Evaluation of Protein Structural Models Using Random Forests

Renzhi Cao[1], Jie Hou[2], Taeho Jo[2,3], Jianlin Cheng[2*]

[1]Department of Computer Science, Pacific Lutheran University,
Tacoma, WA 98447, USA

[2]Department of Computer Science, University of Missouri,
Columbia, MO 65211, USA

[3]Department of Biological Chemistry, University of Michigan,
Ann Arbor, MI, 48109, USA

*Corresponding author: chengji@missouri.edu

Abstract

Protein structure prediction has been a “grand challenge” problem in the structure biology over the last few decades. Protein quality assessment plays a very important role in protein structure prediction. In the paper, we propose a new protein quality assessment method which can predict both local and global quality of the protein 3D structural models. Our method uses both multi and single model quality assessment method for global quality assessment, and uses chemical, physical, geo-metrical features, and global quality score for local quality assessment. CASP9 targets are used to generate the features for local quality assessment. We evaluate the performance of our local quality assessment method on CASP10, which is comparable with two stage-of-art QA methods based on the average absolute distance between the real and predicted distance. In addition, we blindly tested our method on CASP11, and the good performance shows that combining single and multiple model quality assessment method could be a good way to improve the accuracy of model quality assessment, and the random forest technique could be used to train a good local quality assessment model.

1. Introduction

The protein structure prediction has been defined as one of the grand challenges problem in bioinformatics and computational biology, and still not solved over the last several decades [1]. With the development of computer, a huge quantity of computational methods has been generated to predict the protein tertiary structure from the amino acid [2-7]. These methods are mainly divided into the following classes: the template-based methods [4, 7], which uses the known protein structure determined by biology experiment as template to predict the structure of new query protein sequence; the template free methods [5, 6], which try to predict the protein structure from the amino acid sequence without directly using any known protein structures; hybrid methods [2, 3, 8], which takes advantage of the previous two different methods, and generate more accurate protein structures. For all of these different protein structure prediction methods, there is one common important and unsolved problem: which predicted protein structure model is closer to the truth without knowing the native protein structure? That is the protein quality assessment (QA) problem. The model mentioned in this paper represents protein 3D structural model. The protein quality assessment is very useful for selecting the good models from the model pool, to refine the predicted models, and etc [9]. In general, there are two different qualities for the predicted protein model: local and global quality. The global quality score shows how close of the predicted model to the native structure, and the local quality score shows how close of each residues in the predicted model to the native structure. There are two different strategies to evaluate the quality of a predicted model [10]: multi-model methods [9, 11-15] and single-model methods [16-20]. Multi-model methods use pairwise or clustering technique to compare the similarity of each model against all others, and then define the good model as the one which is most similar to all other model. This method works well when a large proportion of the model pool has good quality, such as the case of easy template-based modeling. However, it may fail when a significant portion of low quality modes are dominating the pairwise model comparison since they are very similar to each other [10]. The single-model [16-20] methods predict the predicted protein model's quality without using the information of other models.

In this paper, we introduce a hybrid method to predict the global quality score of the input models, and one random forest based method to predict the local quality score. For the global quality assessment, the pairwise score [10] is generated from the model pool, and an improved version of model evaluator [21] (model check2 score) is also calculated. Either the pair score, or model check2 score is used as the final global quality score. For the local quality assessment, the local features are generated from physical, chemical, and geometrical respective[22] using sliding window size 15 centered on a target residue, and also using the global features from the whole model. Random forest is an ensemble classification which uses tree-structured classifiers. Random forest grows a large number of decision trees, trains them applying the general technique of bootstrap aggregating (bagging). The predictions are determined by majority vote of trees. Because the ensemble reduces variance, random forest is robust to change in data, irrelevant features, and unbalanced class distribution. Random forest showed excellent performance in broad classification tasks [23], which is generally comparable to that of other ensemble classifiers such as AdaBoost [24] or traditional machine learning classification algorithm such as SVM [10, 25] or Deep Learning Networks [26].

The rest of the paper is organized as follows: in the methods section, we will describe the method we use to predict the local and global quality of the input models; in the discussion section, we will evaluate the performance of our method on CASP10, and compare our method with other methods, and then discuss our method's performance; in the conclusion section, we summarize our work.

2. Results and discussion

The global quality assessment is tested on CASP10 targets, and also blindly benchmarked on CASP11 targets. We trained our local quality assessment model based on CASP9 targets, and five cross validation is used for training the random forest model.

We first evaluate the performance of our global quality assessment method, and then evaluate the performance of our local quality assessment.

2.1. Evaluation of global quality predictions

Our global quality assessment method is a hybrid method, and we choose the maximum pairwise GDT-TS score 0.2 as the threshold to decide which method should be used for the input target, either pairwise method or model check2. **Figure 1** shows the average correlation of pairwise method for CASP10 stage1 and stage2 targets with different maximum pairwise score. The x-axis describes maximum pairwise score threshold for all targets. The y-axis shows the average correlation of all targets within the x threshold. This figures shows that as the maximum pairwise score decreases, the performance of pairwise method also decreases for the CASP10 stage1 targets, and the performance decreases a lot between the maximum pairwise score threshold 0.25 to 0.3 on CASP10 stage2 targets. Considering the performance of pairwise method on CASP10 targets, finally we decide to use the threshold 0.2 to decide whether we use pairwise method.

Table 1 and **Table 2** show the average correlation and loss of our global quality assessment method on CASP10 stage1 and stage2 targets respectively. We also include the performance of other top groups' method, such as ModFOLDclust2 method which is based on clustering technique, and ProQ2 method which is one of the best single quality assessment method on CASP10. As the table shows, the pairwise method ModFOLDclust2 performs better than the single quality assessment ProQ2 method from the respective of average correlation, overall correlation and loss. However, the difference between them becomes less on stage2, and the loss between ModFOLDclust2 and ProQ2 is the same on stage2. This tells us that the single QA method has the similar ability to find out the best model out of the model pool comparing with the pairwise method. As our MULTICOM-REFINE server, we take the advantage of both single and pairwise QA method, and the results show that our method gets better performance on both stage1 and stage2 comparing with the state-of-art QA methods, e.g, the average correlation of our method on stage1 is better than ProQ2, and the loss on stage1 is less than ModFOLDclust2, and our method has the biggest average correlation comparing with other two methods on stage2.

As we know that pairwise model assessment methods worked better when a large portion of models in the pool were of good quality, whereas single-model quality assessment methods performed better on some hard targets when only a small portion of models in the pool were of reasonable quality [10], so it would be interest to see the performance of different methods on the human targets of CASP10. **Table 3** and **Table 4** show the performance of our global quality assessment method on stage1 and stage2 of human targets respectively, and we also include the other group's method like ModFOLDclust2 and ProQ2 method. Indeed, we can see from the table, the pairwise and single QA method gets similar performance on the human targets. Moreover, as we can see from **Table 3**, the average correlation of ProQ2 on human targets is 0.58, which is the same as the pairwise method ModFOLDclust2 (the average correlation on stage1 of all targets is 0.68). The average correlation of MULTICOM-REFINE on stage1 is similar to other methods, and the loss is the smallest between ProQ2 and ModFOLDclust2 method. Similar pattern can be found on stage2 for MULTICOM-REFINE. Overall, MULTICOM-REFINE gets better performance on both stage1 and stage2 of CASP10. In addition, our method attends CASP11, and we also show the performance of our method on stage1 and stage2 for all CASP11 targets comparing with ModFOLDclust2 and ProQ2 method at **Table 5** and **Table 6**. We can find out from **Table 5** that our method is better than state-of-art single model QA method ProQ2 based on average correlation or loss, and is better than ModFOLDclust2 based on the average correlation also. **Table 6** shows that our method gets similar performance comparing with ModFOLDclust2 on stage2, and better that ProQ2 method.

2.2. Evaluation of local quality predictions

We evaluate the performance of our local quality assessment method on CASP10 targets on stage1 and stage2. In order to make comparison with other state-of-art local quality assessment methods, we also evaluate the ProQ2 (single model quality assessment tool) and ModFOLDclust2 (multi-model quality assessment tool). In order to evaluate the performance, we calculate the absolute distance difference between real and predicted local distance for each residue. Smaller difference means higher accurate of the predictions. **Figure 2** and **Figure 3** shows the relationship of the real and predicted distance of 98 CASP10 targets on stage1 and stage2 respectively. The x-axis is the real distance between the native and model, which is divided in to 20 bins. The y-axis shows the average of absolute difference between real and predicted distance in each real distance bin. From these two figures, we can see that our MULTICOM-REFINE's new local quality assessment method based on random forest is comparable with the state-of-art local quality assessment method. Especially when the real distance is less than 7 angstrom, the average absolute difference between real and our prediction is very close to the other two methods. Our method even has smaller average absolute difference comparing with the other two methods, e.g, for 98 CASP10 targets on stage2, the average absolute difference when the real distance is 3 angstrom is 0.61, 0.88, and 1.33 for MULTICOM-REFINE, ProQ2, and ModFOLDclust2 respectively. We may also notice that when the real distance is larger than 7, the clustering method ModFOLDclust2 works better than single model quality assessment ProQ2 and our method. There could be two reasons that our method doesn't perform well when the real distance is large. The first reason is that our method tends to predict smaller distance, and we set a threshold 15 for all predictions, so that there is no prediction larger than 15 for our method. The second reason could be in the training data, we involve more accurate models, so that our trained model is more likely to predict small distance for each residue.

3. Conclusions

In this paper, we describe a local and global quality assessment method. Our global quality assessment method takes advantage of pairwise and single-model QA method, and generates better performance comparing with pairwise method and the single-model QA method. We also evaluate the performance of our method on hard targets, it shows that our method consistently gets better performance comparing with two state-of-art quality assessment method ProQ2 and ModFOLDclust2. For the local quality assessment part, we evaluate our method's performance on CASP10 targets, and it shows that our method is comparable with the other two state-of-art model quality assessment method. In the future, we plan to add more training data to improve the accuracy of our local quality assessment method, and consider influence of domains for training the local quality assessment model from the protein structure. Also for the global quality assessment method, we plan to rigorously test it on larger dataset, and find other better way to combine the pairwise and single-model methods. Overall, we believe our method performs well on the CASP10 targets and also has good performance on CASP11, which shows the potential of combining pairwise and single-model method, and also there are a lot of improvements for the local quality assessment method. The web server is built for public use of our method at: http://calla.rnet.missouri.edu/rfqa/.

4. Methods

The global quality assessment of this paper is a hybrid method, and the local quality assessment of this paper is a single model method, which is trained by random forest technique on CASP9 targets. The method to predict the global and local quality scores are introduced in the following sections.

4.1. Global quality assessment method

First of all, the improved version of model evaluator model check2 is used to calculate the score for each input model. Second, while the number of models is larger than one, the pairwise method is applied to the input model pool [10]. The GDT-TS score of each model against all other model is calculated using TM-score [27], and the average GDT-TS score is calculated for each model as the quality of that model. Finally, the maximum of GDT-TS score among the model pool is used to decide which score to be used as the global quality score as the model pool. The pairwise score is used when the maximum GDT-TS score is larger than 0.2, otherwise, model check2 score is used.

4.2. Local quality features preparation

All CASP9 targets are used to generate the local quality features. There are two different types of local quality features, one is global features coming from the quality of the model, and the other is local features coming from the amino acids with sliding window size 15. For each residue, we generate a feature set for making local quality assessment. In total, 4,719,526 feature sets are generated from CASP9 targets. According to the real quality of each feature set, we divide the data into 5 classes. For example, the first class is that all feature sets with the real quality from 0 to 0.2. We randomly select 10,000 feature sets from each class due to the time complexity of training random forest model with large training data set. The RandomForest package in R [28] is used for training the random forest model. The global features includes the difference between secondary structure and solvent accessibility predicted by Spine X [29] and SSpro4 [30] from the protein sequence and that of a model parsed by DSSP [22], the pairwise Euclidean distance score which is calculated by the average Euclidean distance of the model for all pairwise amino acid pairs divided by the same distance of the extended structure for the model, secondary structure penalty score which is calculated from the mismatch of helix and sheet between the predicted secondary structure and the one parsed from the model [20], surface polar score which is calculated by the fractional area of exposed nonpolar residues [20], weighted exposed area score which is the weighted exposed area divide by the whole area [20], total surface area score which is the total surface area divided by the whole area [20]. The local features for each amino acid is coming from the fragment with sliding window size 15, including the amino acids encoded by a 20-digit vector of 0 and 1, secondary structure difference, pairwise Euclidean distance score, secondary structure penalty score, surface polar score, weighted exposed area score, and total area score generated from the fragment.

4.3. Train a model for local quality assessment by random forest

We divided the 10,000 feature sets which were explained above into 10 equal-size subsets for 10-fold cross validation. Nine subsets were used for training and the remaining subset was used for validation. A number of feature sets were randomly selected from each subset for constructing decision trees and standard decision tree training algorithm was applied. After training, the average probability predicted by these trees was calculated as the local quality score. This procedure was repeated 10 times and the sensitivity and specificity were computed across the 10 trials.

5. References

[1] Y. Zhang and J. Skolnick, "SPICKER: a clustering approach to identify near-native protein folds," *J Comput Chem,* vol. 25, no. 6, pp. 865 - 871, 2004.

[2] J. Li, B. Adhikari, and J. Cheng, "An improved integration of template-based and template-free protein structure modeling methods and its assessment in CASP11," *Protein and peptide letters,* vol. 22, no. 7, pp. 586-593, 2015.

[3] J. Li *et al.*, "The MULTICOM protein tertiary structure prediction system," *Protein Structure Prediction,* pp. 29-41, 2014.

[4] M. Källberg, G. Margaryan, S. Wang, J. Ma, and J. Xu, "RaptorX server: a resource for template-based protein structure modeling," *Protein Structure Prediction,* pp. 17-27, 2014.

[5] D. Bhattacharya, R. Cao, and J. Cheng, "UniCon3D: de novo protein structure

prediction using united-residue conformational search via stepwise, probabilistic sampling," *Bioinformatics,* p. btw316, 2016.

[6] A. Leaver-Fay *et al.*, "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules," *Methods in enzymology,* vol. 487, p. 545, 2011.

[7] J. Söding, A. Biegert, and A. N. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic acids research,* vol. 33, no. suppl 2, pp. W244-W248, 2005.

[8] D. Xu and Y. Zhang, "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge−based force field," *Proteins: Structure, Function, and Bioinformatics,* vol. 80, no. 7, pp. 1715-1735, 2012.

[9] R. Cao, D. Bhattacharya, B. Adhikari, J. Li, and J. Cheng, "Large-scale model quality assessment for improving protein tertiary structure prediction," *Bioinformatics,* vol. 31, no. 12, pp. i116-i123, 2015.

[10] R. Cao, Z. Wang, and J. Cheng, "Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment," *BMC structural biology,* vol. 14, no. 1, p. 13, 2014.

[11] L. J. McGuffin, "The ModFOLD server for the quality assessment of protein structural models," *Bioinformatics,* vol. 24, no. 4, pp. 586-587, 2008.

[12] L. J. McGuffin, "Prediction of global and local model quality in CASP8 using the ModFOLD server," *Proteins: Structure, Function, and Bioinformatics,* vol. 77, no. S9, pp. 185-190, 2009.

[13] Q. Wang, K. Vantasin, D. Xu, and Y. Shang, "MUFOLD−WQA: A new selective consensus method for quality assessment in protein structure prediction," *Proteins: Structure, Function, and Bioinformatics,* 2011.

[14] L. J. McGuffin and D. B. Roche, "Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments," *Bioinformatics,* vol. 26, no. 2, pp. 182-188, 2010.

[15] R. Cao, D. Bhattacharya, B. Adhikari, J. Li, and J. Cheng, "Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11," *Proteins: Structure, Function, and Bioinformatics,* 2015.

[16] B. Wallner and A. Elofsson, "Can correct protein models be identified?," *Protein Science,* vol. 12, no. 5, pp. 1073-1086, 2003.

[17] A. Ray, E. Lindahl, and B. Wallner, "Improved model quality assessment using ProQ2," *BMC bioinformatics,* vol. 13, no. 1, p. 224, 2012.

[18] K. Uziela, B. Wallner, and A. Elofsson, "ProQ3: Improved model quality assessments using Rosetta energy terms," *arXiv preprint arXiv:1602.05832,* 2016.

[19] P. Benkert, S. C. Tosatto, and D. Schomburg, "QMEAN: A comprehensive scoring function for model quality assessment," *Proteins: Structure, Function, and Bioinformatics,* vol. 71, no. 1, pp. 261-277, 2008.

[20] R. Cao and J. Cheng, "Protein single-model quality assessment by feature-based probability density functions," *Scientific reports,* vol. 6, 2016.

[21] Z. Wang, A. N. Tegge, and J. Cheng, "Evaluating the absolute quality of a single protein model using structural features and support vector machines," (in eng), *Proteins,* Research Support, Non-U.S. Gov't vol. 75, no. 3, pp. 638-47, May 15 2009.

[22] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers,* vol. 22, no. 12, pp. 2577 - 2637, 1983.

[23] T. Jo and J. Cheng, "Improving protein fold recognition by random forest," *BMC bioinformatics,* vol. 15, no. Suppl 11, p. S14, 2014.

[24] J. Eickholt and J. Cheng, "Predicting protein residue–residue contacts using deep networks and boosting," *Bioinformatics,* vol. 28, no. 23, pp. 3066-3072, 2012.

[25] R. Cao, Z. Wang, Y. Wang, and J. Cheng, "SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines," *BMC bioinformatics,* vol. 15, no. 1, p. 120, 2014.

[26] T. Jo, Hou, J., Eickholt, J. & Cheng, J., "Improving protein fold recognition by deep learning networks," *Sic. Rep,* vol. 5, p. 17573, 2015.

[27] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics,* vol. 57, no. 4, pp. 702-710, 2004.

[28] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news,* vol. 2, no. 3, pp. 18-22, 2002.

[29] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "SPINE X: improving protein

secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," *Journal of computational chemistry,* vol. 33, no. 3, pp. 259-267, 2012.

[30] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, "SCRATCH: a protein structure and structural feature prediction server," *Nucleic Acids Research,* vol. 33, no. suppl 2, pp. W72-W76, 2005.

Acknowledgements

The work was partially supported by an NIH grant (R01GM093123) to JC.

Author Contributions

JC conceived and designed the method and the system. RC, TJ implemented the method, built the system, carried out the CASP experiments. RC, JH, TJ, JC evaluated and analyzed data. RC, JH, TJ wrote the manuscript. All the authors approved the manuscript.

Competing financial interests

The authors declared no competing financial interests.

Figure

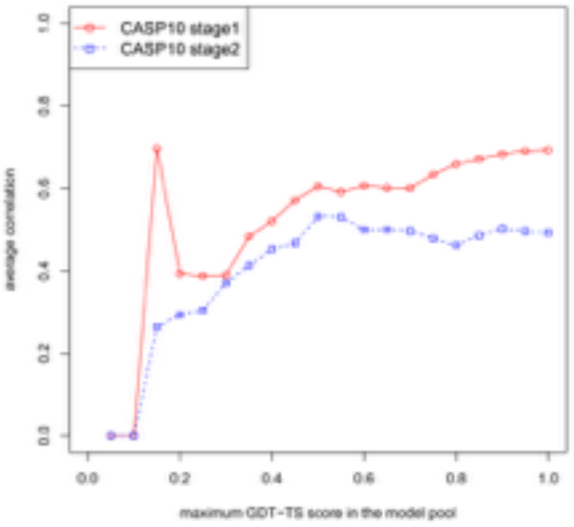

Figure 1. The average correlation of pairwise method for CASP10 stage1 and stage2 targets with different maximum pairwise score

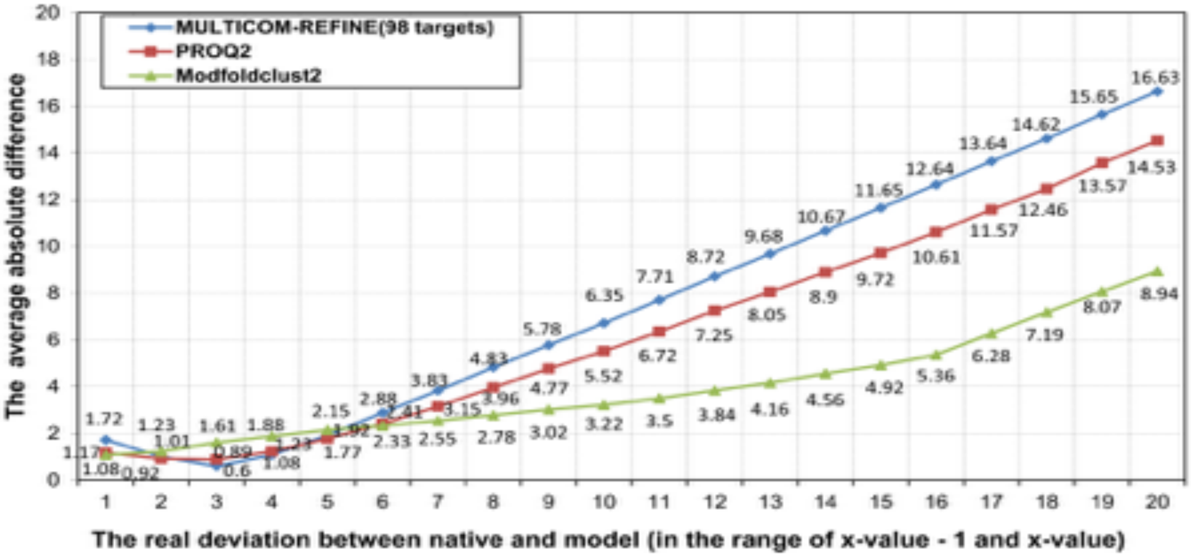

Figure 2. The absolute difference between real and predicted distance against the real distance in 20 bins for 98 CASP10 targets on stage1.

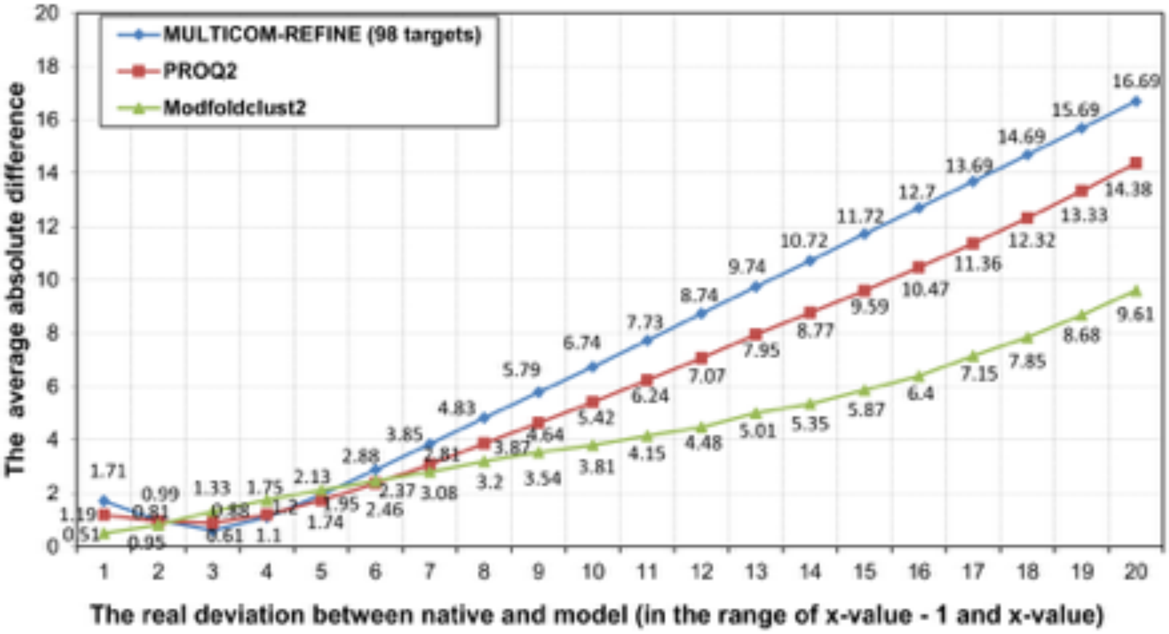

Figure 3. The absolute difference between real and predicted distance against the real distance in 20 bins for 98 CASP10 targets on stage2.

Table 1. The average correlation (Ave. Corr.), overall correlation (Over. Corr.), and average GDT-TS loss (Ave. loss) of MULTICOM-REFINE server, proq2, and ModFOLDclust2 on Stage1 of CASP10.

Stage1 of CASP10	Ave. Corr.	Over. Corr.	Ave. loss
MULTICOM-REFINE	0.68	0.81	0.05
Proq2	0.58	0.61	0.07
ModFOLDclust2	0.68	0.83	0.06

Table 2. The average correlation (Ave. Corr.), overall correlation (Over. Corr.), and average GDT-TS loss (Ave. loss) of MULTICOM-REFINE server, ProQ2, and ModFOLDclust2 on Stage2 of CASP10.

Stage2 of CASP10	Ave. Corr.	Over. Corr.	Ave. loss
MULTICOM-REFINE	0.48	0.83	0.05
ProQ2	0.42	0.60	0.05
ModFOLDclust2	0.45	0.83	0.05

Table 3. The average correlation (Ave. Corr.), overall correlation (Over. Corr.), and average GDT-TS loss (Ave. loss) of MULTICOM-REFINE server, ProQ2, and ModFOLDclust2 on Stage1 of all human targets of CASP10.

Stage1 of CASP10	Ave. Corr.	Over. Corr.	Ave. loss

MULTICOM-REFINE	0.59	0.81	0.06
ProQ2	0.58	0.52	0.08
ModFOLDclust2	0.58	0.86	0.08

Table 4. The average correlation (Ave. Corr.), overall correlation (Over. Corr.), and average GDT-TS loss (Ave. loss) of MULTICOM-REFINE server, ProQ2, and ModFOLDclust2 on Stage2 of all human targets of CASP10.

Stage2 of CASP10	Ave. Corr.	Over. Corr.	Ave. loss
MULTICOM-REFINE	0.50	0.85	0.06
ProQ2	0.41	0.48	0.06
ModFOLDclust2	0.46	0.87	0.05

Table 5. The average correlation (Ave. Corr.), overall correlation (Over. Corr.), and average GDT-TS loss (Ave. loss) of MULTICOM-REFINE server, ProQ2, and ModFOLDclust2 on Stage1 of all human targets of CASP11.

Stage1 of CASP11	Ave. Corr.	Over. Corr.	Ave. loss
MULTICOM-REFINE	0.80	0.93	0.05
ProQ2	0.64	0.79	0.09
ModFOLDclust2	0.74	0.95	0.05

Table 6. The average correlation (Ave. Corr.), overall correlation (Over. Corr.), and average GDT-TS loss (Ave. loss) of MULTICOM-REFINE server, ProQ2, and ModFOLDclust2 on Stage2 of all human targets of CASP11.

Stage2 of CASP11	Ave. Corr.	Over. Corr.	Ave. loss
MULTICOM-REFINE	0.57	0.95	0.07
ProQ2	0.37	0.76	0.06
ModFOLDclust2	0.56	0.95	0.07