored Alignment between the Query and Its Templates



A B

Fig. 2. Two typical structure ensembles of conserved domain clusters.

Experimental determination of domain structures has shown that three-dimensional structure is highly conserved during molecular evolution. We observed that most of the domain clusters in our library have a number of similar protein structures. To highlight the conserved structural regions in each domain, we superimposed these conserved structures using Dali or CE algorithm. Two typical structure ensembles are illustrated in Fig. 2: (A) the structure ensemble of domain cluster IPR000108. It includes 18 structures. All the structures in this cluster are aligned perfectly, the RMSD values of the domains ivolved against the reference are less than 1Å. (B) shows the structure ensemble of PDZ domain cluster (IPR001478) with including 41 structures. The RMSD of the remote structure against the reference is less than 3Å.

Based on the conserved structural ensemble in each domain cluster, a multiple 3D-structural alignment is generated purely from the structure information (the backbone coordinates of residue). Since this structural alignment is independent of the sequence similarity, it provides more sensitive and position-specific signatures than the sequence alignment. In general, residues in secondary structure are more conserved than those in coil regions, thus insertions/deletions are less likely to occur in the secondary structure regions. To align the query sequence to the templates, we firstly predicted the approximate secondary structure of the query sequence by the PHD program[27], and then labeled the templates secondary structures based on their structure information, and fixed the consensus tertiary substructures as anchors to generate the alignment between the query sequence and its template structures. The flowchart is shown in Figure 3.

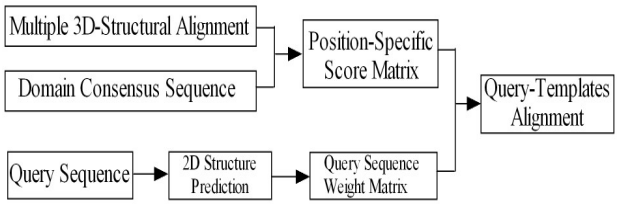


Fig. 3. The Flowchart of structure-anchored query-template alignment.

According to previous studies, the quality of the sequence alignment between a query protein and its templates is an important factor that determines the quality for the structural prediction. Given a query protein sequence, using the corresponding multiple 3D-structural alignment as an anchor, we can obtain more accurate and reliable alignment between the query and its templates, consequently significantly improve the quality of predicted structure for query protein, the details are shown in results and discussions section.

C. Benchmark Selection and Validation

In order to ascertain if using our template library can model more protein structure and our 3D-structure alignments can improve the quality of our structural prediction, we selected the known structures from 1476 domain clusters as benchmarks and compared their original structures with their predicted structures. The flowchart of benchmark selection and validation is shown in Figure 4. For each domain cluster, we selected the structure (domain) as the reference, whose sequence has the highest sequence identity to the consensus, compared the rest structures with the reference and chose the remotest one as the benchmark (query) and others as

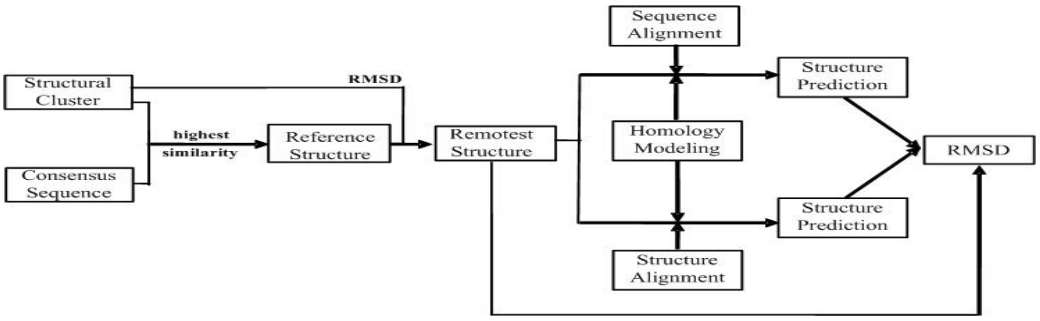


Fig. 4. The Flowchart of benchmark selection and validation.

templates. Then we generated the alignment between query and its templates using the above method.

To validate our query-templates alignment is more accurate and reliable than that obtained by common multiple sequence alignment method, we use the MODELLER program to predict the query structure against the two alignments respectively.

III. Results and Discussions

TABLE I
MAPPING RESULTS OF INTERPRO DOMAINS AND OUR TEMPLATES LIBRARY TO PDB

	Not Find	One Domain	>= 2 Domains
InterPro Domain	2,056 (36.7%)	513 (36.7%)	3,050 (54.2%)
Template Library	2,056 (36.7%)	743 (13.2%)	2,820 (50.1%)

Column 2 is the number and its percentage of the domain cluster cannot be found in PDB. Column 3 shows how many domain clusters only have single structural information. Column 4 provides the number and the percentage of domain clusters, which have more than two individual structures.

A. *Structure information of the templates library*

A summary of the structural information of the conserved domains in PDB is shown in the Table.I. For all the 5,629 domain entries in current InterPro database, we obtain 3,563 (~65%) domain clusters with structural information from PDB, of which 3,050 (~54.2%) clusters contain more than two individual structures with the remaining 513 (~9.1%) containing single structure. We selected the 3,563 domain clusters to construct our template library. In the library, there are 2,820 (~50.1%) clusters can generate a multiple 3D-structural alignments. From the table, we can find that our templates cover about two-thirds of all the 5,629 entries in the current InterPro database, with at least half of all the InterPro domains have multiple 3D-structrual alignment. Since proteins evolve with their structural and functional domains as independent units, and the InterPro domains are dispersed in more than 85% of all proteins [28], we believe we can model more structures using our templates library.

A major advantage of our method is the construction of a multiple 3D-structure alignment for each domain cluster (template) and the use of this alignment as an anchor to improve the sequence alignment with a query and its templates. Since the 3D-structure alignments were created based solely on the residue positions in each domain, the resulting alignments must be most accurate, representing the best profiles and weights, as well as the appropriate gap positions for any coming sequence alignments. This in large part addresses the problem of arbitral sequence alignments especially when the sequence identities are low, including issues of gaps and gap penalties. It is particularly true that the same structure-based multiple alignments can be very different from their sequence multiple alignments in many cases, even in the cases where the sequence identity is higher than 50%. An example is shown in Figure 5, even the sequence identity is very high, and the sequence alignment is still error-prone.

In Figure 5, We chose 5 domain segments from the PDZ domain cluster (InterPro ID: IPR001478), namely 1be9a, 1mfga, 1ihja, 1nf3c and 1l6ob, selected 1l6ob as query and others as templates, then generated the alignment of query and its templates using the different alignment protocols. (1), Part of alignment using common sequence alignment, here ClustalW[29]. (2), Segment of alignment using our

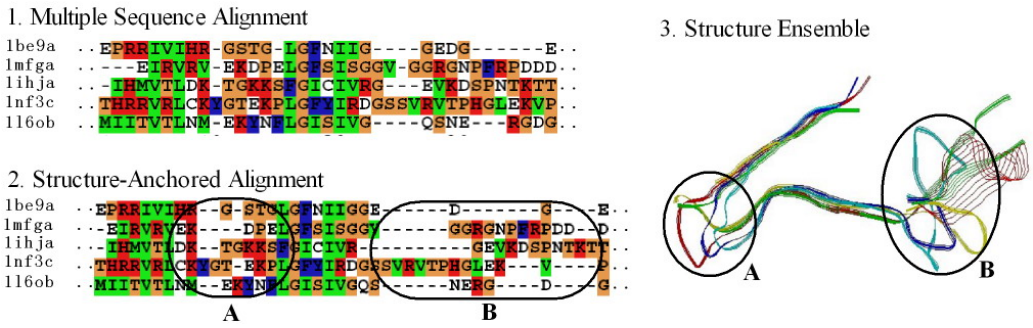


Fig. 5. The different from the sequence and structure-based alignment for some segments in PDZ domains cluster.

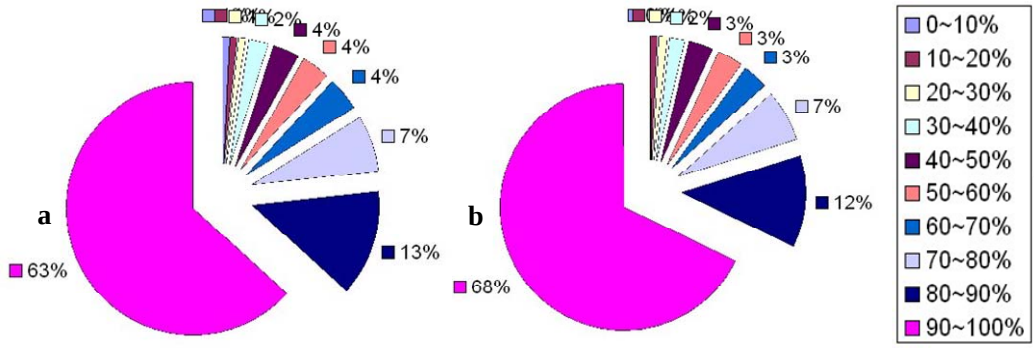


Fig. 6. Distribution of the correlativity of the primary/2D-structural alignment with the corresponding 3D-structural alignment in the templates library.

structure-anchored method. (3), the superposition of these 5 domain segments, the structures are prepared with the program Rasmol[30]. The different regions in these two alignments are labeled A and B, we also point the corresponding structures in (3). The superposed structures showed that the structures in region A and B have some variability, thus causes more gaps inserted into region A and B. Though the structure-anchored alignment has more gaps than the multiple sequence alignment, it shows more accurate and sensitive structural information. So we obtained more accurate predicted structure using the structure-anchored alignment against using the multiple sequence alignment.

Given the fact that accurate alignment of a query sequence with its structural template is the key of successful structural modeling, our methods provide a means to improve the structural prediction of presumably many of the protein sequences.

To further analyze the correlativity between the multiple 3D-structure alignment and the multiple sequence alignment in each of the domain cluster, we compared the multiple sequence alignment, 2D-structural alignment to the 3D-structural alignment, the results are shown in Figure 6. Here, the 2D-structrual alignments were generated by: first predicted the secondary structure of all the sequences in the template by PHD program, and then constructed the alignment with fixing the consensus secondary structures. Figure 6a and 6b show the correlativity of the sequence alignment/2D-structural alignment with the 3D-structural alignment, respectively. The correlativity was calculated by:

$$correlativity = \frac{Score\ of\ 1D\ (or\ 2D\text{-}structure) alignment\ compares\ to\ its\ 3D\text{-}structure\ alignment}{Score\ of\ the\ 3D\text{-}structure\ alignment\ compares\ to\ itself}$$

Here, we use the program *COMPASS*[31] to compare the alignments.

As described above, all the sequences in each domain cluster are conserved with the domain consensus sequence (removing the sequence whose identity to the consensus is less than 30%), for a majority (~63% and ~68%) of the domain clusters, the primary sequence and its 2D-structural alignments are very similar (>90% similarity) to their corresponding 3D-structural alignments. However, for more than 10% of domain clusters, there are less than 50% similarities among their primary, 2d- and 3D-structural alignments. In some cases, primary or 2D-structural alignments are completely different from their corresponding 3D-structural alignments (< 20% similarity). Analyzing the alignments in these domain clusters, we found that there were more gaps in 3D-structure alignments than in corresponding primary and secondary structural alignments. For primary sequence, we hope as few gaps as possible in alignment, however, as to the tertiary structure, we desire that the alignments characterize reliably as much the structural information as possible, just as figure 5 shown. Since the 3D-structural alignment relies only on the tertiary structure information, independently of sequence similarity, figure 6 suggests that even when sequence identity is high, the sequence alignments may still have many errors compared to their 3D-structural alignments. Thus the multiple 3D-structural alignments can be used to generate a more reliable template library for more accurate structure prediction.

B. Comparison of results predicted using different alignment protocols

For the 1,476 benchmarks, we generated the alignment of the query and its templates based on the corresponding 3D-structural alignment, and then predicted the structure of query by homology modeling method (MODELLER). In order to validate that our 3D-structural alignments can improve the quality of structural predictions, we also predicted the benchmark structures based on the common multiple sequence alignment and compared these two predictions with the original structure for each benchmark. A summary of the statistics and analysis for the predicted results is shown in Table II.

Using our structure-anchored alignment protocol, 1,341 (~97.5%) predicted results improved against the results based on common sequence alignment, of which 436 (~31.7%) have significant improvement ($\Delta RMSD > 1$); only 35 (~ 2.5%) predictions are worse than that based on common sequence alignment, of which 28 (75%) results have very little deterioration ($-0.5 < \Delta RMSD < -1$). As was stated earlier, the

TABLE II
MAPPING RESULTS OF INTERPRO DOMAINS AND OUR TEMPLATES LIBRARY TO PDB

Prediction Result				∆ RMSD		Domain Number		Length Distribution		Alignment Similarity		Cluster Granularity (SD)	
Improved	1,341	Significantly improved	436	>1Å	169 (39%)	<= 10	203 (47%)	<=100	154 (36%)	>=70%	357 (82%)	<= 1	346 (80%)
						10~40	156 (35%)	100~200	140 (32%)	60~70%	50 (11%)		
			31.7%	<=1Å	267 (61%)	> 40	77 (17%)	200~300	62 (14%)	50~60%	22 (5%)	> 1	90 (20%)
								300~400	43 (10%)	<50%	7 (2%)		
	97.5%							>400	37 (8%)				
		Not Changed	905			<= 10	501 (64%)	<=100	391 (43%)	>=70%	796 (88%)	<= 1	601 (66%)
						10~40	240 (27%)	100~200	252 (28%)	60~70%	22 (2%)		
			65.8%			> 40	84 (9%)	200~300	116 (13%)	50~60%	87 (10%)	> 1	305 (34%)
								300~400	62 (7%)	<50%	0		
								>400	84 (8%)				
Not Improved		35		-0.5~-1Å	28 (75%)	<= 10	13 (38%)	<=100	3 (10%)				
						10~40	11 (31%)	100~200	11 (31%)	>=70%	7 (25%)	<= 1	11 (31%)
				<-3Å	7 (25%)			200~300	12 (34%)				
		2.5%				> 40	11 (31%)	300~400	4 (11%)	<70%	28 (75%)	> 1	24 (69%)
								>400	5 (14%)				

Column 2 shows the prediction results, compared the predictions using our structure-anchored alignment to those based on sequence alignment. We also show that how many results are significantly improved. Column 3 is the value of ΔRMSD, the RMSD difference between the prediction using our structure-anchored alignment against the original and the prediction based on sequence alignment against the original. Columns 4 and 5 provide the distribution of domains number and their length in the 1,476 clusters. Column 6 shows the correlativity of the structure-anchored alignment to structure-anchored alignment for each domain cluster. Column 7 gives the granularity (standard deviation) of each domain cluster.

accuracy of alignment between query and its templates is an important factor to determine the quality of the structural prediction, the benchmark results suggest that we can construct a more reliable and sensitive sequence alignment between the query and its template, based on our structure-anchored alignment protocol. This better alignment consequently can significantly improve the quality of our structure prediction.

The results in Table II suggest that our 3D-structural alignments can significantly improve the quality of structural prediction. However, the more interesting problem from the results is that there is 35 (~2.5%) benchmark predictions not improved using the 3D-structural alignments. Since in theory, the predictions using our 3D-structural alignments should be more accurate or at least same with, the predictions using common sequence alignment. For each benchmark cluster, we analyzed the distribution of the number of domain involved, the length distribution of the structures involved, the correlativity of the sequence alignment to its 3D-structural alignment and the cluster granularity. The results, listed in Table II, showed that both improved part and not improved part had the similar distribution in domain number and structure length. As to the alignment correlativity, a majority (more than 80%) of the benchmark in improved part has obvious relationship (more than 70% similarity), however only 25% in not improved part have the relationship; as to the cluster granularity, more than two-thirds of the improved benchmarks have small granularity value (<= 1), while less one-third in not improved part. These results suggest that if our 3D-structural alignments have high correlativity to their sequence alignments, and if the domains are even dispersed in the cluster, our 3D-structural alignments can increase the accuracy of the alignment between query and its templates, thus improve the quality of the query structure prediction.

IV. CONCLUSION

In this paper, we present a method to improve structural modeling based on conserved domain clusters. We first searched and mapped all the InterPro domains in the entire PDB, partitioned and clustered homologous domains into structural ensembles. For each of the resulting clusters created, a multiple structural alignment was generated based only on the 3D coordinates for all the residues involved. Then we use these resulting structural alignments as anchors to obtain more accurate and reliable alignment between the query and its templates, thus consequently improve the quality of predicted structure for query protein. Here, we report 1) the construction of such a 3D library for all the protein domains in the InterPro database; 2) the use of structural alignments as anchors to improve the alignment accuracy between a query and its 3D template; and 3) the validation using know structures as benchmarks to assess the modeling outcome. Besides being served as anchors, the structural alignments have also been assessed for sequence-structure correlations as well as biological investigations into regions of both hyper- and hypo-variability (Zhang et al., in preparation).

We constructed the templates library and implemented the method on DAWNING 4000A cluster system. Our preliminary results show that our method can be used for the prediction for a majority of known protein sequences with better qualities. Also, we found that the computing time would be increased, with our template library growth.

Further work to improve the sensitivity of the result and reduce the computing time is being investigated.

V. Acknowledgement

The authors would like to thank the National Research Center for Intelligent Computing System, Chinese Academy of Sciences, for providing the DAWNING 4000A cluster system.

References

[1] Chandonia, J.M., et al, "The ASTRAL Compendium", Nucleic Acids Res. vol32 (Database issue), 2004, pp.D189-92.
[2] Philip E.B, H.W., Structural Bioinformatics. Hoboken, New Jersey: Wiley-Liss, Inc (2003)
[3] Chothia, C. and A.M. Lesk, "The Relation between the Divergence of Sequence and Structure in Proteins", Embo J. 5(4), 1986, pp. 823-6.
[4] Sander C, S.R, "Database of homology-derived protein structures and the structural meaning of sequence alignment", Proteins. 9, 1991, pp. 56-68.
[5] Baker D, S.A, "Protein strucutre prediction and strucutral genomics", Science. 294, 2001, pp.93-96.
[6] Venclovas, C, "Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment", Proteins Suppl.5, 2001, pp. 47-54.
[7] Fiser, A., Do, R.K., Sali, A, "Modeling of loops in protein structures", Protein Sci. 9, 2000, pp.1753-73.
[8] Marti-Renom, M.A., et al, "Comparative protein structure modeling of genes and genomes", Annu Rev Biophys Biomol Struct. 29, 2000, pp. 291-325.
[9] Bates PA, K.L., macCallum RM, "Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM", Proteins Suppl. 5, 2001, pp. 39-46.
[10] Lee, J., et al, "Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing", Proteins. 56(4), 2004, pp.704-14.
[11] Dror, O., et al, "Multiple struct,ural alignment by secondary structures: algorithm and applications", Protein Sci. 12(11), 2003, pp.2492-507.
[12] O'Sullivan, O., et al, "3DCoffee: combining protein sequences and structures within multiple sequence alignments", J Mol Biol. 340(2), 2004, pp.385-95.
[13] Shi, J., T.L. Blundell, Mizuguchi, K, "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties", J Mol Biol. 310(1), 2001, pp. 243-57.
[14] Su, Q.J., et al, "eBLOCKs: enumerating conserved protein blocks to achieve maximal sensitivity and specificity", Nucleic Acids Res. 33, 2005, pp. D178-82
[15] Wang, L., Jiang, T, "On the complexity of multiple sequence alignment", J Comput Biol. 1(4), 1994, pp.337-48.
[16] Thompson, J.D., et al, "Towards a reliable objective function for multiple sequence alignments", J Mol Biol. 314(4), 2001, pp.937-51.
[17] Holm, L., Sander, C, "Dali: a network tool for protein structure comparison", Trends Biochem Sci. 20(11), 1995, pp.478-80.
[18] Shindyalov, I.N., Bourne, P.E, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path", Protein Eng. 11(9), 1998, pp.739-47.
[19] Bateman, A., et al, "The Pfam protein families database", Nucleic Acids Res, 32(Database issue), 2004, pp.138-41.
[20] Andreeva, A., et al, "SCOP database in 2004: refinements integrate structure and sequence family data", Nucleic Acids Res 32(Database issue), 2004, pp. 226-9.
[21] Letunic, I., et al, "SMART 4.0: towards genomic data integration", Nucleic Acids Res, 32(Database issue), 2004, pp.142-4.
[22] Haft, D.H., Selengut, J.D., White, O, "The TIGRFAMs database of protein families", Nucleic Acids Res. 31(1), 2003, pp.371-3.
[23] Bru, C., et al., "The ProDom database of protein domain families: more emphasis on 3D", Nucleic Acids Res, 33(Database Issue), 2005, pp. D212-5.
[24] Attwood, T.K., et al., "PRINTS and its automatic supplement, preprints", Nucleic Acids Res, 31(1), 2003, pp. 400-2.
[25] Eddy, S.R, "Profile hidden Markov models", Bioinformatics 14(9), 1998, pp.755-63.
[26] Zdobnov, E.M., Apweiler,R, "InterProScan--an integration platform for the signature-recognition methods in InterPro", Bioinformatics 17(9), 2001, pp.847-8.
[27] Rost, B., Sander, C, "Combining evolutionary information and neural networks to predict protein secondary structure", Proteins 19(1), 1994, pp.55-72
[28] Mulder, N.J., et al, "The InterPro Database, 2003 brings increased coverage and new features", Nucleic Acids Res. 31(1), 2003, pp.315-8.
[29] Thompson, J.D., Higgins, D.G. and Gibson,T.J, "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", Nucleic Acids Res. 22(22), 1994, pp.4673-80.
[30] Sayle, R.A., Milner-White, E.J, "RASMOL: biomolecular graphics for all", Trends Biochem Sci. 20(9), 1995, pp. 374-8.
[31] Sadreyev, R., Grishin, N, "COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance", J Mol Biol. 326(1), 2003, pp.317-36.