BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 7 2010, pages 882–888
doi:10.1093/bioinformatics/btq058

Structural bioinformatics

Advance Access publication February 11, 2010

MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8

Zheng Wang[1], Jesse Eickholt[1] and Jianlin Cheng[1,2,3,*]

[1]Department of Computer Science, [2]Informatics Institute and [3]C. Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Protein structure prediction is one of the most important problems in structural bioinformatics. Here we describe MULTICOM, a multi-level combination approach to improve the various steps in protein structure prediction. In contrast to those methods which look for the best templates, alignments and models, our approach tries to combine complementary and alternative templates, alignments and models to achieve on average better accuracy.

Results: The multi-level combination approach was implemented via five automated protein structure prediction servers and one human predictor which participated in the eighth Critical Assessment of Techniques for Protein Structure Prediction (CASP8), 2008. The MULTICOM servers and human predictor were consistently ranked among the top predictors on the CASP8 benchmark. The methods can predict moderate- to high-resolution models for most template-based targets and low-resolution models for some template-free targets. The results show that the multi-level combination of complementary templates, alternative alignments and similar models aided by model quality assessment can systematically improve both template-based and template-free protein modeling.

Availability: The MULTICOM server is freely available at http://casp.rnet.missouri.edu/multicom_3d.html

Contact: chengji@missouri.edu

Received on October 30, 2009; revised on February 2, 2010; accepted on February 8, 2010

1 INTRODUCTION

Knowing protein tertiary structure is useful for determining protein–protein interactions, protein function and evolution, and designing drugs. At present, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the two most commonly used experimental methods employed to determine protein structure. However, both methods are far too expensive and time consuming to be used to process the millions of proteins produced by high-throughput genome sequencing (Jaravine *et al*., 2006; Lattman, 2004; Service, 2005). Computer-aided protein structure prediction, in contrast, is less expensive, much faster, and able to generate protein structures on a large scale. As a result, computational protein structure prediction has received much attention in recent years, particularly from those working in computer science, chemistry, molecular biology, and molecular physics, and their efforts have led to steady progress in the area (Kryshtafovych *et al*., 2005, 2007, 2009a).

*To whom correspondence should be addressed.

Recently, the Eighth Critical Assessment of Techniques for Protein Structure Prediction, 2008 (CASP8) (Moult *et al*., 2009) assessed state-of-the-art protein modeling techniques in two categories: template-based modeling (TBM) and template-free modeling (FM). TBM deals with proteins for which suitable templates can be found. It can also be referred to as comparative modeling or fold recognition depending on the availability of sequentially or structurally related proteins. FM deals with proteins for which no suitable templates can be found. This category includes both fragment-based modeling (Simons *et al*., 1997) and purely '*ab initio*', or '*de novo*' modeling, in which predictions are made based solely on chemical and physical principles (Hinds and Levitt, 1994; Sternberg and Thornton, 1978).

TBM typically contains five steps (Baker and Sali, 2001; Cheng, 2008; Zhang, 2008; Zhang and Skolnick, 2005). The first step is to identify templates that have a similar structure to the protein to be modeled (target). Once templates have been selected, the target protein is then aligned with the templates. At this point, models can be built from the alignments and the structural information of each template. After model generation, the models are evaluated and refined.

Here we present a multi-level combination approach to improve protein structure prediction during all facets of TBM. Our approach first attempts to combine complementary templates, alignments, and model generation methods to produce a number of alternative models. It then uses a novel model combination process guided by model quality evaluation (Cheng *et al*., 2009; Wang *et al*., 2008) to refine the models. Five fully automated servers (MULTICOM-CLUSTER, MULTICOM-REFINE, MULTICOM-CMFR, MULTICOM-RANK and MUProt) and one human predictor (MULTICOM) implemented various forms of this approach and participated in CASP8. The CASP8 results show that the multi-level combination approach is effective for the full spectrum of protein modeling, including high-accuracy TBM, hard TBM and FM.

2 METHODS AND IMPLEMENTATION

2.1 A multi-level combination pipeline for protein structure prediction

Our multi-level combination pipeline (Fig. 1) for protein structure prediction is generally comprised of five steps: (i) template identification and ranking, (ii) multi-template combination, (iii) model generation, (iv) model evaluation and (v) model combination and refinement. More specifically, our pipeline first uses a set of fold recognition methods to generate several lists of templates, each one ranked by one of the fold recognition methods employed. Then alignments between the target and one or more of the top templates in

each of the ranked lists are greedily combined into a multiple sequence alignment (Cheng, 2008). This multiple sequence alignment along with the structure for each template are fed into model generation tools which construct models. All the models are then evaluated and ranked by model quality assessment tools. Finally, globally similar models and/or locally similar model fragments are combined using a novel model combination algorithm to generate refined models. These refined models are the end product of our pipeline.

2.2 MULTICOM-CLUSTER

2.2.1 Template identification and ranking To identify and rank templates, we used profile–profile alignments, profile–sequence alignments and a machine learning approach (Cheng and Baldi, 2006). In order to generate profiles, PSI-BLAST, a profile–sequence local alignment method, was used to search a query protein sequence against the NCBI non-redundant protein sequence database to build three different kinds of sequence profiles. These include the position specific scoring matrix (PSSM) of PSI-BLAST, the hidden Markov model (HMM) of hhsearch (Soding, 2005) and the profile of COMPASS (Sadreyev and Grishin, 2003). The PSSM profile, HMM and COMPASS profile were searched against our in-house template sequence database, template HMM database and template COMPASS profile database to identify homologous templates from the output generated by PSI-BLAST, hhsearch and COMPASS, respectively. The query-template alignments generated by PSI-BLAST, hhsearch and COMPASS were kept in three different sets and ranked according to *E*-value. In addition, SPEM (Zhou and Zhou, 2005), a global profile–profile alignment tool, was used to align the query with the top 10 templates found by a sensitive machine learning fold recognition method (Cheng and Baldi, 2006). In this way, we combined the profile–profile alignments with a machine learning approach. This resulted in a fourth set of query-template alignments.

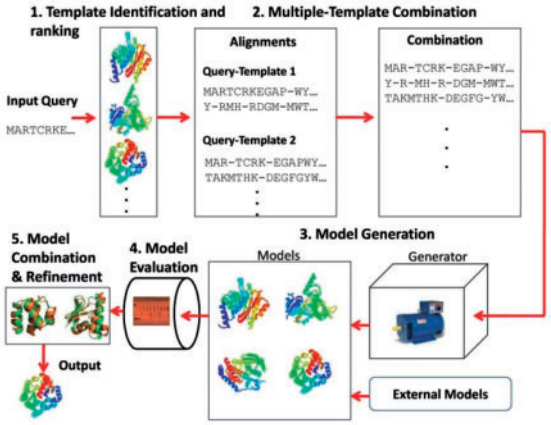

Fig. 1. A multi-level combination pipeline for protein structure prediction.

2.2.2 Multi-template combination It has been studied that combining multiple templates can, in most cases, improve the performance of TBM (Cheng, 2008). This being the case, all three of our predictors which implemented this portion of the pipeline incorporated the multi-template combination algorithm (Cheng, 2008) (Table 1). This algorithm chose and greedily combined the most significant query-template alignment (*E*-value $<10^{-20}$ and cover ratio $>75\%$) in each set with the rest of the alignments from the same set. This resulted in a multiple sequence alignment centered on the query sequence. The most significant alignment was then removed, and the second most significant alignment was combined with the remaining query-template alignments in order to generate a multiple sequence alignment using the same algorithm. The process was repeated up to 10 times to generate up to 10 multiple alignments from each set.

2.2.3 Model generation Models were generated in one of two ways. If query-template alignments existed, then each query-template alignment and its corresponding structure were fed into Modeller 7v7 (Fiser and Sali, 2003), a widely used model generation tool, to generate 10 models. From these, the model with minimum Modeller energy was chosen as a predicted model. If no significant template could be found by hhsearch (*E*-value $<10^{-3}$) and the length of the query protein was <120 residues, ROSETTA was executed to generate 200 models. The 200 models were clustered by ROSETTA, and the centroid of several large clusters were chosen as predicted models. During CASP8, ROSETTA was executed by MULTICOM-CLUSTER to generate models for several hard targets.

2.2.4 Model ranking The previous steps generated a large number of models for each target. To rank all of the models, we used our model quality assessment tool ModelEvaluator (Wang *et al*., 2008), which had been evaluated during CASP8 and considered an efficient and accurate model evaluation tool (Cheng *et al*., 2009). ModelEvaluator compared the secondary structure, solvent accessibility, contact map and beta-sheet topology of a model with those predicted from its primary sequence using the SCRATCH suite (Cheng *et al*., 2005). The comparison resulted in a number of features which were fed into support vector machines (SVM) to predict the GDT-TS score of the model. The predicted GDT-TS scores were used to rank the models and the top five models were submitted to CASP by MULTICOM-CLUSTER.

Table 1. Implementation details of five MULTICOM servers predictors and one MULTICOM human predictor

Steps	Methods	M-CLUSTER	M-RANK	M-CMFR	M-REFINE	MUProt	MULTICOM
(1) Template identification and ranking	PSI-BLAST	√	√	√			
	Hhsearch	√	√				
	COMPASS	√					
	FOLDPro	√	√	√			
(2) Template combination	Greedy algorithm	√	√	√			
(3) Model generation	Modeller	√	√	√	√	√	√
	ROSETTA	√					
	MULTICOM models				√	√	
	CASP8 server models						√
(4) Model evaluation	ModelEvaluator (SVM)	√	√	√	√	√	√
	SPIKER (clustering)					√	
(5) Model combination and refinement	Global-local algorithm				√	√	√

2.3 MULTICOM-RANK and MULTICOM-CMFR

MULTICOM-RANK and MULTICOM-CMFR are two other predictors which also implemented the first four steps of our pipeline. The implementation of MULTICOM-RANK and MULTICOM-CMFR are the same as described above except a few minor differences in the template identification and ranking, and model generation steps (Table 1). More specifically, both used a two-track approach for template identification and ranking. This is to say that for easy targets, MULTICOM-CMFR (or MULTICOM-RANK) used only PSI-BLAST (or PSI-BLAST and then hhsearch) to identify and rank templates according to *E*-value as in MULTICOM-CLUSTER. If fewer than five significant templates could be found, (i.e. when working on relatively hard targets) we used our SVM-based fold recognition method (Cheng and Baldi, 2006) to rank templates and five other alignment tools including MUSCLE (Edgar, 2004), hhsearch, lobster (Edgar and Sjolander, 2003), SPEM and COMPASS to generate additional alignments between the query and each of the top 50 templates. Additionally, when taking the track for hard targets, 250 models were generated as opposed to just 10. This larger number for hard targets was due in part to the fact that more alignment tools were used, and so we generated more query-template alignments, and hence more models ended up being generated. It was also due to our belief that since the templates available were less significant, generating more candidate models, or enlarging the candidate model pool, might result in more good models.

2.4 MULTICOM-REFINE, MUProt and MULTICOM

MULTICOM-REFINE, MUProt, and MULTICOM (our human-expert predictor) implemented the fourth and fifth portions of our general pipeline (model evaluation, model combination and refinement). These predictors made predictions by ranking and combining a number of internal and external models via a novel global–local model combination algorithm. This algorithm works by first attempting a combination of models selected on global similarity. This global model combination procedure worked well for easy targets where many similar models were generated. For harder targets, the algorithm falls back to more a localized approach in which it combines models that have similar fragments. Here, we first describe in detail our global–local model combination algorithm that MULTICOM-REFINE, MUProt and MULTICOM all use. We then go on to describe the differences between the three predictors.

2.4.1 Model combination and refinement For the model combination and refinement step of our pipeline, we used a novel global–local model combination algorithm. This algorithm worked by first selecting a seed model from one of the top five ranked models, and then compared it against all the other models using the structure-comparison tool TM-Score (Zhang and Skolnick, 2004a). Those models in which at least 80% of the model could be aligned to the seed model with a RMSD <4 Å were considered globally similar to the seed model and selected for combination. To combine the seed model and selected models, we fed them into Modeller 7v7, and used them as templates to generate 10 new models for the protein. Of these new models, the one with the minimum Modeller energy was selected as a refined model. This process was repeated up to five times to generate a refined model for each of the top five ranked models.

If no globally similar models were found, which was often the case for hard targets, a *local* model combination algorithm was used to combine the seed model with other locally similar models. To do so, the seed model was first compared against other models using TM-Score. Then long fragments of models that could be aligned with the seed model with a RMSD <3 Å and GDT-TS score >50 were selected. The minimum length of the fragments was initially set to 80 residues and this threshold was repeatedly reduced by five residues if no fragments could be found. The structures for the fragments and the initial seed model were fed into Modeller to generate 10 models, and the model with minimum energy was chosen as a refined model. This process was also repeated up to five times to produce a refined model for each of the top five ranked models.

As mentioned, MULTICOM-REFINE, MUProt and MULTICOM all focus on model combination and refinement, and all implement the global–local model combination algorithm just described. These three predictors differ in which models they consider for combination and refinement, and how those models are initially ranked. MULTICOM-REFINE collected the models predicted by MULTICOM-CMFR, MULTICOM-RANK and MULTICOM-CLUSTER, and used ModelEvaluator to predict the GDT-TS score of each model. The top 50% of the models generated by MULTICOM-CMFR and MULTICOM-RANK in addition to all the models generated by MULTICOM-CLUSTER were selected for model combination and refinement.MUProt also took the same set of models used by MULTICOM-REFINE as input, but before using ModelEvaluator to rank models, it used Spicker (Zhang and Skolnick, 2004b) to cluster models, and the models closest to the centroid of the largest cluster were ranked first. In this way, we combined ModelEvaluator, a SVM-based model quality assessment program (MQAP) with a clustering-based MQAP. For MULTICOM, our human expert predictor, the initial models came from all CASP8 server models (not including models from human predictors). These were ranked using ModelEvaluator according to predicted GDT-TS score.

3 RESULTS AND DISCUSSION

In order to evaluate the performance of our algorithms and predictors, and also compare the various ways of combining different techniques, we evaluated the six MULTICOM predictors on the CASP8 benchmark from two perspectives. First, we developed an automated evaluation pipeline to evaluate the MULTICOM predictors on 120 valid CASP8 targets. For each target, the experimental structures and predicted models were downloaded from the CASP8 website. The sequences extracted from the experimental structures were aligned with the CASP8 target sequences using ClustalW (Larkin and Blackshields, 2007) to identify residues in the target sequences that did not have coordinates in the experimental structures (i.e. potentially disordered regions). These residues were removed from the CASP8 structure models. The filtered models were then compared with the experimental structures using TM-Score. This generated GDT-TS (Zemla, 2003; Zemla *et al*., 1999), TM and Maxsub scores (Siew *et al*., 2000). These scores ranged from 0 to 100, and were used to measure the quality of the predicted models. In order to complement the official CASP8 assessment (Cozzetto *et al*., 2009; Ben-David *et al*., 2009; Keedy *et al*., 2009), our evaluation was based on the entire structure of a target as opposed to only its domains. Second, we downloaded the official GDT-TS and Z-scores for all the CASP8 models and compared the MULTICOM predictors with the other predictors using the official CASP8 results (http://predictioncenter.gc.ucdavis.edu/casp8/results.cgi). Note that the *Z*-score of a model for a target was its GDT-TS score normalized by the mean and standard error of all the models associated with the target (Cozzetto *et al*., 2009).

3.1 Evaluation by our in-house pipeline

CASP allowed a predictor to submit five models for each target, where the first model was believed to be the best prediction (Kryshtafovych *et al*., 2009b). We evaluated each MULTICOM predictor by calculating the average TM, GDT-TS and MaxSub score for the first models and the best-of-five models (the model with the highest score) on 120 CASP8 targets (Table 2). The standard error of the average GDT-TS scores were also calculated by dividing the

Table 2. Evaluation results of MULTICOM predictors for the first and the best-of-five models (inside parentheses) on 120 CASP8 targets

Predictor	Avg. TM	Avg. GDT-TS	Avg. MaxSub	GDT-TS S.E.[a]
MULTICOM	70 (72)	63 (65)	60 (62)	1.99 (1.94)
M-REFINE	67 (69)	60 (62)	56 (58)	2.06 (1.99)
M-CLUSTER	67 (70)	60 (62)	56 (59)	2.03 (1.99)
MUprot	67 (69)	60 (61)	59 (58)	2.04 (1.98)
M-RANK	66 (68)	59 (61)	55 (57)	2.05 (2.02)
M-CMFR	66 (68)	58 (61)	55 (57)	2.04 (2.02)

[a]The standard error of the average of GDT-TS scores of the first/best models.

standard deviation by the square root of the total number of targets (Table 2).

According to the results, MULTICOM, a predictor which made predictions by combining all CASP8 server models, achieved a better performance than MULTICOM-REFINE and MUProt, which made predictions by combining models from only three of the MULTICOM template-based predictors. As the combination and refinement method used was exactly the same in each predictor, this indicates that the quality of the final model increases as the number and quality of candidate models increase. This further proves that our model combination algorithm can detect and combine structural segments with better qualities and refine the final model. The similar performance of MULTICOM-REFINE and MUProt indicates that combining a cluster-based ranking method with ModelEvaluator did not result in much of a change in performance when compared to just using ModelEvaluator. MULTICOM-REFINE performed slightly better than MULTICOM-CLUSTER and notably better than MULTICOM-RANK and MULTICOM-CMFR. This indicates that a well-implemented model combination approach tends to achieve better (or at least similar) performance than (or as) the best base predictors (i.e. those that only implement the first four steps of our pipeline). Also, MULTICOM-CLUSTER performs better than the other two base predictors MULTICOM-RANK and MULTICOM-CMFR, and this indicates that combining more diverse template identification and fold recognition methods can improve structure prediction. Moreover, the fact that MULTICOM-RANK performed slightly better than MULTICOM-CMFR suggests that hhsearch may work slightly better than PSI-BLAST for predicting structures for each target.

To consider the overall quality of our multi-level combination approach, we used TM-score. This tool reports a score between 0 and 1, and measures the absolute quality of a model. A TM-score of 0.40 indicates a moderately accurate model with the correct topology, whereas a score of 0.17 indicates a random prediction (Zhang and Skolnick, 2004a). As we see in Table 2, the average per-target TM-score of all the MULTICOM predictors ranged from 0.66 to 0.70. This indicates that in general the models that our MULTICOM predictors produce are of good quality.

3.2 Comparisons with other CASP8 predictors

We compared the MULTICOM predictors with other CASP8 server and human predictors using the official CASP8 assessment data and measures. CASP8 dissects valid targets into 154 TBM domains, for which at least one template is available, and 13 FM domains, for which no template is available. From the 154 TBM domains, 50 are classified as template-based high accuracy (TBM-HA) domains, for which at least one model with a GDT-TS score >80 was predicted. In our comparison, we took models predicted by 237 predictors (121 server predictors and 116 human predictors), and assessed them from various perspectives (Tables 3–5). Our comparisons serve only as a comparative study of our methods with respect to the state-of-the-art. Readers should refer to the CASP8 articles (Ben-David *et al*., 2009; Cozzetto *et al*., 2009; Keedy *et al*., 2009) for the official CASP8 results accredited by protein structure experts.

Table 3. Top 10 server predictors evaluated on the first models (out of five submissions) of the 50 TBM-HA domains

Predictor	Domain No.	Sum Z-score GDT[a]	Avg. GDT-TS[b]
Zhang-Server[c]	50	27	88
RAPTOR[d]	50	25	87
MULTICOM-REFINE	50	24	87
MUProt	50	24	87
Phyre_de_novo[e]	50	24	87
MULTICOM-CLUSTER	50	23	87
HHpred5[f]	50	23	85
MULTICOM-RANK	50	22	86
HHpred2[f]	50	22	86
pro-sp3-TASSER[g]	50	21	86

The standard error of the average GDT-TS scores of MULTICOM-REFINE, MUProt, MULTICOM-CLUSTER and MULTICOM-RANK are 0.84, 0.81, 0.81 and 0.96, respectively. This analysis only considers predictors that predicted more than 46 domains. For details about each CASP8 server, please refer to CASP8 meeting abstract (http://www.predictioncenter.org/casp8/doc/CASP8_book.pdf).
[a]Sum of the Z-scores of GDT-TS.
[b]Average GDT-TS score.
[c]Zhang (2009).
[d]Xu *et al*. (2009).
[e]Kelley *et al*. (2008).
[f]Hildebrand *et al*. (2009).
[g]Zhou *et al*. (2009).

Table 3 reports the top 10 server predictors on 50 TBM-HA domains. Predictors were evaluated by the cumulative GDT-TS Z-scores and the average GDT-TS scores of the first models. The GDT-TS Z-score of a domain was calculated by $(x-\mu)/\sigma$, where x is the GDT-TS score of the target model, μ and σ, respectively, are the mean and standard deviation of the GDT-TS scores of all the models of the domain from the various predictors (Kryshtafovych *et al*., 2009b). On 50 TBM-HA domains, MULTICOM, our human predictor, achieved a cumulative GDT-TS Z-score of 27.41 and average GDT-TS score of 88.80, which is higher than the best server predictor. As MUTICOM generates models by combining all the CASP8 server models, this clearly shows the value and contribution of our model combination algorithm, at least on the easy targets. Furthermore, four of the five MULTICOM server predictors were ranked within the top 10 of the 121 server predictors (Table 3). This demonstrates that the multi-level combination approach works competitively well on high accuracy easy targets.

Table 4 shows the top 10 predictors on the 64 human/server TBM domains. Most of these domains were considered hard template-based cases as only weak templates could be found. MULTICOM was ranked within the top 10 predictors in terms of cumulative Z-scores, but ranked below the best server. This may indicate that the model selection component of MULTICOM was not able to

Table 4. Top 10 human and server predictors and MULTICOM predictors on the first models (of the five possible submissions) of 64 TBM domains from human/server targets

Predictor	Domain number	Sum *Z*-score GDT	Average GDT-TS
IBT_LT[a]	64	67	65
DBAKER[b]	64	64	64
Zhang[c]	64	56	64
fams-ace2[d]	64	52	63
Zhang-server[e]	64	52	63
TASSER[f]	64	51	63
SAM-T08-human[g]	62	51	62
ZicoFullSTP[h]	64	50	61
Zico[i]	64	48	61
MULTICOM	64	48	61

The standard error of the average GDT-TS scores of MULTICOM on these domains is 2.37. This analysis only considers predictors that predicted more than 60 domains.
[a]Venclovas and Margelevicius (2009).
[b]Raman *et al*. (2009).
[c]Zhang (2009).
[d]Terashi *et al*. (2008).
[e]Indicates a server predictor; otherwise, it is a human predictor.
[f]Zhou *et al*. (2009).
[g]Karplus (2008).
[h]Girgis *et al*. (2008).
[i]Girgis and Fischer (2008).

Table 5. Top 10 CASP8 server predictors on the first models (of five possible submissions) of 154 TBM domains

Predictor	Domain No.	Sum *Z*-score GDT	Avg. GDT-TS
Zhang-server	154	104	71
RAPTOR	154	86	69
Pro-sp3-TASSER	154	81	68
Phyre_de_novo	154	79	68
HHpred5	154	79	66
BAKER-ROBETTA[a]	154	76	67
METATASSER[b]	154	75	67
HHpred4[c]	154	75	67
MULTI-CLUSTER	154	73	67
MULTI-REFINE	154	71	67

The standard error of the average GDT-TS scores of MULTICOM-CLUSTER and MULTICOM-REFINE are 1.66 and 1.69, respectively.
This analysis only considers predictors that predicted more than 150 domains.
[a]Raman *et al*. (2009).
[b]Zhou *et al*. (2009).
[c]Hildebrand *et al*. (2009).

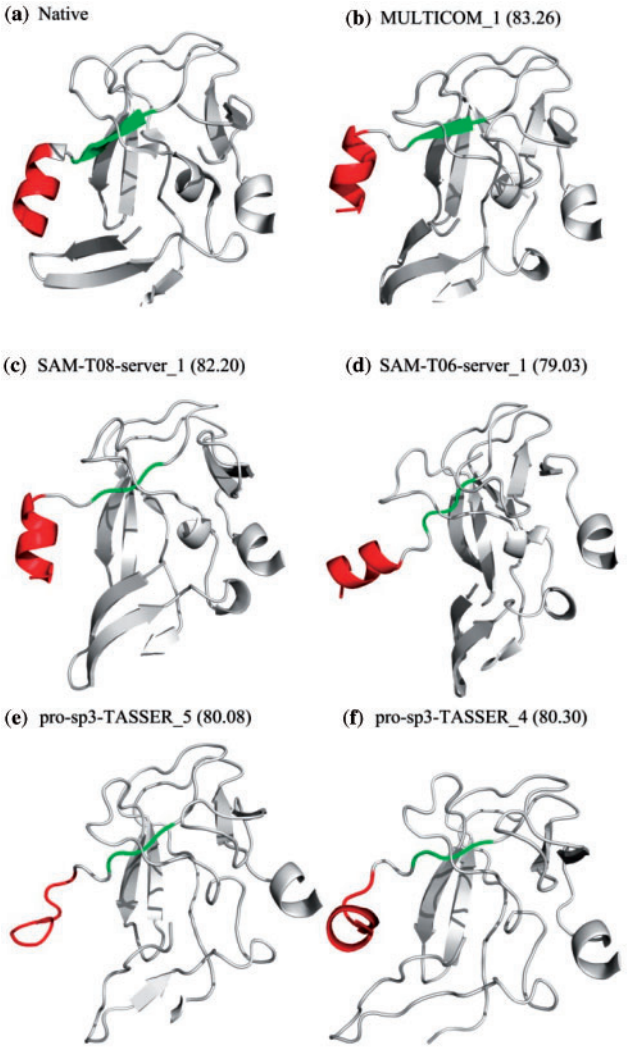

Fig. 2. Comparisons between the experimental structure (**a**) of domain 1 of T0435, the first model of MULTICOM (**b**), and four models combined by MULTICOM (**c–f**). The GDT-TS scores are listed inside parentheses. MULTICOM model (b) is the best model among all the server and human models for this domain. (c) The second best sever model and the best model selected and combined by MULTICOM. MULTICOM correctly predicted a beta-strand [green in (b)], which was not correctly predicted by any of the four models it combined [green, (c–f)]. Furthermore, a helix (red) was correctly modeled in (b) and (c), but not in any of the other models [red, (d–f)]. This indicates that the model combination algorithm can detect and combine portions with good qualities, and further refine structural portions to achieve a better overall quality.

select the top models for the hard TBM targets. Table 5 reports the results of the top 10 server predictors on 154 template-based domains. Two of our server predictors (MULTICOM-CLUSTER, MULTICOM-REFINE) were ranked within top 10 in terms of both *Z*-scores and average GDT-TS scores. Their performance in terms of average GDT-TS is close to the second best predictor, indicating the multi-level combination is competitive in this category.

In the category of template free modeling, the CASP8 official assessment (Ben-David, 2009) mainly used two measures to evaluate predictors in 13 FM domains: scoring scheme A and M. Scoring scheme A indicates the total number of models with high quality (within top 3), whereas scoring scheme M highlights the number of targets, in which a group generated high quality models. The scheme A and M scores of MULTICOM on 13 FM domains are 24 and 7, respectively (Ben-David, 2009). Both of them were ranked first among all human and server predictors, which clearly indicates its strength, and further highlights that the evaluation guided model combination approach can effectively select and refine low-resolution models generated on hard FM targets. The average GDT-TS scores of our server predictors ranged from ~ 29 to 31 (data

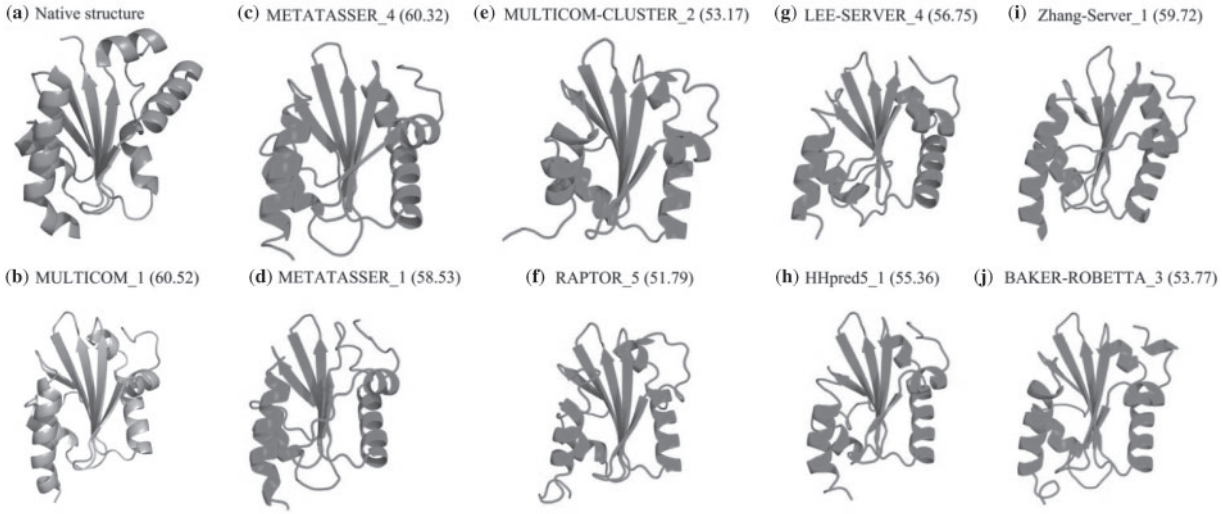

Fig. 3. Comparisons between the experimental structure (**a**) of domain 2 of T0501, the first MULTICOM model (**b**), and eight of the 20 models MULTICOM combined (**c**–**j**). The GDT-TS scores are listed inside parentheses. (b) The best model among all the server and human models for this domain. (c) The best server model. METATASSER did not rank its best model (c) as the top one model, but this model was included into the combination process of MULTICOM. In this case, the combined model (b) achieved a better quality than all the models it did or did not combine.

not shown), lower than 40 of the MULTICOM human predictor. The reason is that the human predictor used a large pool of input models, which contained some good quality third-party models for the FM targets.

3.3 A deeper look into model combination

The CASP8 official evaluations have statistically shown the good performance of our model combination approach. To delve further into the effectiveness of our approach, we examined several of the models generated by MUTLICOM and the source of these models (i.e. those models it chose to combine). We found that multi-model combination can improve structure prediction in two ways. First, it can combine complementary good regions from multiple models to generate a model that is better than all the models it combined (see Fig. 2 for an example). Second, it can include good models that were not originally ranked as the first model and combine these models or portions of them to generate a model that is better than the first model (see Fig. 3 for an example). In general, the model combination process is a selective averaging process, which can produce a model that is on average better than or as good as the top model among combined models. On 11 CASP8 domains, for instance, the combined models generated by MULTICOM achieved the best qualities among all the server and human models. However, the performance of the approach relies on the selection of the good models for combination. This explains why MULTICOM achieved better performance than the best server on high-accuracy and free-modeling targets, but not on hard template-based targets, whose models often contain both a good structure core and bad local regions (e.g. unfolded tails) that may make ModelEvalutor to underestimate their quality. Our CASP8 experiments demonstrate the overall success of MULTICOM although some parts of it, such as its model selection abilities on hard template-based targets may need improvement.

4 CONCLUSIONS

We described a comprehensive and effective approach to combine multiple templates, alternative alignments, and similar models under the guidance of model quality assessment. This approach was successfully applied to protein structure modeling during the recent CASP8 experiments. Our results show that our approach is effective for the full spectrum of protein modeling, particularly for high-accuracy TBM and FM. Compared with most existing protein structure prediction systems, our approach contains a unique and novel model combination step that can refine protein models by averaging complementary good models or fragments. The general combination approach can be further improved at each modeling step (e.g. model ranking) or by integrating complementary techniques. We are currently improving the performance of the method for hard template-based targets by increasing the accuracy of model ranking and integrating *ab initio* modeling with TBM to enhance model generation. We plan to test our improved systems that are largely based on MULTICOM-CLUSTER, MULTICOM-REFINE and MULTICOM in the CASP9 experiment.

Funding: This work is partially supported by an University of Missouri (UM) research board grant and an University of Missouri, Columbia (MU) research council grant to J.C.

Conflict of Interest: none declared.

REFERENCES

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Ben-David,M. *et al*. (2009) Assessment of CASP8 structure predictions for template free targets. *Proteins*, **77**, 50–65.

Cheng,J. (2008) A multi-template combination algorithm for protein comparative modeling. *BMC Struct. Biol.*, **8**, 18.

Cheng,J. and Baldi,P. (2006) A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, **22**, 1456–1463.

Cheng,J. *et al*. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

Cheng,J. *et al*. (2009) Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins*, **77**, 181–184.

Cozzetto,D. *et al*. (2009) Evaluation of template-based models in CASP8 with standard measures. *Proteins*, **77**, 18–28.

Edgar,R. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Edgar,R. and Sjolander,K. (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics*, **19**, 1404–1411.

Fiser,A. and Sali,A. (2003) Modeller: generation and refinement of homology-based protein structure models. *Meth. Enzymol.*, **374**, 461–491.

Girgis,H.Z. and Fischer,D. (2008) Hierarchy of general linear models for selecting and ranking the best predicted protein structures. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 120–121.

Girgis,H.Z. *et al*. (2008) On-line hierarchy of general linear models for selecting and ranking the best predicted protein structures. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 122–123.

Hildebrand,A. *et al*. (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins*, **77**, 128–132.

Hinds,D.A. and Levitt,M. (1994) Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, **243**, 668–682.

Jaravine,V. *et al*. (2006) Removal of a time barrier for high-resolution multidimensional NMR spectroscopy. *Nature Methods*, **3**, 605–607.

Karplus,K. (2008) SAM-T08-human. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 95.

Keedy,D. *et al*. (2009) The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. *Proteins*, **77**, 29–49.

Kelley,L.A. *et al*. (2008) From comparative modeling to de novo folding with Phyre, Poing and Phragment. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 111–112.

Kim,D.E. *et al*. (2008) Robetta de novo and homology modeling in CASP8. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 7–8.

Kryshtafovych,A. *et al*. (2005) Progress over the first decade of CASP experiments. *Proteins*, **7**, 225–236.

Kryshtafovych,A. *et al*. (2007) Progress from CASP6 to CASP7. *Proteins*, **69**, 194–207.

Kryshtafovych,A. *et al*. (2009a) CASP8 results in context of previous experiments. *Proteins*, **77**, 217–228.

Kryshtafovych,A. *et al*. (2009b) Protein Structure Prediction Center in CASP8. *Proteins*, **77**, 5–9.

Larkin,M.A. and Blackshields,G. (2007) ClustalW and ClustalX version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lattman,E. (2004) The state of the protein structure initiative. *Proteins*, **54**, 611–615.

Moult,J. *et al*. (2009) Critical assessment of methods of protein structure prediction (CASP)-round *VIII. Proteins*, **77**(Suppl. 9) 1–4.

Pandit,S.B. *et al*. (2008) METATASSER: a 3D-jury threading approach with TASSER model assembly/refinement. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 63–64.

Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.

Service,R. (2005) STRUCTURAL BIOLOGY: structural genomics, round 2. *Science*, **307**, 1554–1558.

Siew,N. *et al*. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

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

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Sternberg,M. and Thornton,J. (1978) Prediction of protein structure from amino acid sequence. *Nature*, **271**, 15–20.

Terashi,G. *et al*. (2008) Structure evaluation program using the local consensus-based similarity and circle quality assessment method. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp 27–28.

Thompson,J. *et al*. (2008) Comparative modeling of protein structures in CASP8 using full-atom Rosetta refinement and manual alignment selection. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 21–22.

Venclovas,C. and Margelevicius,M. (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins. *Proteins*, **77**, 81–88.

Wang,Z. *et al*. (2008) Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins*, **75**, 638–647.

Xu,J. *et al*. (2009) Template-based and free modeling by RAPTOR++ in CASP8. *Proteins*, **77**, 133–137.

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

Zemla,A. *et al*. (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, **37**, 22–29.

Zhang,Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.*, **18**, 342–348.

Zhang,Y. and Skolnick,J. (2004a) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhang,Y. and Skolnick,J. (2004b) SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem.*, **25**, 865–871.

Zhang,Y. and Skolnick,J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci.*, **102**, 1029–1034.

Zhang,Y. (2009) I-TASSER: fully automated protein structure prediction in CASP8. *Proteins*, **77**, 100–113.

Zhou,H. *et al*. (2009) Performance of the Pro-sp3-TASSER server in CASP8. *Proteins*, **77**, 123–127.

Zhou,H. and Zhou,Y. (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, **21**, 3615–3621.

Zhou,H. *et al*. (2008) TASSER-based protein structure prediction in CASP8. In *proceedings of the critical assessment of techniques for protein structure prediction - eighth meeting, Cagliari, Sardinia, Italy*, pp. 115–116.