An iterative self-refining and self-evaluating approach for protein model quality estimation

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

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

Received 29 June 2011; Revised 30 September 2011; Accepted 31 October 2011
DOI: 10.1002/pro.764
Published online 4 November 2011 proteinscience.org

Abstract: Evaluating or predicting the quality of protein models (i.e., predicted protein tertiary structures) without knowing their native structures is important for selecting and appropriately using protein models. We describe an iterative approach that improves the performances of protein Model Quality Assurance Programs (MQAPs). Given the initial quality scores of a list of models assigned by a MQAP, the method iteratively refines the scores until the ranking of the models does not change. We applied the method to the model quality assessment data generated by 30 MQAPs during the Eighth Critical Assessment of Techniques for Protein Structure Prediction. To various degrees, our method increased the average correlation between predicted and real quality scores of 25 out of 30 MQAPs and reduced the average loss (i.e., the difference between the top ranked model and the best model) for 28 MQAPs. Particularly, for MQAPs with low average correlations (<0.4), the correlation can be increased by several times. Similar experiments conducted on the CASP9 MQAPs also demonstrated the effectiveness of the method. Our method is a hybrid method that combines the original method of a MQAP and the pair-wise comparison clustering method. It can achieve a high accuracy similar to a full pair-wise clustering method, but with much less computation time when evaluating hundreds of models. Furthermore, without knowing native structures, the iterative refining method can evaluate the performance of a MQAP by analyzing its model quality predictions.

Keywords: protein model quality assessment; protein structure prediction; iterative Algorithm; model ranking

Introduction

Nowadays, computer programs can generate a large number of protein models in a relatively short time, which makes protein model quality evaluation/assessment indispensible. Protein model quality assessment programs (MQAPs) can predict the qualities of protein models before knowing the experimental structures, which is essential to the proper usage of the models.[1–3] Current model quality assessment programs can predict both global and local qualities of one or multiple models. The methods used to predict global qualities can be categorized as multiple-model (clustering) methods and single-model methods.

Multiple-model methods assess the quality of a model by assessing its similarity with other models for the same protein target through full pair-wise structural comparisons.[4–10] Single-model methods directly predict the quality of a model from its structural features using machine learning, statistical, or

ZW is supported in part by the Paul K. and Diane Shumaker Endowment in Bioinformatics.

Grant sponsor: National Institutes of Health (NIH); Grant number: 1R01GM093123.

*Correspondence to: Jianlin Cheng, Department of Computer Science, University of Missouri, Columbia, 65211.
E-mail: chengji@missouri.edu

Table I. *Average Correlation, Overall Correlation, and Average Loss of CASP8 MQAPs Before and After Iterative Refinements*

	Average correlation		Overall correlation		Average loss	
	T-test P-value < 0.0001		T-test P-value < 0.0001		T-test P-value < 0.01	
	Bef. Refine	Aft. Refine	Bef. refine	Aft. Refine	Bef. Refine	Aft. Refine
qa-ms-torda-server	0.012	**0.767**	0.110	**0.730**	0.483	**0.149**
ProtAnG_s	0.145	**0.823**	0.100	**0.878**	0.130	**0.070**
MODCHECK-HD	0.284	**0.826**	0.501	**0.858**	0.141	**0.081**
Fiser-QA-COMB	0.476	**0.836**	0.484	**0.856**	0.214	**0.092**
Fiser-QA-FA	0.485	**0.822**	0.287	**0.834**	0.183	**0.105**
Fiser-QA	0.523	**0.857**	0.506	**0.879**	0.176	**0.063**
ModFOLD	0.597	**0.835**	0.681	**0.868**	0.132	**0.076**
SELECTpro	0.608	**0.805**	0.432	**0.844**	0.138	**0.093**
SIFT_SA	0.623	**0.840**	0.459	**0.858**	0.102	**0.074**
MUFOLD-QA	0.633	**0.832**	0.576	**0.872**	0.108	**0.067**
Pcons_ProQ	0.652	**0.860**	0.652	**0.882**	0.114	**0.055**
SIFT_consensus	0.658	**0.850**	0.673	**0.869**	0.097	**0.068**
MULTICOM-RANK	0.665	**0.838**	0.705	**0.867**	0.069	**0.061**
QMEANfamily	0.678	**0.847**	0.733	**0.869**	0.080	**0.058**
GS-MetaMQAP	0.681	**0.843**	0.771	**0.856**	0.124	**0.079**
Circle	0.683	**0.862**	0.658	**0.881**	0.098	**0.055**
QMEAN	0.699	**0.859**	0.740	**0.877**	0.081	**0.060**
MULTICOM-REFINE	0.710	**0.848**	0.772	**0.871**	0.085	**0.061**
MULTICOM-CMFR	0.721	**0.836**	0.734	**0.869**	0.075	**0.066**
Mariner2	0.730	**0.813**	0.877	**0.889**	0.126	**0.068**
FAMSD	0.825	**0.856**	0.661	**0.880**	0.060	**0.058**
selfQMEAN	0.833	**0.842**	0.893	0.892	0.071	**0.063**
GS-MetaMQAPconsII	0.838	**0.866**	0.829	**0.882**	0.074	**0.053**
GS-MetaMQAPconsI	0.860	**0.870**	0.855	**0.883**	0.072	**0.051**
MULTICOM-CLUSTER	0.865	0.847	0.878	0.871	0.064	0.066
LEE-SERVER	0.866	**0.882**	0.778	**0.878**	0.062	**0.056**
MULTICOM	0.879	0.869	0.891	0.886	0.050	**0.049**
QMEANclust	0.886	0.864	0.919	0.909	0.062	**0.056**
ModFOLDclust	0.894	0.856	0.891	0.878	0.053	**0.049**
Pcons_Pcons	0.900	0.840	0.886	0.870	0.055	0.057

Bold fonts denote improvements. According to t-tests, the P-values of observing differences in average correlation, overall correlation, and average loss are less than 0.0001, 0.0001, and 0.01, respectively. The method "ModFOLDclust"[10] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose. Our refinement method improved the performance of some single-model MQAPs, such as QMEAN, to a level close to that of ModFOLDclust.

physical methods.[11–16,21,22] According to recent CASP experiments,[17] multiple-model methods are currently more accurate than single-model methods, although they do not work well if only a small number of models are available or the structures of input models are largely different. Another drawback is that clustering methods usually need relatively long computational time that makes it less efficient and less feasible to be used in daily research. To address these problems, recently a hybrid quality assessment method[18] was developed to integrate the strengths of the two approaches. The hybrid method at first uses a single-model quality assessment method[16] to generate initial quality scores of input models, and then compares the structure of each model with those of the top ranked models. It uses the average structural similarity score with the top ranked models as predicted quality score.

Here we generalize the hybrid approach and use it to refine the quality scores predicted by any MQAPs. The iterative self-refining approach can consistently improve single-model MQAPs in almost all situations in just three iterations. Our results showed that instead of performing full pair-wise comparisons between models, partial pair-wise comparisons against a few top models can achieve similarly high accuracy, but with much less computational time. Moreover, for the first time, the iterative method can help evaluate the performance of a MQAP before knowing the experimental structures. Although our algorithm can also generate local quality scores, in this article, we mainly focus on discussing its performances in improving global quality assessment.

Results and Discussion

We applied our iterative refinement approach to each of the MQAPs that participated in the Eighth Critical Assessment of Techniques for Protein Structure Prediction (CASP8, 2008) and the Ninth Critical Assessment of Techniques for Protein Structure Prediction (CASP9, 2010). Taking CASP8 as an example, we downloaded the predicted quality scores of more than 50,000 tertiary structure (TS) models

Table II. *Average Correlation, Overall Correlation, and Average Loss of CASP9 MQAPs Before and After Iterative Refinements*

	Average correlation		Overall correlation		Average loss	
	T-test P-value < 0.1		T-test P-value < 0.1		T-test P-value < 0.05	
	Bef. Refine	Aft. Refine	Bef. refine	Aft. refine	Bef. refine	Aft. Refine
PconsR	0.052	**0.629**	0.026	**0.743**	0.155	**0.102**
PconsD	0.119	**0.649**	−0.158	**0.605**	0.168	**0.120**
PRECORS-QA	0.260	**0.676**	0.065	**0.694**	0.155	**0.124**
ProQ	0.415	**0.777**	0.665	**0.684**	0.140	**0.092**
MetaMQAP	0.583	**0.783**	0.744	**0.883**	0.143	**0.098**
Baltymus	0.586	**0.810**	0.573	**0.888**	0.117	**0.085**
Distill_NNPIF	0.601	**0.757**	0.626	**0.833**	0.128	**0.096**
ProQ2	0.627	**0.798**	0.781	**0.901**	0.074	**0.072**
ConQuass	0.656	**0.837**	0.722	**0.853**	0.134	**0.093**
MULTICOM-NOVEL	0.662	**0.795**	0.767	**0.890**	0.101	**0.082**
QMEAN	0.685	**0.777**	0.808	**0.889**	0.108	**0.097**
QMEANfamily	0.697	**0.805**	0.805	**0.904**	0.111	**0.088**
Modcheck-J2	0.730	**0.799**	0.820	**0.884**	0.145	**0.093**
Gws	0.769	**0.772**	0.868	**0.893**	0.110	**0.100**
MQAPsingle	0.810	0.766	0.926	0.906	0.100	**0.097**
Splicer_QA	0.818	**0.827**	0.885	**0.914**	0.079	**0.073**
MULTICOM-CONSTRUCT	0.832	0.806	0.903	0.898	0.078	0.078
ModFOLDclustQ	0.832	**0.849**	0.929	0.898	0.062	0.066
QMEANdist	0.833	**0.854**	0.788	**0.863**	0.066	0.071
MULTICOM-REFINE	0.866	0.821	0.929	0.918	0.086	**0.083**
Pcomb	0.870	0.862	0.929	0.892	0.063	**0.061**
MULTICOM	0.885	0.860	0.933	0.925	0.060	**0.059**
PconsM	0.885	0.838	0.930	0.893	0.066	0.066
IntFOLD-QA	0.887	0.870	0.940	0.912	0.060	**0.058**
ModFOLDclust2	0.888	0.863	0.944	0.915	0.061	**0.058**
Pcons	0.893	0.851	0.933	0.881	0.069	**0.066**
MQAPmulti	0.895	0.855	0.932	0.920	0.064	**0.061**
MetaMQAPclust	0.896	0.835	0.936	0.919	0.064	0.065
MULTICOM-CLUSTER	0.916	0.872	0.947	0.912	0.059	0.060
MUFOLD-QA	0.920	0.874	0.941	0.914	0.062	0.062
MUFOLD-WQA	0.920	0.865	0.896	0.888	0.057	0.058
QMEANclust	0.921	0.865	0.950	0.917	0.059	0.061

Bold fonts denote improvements. According to t-tests, the P-values of observing differences in average correlation, overall correlation, and average loss are less than 0.1, 0.1, and 0.05, respectively. The method "MULTICOM-CLUSTER"[23] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose.

associated with 120 CASP8 targets from the CASP8 web site. We also downloaded all the TS models and compared each of them with its true experimental structure using the tool TM_Score.[19] The GDT-TS[20] score resulted from comparison is considered as the real quality score of the model. The real quality scores were used to evaluate whether the iterative quality assessment method improved the initial quality scores predicted by CASP8 MQAPs.

We evaluated the iterative quality assessment method using the following criteria: average and overall correlations of predicted and real GDT-TS scores, and average loss of the GDT-TS scores on top one ranked models. The average correlation is the average of the per-target Pearson correlations between predicted quality scores and real GDT-TS scores. The overall correlation is the Pearson correlation of predicted quality scores and real GDT-TS scores of all models of all CASP8 or CASP9 targets. The loss on a target is the difference between the real GDT-TS score of the top one ranked model and the real GDT-TS score of the best model. The average loss over all targets measures the ranking ability of a MQAP, which ideally equals to zero indicating the program can always rank the best model as the top one model.

Table I reports the average correlation, overall correlation, and average loss of 30 CASP8 MQAPs before and after applying our refinement algorithm. The average (overall) correlations of 25 (24) out of 30 MQAPs were increased. The average losses of 28 MQAPs were reduced. According to t-tests, the P-value of observing the difference before and after refinements for average correlation, overall correlation, and average loss is less than 0.0001, 0.0001, and 0.01, respectively. The correlations of MQAPs with low initial correlation scores (<0.4), such as qa-ms-torda-server and ProtAnG_s, were increased by several times. After refinement, the correlations of all MQAPS except one were improved to above 0.80; and the average losses of all the MQAPs except two were reduced to below 0.10. One extreme example is

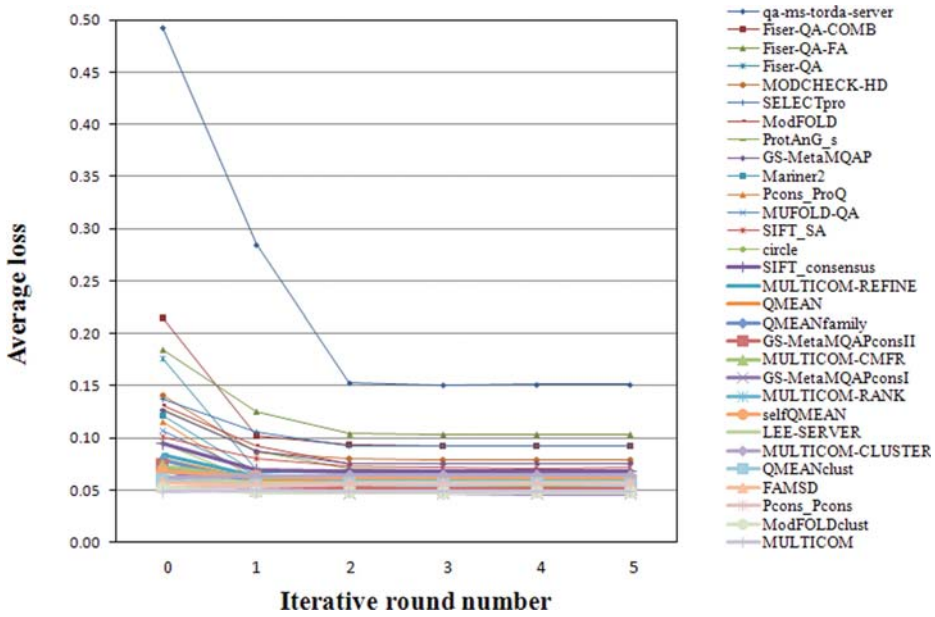

Figure 1. The plot of the average losses against iterations for CASP8 MQAPs. The method "ModFOLDclust"[10] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose.

qa-ms-torda-sever, whose average correlation was improved from 0.012 to 0.767. However, we noticed that the refinement method did not improve the correlations of several clustering-based methods probably because they had already used structural comparisons in their model evaluation process. In contrast, all the single-model methods that do not utilize structural comparisons were improved by the iterative refinement method. The same experiment was performed on 107 valid CASP9 targets (Table T2 II). Our method improved the average correlation, overall correlation, and average loss of almost all CASP9 MQAPs that did not use structural comparisons, such as PconsR, PconsD, PRECORES-QA, ProQ, MetaMAQP, Batymus, DistillNNPIE, ProQ2, ConQuass, MULTICOM-NOVEL, and QMEAN. However, our method rarely improved clustering-based MQAPs that used structural comparisons, such as MULTICOM-CLUSTER, MUFOLD-QA, and QMEANclust, although it slightly reduced the average loss of ModFOLDclust2 and Pcons, two of the top pair-wise comparison methods. According to *t*-test, the *P*-value of the improvements on average correlation and overall correlation is 0.1 for all

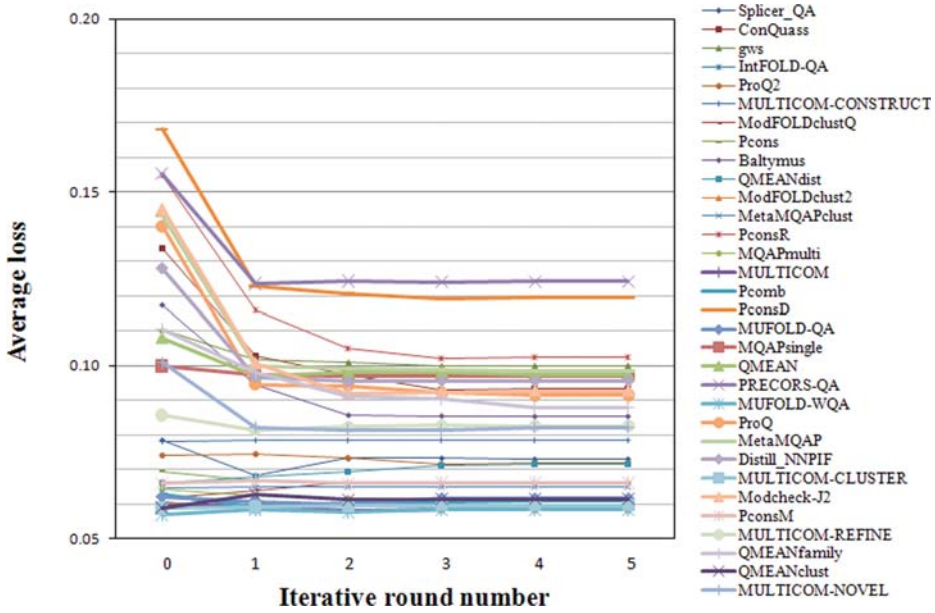

Figure 2. The plot of the average losses against iterations for CASP9 MQAPs. The method "MULTICOM-CLUSTER"[23] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose.

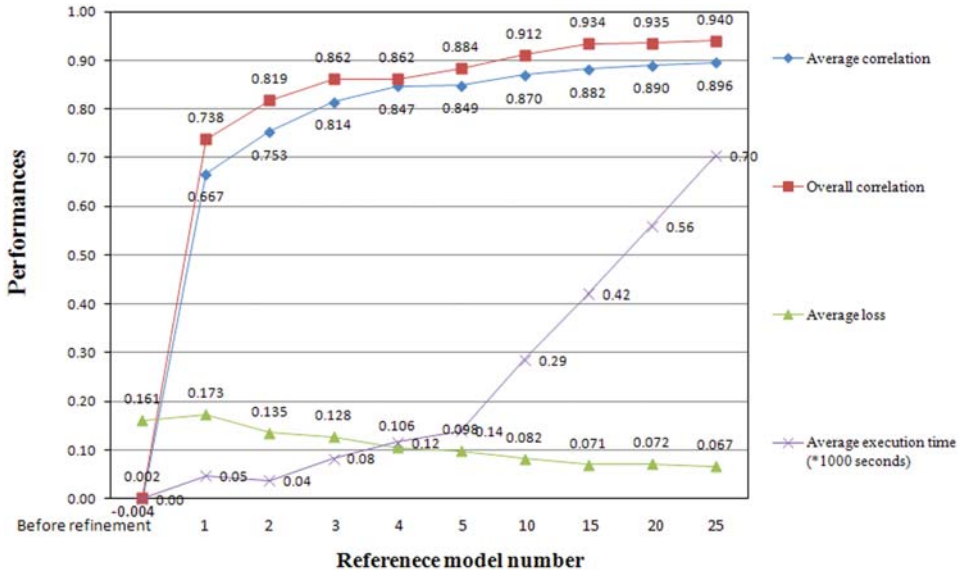

Figure 3. The average correlation, overall correlation, average loss, and average computational time under different numbers of reference models. This experiment was conducted on a MQAP whose predicted quality scores of CASP9 models were randomly generated. The predicted model quality scores had an average correlation of −0.0036 with the true model quality scores. Different numbers of reference models were tested under a single round of refinement.

CASP9 MQAPs, which is less significant than the ones on CASP8 data. This may be because a larger portion of CASP9 MQAPs used structural comparisons. However, the *P*-value of the improvements on loss is still at a significant level 0.05.

To investigate how fast the iterative QA method converged, we plotted the average loss against iterations for each CASP8 MQAP (Fig. 1) and CASP9 MQAP (Fig. 2). Most methods converged in the first one or two iterations (Figs. 1 and 2). On average, it takes up to about five iterations to converge. The number of iterations depends on the quality of initial ranking. Better initial rankings require fewer iterations of refinement.

To investigate how "the number of reference models" influences the refinement performance and also the efficiency of our method, we created a random MQAP on CASP9 dataset (Fig. 3). The predicted model quality scores of this random MQAP were randomly generated, which had an average correlation −0.00357, an overall correlation 0.0021, and a loss 0.161 compared with the true model quality scores. Models were then initially ranked by these randomly generated quality scores. After a

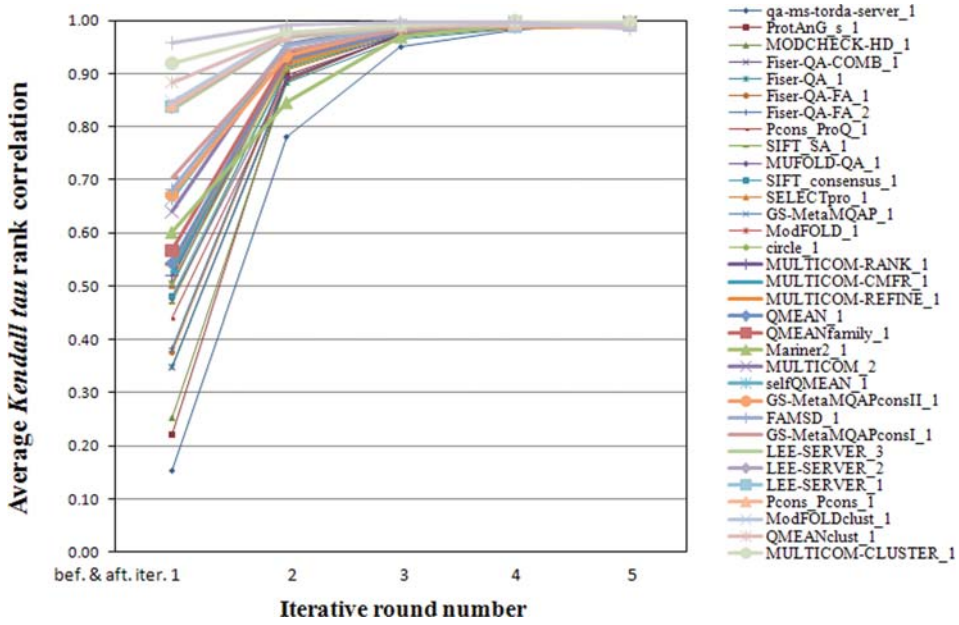

Figure 4. The Kendall tau rank correlations of the rankings before and after each round of refinement for CASP8 MQAPs.

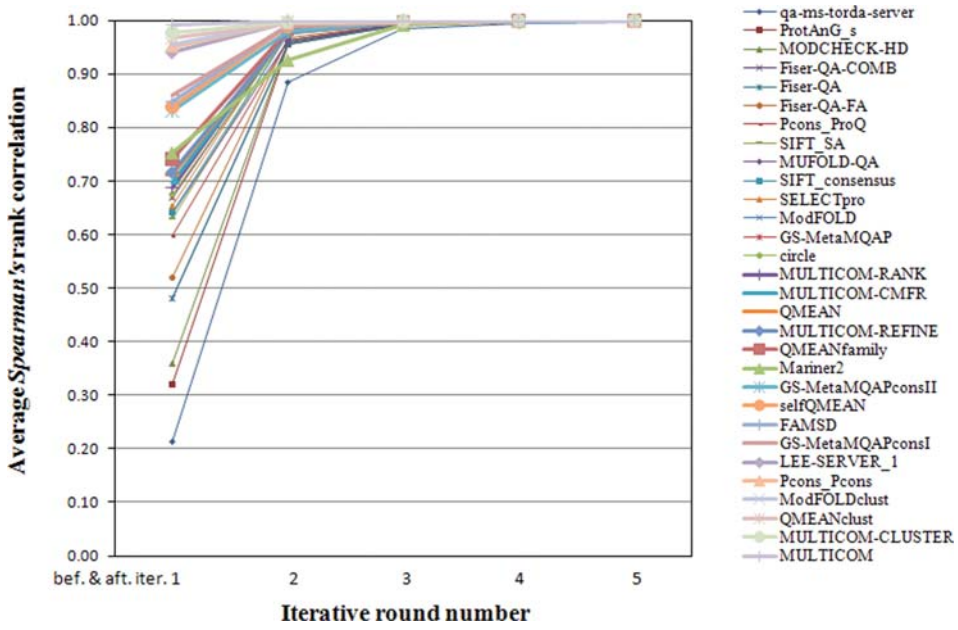

Figure 5. The Spearman's rank correlations of the rankings before and after each round of refinement for CASP8 MQAPs.

single iteration of refinement using top 1 ranked model as reference model, our method substantially improved the average correlation to 0.667 and the overall correlation to 0.738. Moreover, when top 3 ranked models were used as reference models, both the average correlation and overall correlation were improved to 0.814 and 0.862, respectively, after only one round of iteration. The improvement continued as the number of reference models increased and started to saturate after using 15–25 reference models. When 25 top models were used as reference models, the average correlation, overall correlation, and average loss were improved to 0.896, 0.940, and 0.067 respectively, which were much better than the initial ranking generated by the random MQAP. This performance was also close to the average correlation 0.916, overall correlation 0.947, and average loss 0.059 of a full pair-wise comparison method MULTICOM-CLUSTER,[23] which was developed by our group and was ranked as one of the top MQAPs in CASP9 (see Table II).

We studied some cases in which our refinement method worked well or failed in the experiment on the random MQAP mentioned above. We found that it worked well on template-based modeling (TBM) targets whose models are largely of good quality. For

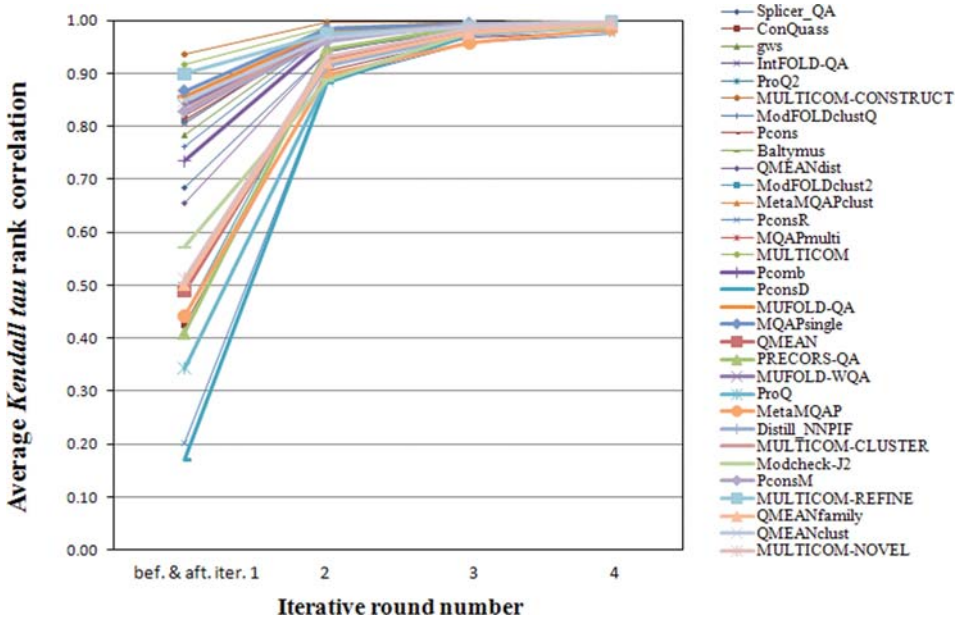

Figure 6. The Kendall tau rank correlations of the rankings before and after each round of refinement for CASP9 MQAPs.

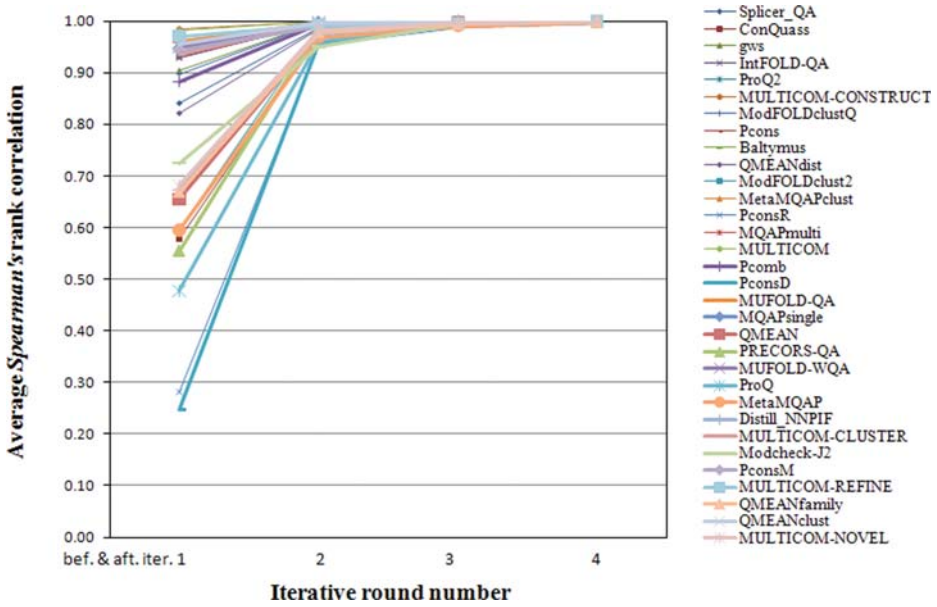

Figure 7. The Spearman's rank correlations of the rankings before and after each round of refinement for CASP9 MQAPs.

example, the predictions of the random MQAP had an average correlation −0.071 on an easy TBM target T0522; and 218 out of 371 models of that target have true GDT-TS[20] scores > 0.9. A GDT-TS score is a structural similarity score that ranges from 0 to 1, whereas 1 indicates the model is the same as the native structure and 0 completely different. After one round of refinement using the top one ranked model as reference model, the average correlation was improved to 0.985. In contrast, our refinement

Table III. *The Average Kendall Tau Ranking Correlation and Average Spearman's Ranking Correlation Before and After the First and Last Iteration Tested on CASP8 MQAPs*

	Average *Kendall tau rank* correlation		Average *Spearman's rank* correlation	
	Bef. & aft. first iter.	Bef. & aft. last iter.	Bef. & aft. first iter.	Bef. & aft. last iter.
qa-ms-torda-server	0.153	0.991	0.214	0.999
ProtAnG_s	0.221	0.993	0.320	0.999
MODCHECK-HD	0.254	1.000	0.358	1.000
Fiser-QA-COMB	0.348	0.998	0.482	1.000
Fiser-QA	0.348	0.989	0.482	0.999
Fiser-QA-FA	0.375	0.991	0.521	0.999
Pcons_ProQ	0.438	0.992	0.597	0.999
SIFT_SA	0.468	0.987	0.630	0.998
MUFOLD-QA	0.476	0.996	0.638	1.000
SIFT_consensus	0.479	0.989	0.642	0.998
SELECTpro	0.499	0.987	0.654	0.999
GS-MetaMQAP	0.503	0.988	0.670	0.999
ModFOLD	0.504	0.991	0.668	0.999
circle	0.506	0.988	0.676	0.998
MULTICOM-RANK	0.518	0.990	0.689	0.998
MULTICOM-CMFR	0.528	1.000	0.696	1.000
MULTICOM-REFINE	0.541	0.983	0.717	0.997
QMEAN	0.542	0.992	0.713	0.999
QMEANfamily	0.567	0.984	0.741	0.998
Mariner2	0.601	0.990	0.753	0.998
selfQMEAN	0.665	0.992	0.830	0.998
GS-MetaMQAPconsII	0.671	0.986	0.838	0.999
FAMSD	0.681	0.991	0.848	0.999
GS-MetaMQAPconsI	0.705	0.999	0.862	1.000
LEE-SERVER	0.837	1.000	0.941	1.000
Pcons_Pcons	0.840	0.989	0.951	0.998
ModFOLDclust	0.847	0.969	0.955	0.995
QMEANclust	0.884	0.961	0.967	0.994
MULTICOM-CLUSTER	0.919	0.978	0.977	0.999
MULTICOM	0.958	0.970	0.993	0.998

Table IV. *The Average Kendall Tau Ranking Correlation and Average Spearman's Ranking Correlation Before and After the First and Last Iteration Tested on CASP9 MQAPs*

	Average *Kendall tau rank* correlation		Average *Spearman's rank* correlation	
	Bef. & aft. first iter.	Bef. & aft. last iter.	Bef. & aft. first iter.	Bef. & aft. last iter.
PconsD	0.170	0.992	0.246	0.999
PconsR	0.201	0.975	0.282	0.994
ProQ	0.343	0.989	0.478	0.998
PRECORS-QA	0.411	0.996	0.555	1.000
ConQuass	0.424	0.981	0.578	0.997
Baltymus	0.438	0.984	0.595	0.997
ProQ2	0.441	0.987	0.596	0.998
MetaMQAP	0.442	0.984	0.595	0.998
QMEAN	0.489	0.998	0.655	1.000
QMEANfamily	0.500	0.995	0.669	0.999
Distill_NNPIF	0.505	0.989	0.673	0.998
MULTICOM-NOVEL	0.512	0.996	0.682	1.000
Modcheck-J2	0.571	0.990	0.725	0.999
QMEANdist	0.654	0.999	0.822	1.000
Splicer_QA	0.684	0.996	0.842	0.999
Pcomb	0.734	0.995	0.884	0.999
ModFOLDclustQ	0.763	0.997	0.897	1.000
Gws	0.783	0.999	0.905	1.000
IntFOLD-QA	0.806	1.000	0.929	1.000
ModFOLDclust2	0.810	0.999	0.932	1.000
MQAPmulti	0.812	0.996	0.932	1.000
Pcons	0.820	0.996	0.941	0.999
PconsM	0.828	0.996	0.944	1.000
MULTICOM-CLUSTER	0.830	0.999	0.936	1.000
MUFOLD-WQA	0.839	0.998	0.951	1.000
MetaMQAPclust	0.847	0.996	0.951	1.000
QMEANclust	0.848	0.998	0.954	1.000
MUFOLD-QA	0.856	0.999	0.961	1.000
MQAPsingle	0.867	0.991	0.948	0.999
MULTICOM-REFINE	0.899	0.998	0.970	1.000
MULTICOM	0.918	1.000	0.983	1.000
MULTICOM-CONSTRUCT	0.936	1.000	0.984	1.000

method did not work well on some of the hard targets whose models are mostly of low quality. For example, the random MQAP has an initial correlation −0.013 on the models of target T0537, which is a hard target that contains two free modeling (FM) domains. The best CASP9 model of the target has a GDT-TS score 0.32 whereas all other models have a GDT-TS score <0.3. After one round of refinement using the top-one ranked model as reference model, the correlation became −0.067. These two extreme examples may suggest that, similarly as clustering method, the iterative refinement method works better if a large portion of input models have reasonable qualities.

Moreover, to investigate how model rankings are changed during the refinement process, we calculated the average Kendall tau rank correlation and Spearman's rank correlation. Kendall tau rank correlation coefficient is defined as

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where n_c is the number of concordant pair of models whose ranking orders are not changed in two rankings, n_d is the discordant pairs, and n is the total number of models in the ranking. Kendall tau rank correlation measures the agreement level between two rankings and ranges from −1 and 1, while 1

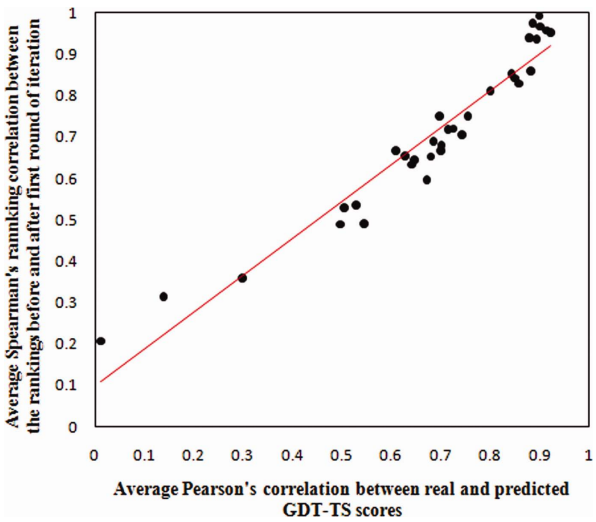

Figure 8. The plot of the Spearman's RCBAF values against the average per-target correlation of the 30 CASP8 MQAPs. Their Pearson's correlation is 0.965. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

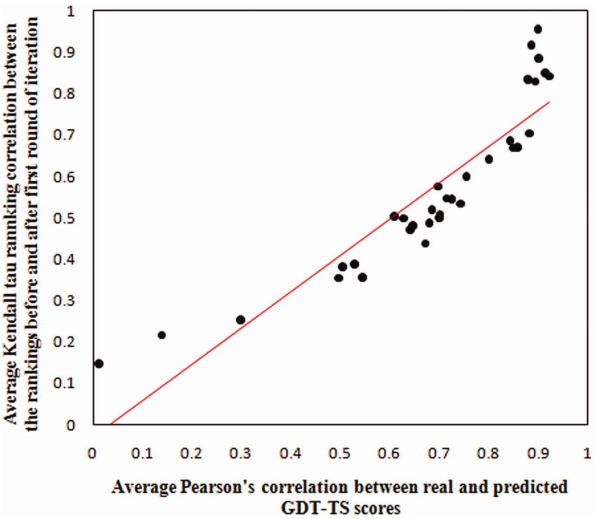

Figure 9. The plot of the Kendall tau RCBAF values against the average per-target correlation of the 30 CASP8 MQAPs. Their Pearson's correlation coefficient is 0.911. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

indicates the two rankings are the same, −1 one ranking is the reverse of the other, and 0 the two rankings are completely independent. The Spearman's rank correlation coefficient is defined as

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i = x_i - y_i$, which is the difference between the ranking orders of a model in two rankings; and n is the number of models in the rankings.

The average Kendall tau and Spearman's rank correlations were plotted against iteration numbers in Figures 4 and 5 for CASP8, and Figures 6 and 7 for CASP9. Similarly as for the average correlation, it took about three iterations to converge on average. For almost all the cases, the biggest increase happened after the first iteration of refinement. The "rank correlations between the rankings before and after the first iteration of refinement" (RCBAF) is particularly interesting since it reports the degree a ranking is changed by the refinement. The RCBAF of initially less accurate MQAPs (e.g., qa-ms-torda-server) is much lower than that of initially more accurate MQAPs (e.g., Pcons-Pcons, ModFOLDclust, QMEANclust, and MULTICOM). Tables III and IV report the average Spearman's and Kendall tau rank correlations before and after the fist and last iteration. The RCBAF for a less accurate MQAP is relatively low (e.g., < 0.5 for Spearman's and < 0.4 for Kendall tau). These suggest that the RCBAF can be used to assess the performance of a MQAP.

To further verify this, we plotted the RCBAF values against average per-target correlations of 30 CASP8 MQAPs (Figs. 8 and 9). The average per-target correlation of a MQAP indicates its actual performance or accuracy. Figures 8 and 9 show that RCBAF have strong correlations with actual accuracies of a MQAP, which are 0.965 and 0.911, respectively. These results indicate that the iterative refinement method can be used to estimate the performance of a MQAP or the accuracy of a model ranking list, without knowing real quality scores of the models. This could be a useful pre-assessing procedure for a MQAP without any other data sources but its own ranking.

Another finding that will contribute to the community is that instead of performing full pair-wise comparisons, partial pair-wise comparisons against a few top models can achieve a similarly high accuracy. This decreases the computational complexity from $O(n^2)$ as of full pair-wise comparisons to linear $O(n)$. This computational efficiency makes our method a fast and accurate alternative to full pair-wise comparison methods, particularly when evaluating a large number of models.

Conclusions

We described an iterative refinement method to improve the initial ranking quality and prediction accuracy of a MQAP. The method can improve the performances of MQAPs in terms of average correlation, overall correlation, and loss. It is particularly effective for single-model MQAPs. Moreover, the iterative refinement method can be used to estimate the performance and accuracy of a MQAP by analyzing how much the initial ranking is changed during the refinement process. Since in reality the real structures are mostly unknown, this unique property makes it a useful tool to self-assess a MQAP.

Materials and Methods

The iterative quality assessment (IQA) method starts from the initial quality scores of a set of protein models. In the first round of refinement, the initial scores are used to rank all models. The top n models are selected as reference models and used to compare with every model by a structural comparison tool TM-Score,[19] which generates a GDT-TS score[20] for each comparison. The average GDT-TS score over the n reference models is used as the refined global quality score of a model. The new, presumably better, quality scores are then used to generate a new ranking of the models for the next round of refinement. The same refinement process is executed iteratively until it converges, that is, the ranking of models does not change any more. The average GDT-TS scores generated in the last round are used as the final global quality scores. When comparing a model to each of the n reference models in each round, TM-Score superimposes two models and outputs the superimposed coordinates of each pair of residues. These coordinates are used to calculate the residue-specific distances. The averaged

residue-specific distances over the n reference models are used as the refined local quality scores. The average residue-specific distances generated in the last round are used as the final local quality scores. The only parameter of the iterative quality assessment is n, the number of reference models, which is set to five during most of our experiments except for Figure 3.

Acknowledgments

The authors thank Dr. Anna Tramontano for suggesting using this approach to self-evaluate the performance of a MAQP.

References

1. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96.
2. Cheng J (2008) A multi-template combination algorithm for protein comparative modeling. BMC Struct Biol 8:18.
3. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18:342–348.
4. McGuffin L (2007) Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinform 8:345.
5. McGuffin L (2008) The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 24:586.
6. Paluszewski M, Karplus K (2008) Model quality assessment using distance constraints from alignments. Proteins 75:540–549.
7. Wallner B, Elofsson A (2007) Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 69:184–193.
8. Zhang Y, Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. J Comp Chem 25:865–871.
9. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015–1018.
10. McGuffin L (2009) Prediction of global and local model quality in CASP8 using the ModFOLD server. Proteins 77:185–190.
11. Archie J, Karplus K (2009) Applying undertaker cost functions to model quality assessment. Proteins 75: 550–555.
12. Benkert P, Tosatto S, Schomburg D (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71:261–277.
13. Cline M, Hughey R, Karplus K (2002) Predicting reliable regions in protein sequence alignments. Bioinformatics 18:306–314.
14. Qiu J, Sheffler W, Baker D, Noble W (2008) Ranking predicted protein structures with support vector regression. Proteins 71:1175.
15. Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086.
16. Wang Z, Tegge A, Cheng J (2008) Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins 75: 638–647.
17. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A (2007) Assessment of predictions in the model quality assessment category. Proteins 69:175–183.
18. Cheng J, Wang Z, Tegge A, Eickholt J (2009) Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 77:181–184.
19. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710.
20. Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31: 3370–3374.
21. Yinghao Wu, Mingyang Lu, Mingzhi Chen, Jialin Li, Jianpeng Ma (2007). OPUS-Ca: a knowledge-based potential function requiring only Ca positions. Protein Sci 16:1449–1463.
22. Randall A, Baldi P (2008) SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. BMC Struct Biol 8:52.
23. Wang Z, Eickholt J, Cheng J (2011) APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics 27:1715–1716.