

Application of PROSPECT in CASP4: Characterizing Protein Structures with New Folds

Dong Xu*, Oakley H. Crawford, Philip F. LoCasio, and Ying Xu

Computational Biology Section, Life Sciences Division

Oak Ridge National Laboratory, Oak Ridge, TN 37830-6480, USA

Running title: Prediction Experiment by PROSPECT in CASP4

Key words: protein structure prediction, threading, fold recognition, new fold, CASP.

*Correspondence to: Dong Xu, Computational Biology Section, Oak Ridge National Laboratory, 1060 Commerce Park Drive, Oak Ridge, TN 37830-6480. Email: xud@ornl.gov. Tel: 865-574-8934. Fax: 865-241-1965.

Abstract In the Fourth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP4), we predicted all 43 targets using our threading application PROSPECT. PROSPECT guarantees to find an optimal alignment between a protein sequence and a structural fold for a general energy function with pairwise contact potential. For each prediction, it gives a reliability assessment based on a neural network approach. Additionally, PROSPECT has been added to the Genomic Integrated Supercomputing Toolkit (GIST), and is deployed on terascale computing resources. Structural predictions in CASP4 included three categories, i.e., comparative modeling, fold recognition, and prediction for structures with new folds. In the fold recognition category, PROSPECT correctly identified 8 folds out of 22 in total and finished the sixth in the total scores among 127 assessed groups. In the *new fold* category, it found important structural features for most targets, and its overall performance is among the best of all prediction methods. Our CASP4 performance demonstrates that PROSPECT is a powerful tool to quickly characterize structures with new folds, and it may provide useful structural restraints for *ab initio* prediction methods.

1 INTRODUCTION

The Fourth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP4), like the previous three CASP experiments [1, 2, 3], offered an excellent opportunity for a blind test of various protein structure prediction methods. We predicted and submitted structures for all of the protein targets (43 in total) given by CASP4, mainly using our threading program PROSPECT (PROtein Structure Prediction and Evaluation Computer Toolkit) [4]. We took this opportunity to systematically assess PROSPECT in all three categories of protein structure predictions, i.e., comparative modeling, fold recognition, and structure prediction for *new fold*, in which the structure of a protein does not share significant structural similarity to any known protein structure. The focus of this paper is to analyze and assess our structural predictions in the *new fold* category, as the CASP4 organizers invited us to write in the “New Fold Methods” section.

Structural prediction for proteins in the *new fold* category is much more challenging than predictions in the other two categories. In the categories of comparative modeling and fold recognition, one can use information of protein family or function to help structural prediction. For example, as more protein sequences become available, sequence comparison methods based on multiple-sequence profiles, such as PSI-BLAST [5], could be highly effective at recognizing native-like folds and thus achieve accurate alignments. Some manual predictions by human experts, most notably the predictions by Murzin *et al.* [2, 6], are also successful, since an expert can infer structural information from the known function of a protein. In the *new fold* category, however, neither the multiple-sequence profile nor human knowledge of protein function is effective for structure prediction. This is also the case for a significant portion of new proteins in a genome that do not have sequence

homologs to known protein sequences (so-called “orphan genes”).

New folds are typically predicted by *ab initio* prediction methods, which derive a structural model by optimizing an energy function based on the physical properties or statistical preferences of amino acids. Although some progress has been made in recent years [6], *ab initio* prediction methods typically require a huge amount of computing time and extensive human intervention, and thus these methods are far from suitable for large-scale applications. On the other hand, threading requires much less computing time or human intervention. Although by definition, a protein with a new fold does not have any known protein structure with the same fold in PDB [7] for threading to recognize, a close fold to the new fold is often available. A structural model built from the close fold can provide useful functional information and a good starting point for *ab initio* prediction methods to carry out further structural refinement. We will show in this paper that threading can provide a useful and scalable method to find close folds and to provide useful structural information for a protein with a new fold.

2 METHODS

The CASP4 contest is our second time participating in the CASP series. We employed PROSPECT in CASP3 [8], when the initial development of PROSPECT had just been completed. Significant improvements to PROSPECT have been made since then. Given the same number of targets (43 each in CASP3 and CASP4), we spent much less time with less manual intervention in CASP4 than in CASP3. We depended heavily on the PROSPECT outputs for all targets.

2.1 Protein Threading by PROSPECT

PROSPECT is a computer package for finding an optimal alignment between a protein sequence and a protein structural fold. A detailed description of PROSPECT can be found in [4]. Two unique features of PROSPECT are (1) that it guarantees to find the globally optimal sequence-structure alignment and does so in an efficient manner, when considering both alignment gap penalty and pairwise potential between residues that are spatially close [9]; and (2) that it allows a user to use various structural information (e.g., disulfide bonds and secondary structures) of the target as constraints in threading and guarantees to find the globally-optimal alignment under those constraints [10]. These constraints often improve the prediction accuracy. The global optimality in both cases is achieved using a divide-and-conquer algorithm [9], which avoids explicit examination of most of the search space that is mathematically proven not to contain the global optimum.

Several new features have been added to the PROSPECT system since CASP3. These new additions have significantly improved the performance of PROSPECT in both fold recognition and sequence-structure alignment. One of the main recent technical developments in PROSPECT, is the capability of assessing the prediction reliability for each threading alignment, by mapping the threading score to a value in the range of [0, 1]. The closer a mapped score to 1, the higher the probability that the prediction gives a correct fold recognition and a better sequence-fold alignment. We have accomplished this mapping using a neural network approach [11]. For details, we refer readers to reference [12]. The reliability assessment played an important role in our CASP4 predictions.

PROSPECT has been ported to the Genomic Integrated Supercomputing Toolkit (GIST) [13], which is a framework for large scale biological applications, that has been deployed on the teras-

cale supercomputing resources of the Center for Computational Sciences at Oak Ridge National Laboratory. The method used by GIST for MPP-PROSPECT (the massively parallel version), is to distribute alignments between a target sequence and template to different tasks, and it achieves a near linear speedup. The application now has the capability of threading hundreds of protein sequences against our template database of 2177 structures defined by FSSP [14] (release of May 2000). A Web interface has been built for running the PROSPECT program on the supercomputer, which is available at http://compbio.ornl.gov/structure/prospect_server/. An obvious advantage of running PROSPECT on a supercomputer was that we could carry out many runs for the same targets: (a) for different possible domains in the sequence, (b) with different energy functions (e.g., with or without secondary structure prediction), and (c) using different homologous sequences to the target as queries for PROSPECT. The consensus among different runs generally provides higher confidence in a prediction. For example, we threaded 6 sequences that are in the same Pfam [15] family of T0100, and all of them ranked the β -helix fold on the top. This gave us very high confidence of predicting T0100 to be in the β -helix fold, although there is no significant sequence similarity between T0100 and any of the known β -helix templates.

2.2 Prediction Protocol

A standard protocol for predicting each CASP4 target is described as follows. We first run PSI-BLAST [5] to see if there is any obvious homolog in PDB [7]. When there is no homolog in PDB, or the alignment is ambiguous, we carry out threading using PROSPECT. We collect structural and functional information from the SWISS-PROT database [16] and MedLine (<http://www.ncbi.nlm.nih.gov/PubMed/>). The information is used as a potential constraint during the threading process. In the *new fold* category, such information generally does not help.

Secondary structures are predicted using PHD [17] as possible inputs for threading. Possible domains in a target sequence are determined using *ProDom* [18]. Each possible domain and the whole sequence of the target are then submitted to PROSPECT. Based on the neural network assessment, we choose up to 5 templates to build 5 models. A confidence level (“high”, “medium”, or “low”) is assigned for each model according to its neural network assessment score. When a template is given a “high” confidence level, we build multiple models for the same template. The three-dimensional (3D) atomic structures are constructed using MODELLER [19], and ten structures are generated for each alignment. We then use structural assessment tools WHATIF [20] and PROCHECK [21] to evaluate the packing and backbone conformations, the inside/outside occupancies of hydrophobic and hydrophilic residues, and stereochemical quality of the predicted structures. We also check the consistency between the predicted secondary structures and secondary structure assignments of a predicted structure. Based on the structural assessment, we pick the best among the ten structures for each model. When none of the ten structures are acceptable, we adjust the alignment by changing threading parameters (e.g., the weight for gap penalties) and rebuild the model. We also visually check the structures and reject poor models. When the neural network assessment does not give a high confidence level, we also use structural assessments to rank the 5 models.

The protocol currently involves manual procedures, but we believe most of them can be automated. For example, we use the Enzyme Classification (EC) number of the target to give favorable treatment for the templates with the same EC number. If a template with the same EC number of the target ranks among the top 100, we choose it as one of the 5 models. Currently we use the Enzyme Structure Database (<http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html>) to locate the structures of the given EC number. We plan to construct a server to automatically look up the EC number of each template and retrieve a match of template with the same EC number as the

target, as part of an automated pipeline for structural prediction.

2.3 Post-prediction assessment

For the assessment of prediction accuracy, we used the evaluations provided in the CASP4 proceedings [22] and the CASP4 Web page (<http://predictioncenter.llnl.gov/casp4/>). We have also conducted additional comparisons between the experimental structures and the predicted structures using the sequence independent superposition program SARF [23]. SARF is particularly suitable to evaluate the predictions in the *new fold* category. The quality of the predicted structures in this category is generally poor, while SARF finds alignable portions between two structures with some tolerance of errors. In addition, SARF also gives the detailed alignment between two structures. The schematic diagrams of the proteins were made by Rasmol [24] (for Figures 2 and 4) and VMD [25] (for Figure 1 (b)).

3 RESULTS

3.1 Overview of Our CASP4 Predictions

There were 43 targets (T0086-T0128) for prediction in CASP4. At the time of writing this paper, the structures of eight targets have not been made available to the predictors so that we cannot assess them. The CASP4 organizers/assessors classified the targets into six categories: (1) comparative modeling (CM), (2) comparative modeling/fold recognition (CM/FR), (3) homologous fold recognition (FR/H), (4) analogous fold recognition (FR/A), (5) fold recognition/new fold (FR/NF), and (6) new fold (NF). When the structure of a target cannot be aligned to a single template, the target is divided into different domains, each as an individual prediction. In total, we have 47

predictions for 35 targets. Table 1 shows a summary of our CASP4 predictions for *model 1* (the most confident one among the five submitted models).

In the comparative modeling category (CM and CM/FR), PROSPECT used the correct templates (folds) in *model 1* for every target. In the CM category (8 in total), like some other sequence-sequence alignment methods, PROSPECT alignments are generally “perfect” (the same as the structure-structure alignment between the target and the template). This is confirmed by the large numbers of alignable residues and small RMSDs for “ C_α -set-2Å” and “ C_α -set-6Å” in Table 1. In the CM/FR category (7 in total), the alignments are typically ambiguous using different sequence-sequence alignment methods. PROSPECT alignments are either “perfect” or “good”, with the exception of target T0103, where extensive alignment gaps exist. Basically, all the predictions in this category were done with little human intervention. The difference between our model and the best model in CASP4 is typically small [22].

The fold recognition category includes FR/H and FR/A. Table 1 lists the Sippl’s evaluation scores [26], where a score of 0 indicates an incorrect model and a score of 1-4 shows a different level of structural relationship between the model and the experimental structure: 1 for having structural similarity, 2 for finding correct fold, 3 for finding correct fold and having a medium alignment quality, and 4 for finding correct fold and having a good alignment quality. Our performance in this category in CASP4 has been substantially improved compared with CASP3 [8]. PROSPECT recognized correct folds (with 2 points or more) in *model 1* for 8 predictions out of 22 in total (4 of which are not listed in Table 1 since we do not have the structures) [26]. In the FR/H category (8 in total), we have obtained “reasonable” alignments (with 3 points or more) for four targets. Even in the FR/A category (10 in total), where threading methods typically fail, we recognized the correct fold for two targets, although the alignments are poor.

The *new fold* category includes FR/NF and NF. For FR/NF, a target has weak structural similarity to a known protein structure in PDB. In this case, the CASP4 target and its closest structure in PDB have some similar secondary structures in the same arrangement and with similar topological connections, but it is difficult to say whether the two proteins belong to the same fold due to many other differences. Therefore, Sippl’s evaluation score for the fold recognition category is too strict to assess the prediction quality in the *new fold* category. A widely used assessment is the RMS/Coverage graph [27], as shown in Figure 1 (a). A “good” model is defined to have a large number of residues that can be superimposed on the experimental structure for a relatively low RMSD, and thus represented by lines closer to the x-axis than almost all the other predictions. We consider a prediction “reasonable” if it is better than at least 70% of other predictions; otherwise, it is labeled “poor”. As shown in Figure 1 (a) and listed in Table 1, the qualities of most of our predictions are either “good” or “reasonable”. By visualizing our “good” and “reasonable” predictions, we found that they all have captured some non-trivial features of the experimental structures, and appear significantly better than most models submitted by other groups.

3.2 Prediction Examples in the New Fold Category

The “good” models (T0087_1, T0094, T0105, and T0124), in the *new fold* category, PROSPECT generally covers significant portions of the target’s structure “correctly”. Here, we use T0087_1 (pyrophosphatase) as an example to illustrate how we selected templates. By running PROSPECT and the reliability assessment of the prediction, we found 5 templates that met the criterion of ranking top 50 using raw scores and top 10 using reliability assessment scores (a criterion typically giving the best chance of being the native-like fold). They are 1wod (13, 2), 1cyd-A (27, 3), 1enp

(47, 5), 2cmd (39, 7), and 1qpz-A (28, 8), where the first number in a bracket indicates the rank of raw score and the second number shows the rank of reliability assessment score. The reliability assessment scores indicate that these templates have 40-50% probability of being the native-like fold. Interestingly, all 5 templates are structurally alignable with each other with a Z-score of at least 4.3 in FSSP. This observation convinced us that T0087_1 must share some common structural features with the 5 templates. Then we generated structural models for all 5 templates and ran WHATIF to check the quality of the models. The model derived from 1qpz-A has the best quality. Hence, we chose this model as our *model 1*. It turned out that we predicted “correctly” the portion 1-116 in *model 1*. We searched the experimental structure of T0087 against PDB using the DALI server [28] and located its closest structure, 1moq (glutamine amidotransferase). The template 1moq can be aligned to 2-133 in T0087_1 (2-194), i.e., 61 residues in the compact domain T0087_1 are not alignable. Hence, our prediction achieved what is close to the limit that threading can offer in this case. In fact, 1moq and the template we used, 1qpz-A (purine nucleotide synthesis repressor), are structurally alignable in the region that is structurally similar to T0087_1. Figure 1 (b, 87_1) shows that the β -strand and helices of T0087_1 are connected in the correct topology. The main difference is the position of the helix to the right side of the experimental structure. Another example of a “good” prediction is our *model 1* for T0124 (phospholipase C beta C-terminus). A native-like fold of T0124 is not available in PDB, but PROSPECT was still able to find a template, 1dg3-A (guanylate-binding protein 1), which is similar to the experimental structure of T0124. The overall shape and the major topology between helices are well characterized in this prediction, although the detailed arrangement is inaccurate.

Our “reasonable” models (T0087_2, T0089_2, T0096_2, T0106, and T0120_1) all have some interesting features. For example, Figure 2 shows a comparison between our *model 1* and the

experimental structure for T0096 (FadR). The prediction captures a very similar fold to FadR. The secondary structure elements are basically predicted correctly, and connections between some of the adjacent helices on the sequence are also good. There are 60 (about half of the total number of residues) structurally alignable residues between the model and the experimental structure, although parts of the alignment are off. Hence, it does not look outstanding in the RMS/Coverage graph in Figure 1 (a). To understand why PROSPECT can capture the structural features on a fairly consistent basis, we have calculated the number of structurally alignable residues (defined by SARF) between the experiment structure and the template *versus* the rank of a template’s raw score for T0096_2. The template for our *model 1* of T0096_2, 2vhh-A (hemoglobin), is ranked number 1 by PROSPECT and has 63 SARF-alignable residues. As the threading rank goes down, generally the number of structurally alignable residues between the target and the template decreases, as shown in Figure 3. Most templates ranked in the bottom 1/3 have no SARF-alignable residues. This shows that PROSPECT’s ranking is quite consistent with the number of structurally alignable residues, i.e., structural similarities, even when the native-like fold of the target is unavailable in the template database.

Our “poor” predictions (T0097, T0098, T0115_2, and T0116_3) did not realize the potential of threading methods. We use T0097 as an example to show how we failed. Unlike the case in T0087_1, the top template hits using either raw scores or reliability assessment scores did not show any common structural pattern. No template had higher than 30% probability of being the native-like fold. The only template that brought our attention was 1lis, which ranked 6 using raw scores and 19 using reliability assessment scores. We found that 1lis was a PSI-BLAST hit for T0097. Although the confidence level of the PSI-BLAST hit was very low (with an expectation value of 2.6), given no other hints, we selected 1lis as the template for *model 1*. Both *model 1* and the

experimental structure of T0097 have similar secondary structures, but their structural folds are quite different. There were much more similar folds to T0097 than 1lis among the 2177 templates we used, e.g., 1nsg-B (50, 47) and 1dd5-A (17, 36), where the two numbers in a bracket give the ranks of raw scores and reliability assessment scores, respectively.

An interesting feature among the “poor” predictions is that the predicted models, albeit with wrong folds, often have some structural similarities to the experimental ones in the correct classes. For example, the experimental structure of T0115 (homoserine kinase) and our *model 1* both are all- α proteins, and some segments are structurally alignable, as shown in Figure 4. For T0098 (Spo0A), we failed to place a better template (used in *model 2*) as the top choice (*model 1*). The *model 2* of T0098 used the template 1qgr-A (importin β subunit) and has 60 SARF-alignable residues to the experimental structure (with 119 residues in total). As shown in Figure 1 (b), the model and the experimental structure, for the alignable portions, have similar secondary structure elements and topology, although the threading alignment is inaccurate. One reason why 1qgr-A was not placed number 1, was that its structural model is not compact, as the N-terminal helix does not interact with the rest of the protein. This problem might be solved if we could carry out *ab initio* refinement to the threading model.

4 Discussion

4.1 What Went Right

We have made substantial progress in PROSPECT since CASP3. Both CASP official assessments and our own assessments indicate that our CASP4 performance is much better than our CASP3 performance. PROSPECT has consistently given good performance for many CASP4 targets in

all prediction categories. Our alignments in the comparative modeling category are basically all “perfect” or “good”. Eight folds in the fold recognition category were identified correctly. In the *new fold* category, PROSPECT has often identified important structural features, and its overall performance is among the best of all the presented prediction methods. This demonstrates that PROSPECT can be a powerful tool to quickly characterize structures with new folds.

We believe our success is mainly due to two reasons: (1) PROSPECT guarantees to find the globally-optimal sequence-structure alignment with pairwise contact potential; (2) it has a reliability assessment for a threading alignment. Our study shows that pairwise contact potential is very important in recognizing native-like folds. In a test set that does not have significant sequence similarity between a target and any template, PROSPECT has recognized the native-like folds for 69% of the targets with pairwise contact potential, compared with 55% of targets without pairwise contact potential [4]. Our CASP4 experience reconfirms that having a mathematically rigorous algorithm to find the globally optimal solution of a general energy function is a good approach for protein structure prediction. Furthermore, the reliability assessment makes better use of the information content in an energy function than raw threading score, by decreasing the background bias of different templates. Our reliability assessment is especially useful to recognize the folds with weak structural similarities to the target [12], and hence, it is more important in the *new fold* category.

4.2 What Went Wrong and What Can Be Improved

While some success was achieved by our threading method in CASP4, we have learned a number of lessons from some of the outstanding predictions by other groups. Although we identified correct folds for 8 targets in the fold recognition category, our alignment quality and structural refinement

are not as good as a few other groups, for some targets. In the *new fold* category, threading noticeably has its limitations. In particular, since no native-like fold is available for a target in this category, the predicted structure based on a threading alignment can only agree with the experimental structure partially. The CASP contests have shown that the sequence profile method, threading, and *ab initio* prediction each have their own strengths and limitations. We have started using sequence profiles in threading, and will continue to improve the alignment accuracy, by applying sequence profiles more effectively. More importantly, a major weakness in our current prediction approach is the lack of good structural refinement after the threading alignment.

Clearly a more suitable refinement is needed for building 3D structures, after PROSPECT produces sequence-structure alignments. In CASP4, we used MODELLER to generate a 3D structure after getting the threaded alignment. MODELLER works reasonably well when a target and its template have significant sequence similarity, but it was not designed for a target with no significant sequence similarity to its template. We have found that in several cases, other groups achieved higher prediction accuracy mainly because a better refinement was performed on the regions which are unalignable to the templates. For example, Baker's group did a better job at refining the structural models for T0100 than we did using MODELLER [6]. This is a more noticeable problem in the *new fold* category than in the fold recognition category, since the structural difference between a target and the template in the former is much larger.

As we have shown, our threading method can retrieve structural information for much longer segments than the short fragment (typically 9 residues) in the mini-threading method that Baker's group used [29]. We expect that significant structural features of long segments, captured in the threading alignment, would provide further help in building 3D models. To facilitate using structural features in threading alignments for *ab initio* prediction or refinement, we need a capacity

of assessing the prediction accuracy for each local region. Our current reliability assessment provides only a global assessment for the entire alignment. For example, our *model 1* of target T0087 got a “medium” confidence overall, but we have no good way to determine the information that segment 1-116 of the model is the alignable portion of the experimental structure. A capacity of local reliability assessment may tell us which segments are reliable, and hence, their structural information can be used in *ab initio* prediction. Another improvement that we are currently considering is to use the population distribution of local structures in predicted models, in a similar way used by Baker’s group [29]. Many top hits in threading may have some alignable structural segments, as shown in Figure 3. It is possible to use them together by clustering conformations in a given segment, and restraints can be weighed according to the occurrence of a conformation. Such a method may provide a more effective way to define the global fold than the mini-threading method, which only uses restraints derived from short fragments.

4.3 Availability of Software

PROSPECT, including its database, energy function, and executables, is freely available to academic users. Further details can be found at <http://compbio.ornl.gov/structure/prospect/>.

Acknowledgments

We thank the CASP4 organizers and assessors as well as the groups who provided prediction targets and their structures. In particular the extensive evaluation data provided by the organizers and assessors are very helpful. We also thank our colleague Doug Hyatt for a critical reading of this manuscript. Part of the computation was carried out using the resources at the Center for Computational Sciences at Oak Ridge National Laboratory. This research was sponsored by the

Office of Health and Environmental Research, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.

References

- [1] CASP. Protein structure prediction issue. *Proteins: Struct. Funct. Genet.*, 1995, 23:295–462.
- [2] CASP. Protein structure prediction issue. *Proteins: Struct. Funct. Genet.*, 1997, Suppl. 1. 29:1–230.
- [3] CASP. Protein structure prediction issue. *Proteins: Struct. Funct. Genet.*, 1999, Suppl. 3. 37:1–237.
- [4] Xu, Y. and Xu, D. Protein threading using PROSPECT: Design and evaluation. *Proteins: Struct. Funct. Genet.*, 2000, 40:343–354.
- [5] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 1997, 25:3389–3402.
- [6] CASP. Protein structure prediction issue. *Proteins: Struct. Funct. Genet.*, 2001, Suppl. 4.
- [7] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, 1977, 112:535–542.
- [8] Xu, Y., Xu, D., Crawford, O. H., Einstein, J. R., Larimer, F., Uberbacher, E. C., Unseren, M. A., and Zhang, G. Protein threading by PROSPECT: A prediction experiment

in CASP3. *Protein Eng.*, 1999, 12:899–907.

- [9] Xu, Y., Xu, D., and Uberbacher, E. C. An efficient computational method for globally optimal threading. *J. Comp. Biol.*, 1998, 5(3):597–614.
- [10] Xu, Y., Xu, D., Crawford, O. H., and Einstein, J. R. A computational method for NMR-constrained protein threading. *J. Comp. Biol.*, 2000, 7:449–467.
- [11] Jones, D. T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, 1999, 287:797–815.
- [12] Xu, Y., Xu, D., and Olman, V. A practical method for interpretation of threading scores: An application of neural network. *Statistica Sinica*, 2001. Invited publication.
- [13] Locascio, P. Genomic integrated supercomputing toolkit. Oak Ridge National Laboratory, 2001.
- [14] Holm, L. and Sander, C. Mapping the protein universe. *Science*, 1996, 273:595–602.
- [15] Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, F. D., and Sonnhammer, E. L. L. Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucleic Acids Research*, 1999, 27:260–262.
- [16] Bairoch, A. and Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Research*, 1999, 27:49–54.
- [17] Rost, B. and Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 1993, 232:584–599.

- [18] Corpet, F., Gouzy, J., and Kahn, D. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Research*, 1999, 27:263–267.
- [19] Sali, A. and Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 1993, 234:779–815.
- [20] Vriend, G. WHAT IF: a molecular modelling and drug design program. *J. Mol. Graphics*, 1990, 8:52–56.
- [21] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 1993, 26:283–291.
- [22] Moulton, J., Hubbard, T., Fidelis, K., and Zemla, A. Fourth meeting on the critical assessment of techniques for protein structure prediction. December 3-7, Asilomar Conference Center, California, 2000.
- [23] Alexandrov, N. N. SARFing the PDB. *Protein Eng.*, 1996, 9:727–732.
- [24] Sayle, R. A. and Milner-White, E. J. RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences*, 1995, 20:374–376.
- [25] Humphrey, W. F., Dalke, A., and Schulten, K. VMD – visual molecular dynamics. *J. Mol. Graphics*, 1996, 14:33–38.
- [26] Sippl, M. CASP4 fold recognition assessment. *Proteins: Struct. Funct. Genet.*, 2001, Suppl. 4.

- [27] Hubbard, T. J. RMS/coverage graphs: A qualitative method for comparing three-dimensional protein structure predictions. *Proteins: Struct. Funct. Genet.*, 1999, 37:15–21.
- [28] Holm, L. and Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 1993, 233:123–138.
- [29] Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 1997, 268:209–225.

Table 1. Summary of Our Model-1 Predictions

Target	Domain	Category	Template	Length (a.a.)	SARF (a.a./Å)	C_{α} -set-2Å (a.a./Å)	C_{α} -set-6Å (a.a./Å)	Overall Quality	
T0099	1-240	CM	1nyf	56	43/2.43	25/1.43	48/2.85	perfect alignment	
T0111		CM	1pdz	429	414/1.15	382/0.85	424/1.41	perfect alignment	
T0113		CM	2hsd-A	348	223/1.56	201/0.87	234/1.97	perfect alignment	
T0121_1		CM	1b0u-A	372	173/2.14	122/1.24	165/2.11	perfect alignment	
T0122		CM	2tys-A	241	216/1.63	182/1.13	232/2.37	perfect alignment	
T0123		CM	1b0o	160	120/1.99	89/1.12	143/3.01	perfect alignment	
T0125		CM	3lyn-A	137	107/2.54	70/1.22	120/2.90	perfect alignment	
T0128		CM	1ar5-B	211	198/0.93	191/0.67	201/1.28	perfect alignment	
T0089_1	7-85/167-199	CM/FR	3hsc	187	86/1.86	45/1.18	70/2.87	good alignment	
T0089_3	200-390	CM/FR	3hsc	191	123/2.37	17/0.48	26/3.88	good alignment	
T0090_2	58-209	CM/FR	1tum	152	85/2.69	20/1.37	79/4.05	good alignment	
T0092	543-765	CM/FR	1xva-A	227	148/2.28	58/1.52	138/3.34	good alignment	
T0103		CM/FR	1sbh	368	209/2.09	16/0.94	54/4.57	poor alignment	
T0112		CM/FR	1teh-A	348	294/2.24	162/1.25	318/2.89	perfect alignment	
T0117		CM/FR	5tmp-A	197	135/2.26	68/1.28	148/3.43	perfect alignment	
T0096_1		6-79	FR/H	1bia	74	57/2.49	31/1.21	56/3.06	4
T0100		241-372	FR/H	1pcl	342	177/2.53	68/1.17	126/3.17	4
T0101	FR/H		1dbg-A	400	195/2.10	44/1.15	76/4.28	2.5	
T0109	FR/H		1xwl	182	90/2.50	15/1.11	35/4.25	2	
T0110	FR/H		1b3t-A	95	48/2.63	18/0.84	33/3.45	0	
T0116_4	543-765		FR/H	1ngl	223	70/2.79	21/1.16	44/4.51	1
T0121_2	241-372		FR/H	1b9m-A	132	94/2.13	68/1.06	95/2.33	3.5
T0127_1	19-266		FR/H	1a5t	246	105/2.53	43/1.43	118/3.41	3
T0102	267-350		FR/A	1imp	70	31/2.37	15/1.10	30/3.60	2
T0107		FR/A	1eut	188	46/2.69	19/1.18	42/4.00	1	
T0108		FR/A	2nlr-A	179	69/2.84	11/1.24	35/4.47	1	
T0114		FR/A	1qfm-A	87	33/2.71	10/1.14	24/4.80	0	
T0115_1		5-168	FR/A	1zin	164	46/2.34	16/0.64	33/4.37	0
T0116_1		1-128	FR/A	1a53	128	44/2.98	15/1.08	36/4.43	0
T0116_2		129-249	FR/A	1a53	121	55/2.81	14/1.58	38/3.85	1
T0120_2		117-204	FR/A	1quu-A	88	25/1.87	31/1.12	68/3.79	2
T0126		FR/A	2cbl-A	162	27/2.77	17/1.33	31/3.80	0	
T0127_2		FR/A	1a5t	84	37/2.18	19/1.47	41/3.40	0.5	
T0087_1		2-194	FR/NF	1qpz-A	193	74/2.72	24/1.34	70/3.93	good (1.5)
T0087_2		195-310	FR/NF	1qpz-A	116	40/2.71	16/0.97	41/3.77	reasonable (0.5)
T0089_2	86-166	FR/NF	3hsc	81	43/2.65	26/1.03	35/3.07	reasonable (0)	
T0090_1	1-57	FR/NF	-	57	-	-	-	no coordinates	
T0094	80-227	FR/NF	1tyf-A	177	28/2.67	21/0.80	41/3.54	good (1)	
T0096_2		FR/NF	2vhb-A	148	60/2.44	22/0.82	41/3.92	reasonable (0)	
T0097		FR/NF	1lis	105	35/2.78	17/1.15	34/3.78	poor (0)	
T0105		FR/NF	1d7q-A	94	25/1.83	12/1.28	36/4.11	good (0)	
T0106		FR/NF	1b3u-A	125	31/2.75	17/1.27	42/3.89	reasonable (0)	
T0115_2		169-300	FR/NF	1zin	132	37/2.56	13/0.76	32/3.87	poor (0)
T0098	250-542	NF	1qgr-A	119	31/2.87	16/0.81	30/4.16	poor (0)	
T0116_3		NF	1d9x-A	293	37/2.76	17/0.97	31/3.67	poor (0)	
T0120_1		1-116	NF	1quu-A	116	24/2.34	15/1.46	31/4.20	reasonable (0)
T0124		NF	1dg3-A	242	96/2.32	41/1.24	97/3.75	good (0.5)	

“Target” gives the CASP4 targets with available experimental structures, where the suffix indicates the domain number. “Domain” gives the residue range of a target’s domain. “Length” is the number of amino acids (a.a.) in the experimental structure. “SARF” shows the number of alignable residues and their RMSDs for the structurally alignable portion between the experimental structure and the predicted one in a sequence independent superposition using SARF. “ C_{α} -set-2Å” and “ C_{α} -set-6Å” represent the largest set of residues that can fit by the sequence dependent superposition under 2 Å and 6 Å, respectively, together with their RMSDs. The Sippl’s evaluation scores of the overall quality in FR/H and FR/A, and in the brackets in FR/NF and NF, are provided by the CASP4 assessors [26]. The values for “ C_{α} -set-2Å” and “ C_{α} -set-6Å” were taken from the CASP Web page at <http://predictioncenter.llnl.gov/>. Our *model 1* for T0090 did not include the domain 1-57, since we believed it was a new fold.

Figure legends:

Figure 1. Summary of our performance in the *new fold* category. The 16 plots in (a) show the RMS/Coverage graphs, where the x-axis (from 0 to 100%) is the maximum percentage of residues in a predicted structure that can be superimposed on the experimental structure for a given RMSD (root mean square deviation) threshold (represented along y-axis, from 0 to 10 Å). The graphs were taken from the CASP Web page available at <http://predictioncenter.llnl.gov/>, with axis labels omitted. Blue lines represent our *model 1*; cyan lines represent other models of our predictions; brown lines represent models submitted by other groups. For the assessment of domains, only *model 1* is included, since the evaluation of other models are not available from the CASP4 Web page. The structures of targets T0086, T0091, and T0104 are available to the CASP4 assessors but not predictors. The three sets of protein structure pictures at the bottom in (b) show comparisons between the experimental structures (left) and our predicted ones (right), where the cylinders indicate α -helices and the strands indicate β -sheets. The ribbons in target 98 indicate portions that are not structurally alignable between the experimental structure and the predicted one. The red to blue shows the N-terminus to the C-terminus.

Figure 2. A comparison between the experimental structure (left) and our *model 1* (right) for T0096_2.

Figure 3. The number of SARF-alignable residues between the experiment structure of T0096_2 and the template *versus* the rank of a template.

Figure 4. A comparison between the experimental structure (left) and *model 1* (right) for T0115_2. The thick ribbons show structurally SARF-alignable regions between the two.

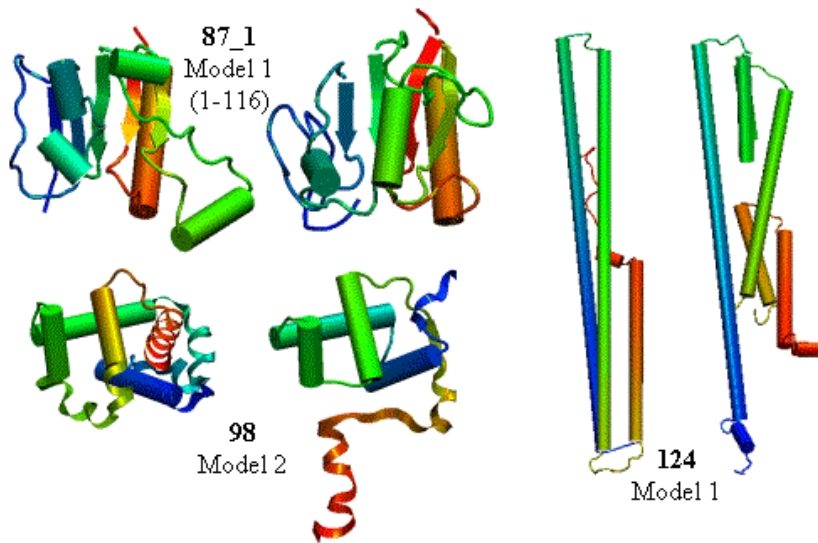
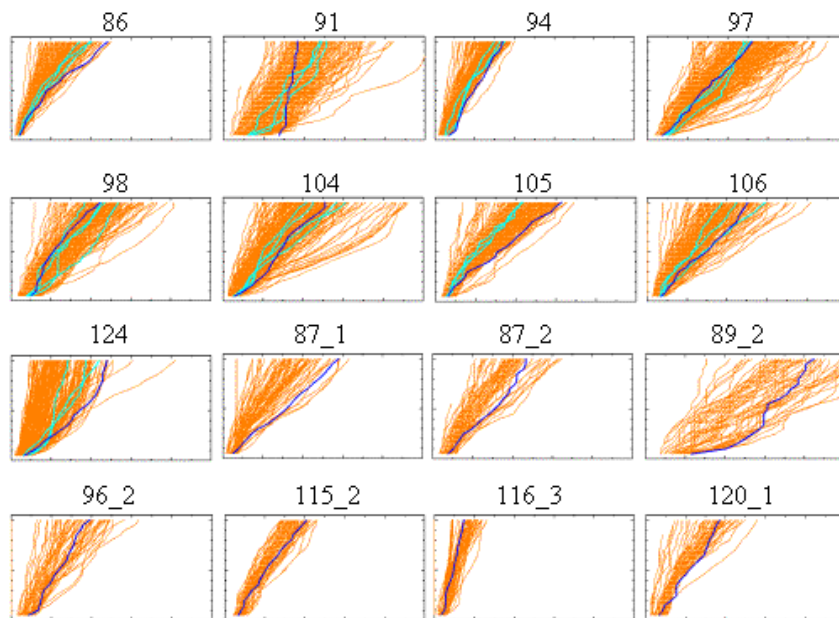


Figure 1:

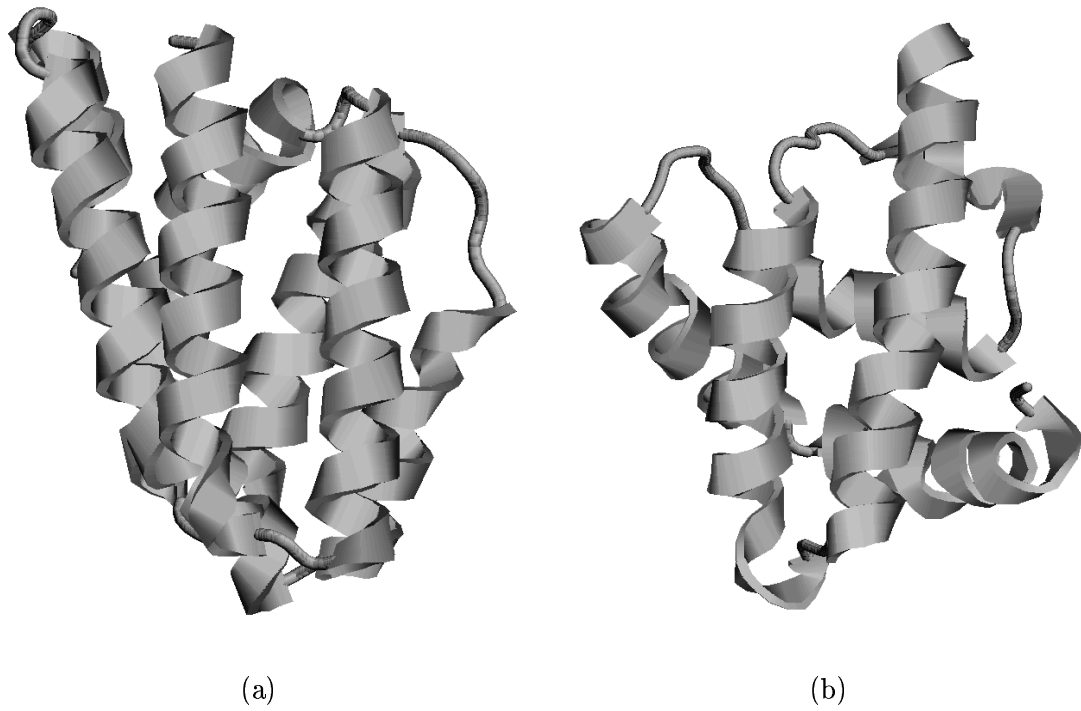


Figure 2

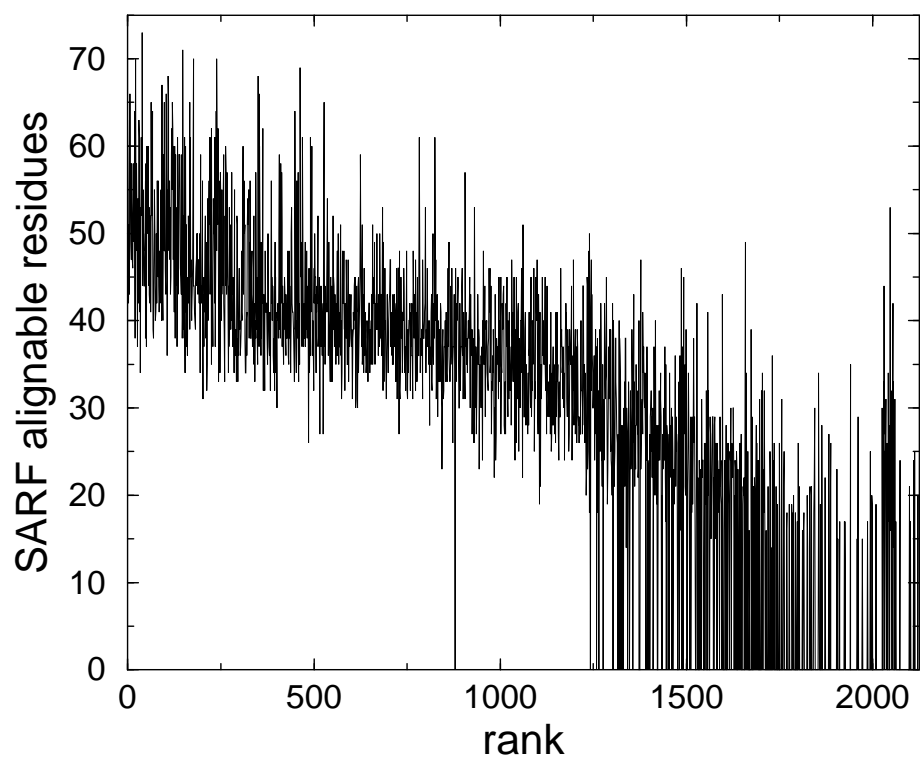


Figure 3

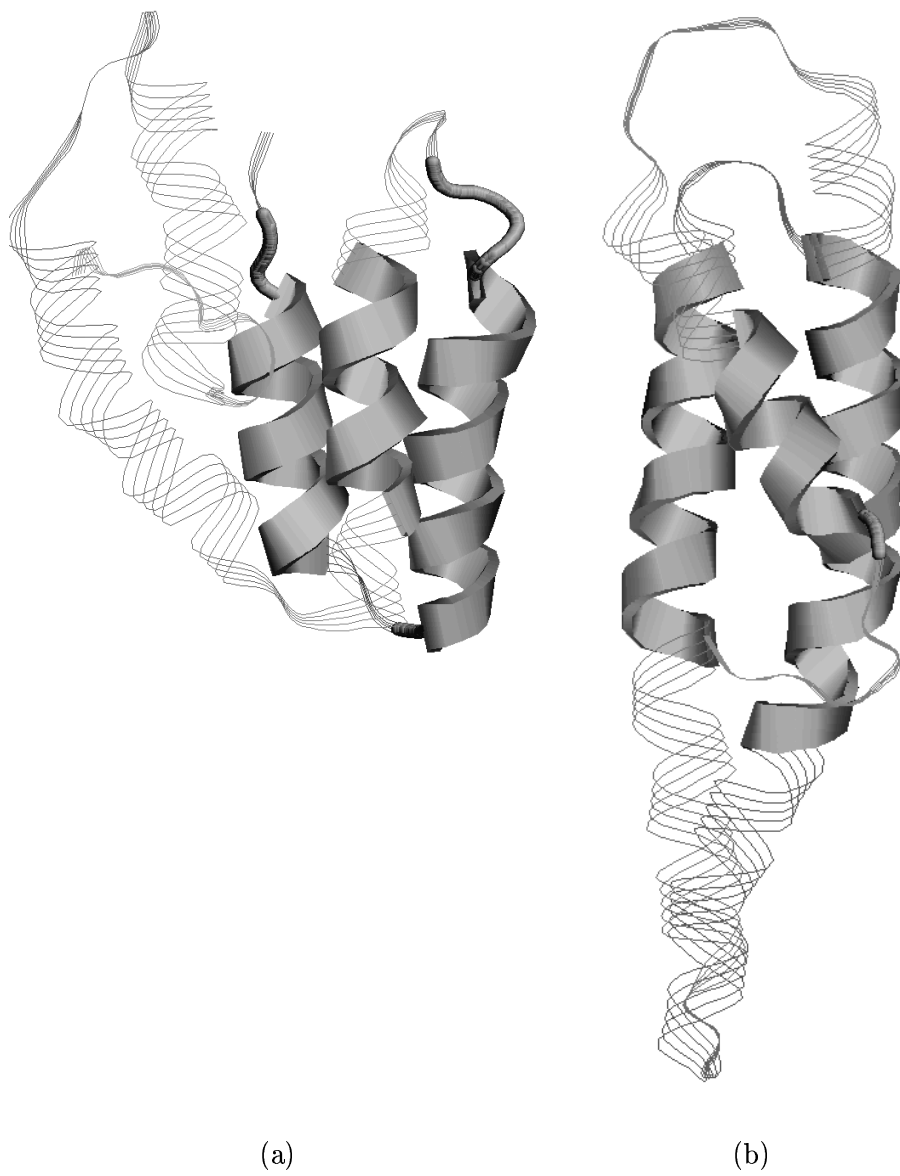


Figure 4