Recursive protein modeling: a divide and conquer strategy for protein structure prediction and its case study in CASP9

Jianlin Cheng
Department of Computer Science, Informatics Institute, C.Bond Life Science Center
University of Missouri, Columbia
MO 65211, USA
chengji@missouri.edu

Jesse Eickholt
Department of Computer Science
University of Missouri, Columbia
MO 65211, USA
jlec95@mail.mizzou.edu

Zheng Wang
Department of Computer Science
University of Missouri, Columbia
MO 65211, USA
zwyw6@mail.mizzou.edu

Xin Deng
Department of Computer Science
University of Missouri, Columbia
MO 65211, USA
xd9d3@mail.mizzou.edu

***Abstract*--After decades of research, protein structure prediction remains a very challenging problem. In order to address the different levels of complexity of modeling structure, two types of modeling techniques – template-based modeling and template-free modeling – have been developed. Template-based modeling can often generate a moderate to high resolution model when a similar, homologous template structure is found for a query protein but fails if no template or only incorrect templates are found. Template-free modeling such as fragment-based assembly may generate models of moderate resolution for small proteins of low topological complexity. Seldom have the two techniques been integrated together to improve protein modeling. Here we develop a recursive protein modeling approach to selectively and collaboratively apply template-based and template-free modeling methods to model template-covered (i.e., certain) and template-free (i.e., uncertain) regions of a protein. A preliminary implementation of the approach was tested on a number of hard modeling cases during the 9th Critical Assessment of Techniques for Protein Structure Prediction (CASP9) and successfully improved the quality of modeling in most of these cases. Recursive modeling can significantly reduce the complexity of protein structure modeling and integrate template-based and template-free modeling to improve the quality and efficiency of protein structure prediction.**

Keywords-recursive protein modeling; template-free modeling; template-based modeling; CASP9; protein structure prediction

I. INTRODUCTION

Predicting protein tertiary structure from protein sequence is important for protein engineering, protein design and protein function analysis [1]. It is becoming more and more important in the post-genomic era as millions of protein sequences are being generated by high-throughput and next-generation sequencing projects and the vast majority of these sequences do not have known structure. Currently there are more than 100 million protein sequences in GenBank [2], whereas only about 65 thousand of them have known structures in the Protein Data Bank [3].

In order to address this challenge, two major types of protein structure modeling methods have been developed to model protein structure from sequence – template-based modeling and template-free modeling. Template-based modeling (e.g. comparative modeling or homology modeling) builds the structure of a query protein from the structure of proteins with known structure (i.e., templates) which are homologous to the query [4-6]. Template-free (e.g. ab initio) modeling folds the structure of a query protein from scratch without explicitly referring to specific structural templates [7-8]. Template-based methods work well if a good template structure (e.g. a close homolog) can be found, but fails to produce an accurate structure if no template is available or only incorrect templates are used. At present, template-free modeling can generate low resolution models for small proteins with simple topologies. This is due to the difficulty of efficiently exploring the huge conformation space.

Although a variety of methods have been developed and tested for template-based and template-free modeling, only a few have integrated the two methodologies together to improve protein structure prediction. Initial efforts at combining both approaches were aimed at modeling relatively small local regions and included the application of ab initio methods to model loops [5] or N-/C- terminal tail regions of existing models [9]. Inspired by these initial attempts and the hierarchical protein folding process [10-11], we designed a general, iterative, recursive protein folding procedure to seamlessly integrate the complementary strengths of both template-based and template-free methods to effectively and efficiently predict the structure of any protein. The approach can reduce the complexity of protein modeling by dividing the modeling problem into certain (i.e., template-based) and uncertain (i.e., template-free) regions. The regions are then modeled recursively and collaboratively using the appropriate techniques and the most useful information. The approach was implemented in our MULTICOM protein structure prediction system [12], which was blindly tested on a number of hard protein targets in the ninth Critical Assessment of Techniques for Protein Structure Prediction (CASP9) (http://predictioncenter.org/casp9/). The approach successfully improved the accuracy of predicted models in a majority of cases. The experiment demonstrated that the recursive protein modeling approach can integrate template-based and template-free information together in a collaborative and reinforcing way to address a full spectrum of protein modeling problems.

II. METHODS

A. General recursive modeling procedure

In the recursive protein modeling procedure, a query protein is first searched against a template protein library using a sequence or profile alignment method. A query-template sequence alignment will be generated if some seemly homologous / analogous templates or template fragments are found. The sequence of the query protein is then initially decomposed into certain and uncertain regions based on its alignment with the significant homologous template hits. Certain regions correspond to portions of the query sequence which align well with any one of significant homologous templates (e.g. low PSI-BLAST e-value < 0.001) [13], and uncertain regions are the long query regions (e.g. >= 20 residues) that are not covered by a template or aligned with low confidence. The short unaligned regions in the query sequence are not considered as uncertain regions for special treatment. Instead, they are treated as loops in the certain regions to be handled by template-based modelling. Therefore, the uncertain regions in the decomposition usually correspond to one or more domains or a large portion of a domain composed of different kinds of secondary structures rather than a single loop, which distinguishes our approach from traditional protein loop modelling. After the decomposition, the conformations of the certain regions are generated by template-based modeling using the alignments and the corresponding template structures while leaving the uncertain regions alone. It is worth noting that in a complicated situation, one query may have multiple disjoint certain regions covered by one template or multiple templates. In practice, this situation does not pose any difficulties as such regions can be handled altogether by current template-based modeling tools. While keeping the conformation of the certain regions which usually form the core of the structure fixed or rigid, template-free modeling methods are applied to sample the conformations of uncertain regions. This template-free sampling is different from an independent, free sampling of uncertain regions because of the influence of the certain regions (e.g. core) is taken into account in both conformation sampling and energy assessment. The core-restrained sampling can often improve the effectiveness and efficiency of template free sampling by dragging the “wild”, free conformation toward the core region.

After a round of sampling, the quality of each certain and uncertain region is assessed using global /local protein model quality assessment methods [14-19]. The conformations of certain regions and some well modeled uncertain regions are combined into larger certain regions, leaving a smaller set of uncertain regions. The same modeling process is applied to model the newly defined certain and uncertain regions by using the conformations generated in the last iteration as templates. The process continues until no uncertain regions remain or the quality of the entire query protein is acceptable. The entire procedure is described in Fig. 1.

It is worth pointing out that the term “region” here may refer to any level of protein structure, such as a loop, a part of a domain, an entire domain, or even multiple domains. It is different from the ab initio loop modeling that is exclusively used to build a loop joining two parts of protein structure. Here, conceptually the recursive modeling procedure aims to build a protein structure from smaller components in a bottom up, hierarchical way. On one hand, it somewhat conceptually mimics or is in accordance with the physical, hierarchical protein folding process where local regions fold first and then interact to fold into larger protein conformations [10-11], although each decomposed region may not actually correspond to a physical folding unit. On the other hand, the procedure is in accordance with the “divide and conquer strategy” widely used in computer science, where a complicated problem is divided into smaller, easier to solve problems and the solutions to the smaller problems are combined recursively in order to solve the larger problem. In the protein modeling context, the procedure improves template-based modeling by better packing of long un-aligned regions (e.g., loops, tails, and small domains) and enhances template-free modeling by utilizing the template core as restraints. The protocol can not only integrate template-base and template-free modeling seamlessly and collaboratively, but also improve the quality and speed of protein modeling. One conceptual difference between our method and “threading” based methods, such as *TASSER* developed by Zhang [5], is that our method synergistically models certain and uncertain regions using both template-based and template-free modeling alternatively in order to shrink uncertain regions gradually to reach an optimal solution. During each iteration, template-base and template-free modeling influence / improve each other through expansion of certain regions.

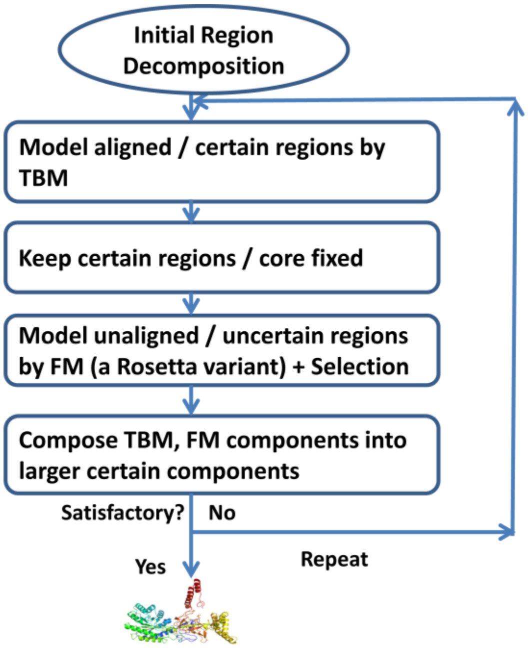

Initial Region Decomposition
Model aligned / certain regions by TBM
Keep certain regions / core fixed
Model unaligned / uncertain regions by FM (a Rosetta variant) + Selection
Compose TBM, FM components into larger certain components
Satisfactory?
No
Yes
Repeat

Figure 1. Flowchart of the recursive protein modeling procedure.TBM and FM denote template-based modeling and template-free modeling, respectively.

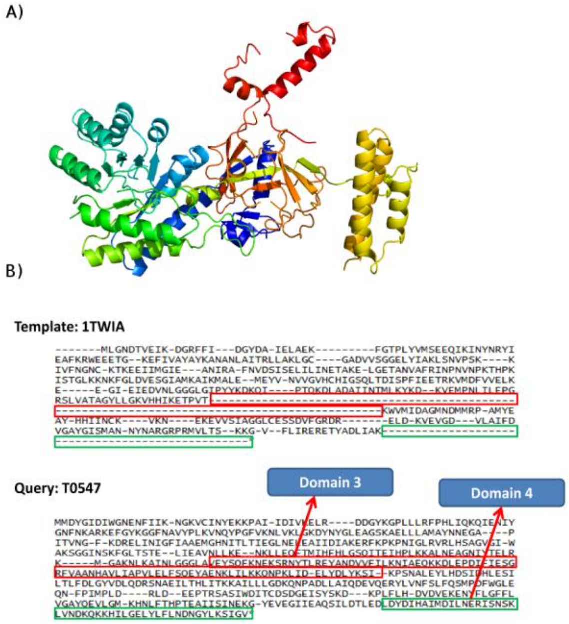
B. *A specific implementation*

The general recursive modeling process can be implemented in a number of ways depending on the specific tools and methods selected for each step. For the recursive modeling in our MULTICOM protein modeling system, we use a profile-profile alignment tool HHSearch [20] to search a query protein against our in-house template library and generate alignments for the initial region decomposition. The e-value threshold for selecting significant templates was set to 0.001. We used the simplest protocol of Modeller 9v7 [21] – automodel to do template-based modelling for certain regions with default parameter settings. The automodel protocol constructed structural conformations for aligned residues in certain regions and also automatically built loops for unaligned residues if there exist. If certain regions were covered by multiple significant templates, all of them were used by automodel to generate models according to their alignments with the same certain regions. Ten models were generated and the model with the minimum Modeller energy was chosen as the model of the certain regions. We used a modified Rosetta variant [8] to do template-free modeling. We modified Rosetta 3.0 such that one part of the input conformation for a query protein could be kept fixed while the conformations of other regions were sampled by a fragment assembly approach. This was accomplished by restricting fragment replacements to specified ranges of the protein sequence. By specifying and limiting fragment replacements to uncertain regions, the fragment assembly of the other regions is influenced by the rigid regions because their conformations are considered during the assembly of fragments for uncertain regions. For instance, fragment insertions for uncertain regions that are energetically favored by certain regions are more likely to be accepted. Usually several hundred models for uncertain regions were generated. We used ModelEvalutor [19], a single model quality assessment tool, and a pairwise comparison-based model evaluation method [15] to assess the quality of the conformations for the query protein and its regions in order to select conformations. The entire recursive modeling process is automated.

III. RESULTS

The recursive modeling approach was implemented within our MULTICOM system as four automated protein structure prediction servers (i.e. MULTICOM-CLUSTER, MULTICOM-REFINE, MULTICOM-NOVEL, and MULTICOM-CONSTRUCT), which mainly differ in model ranking and combination [23]. The MULTICOM system was blindly tested during the 9th Critical Assessment of Techniques for Protein Structure Prediction (CASP9), 2010. It showed its promise by improving the quality of protein modeling in the majority of hard cases where both template-based and template-free modeling could be applied. Here we discuss how recursive modeling improved structure prediction in three typical situations.

Case 1: recursive modeling enhances the modeling of large, complicated, multi-domain proteins. Instead of improving the uncertain regions of a single domain, here, recursive modeling can synergistically model several template-based and template free domains entangled together. The decomposition of a query protein into multiple regions can help solve the complicated domain architecture involving discontinuous segments and domain insertions.



```
-------MLGNDTVEIK-DGRFFI---DGYDA-IELAEK---------FGTPLYVMSEEQIKINYNRYI
EAFKRWEEETG--KEFIVAYAYKANANLAITRLLAKLGC----GADVVSGGELYIAKLSNVPSK----K
IVFNGNC-KTKEEIIMGIE---ANIRA-FNVDSISELILINETAKE-LGETANVAFRINPNVNPKTHPK
ISTGLKKNKFGLDVESGIAMKAIKMALE--MEYV-NVVGVHCHIGSQLTDISPFIEETRKVMDFVVELK
E-----E-GI-EIEDVNLGGGLGIPYYKDKQI----PTQKDLADAIINTMLKYKD--KVEMPNLILEPG
RSLVATAGYLLGKVHHIKETPVT----------------------------------------------
-------------------------------------------------KWVMIDAGMNDMMRP-AMYE
AY-HHIINCK----VKN----EKEVVSIAGGLCESSDVFGRDR-------ELD-KVEVGD---VLAIFD
VGAYGISMAN-NYNARGRPRMVLTS--KKG-V--FLIRERETYADLIAK--------------------
-----------------------------
```

```
MMDYGIDIWGNENFIIK-NGKVCINYEKKPAI-IDIVKELR----DDGYKGPLLLRFPHLIQKQIENIY
GNFNKARKEFGYKGGFNAVYPLKVNQYPGFVKNLVKLGKDYNYGLEAGSKAELLLAMAYNNEGA----P
ITVNG-F-KDRELINIGFIAAEMGHNITLTIEGLNEKEAIIDIAKERFKPKPNIGLRVRLHSAGVGI-W
AKSGGINSKFGLTSTE--LIEAVNLLKE--NKLLEQETMIHEHLGSQITEIHPLKKALNEAGNIYTELR
K-----M-GAKNLKAINLGGGLAVEYSQFKNEKSRNYTLREYANDVVEILKNIAEQKKDLEPDIFIESG
REVAANHAVLIAPVLELFSQEYAENKLILKKQNPKLID-ELYDLYKSI--KPSNALEYLHDSIDHLESI
LTLFDLGYVDLQDRSNAEILTHLITKKAILLLGDKQNPADLLAIQDEVQERYLVNFSLFQSMPDFWGLE
QN-FPIMPLD----RLD--EEPTRSASIWDITCDSDGEISYSKD---KPLFLH-DVDVEKENYFLGFFL
VGAYQEVLGM-KHNLFTHPTEAIISINEKG-YEVEGIIEAQSILDTLEDLDYDIHAIMDILNERISNSK
LVNDKQKKHILGELYLFLNDNGYLKSIGV
```

Figure 2. Domain architecture of CASP target T0547. (A) The experimental structure of T0547; (B) The region decomposition of target T0547 based on its sequence alignment with a template. T0457 was aligned with template 1TWI (chain A) by HHSearch. Red and green rectangles delineate the unaligned regions, which correspond to template-free domains 3 and 4 of T0547. 1TWI covers domains 1 and 2 of T0547.

The CASP9 target T0547 is a good example and illustrates this case. This protein has a very complicated domain architecture composed of four domains as illustrated in Fig. 2 (A). The first template-based domain has three discontinuous segments interrupted by two inserted domains – one template-based domain (i.e. domain 2) and one template-free domain (i.e. domain 3). The third fragment of domain 1 is joined by the fourth template-free domain. Traditional template-based modeling alone will fail on the two template-free domains and template-free modeling alone simply cannot handle such a large protein with such a complicated domain architecture. However the region decomposition approach used by recursive modeling can successfully identified the two template-based domains and template-free domains and compose them together. Fig. 3 shows that two disjoint fragments of T0547 were aligned with one template 1TWI (chain A), which is considered a certain region. The entire aligned region was modeled based on the structure of template 1TWI using template-based modeling, which is better than modeling the two disjoint parts separately using a traditional domain-cutting strategy. The latter would not be able to model the three disjoint fragments of the first domain. Thus the "region" concept used in recursive modeling is a broader modeling-oriented concept, which may correspond to a part of a domain, one domain or even multiple domains. The two unaligned /uncertain regions were modeled by the template-free method. Then the three components were composed into one model using the template-based modeling again.

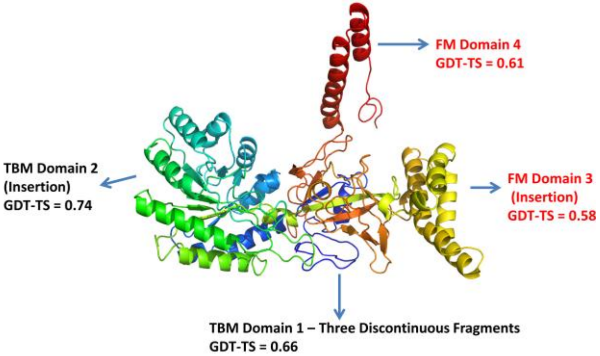


Figure 3. The model predicted by MULTICOM-REFINE for target T0547

The template-based domain 1 has three discontinuous segments flanked by template-based domain 2 and template-free domain 3. Domain 4 is a template-free domain. According to the GDT-TS scores, these four domains are among the top server models in CASP9.

It turned out that all four domains generated by MULTICOM-REFINE's recursive modeling procedure were ranked among the high-quality server models in CASP9. This example clearly demonstrates that the recursive modeling protocol can effectively decompose a large protein to reduce modeling complexity, resulting in better modeling quality. In addition to this example, we found that recursive modeling can also improve modeling on other targets composed of multiple template-based and template-free domains (e.g. T0543, T0571).

Case 2: recursive modeling improves ab initio modeling by starting from a very weak, largely incorrect template that contains a few fragments close to the native structure. For some very hard targets, only a number of highly uncertain templates can be found and these templates may only have partially correct template conformations (e.g. just one or two secondary structure elements). In this case, a template-free extension from the partially correct core secondary element(s) may still improve the quality of modeling. Target T0616 (107 residue long) is a TBM/FM example for which some analogous templates exists but are not likely be found or used by any server predictor. Our server MULTICOM-REFINE found a partial template covering about the last 80 residues, which at most has part of a helix matching the native structure. As shown in Fig. 4, starting from the partial central helix, the template-free modeling on the first 31 residues is able to extend the partially correct region to a structure closer to the native structure. The model is the best CASP9 server model submitted for this target.

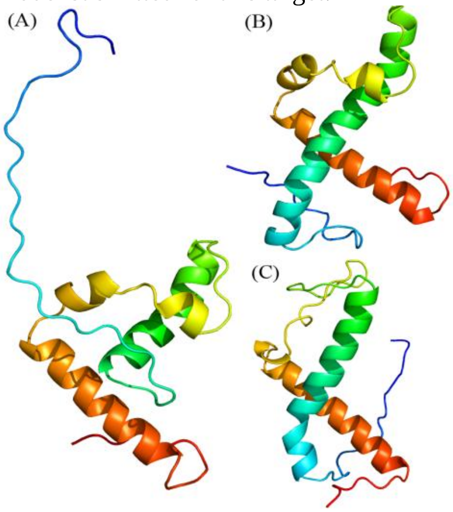


Figure 4. An example of recursive modeling on CASP target T0616. (A) a model generated solely by template-based modeling (GDT-TS = 0.34); (B) a model integrating both template-based and template-free modeling (GDT-TS = 0.39); (C) the native experimental structure.

Case 3: the recursive modeling procedure improves template-based modeling by fixing uncertain terminal regions. In this case, a large portion of a query protein can be aligned confidently to one or more partial templates, while leaving some parts of the query unaligned (e.g. front / back tails, partially unfolded internal helices / strands / loops). Recursive modeling will model template-based regions first and use them as additional restraints for template-free modeling to improve the modeling of unaligned / uncertain regions iteratively. The CASP9 target T0539 is a good example, where the whole target except for the ~20 N-terminal residues can be aligned to a few templates. Using the conformation core generated from the template information as restraints, the recursive modeling method in the MULTICOM server correctly reconstructed the loop-helix-loop structure of the uncertain front region and its interaction with the core as shown in Fig. 5. The GDT-TS score [22] was increased by 14% from 0.64 to 0.73. There are quite a few other similar CASP9 targets (e.g. T0568, T0574, T0592, T0593, T0596, T0597, T0632, and T0636) whose uncertain regions can be improved by the recursive modeling procedure. However, the improvement may not always be reflected in the GDT-TS scores according to CASP9 assessment because in CASP some uncertain regions are often removed before the assessment. Overall, according to our assessment recursive modeling generally improves the quality of modeling in this situation.

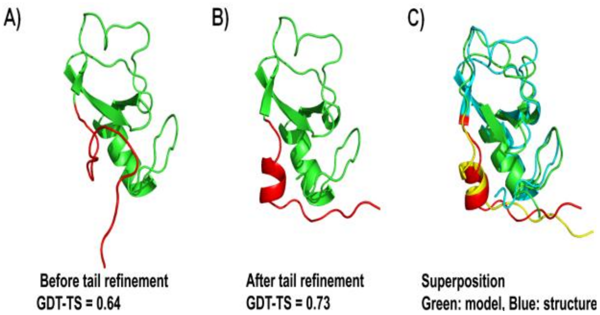


Figure 5. An example of applying recursive modeling to CASP9 target T0539 by the server MULTICOM-CONSTRUCT. (A) Template-based modeling is used to model the aligned / certain part (green) of T0539 while leaving the unaligned / uncertain region (red) free; the GDT-TS score of the model is 0.64. (B) Fragment assembly is used to model the uncertain region (red) while keeping template-based core (green) fixed; the GDT-TS score of the model composed of both template-based and template-free components is 0.73, 14% higher than the model in (A). (C) The superposition of the composed model (green + red) with the experimental structure (blue + yellow), showing that the uncertain tail (i.e. loop-helix-loop) is well packed with the template-based core in the model. Particularly the helix-helix interaction is reproduced in the composed model, which may not be possible by using either template-based modeling or template-free modeling independently.

The three typical cases above demonstrate that recursive modeling can readily integrate template-based and template-free modeling to improve protein structure prediction. It can also be easily implemented using existing or slightly modified alignment and model generation tools. However, it is worth pointing out that recursive modeling may not realize its best potential if region decomposition deviates too far away from the true boundaries between certain and uncertain regions. For instance, modeling one half of an uncertain region (e.g. template-free / ab initio domain) using template-free modeling and the other half by an incorrect template usually leads to a poor prediction as evidenced by our predictions for target T0534. In this situation, the GDT-TS score of the template-free region is often low (e.g. ~ 0.2). Nevertheless, the alignment-based region decomposition is generally robust to some residue shifts. A slightly more conservative region decomposition approach, that is to say only classifying very confident regions into certain regions at the beginning, seems to work better.

IV. CONCLUSIONS

In summary, we have described a general recursive protein modeling approach which can effectively integrate template-based and template-free modeling to improve protein modeling quality as demonstrated by its successful experiment in CASP9. The approach can often decompose a large, complicated modeling problem into several smaller and simpler modeling problems, which can be more readily addressed by synergistically integrating template-based and template-free modeling. Furthermore, the solutions to the smaller problems can be composed together to solve a larger, more complex modeling problem. In general, this divide and conquer strategy can improve both the quality and speed of protein structure modeling. According to this strategy, it is not necessary to divide protein modeling into two distinct approaches; instead, it can be viewed as a full spectrum of modeling based on an arbitrary percentage of template-based or template-free modeling. In the future we plan to improve the modeling process by designing more robust methods for the detection of certain and uncertain regions based on sequence alignments or the local quality of a model. We also plan to implement more effective ways to use template information to guide template-free modeling or to use template-free modeling to extend template-based regions.

ACKNOWLEDGEMENT

The work was partially supported by an NIH grant R01GM093123 to JC, a NLM fellowship to JE and the Shumaker fellowship to ZW.

REFERENCES

[1] Baker D, Sali A: Protein structure prediction and structural genomics. Science 2001, 294:93-96.

[2] Benson D, Boguski M, Lipman D, Ostell J, Ouellette B, Rapp B, Wheeler D: GenBank. Nucleic Acids Research 1999, 27:12-17.

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

[4] Cheng J: A multi-template combination algorithm for protein comparative modeling. BMC Structural Biology 2008, 8:18.

[5] Zhang Y, Skolnick J: Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences 2004, 101:7594-7599.

[6] Sali A, Blundell T: Comparative protein modelling by satisfaction of spatial restraints. 1994.

[7] Jones D, McGuffin L: Assembling novel protein folds from super-secondary structural fragments. Proteins: Structure, Function, and Bioinformatics 2003, 53:480-485.

[8] Simons K, Kooperberg C, Huang E, Baker D: Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of Molecular Biology 1997, 268:209-225.

[9] Yang Y, Zhou Y: Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 2008, 17:1212-1219.

[10] Boczko E, Brooks 3rd C: First-principles calculation of the folding free energy of a three-helix bundle protein. Science 1995, 269:393-396.

[11] Dill K: Dominant forces in protein folding. Biochemistry 1990, 29:7133-7155.

[12] Wang Z, Eickholt J, Cheng J: MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 2010, 26:882-888.

[13] Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997, 25:3389-3402.

[14] Benkert P, Tosatto S, Schomburg D: QMEAN: a comprehensive scoring function for model quality assessment. Proteins 2008, 71.

[15] Cheng J, Wang Z, Tegge A, Eickholt J: Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 2009, 77:181-184.

[16] Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A: Pcons: A neural-network-based consensus predictor that improves fold recognition. Protein Science 2001, 10:2354-2362.

[17] McGuffin L: Prediction of global and local model quality in CASP8 using the ModFOLD server. Proteins: Structure, Function, and Bioinformatics 2009, 77:185-190.

[18] Pettitt C, McGuffin L, Jones D: Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 2005, 21:3509-3515.

[19] Wang Z, Tegge A, Cheng J: Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins 2008, 75:638-647.

[20] Soding J, Biegert A, Lupas A: The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research 2005, 33:W244-W248.

[21] Fiser A, Sali A: Modeller: generation and refinement of homology-based protein structure models. Methods in enzymology 2003, 374:461-491.

[22] Zemla A: LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research 2003, 31:3370-3374.

[23] Cheng J., Wang Z., Eickholt J: Integrated prediction of protein tertiary structure by MULTICOM predictors. CASP9 proceeinding 2009, 9:173-175.