Bioinformatics, 31, 2015, i116–i123
doi: 10.1093/bioinformatics/btv235
ISMB/ECCB 2015

OXFORD

Large-scale model quality assessment for improving protein tertiary structure prediction

Renzhi Cao[1], Debswapna Bhattacharya[1], Badri Adhikari[1], Jilong Li[1], and Jianlin Cheng[1,2,3,*]

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

*To whom correspondence should be addressed.

Abstract

Motivation: Sampling structural models and ranking them are the two major challenges of protein structure prediction. Traditional protein structure prediction methods generally use one or a few quality assessment (QA) methods to select the best-predicted models, which cannot consistently select relatively better models and rank a large number of models well.

Results: Here, we develop a novel large-scale model QA method in conjunction with model clustering to rank and select protein structural models. It unprecedentedly applied 14 model QA methods to generate consensus model rankings, followed by model refinement based on model combination (i.e. averaging). Our experiment demonstrates that the large-scale model QA approach is more consistent and robust in selecting models of better quality than any individual QA method. Our method was blindly tested during the 11th Critical Assessment of Techniques for Protein Structure Prediction (CASP11) as MULTICOM group. It was officially ranked third out of all 143 human and server predictors according to the total scores of the first models predicted for 78 CASP11 protein domains and second according to the total scores of the best of the five models predicted for these domains. MULTICOM's outstanding performance in the extremely competitive 2014 CASP11 experiment proves that our large-scale QA approach together with model clustering is a promising solution to one of the two major problems in protein structure modeling.

Availability and implementation: The web server is available at: http://sysbio.rnet.missouri.edu/multicom_cluster/human/.

Contact: chengji@missouri.edu

1 Introduction

Protein tertiary structure prediction has been an important scientific problem for few decades, especially in bioinformatics and computational biology (Eisenhaber *et al.*, 1995). Despite more and more native structures are included in protein data bank (PDB) (Berman *et al.*, 2000) database, the gap between the sequenced proteins and the native structures is still enlarging due to the exponential increase of protein sequences produced by large-scale genome and transcriptome sequencing. It is estimated that <1% of protein sequences have the native structures in PDB database (Rigden, 2009). Therefore, accurate computational methods for protein tertiary structure prediction that are much cheaper and faster than experimental structure determination techniques are needed to reduce this large sequence-structure gap. Furthermore, computational structure prediction methods are important for obtaining the structures of membrane proteins whose structures are hard to be determined by experimental techniques such as X-ray crystallography (Yonath, 2011).

The two major problems of protein structure prediction are *model sampling* and *model ranking*. The former is to generate a number of structural models (conformations) for a protein target, and the latter is to rank these models and to select the presumably best ones as final predictions. The two main ways of generating protein models are *template-based modeling* and *template-free modeling*. Template-based modeling methods use the known structures (templates) of the proteins that are homologous or analogous to a target protein to construct structural models for the target (Bowie *et al.*, 1991; Jones, *et al.*, 1992; Zhang, 2008a, b). For instance,

Table 1. All 14 QA methods with the details

Methods	Type	Features
MULTICOM-NOVEL	Single	Structural, physical, chemical features
OPUS-PSP	S	Contact potentials based on side chain functional groups
ProQ2	S	Structural features
RWplus	S	Side-chain orientation dependent potential
ModelEvaluator	S	Structural features, contacts
Modelcheck2	S	Structural features, contacts, disorder, conservation
RF_CB_SRS	S	Distance dependent statistical potential
SELECTpro	S	Energy-based (h-bond, angle, electrostatics, vdw)
Dope	S	Statistical potential
DFIRE2	S	Energy-based potential
ModFOLDclust2	Multi	Pairwise model similarity (geometry)
APOLLO	M	Pairwise model similarity
Pcons	M	Pairwise model similarity
QApro	M+S	Weighted pairwise model similarity
MULTICOM (human)	Consensus	Average ranking

The highlighted methods are built in house. S: single-model method; M: multi-model method.

during 2014 CASP11 experiment, almost all the structure prediction servers such as I-TASSER (Zhang, 2008a, b; Zhang, 2014), MULTICOM (Cheng *et al.*, 2012; Li *et al.*, 2013), MUFOLD (Zhang *et al.*, 2010) and RaptorX (Källberg *et al.*, 2012) used the template-based model technique to predict structures of some CASP11 targets for which some homologous template structures could be found. Template-free modeling methods predict the protein tertiary structure from scratch without using template information. This is especially important when there are no structural homologs existing in the database or the template identification techniques cannot find good templates (Zhang, 2008b). Some CASP11 prediction servers such as ROSETTA (Simons *et al.*, 1997), QUARK (Xu and Zhang, 2012) and FALCON (Li *et al.*, 2008) used template-free modeling method to generate structural models for some hard CASP11 targets.

Once some structural models are generated for a protein, the remaining challenge is to assess the quality of these models and select the most accurately predicted models. There are generally two main kinds of quality assessment (QA) methods: single-model QA methods (Cao *et al.*, 2014a, b; Randall and Baldi, 2008; Ray *et al.*, 2012; Shen and Sali, 2006; Wang *et al.*, 2009; Zhang and Zhang, 2010), which evaluate the quality of one single model without using the information of other models; and multi-model QA methods (Cao *et al.*, 2014a, b; McGuffin and Roche, 2010; Wallner and Elofsson, 2006; Wang *et al.*, 2011), which use the structural similarity between one model and other models of the same protein to assess its quality. The multi-model quality prediction methods generally perform better than the single-model quality prediction methods given the pool of models is sampled by independent structure predictors. However, multi-model QA method is largely influenced by the proportion of good models in the pool or the average quality of the largest model cluster in the pool, whereas single-model QA methods may work better in assessing a small number of models of wide-range quality usually associated with a hard target or a pool of models with very low proportion of good ones (Cao *et al.*, 2014a).

Currently, most protein structure prediction methods use one or at most a few QA methods to rank and select models, generally leading to the poor performance in selecting models of good quality due to the extreme difficulty of ranking models and intrinsic limitations of individual QA methods. Some structure prediction methods also apply clustering techniques to group models into different clusters whose center is considered as the best model in each cluster based on the structural similarities. The hypothesis behind it is that near-native structures are more likely clustered in a large free-energy basin in the free-energy landscape (Dobson *et al.*, 1998; Shortle *et al.*, 1998). The clustering based approaches generally select an average model rather than the best model and cannot work well if the quality of the largest cluster is not good. Therefore, although numerous methods have been developed to assess, rank and select models, protein model ranking is still largely an unsolved problem.

In order to address this challenge, we developed a large-scale consensus QA method (MULTICOM) to combine 14 complementary model QA methods to improve the reliability and robustness of protein model ranking. The general model ranking is also synergistically integrated with model clustering techniques to increase the diversity and quality of the final selected models. On the very competitive 2014 CASP11 benchmark, this new method substantially outperform any single QA method, suggesting its unique value in addressing one major problem of protein structure prediction.

2 Materials and methods

2.1 Large-scale protein model QA for protein tertiary structure prediction

Given a pool of structural models generated for a target protein (e.g. hundreds of models generated for a CASP11 target), the MULTICOM method used unprecedentedly 14 complementary model QA methods to predict the quality score of each model first (Table 1). These QA methods include both single-model and multi-model QA methods. The single-model methods include our new single-model global QA method MULTICOM-NOVEL based on the difference between secondary structure and solvent accessibility predicted by Spine X (Faraggi *et al.*, 2012) and SSpro4 (Cheng *et al.*, 2005) from the protein sequence and those of a model parsed by DSSP (Kabsch and Sander, 1983), physical-chemical features (i.e. surface polar score, weighted exposed score, and etc.) (Mishra *et al.*, 2013), the normalized quality score generated by ModelEvaluator (Wang *et al.*, 2009), RWplus score (Zhang and Zhang, 2010), dope score (Shen and Sali, 2006) and RF_CB_SRS_OD score (Rykunov and Fiser, 2007); ProQ2

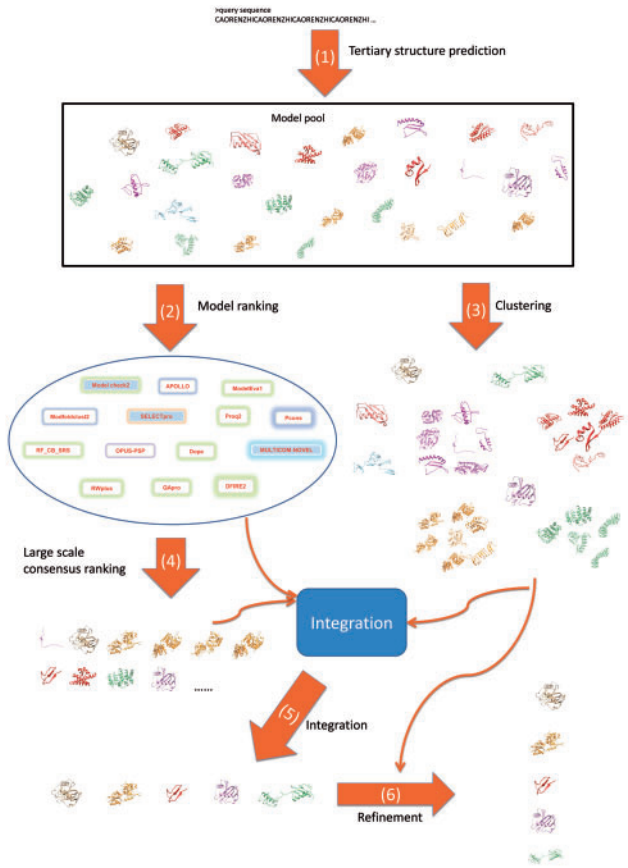

Fig. 1. The workflow of the MULTICOM method comprised of six steps. (1) A pool of tertiary structure models is predicted for a target protein. (2) Models are scored and ranked by different QA methods. (3) Models are clustered into groups based on structural similarity. (4) The consensus of individual QA rankings and other information are synthesized to generate the final ranking of all the models. (5) The final ranking and the clustering results are integrated to select top five diverse models for submission. (6) The top five models are combined to generate five refined models to be submitted to CASP11

(Ray *et al.*, 2012); Model check2 method produced by an improved version of ModelEvaluator (Wang *et al.*, 2009); a recalibrated SELECTpro energy (Randall and Baldi, 2008); Dope (Shen and Sali, 2006); DFIRE2 (Yang and Zhou, 2008); OPUS_PSP (Lu *et al.*, 2008); Rwplus (Zhang and Zhang, 2010); ModelEvaluator (Wang *et al.*, 2009) and RF_CB_SRS_OD (Rykunov and Fiser, 2007). The multi-model QA methods include ModFOLDclust2 (McGuffin and Roche, 2010); Pcons (Wallner and Elofsson, 2006); APPOLLO (Wang *et al.*, 2011); QApro—a weighted combination of ModelEvaluator and APOLLO (Cao *et al.*, 2014a). The details of each method are described in Table 1.

During the 2014 CASP11 experiment, MULTICOM used two different combinations of the QA scores produced by 14 QA methods to generate consensus rankings to rank all models of each target. The first one is the complete combination, in which each of 14 QA methods was applied to all the models of a target and generated a ranking for them based on their QA scores, and the average rank of 14 ranks of each model assigned by the 14 QA methods was used as its final rank. The second one is the consensus rankings based on the same average ranks produced by only six QA methods including (MULTICOM-NOVEL QA score, QApro score, Pcons score, Modelcheck2 score, Dope score, OPUS_PSP score). These six methods were selected because their combination performed best on all the models of 46 CASP10 when all possible combinations were benchmarked before CASP11 experiment started. On these CASP10 models, the average loss score of top one model based on 6 QA methods is 0.037, lower than 0.057 of all 14 QA methods. However, considering that the optimization process in benchmarking could over fit the data, we let MULTICOM use the consensus rankings of both the 6 selected QA methods and all 14 QA methods.

During the modeling ranking process, if the same top one model was selected by the two consensus rankings, which happened in >50% cases, the consensus ranking of the six QA methods were used as the final ranking of all the models. But if they disagreed with each other, the score of top one model selected by the pairwise QA method APOLLO was used to break the tie as follows. On one hand, if the score of APOLLO's top one model was >0.3, which generally meant quite some models in the model pool were of good quality due to relatively high pairwise similarity between them, the final ranking was set as the consensus ranking of the 6 QA methods or all 14 QA methods depending on whose top one model was more similar to the top one model of APOLLO than the other. Furthermore, the top one models of the two consensus rankings and of the top predictors (e.g. MULTICOM-CLUSTER and Zhang-Server) were compared with the top one model of APOLLO, and the model most similar to the top one model of APOLLO was used the top one model in the final ranking without changing the ranking of all other models. On the other hand, if the score of the top one model selected by APOLLO was <=0.3, which only occasionally happened and suggested that the target was hard and most models were of bad quality, MULTICOM calculated the percent of matching between the secondary structures extracted from the top one model selected by either 6- or 14-QA consensus ranking with the secondary structure of the target predicted from its sequence. The final ranking was one of 6 or 14 consensus ranking whose top one model had the higher percentage of matching of secondary structures.

Since the top five models selected by the final ranking above sometime could be very similar to each other, the risk for all of them to fail altogether was high for hard targets. To reduce this risk, MULTICOM only kept the top two models of the final ranking as the two predicted structures. And then, in order to increase the diversity of top five models selected as final predictions for each target, MULTICOM used MUFOLD_CL (Zhang and Xu, 2013) to cluster models, and then selected the other three models according to the final ranking in separate clusters different from those of the top two models. MUFOLD-CL (Zhang and Xu, 2013) is a model clustering method based on the comparison of the protein distance matrices. Comparing with other clustering techniques based on structural distance such as root-mean-square deviation (RMSD) (Kabsch, 1976), it is much faster, but yields similar accuracy, which is desirable for clustering a large number of protein models. During the selection of the other three models from different clusters, MULTICOM also skipped the models ranked at bottom 10% according to our newly developed MULTICOM-NOVEL QA method. This guaranteed that the top five selected structures were largely different, which indeed improved the score of the best of top five models.

Finally, MULTICOM used a model combination approach (Wang *et al.*, 2010) to integrate each selected model with other similar models in the model pool to generate its refined model. The workflow of our MULTICOM method described earlier is illustrated in Figure 1.

2.2 Evaluation of top-ranked models

We downloaded publically available native structures for 42 CASP11 human targets from the CASP's website (http://www.predictioncenter.org/casp11/index.cgi). During CASP11, our MULTICOM method was blindly benchmarked on these targets together with 142 human and server predictors. The predicted structural models were assessed on 55 domains of the 42 targets. For comparison, we downloaded both the other predictors' predictions and our submitted predictions from the CASP11's website. During CASP11, each predictor submitted up to five predicted model with the first one (TS1) designated as the best model. We evaluated the performance of each predictor's first model by calculating the GDT-TS score between it and its native structure. The TM-score (Zhang and Skolnick, 2004) was used for calculating the global distance test - total score (GDT-TS). The Z-score of a model was calculated as the model's GDT-TS score minus the average GDT-TS score of all the models in the model pool of a target divided by the standard deviation of all GDT-TS scores. The negative Z-score was converted to 0 during summation of Z-scores. The sum of the Z-scores of the first models predicted by a predictor for the 42 targets was used to measure its overall performance. Similarly, the sum of the Z-scores of the best of the five submitted models predicted by a predictor for the 42 targets was used to measure its performance if the best of all five submitted models was considered.

3 Results and discussion

We evaluated the performance of MULTICOM human predictor along with 44 CASP11 server predictors on 42 CASP11 human targets. The sum of Z-scores of all first (i.e. TS1) models or the best of five submitted models predicted by these predictors was reported in Table 2. Other human server predictions were not considered in the analysis here since they were not publicly available. It is shown that MULTICOM performs better than all server predictors. Its total Z-score of first models is around 4 points higher than the best server predictor Zhang-Server, and its total Z-score of the best of five models is >6 points higher than the best server predictor QUARK. These results demonstrate MULTICOM's ability to rank a large pool of models for selecting top one or five models. According to CASP11's official evaluation of all 143 human and server predictors, MULTICOM was ranked third based on the sum of Z-score of the first model and second based on the sum of Z-score of best of the five submitted models. The MULTICOM's outstanding performance in the extremely competitive CASP11 experiment demonstrates that our large-scale model QA is powerful for ranking and selecting good models from a pool of models of different quality.

Table 2. The top 10 tertiary structure predictors ranked based on the summation of the Z-scores of the first models, and their summation of the Z-scores of best of the five submitted models

Server name	Sum of Z/rank	Sum of Z of best of five/rank
MULTICOM (human)	57.49/1	78.42/1
Zhang-Server	53.62/2	70.57/3
QUARK	51.90/3	71.93/2
Nns	35.07/4	51.79/6
Myprotein-me	34.11/5	52.73/5
MULTICOM-CLUSTER	31.39/6	39.03/10
MULTICOM-CONSTRUCT	31.33/7	38.65/11
RBO_Aleph	30.77/8	40.65/9
BAKER-ROSETTASERVER	28.80/9	63.64/4
MULTICOM-NOVEL	25.71/10	43.43/7

In order to investigate types of the models selected by MULTICOM and the contribution of individual structure predictors, we calculated the number of times that the models predicted by each predictor were ranked within top five by MULTICOM. Table 3 shows that the contribution of top 10 server predictors whose models were selected by MULTICOM to refine to generate the final predictions. It shows that a diverse set of server predictors including Zhang-Server made significant contributions to the final prediction, suggesting the large-scale QA used by MULTICOM can reliably assess a very diverse set of models generated by different tertiary structure predictors in the field.

To study how our large-scale model QA method improves model ranking, we compared its performance with that of each individual QA method and the two other simple consensus methods (one based on the sum of 14 original QA scores and another based on the sum of 14 Z-scores calculated from original scores). The first two columns in Table 4 reports the average GDT-TS score of the first models selected by these QA methods for all 42 human targets and a subset of 30 template-based human targets, respectively. The results show that MULTICOM performs better than every individual QA method, and sometime the improvement is substantial. And not surprisingly, the multi-model QA methods outperformed single-model QA methods on template-based human targets whose model pool was often of good quality. For instance, a multi-model QA method APOLLO ranks sixth on all human targets, but third on template-based human targets. The third columns in Table 4 shows the average Z-score of the first models selected by different QA methods. It is interesting to notice that the single-model QA methods tend to have higher Z-score than the multiple-model QA methods. For example, the multiple QA method APOLLO has a relatively high average GDT-TS score (0.338) of the first selected models, however, its

Table 3. The top 10 predictors ranked based on the total number times their models were selected by our MULTICOM predictor on all the human targets or template-based (TBM) human targets only

Rank	Servers on all human targets	Num. on all	Servers on TBM	Num. on TBM
1	Zhang-Server	58	Zhang-Server	43
2	BAKER-ROSETTASERVER	36	BAKER-ROSETTASERVER	27
3	QUARK	29	QUARK	22
4	RBO_Aleph	29	myprotein-me	20
5	myprotein-me	28	Nns	19
6	Nns	21	Seok-server	14
7	Seok-server	17	RBO_Aleph	13
8	MULTICOM-REFINE	10	MULTICOM-REFINE	8
9	FUSION	7	RaptorX	4
10	RaptorX	5	FUSION	4

average Z-score of the first selected models is lower than most single QA methods. The reason is probably because the multiple-model QA methods tend to work well on easy targets whose models have similarly good quality and thus low Z-scores, whereas single-model QA methods may select some good models for some hard targets whose models are mostly bad, resulting in a high Z-score.

Considering average ranking is just one way of combining different QA scores, we tested another two ways to combine QA scores for comparison. The first one simply calculated the average of original 14 QA scores to rank models. The second one first converted all original QA scores of each method into Z-scores, and then used the average of 14 Z-scores to rank models. Table 4 shows that consensus of 14-QA Z-scores performed best in terms of the average Z-score of the top one models, whereas MULTICOM performed best in terms of the average GDT-TS score of the top one models. The results demonstrate that the way of integrating different QA scores influences the quality of the final ranking.

Moreover, we compared MULTICOM with a simple combination approach that used a good single-model QA method (i.e. ProQ2) to rank models of very hard targets and a good clustering method (APOLLO) to rank the models of other targets. If the maximum APOLLO pairwise score of the models of a target is <0.2, it is considered a hard target, otherwise an easy target. The average Z-score and GDT score of the top one model selected by this simple combination method is 0.980 and 0.350, respectively, which is higher than that (0.584 and 0.338) of APOLLO, but substantially lower than that (1.364 and 0.374) of MULTICOM.

Furthermore, compared with the two other top-ranked consensus methods participating in CASP11 experiment—TASSER (ranked ninth in CASP11) and keasar (ranked 27th) that used several QA methods according to the official CASP11 experiment, MUTLICOM was rank third, demonstrating its effectiveness and robustness.

We also used Wilcoxon signed ranked sum test to assess the significance of the difference between MULTICOM and each individual QA method. The fifth column of Table 4 shows *P*-value of the top one model's Z-score difference between MULTICOM and each QA method. According to 0.05 threshold, MULTICOM performed significantly better than any individual QA method.

In addition, in order to test the impact of each single-model QA method on the performance of the consensus approach, we tested how removing each QA method may change the average Z-score of top one model selected by the consensus ranking of the remaining 13 QA methods. The results were in column 6 in Table 4. According to the results, the removal of MULTICOM-NOVEL caused the biggest decrease in the average Z-score of top one models selected by the consensus method.

Moreover, we counted the total number of times one QA method selected better models than all other QA methods. In the *cases* where more than one QA method selected the same better model, all of them were counted as better than others methods once. Table 5 shows that MULTICOM consistently selected better top models more frequently than any other QA method. Interestingly, SELECTpro only selected better model once (Table 5), yet it had the higher average GDT-TS scores for all the top one models than the other 13 individual QA methods (Table 4), suggesting that SELECTpro selected top models with relatively higher GDT-TS score for most targets, but not necessarily the best models compared with other individual QA methods.

In addition to assessing the overall performance, we specifically investigated two examples to illustrate how MULTICOM assessed the quality of the models of the following two targets. The first case is T0783-D2 (domain 2 of Target T0783). Figure 2A illustrates the distribution of the GDT-TS scores of the models of this domain, where most of the models actually have the true GDT-TS score less than 0.2 (i.e. very low quality), some models have the GDT-TS score around 0.4 (medium quality), and a few models have GDT-TS score 0.6 (relatively good quality). Figure 2B is the plot of true GDT-TS scores of these models against their ranking predicted by MULTICOM. It is shown that MULTICOM ranked the best model with the highest GDT-TS score (e.g. nns_TS1) as no. 1. In this case, all the individual single QA methods ranked this model within top five, but a pairwise method ranked it at no. 19. Combining these individual rankings, the consensus ranking predicted by MULTICOM

Table 4. Comparison of MULTICOM with each QA method and the two different consensus methods (one based on 6 QA methods and another one based on 14 QA methods) on the average GDT-TS score and Z-score of the top models selected, and the significance of difference between each QA method and MULTICOM

QA method	Ave. GDT-TS score on all	Ave. GDT-TS score on TBM	Ave. Z-score on all	*P*-value of Z-score diff.	Ave. Z-score removed
MULTICOM	0.374	0.425	1.364	–	–
Consensus of 14 QA scores	0.369	0.420	1.217	–	–
Consensus of 14 Z-scores	0.357	0.402	1.406	–	–
SELECTpro	0.351	0.407	0.893	1.831e-05	1.338
ProQ2	0.343	0.387	0.887	1.19e-02	1.365
MULTICOM-NOVEL	0.340	0.383	0.861	5.612e-03	**1.321**
ModFOLDclust2	0.339	0.399	0.734	2.074e-04	1.356
APOLLO	0.338	0.403	0.584	9.331e-05	1.379
Dope	0.334	0.382	0.819	1.861e-03	1.360
Pcons	0.333	0.397	0.565	1.831e-05	1.325
ModelEva	0.333	0.378	0.870	9.840e-03	1.334
Dfire2	0.329	0.367	0.826	1.662e-03	1.360
QApro	0.328	0.371	0.783	2.889e-02	1.430
RWplus	0.327	0.373	0.752	5.193e-04	1.365
OPUS-PSP	0.326	0.366	0.793	5.784e-03	1.356
RF_CB_SRS	0.300	0.343	0.372	7.13e-05	1.365
Modelcheck2	0.297	0.347	0.559	1.192e-02	1.340

Italic font denotes single-model methods.

Table 5. The total number times that each QA method performed better than other QA methods on all human targets or all template-based (TBM) human targets only

QA methods	Frequency on all targets	QA methods	Frequency on TBM
MULTICOM	17	MULTICOM	11
QApro	12	QApro	8
ProQ2	11	*ModelEva*	7
ModelEva	9	*ProQ2*	7
Dfire2	9	*Dope*	7
Dope	9	*RWplus*	6
RWplus	8	*Dfire2*	6
MULTICOM-NOVEL	8	*MULTICOM-NOVEL*	6
OPUS-PSP	8	*OPUS-PSP*	6
Modelcheck2	4	*APOLLO*	4
RF_CB_SRS	4	*Modelcheck2*	3
APOLLO	4	*RF_CB_SRS*	3
ModFOLDclust2	3	ModFOLDclust2	3
Pcons	2	Pcons	2
SELECTpro	1	*SELECTpro*	1

Italic denotes single-model methods.

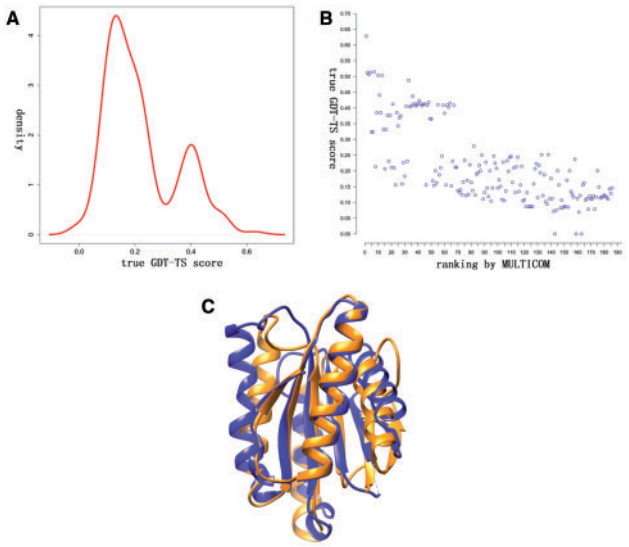

Fig. 2. Tertiary structure prediction of domain 2 of T0783 (T0783-D2). (**A**) The superposition of the MULTICOM human TS1 model on domain 2 with the native structure. (**B**) The distribution of 191 models in the model pool. (**C**). The plot of the true GDT-TS scores of models against their predicted ranking

was able to select this model to combine with other three similar models (nns_TS3, nns_TS2, and FFAS-3D_TS1) to generate a refined model as final prediction. Figure 2C is the superposition of this model with the native structure, which is an alpha-best-alpha protein. Our final model has a well-predicted four-strand beta-sheet in the middle and two well-positioned alpha helices in periphery. The final GDT-TS score of this model is 0.625.

The second case is T0767-D1 (domain 1 of Target T0767). Figure 3A shows the distribution of the true GDT-TS score for the whole model pool. Most models are of low quality (i.e. the true

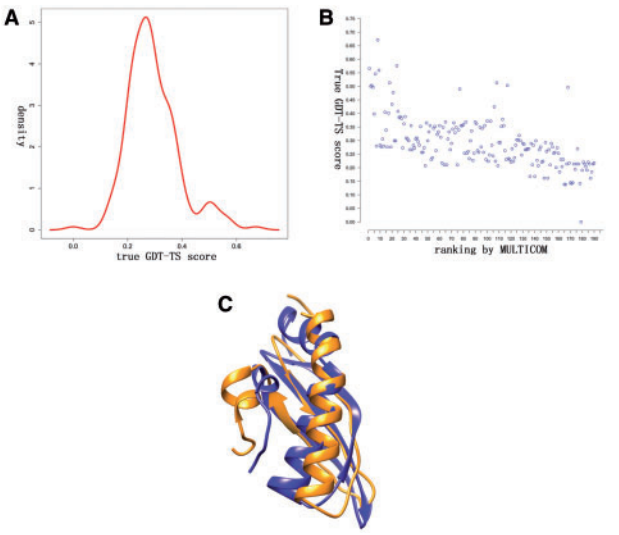

Fig. 3. Tertiary structure prediction of domain 1 of T0767 (T0767-D1). (**A**) The superposition of the MULTICOM human TS1 model on domain 1 with the native structure. (**B**) The distribution of 195 models in the model pool. (**C**) The plot of the true GDT-TS scores of models against their predicted ranking

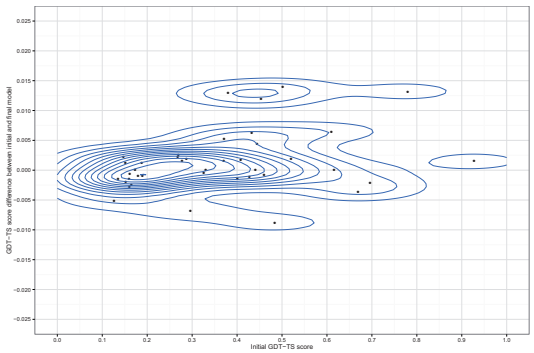

Fig. 4. The plot of the difference between the initial GDT-TS scores before model combination and the GDT-TS scores after model combination against the initial GDT-TS scores of top one models of 42 targets

GDT-TS score around 0.25), which makes model QA difficult. Therefore, three pairwise QA methods (APOLLO, Pcons and ModFOLDclust2) failed to rank the models of good quality at or near the top, whereas some single-model QA methods ranked them higher. Figure 3B is the plot of the true GDT-TS scores of these models against their ranking predicted by MULTICOM. It is shown that our large-scale model QA combining both single- and multi-model QA methods was able to rank the third best model at the top, even though it missed the best model BAKER-ROSETTASERVER_TS2 in the model pool. The initial model selected by MULTICOM was Zhang-Server_TS5 with GDT-TS score 0.5658. Figure 3C visualizes the superposition of the predicted model and the native structure. It is shown that the beta sheet was predicted rather accurately, whereas the alpha helices were only partly correctly predicted.

Finally, we investigated if the model combination could refine and improve the quality of the selected models. Figure 4 shows the difference between the initial GDT-TS scores of the models before refinement and the GDT-TS scores of the final models after the refinement process on 42 CASP11 human targets. The GDT-TS scores of the

models of 19 targets were increased by the model combination, those of another 19 targets were decreased, and those of the remaining four targets stayed the same. The average change of GDT-TS scores of all 42 targets was 0, suggesting the refinement process did not improve the global quality of the models on average, which is consistent with the observation on the performance of most current model refinement protocols (Bhattacharya and Cheng, 2013).

4 Conclusion

We developed a large-scale model QA technique in conjunction with model clustering and refinement to improve protein tertiary structure prediction. Inspired by the previous work (Pawlowski *et al.*, 2008) that integrated several primary QA methods, our method that combined a large number of protein model QA methods reliably and consistently improved protein model ranking—one of the major challenges of protein structure prediction. For the first time, we demonstrate that this large-scale consensus QA approach is more robust and accurate than any individual quality method by integrating their strength together. Our tertiary structure prediction based on this method outperformed all the server predictors during the very competitive CASP11 experiment in 2014. The CASP11 official assessment also ranked our method as one of the top three best tertiary structure prediction methods on all the CASP11 human targets. This outstanding performance demonstrates our large-scale model QA approach is a promising direction to advance the state of the art of protein model ranking and selection. Moreover, our approach adopts an open QA system, into which, adding more complimentary methods may potentially improve the ranking, but incorporating redundant methods does not necessarily lead to an improvement. However, our general combination approach demonstrates the importance of developing more individual QA methods and the possibility of optimally combining them together to advance the field of protein structure prediction.

Funding

The research was partially supported by an NIH grant (R01GM093123) to J.C. We thank anonymous reviewers for valuable comments.

Conflict of Interest: none declared.

References

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhattacharya,D. and Cheng,J. (2013) 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. *Proteins Struct. Funct. Bioinform.*, **81**, 119–131.

Bowie,J.U. *et al.* (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Cao,R. *et al.* (2014a) Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. *BMC Struct. Biol.*, **14**, 13.

Cao,R. *et al.* (2014b) SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics*, **15**:120.

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

Cheng,J. *et al.* (2012) The MULTICOM toolbox for protein structure prediction. *BMC Bioinformatics*, **13**, 65.

Dobson,C.M. *et al.* (1998) Protein folding: a perspective from theory and experiment. *Angewandte Chemie International Edition*, **37**, 868–893.

Eisenhaber,F. *et al.* (1995) Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 1–94.

Faraggi,E. *et al.* (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comput. Chem.*, **33**, 259–267.

Jones,D.T. *et al.* (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sec. A*, **32**, 922–923.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Källberg,M. *et al.* (2012) Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.*, **7**, 1511–1522.

Li,J. *et al.* (2013) Designing and benchmarking the MULTICOM protein structure prediction system. *BMC Struct. Biol.*, **13**, 2.

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

Lu,M. *et al.* (2008) OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.*, **376**, 288–301.

McGuffin,L.J. and Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, **26**, 182–188.

Mishra,A. *et al.* (2013) Capturing native/native like structures with a physicochemical metric (pcSM) in protein folding. *Biochim. Biophys. Acta (BBA) Proteins Proteomics*, **1834**, 1520–1531.

Pawlowski,M. *et al.* (2008) MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics*, **9**, 403.

Randall,A. and Baldi,P. (2008) SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Struct. Biol.*, **8**, 52.

Ray,A. *et al.* (2012) Improved model quality assessment using ProQ2. *BMC Bioinformatics*, **13**, 224.

Rigden,D.J. (2009) *From Protein Structure to Function with Bioinformatics*. Springer, Dordrecht.

Rykunov,D. and Fiser,A. (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. *Proteins Struct. Funct. Bioinform.*, **67**, 559–568.

Shen,M.Y. and Sali,A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, **15**, 2507–2524.

Shortle,D. *et al.* (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci.*, **95**, 11158–11162.

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

Wallner,B. and Elofsson,A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci.*, **15**, 900–913.

Wang,Z. *et al.* (2009) Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins Struct. Funct. Bioinform.*, **75**, 638–647.

Wang,Z. *et al.* (2010) MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics*, **26**, 882–888.

Wang,Z. *et al.* (2011) APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics*, **27**, 1715–1716.

Xu,D. and Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinform.*, **80**, 1715–1735.

Yang,Y. and Zhou,Y. (2008) Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci.*, **17**, 1212–1219.

Yonath,A. (2011) X-ray crystallography at the heart of life science. *Curr. Opin. Struct. Biol.*, **21**, 622–626.

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

Zhang,J. and Xu,D. (2013) Fast algorithm for population-based protein structural model analysis. *Proteomics*, **13**, 221–229.

Zhang,J. and Zhang,Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS One*, **5**, e15386.

Zhang,Y. (2008a) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**, 40.

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

Zhang,Y. (2014) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins Struct. Funct. Bioinform.*, **82**, 175–187.

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