PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

QUALITY: PREDICTIONS

Prediction of global and local model quality in CASP8 using the ModFOLD server

Liam J. McGuffin*

School of Biological Sciences, University of Reading Whiteknights, Reading RG6 6AS, United Kingdom

ABSTRACT

The development of effective methods for predicting the quality of three-dimensional (3D) models is fundamentally important for the success of tertiary structure (TS) prediction strategies. Since CASP7, the Quality Assessment (QA) category has existed to gauge the ability of various model quality assessment programs (MQAPs) at predicting the relative quality of individual 3D models. For the CASP8 experiment, automated predictions were submitted in the QA category using two methods from the ModFOLD server–ModFOLD version 1.1 and ModFOLDclust. ModFOLD version 1.1 is a single-model machine learning based method, which was used for automated predictions of global model quality (QMODE1). ModFOLDclust is a simple clustering based method, which was used for automated predictions of both global and local quality (QMODE2). In addition, manual predictions of model quality were made using ModFOLD version 2.0—an experimental method that combines the scores from ModFOLDclust and ModFOLD v1.1. Predictions from the ModFOLDclust method were the most successful of the three in terms of the global model quality, whilst the ModFOLD v1.1 method was comparable in performance to other single-model based methods. In addition, the ModFOLDclust method performed well at predicting the per-residue, or local, model quality scores. Predictions of the per-residue errors in our own 3D models, selected using the ModFOLD v2.0 method, were also the most accurate compared with those from other methods. All of the MQAPs described are publicly accessible via the ModFOLD server at: http://www.reading.ac.uk/bioinf/ModFOLD/. The methods are also freely available to download from: http://www.reading.ac.uk/bioinf/downloads/.

Proteins 2009; 77(Suppl 9):185–190.

Key words: model quality assessment program; protein structure prediction; comparative modelling; fold recognition; quality prediction; clustering; metaserver; consensus.

INTRODUCTION

A plethora of algorithms for protein tertiary structure (TS) prediction have been developed over the successive CASP experiments. Often numerous alternative three-dimensional (3D) models can be generated for a given protein target and choosing between them is a considerable challenge. Model quality assessment programs (MQAPs) have been developed that allow us predict the global and/or per-residue accuracy of 3D models of proteins. Thus, using MQAPs we are able to discriminate between low and high quality models, given multiple alternatives. An accurate model quality assessment (QA) strategy has therefore often been a key component of successful TS prediction methods. Since CASP7,[1] QA has been treated as a separate prediction category, and this has allowed to identify the optimal techniques that purely focus on the prediction of model quality. In general, model quality prediction methods can be separated into two broad categories—the single-model based methods and the clustering based methods. However, the boundaries between these categories are not always distinct.

The single-model based methods, as the name suggests, only require 1 model (and possibly the target sequence) as an input, to produce global or local (per-residue) model quality scores. Numerous single-model based MQAPs have been developed in recent years and the most successful of these have been those that combine a number of input features from a single model, either using

The author states no conflict of interest.
Grant sponsor: RCUK Academic Fellowship.
*Correspondence to: Liam J. McGuffin, School of Biological Sciences, University of Reading Whiteknights, Reading RG6 6AS, United Kingdom.
E-mail: l.j.mcguffin@reading.ac.uk
Received 12 March 2009; Revised 23 April 2009; Accepted 12 May 2009
Published online 2 June 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22491

various machine learning methods[2–5] or multivariate linear regression.[6,7] These single-model based methods are often the best option if only a few models are available from an individual server. However, in the case of the CASP experiment, hundreds of models are produced from a variety of servers. In a real world scenario, a biologist may also have access to many alternative models, which they may have obtained via a meta-server.

It was clear from the results of CASP7[1] that the optimal methods for model QA were those that incorporated some variety of clustering, when hundreds of alternative models were available for each target. Clustering based methods work by comparing multiple models against one another using a structural comparison method, to determine individual model quality scores. In general, the models with the least average structural separation from all other models are selected as the highest quality models, although more complex scoring methods and clustering approaches are sometimes also used.[8]

The ModFOLD server[9] was developed to provide users with the option of carrying out model quality predictions using either the single-model based method, ModFOLD (v1.1), or the clustering based method, ModFOLDclust. The predictions from both methods were entered as automated servers in the CASP8 experiment. In addition, a newer method, ModFOLD v2.0, was trialled, which attempted to combine the output scores from ModFOLDclust and ModFOLD v1.1, blurring the boundary between clustering and single-model based methods.

METHODS

ModFOLD (Group 199) and ModFOLDclust (Group 031)

The original version of the ModFOLD method[4] was originally tested at CASP7 in the QMODE1 prediction category. The method performed reasonably well in terms of the global correlation between predicted and observed model quality scores, but not so well on per-target basis. The original ModFOLD protocol combined scores obtained from the ModSSEA method,[4] the MODCHECK method[10] and the two ProQ methods[3] using a neural network trained with the TM-score[11] as a measure of the observed model quality. An early version of the ModFOLDclust method for global model quality, which was based on a simplified version of the 3D-Jury method,[12] was previously shown to significantly outperform every MQAP method tested using the CASP7 data.[4] The ModFOLDclust global score was calculated as below:

$$Q = \frac{1}{N-1}\sum_{m \in M} T_m$$

where Q was the global quality score for a model, N was the number of models for the target, $N-1$ was the number of pairwise structural alignments carried out for each model (i.e. models were not aligned with themselves), M was the set of alignments and T_m was the TM-score for each pairwise alignment of models. A TM-score cut-off was implemented so that alignments with scores <0.2 were not included in the calculation. Therefore, the size of set M was equal to the number of alignments with TM-scores ≥ 0.2.

Both the ModFOLD and ModFOLDclust methods have since been updated and integrated into a web server,[9] which was then tested at CASP8. The server version of the ModFOLD method (version 1.1) was retrained to include two additional secondary structure scores as inputs to the neural network, similar to those used by Eramian *et al.*[2] In addition, to the global clustering score, the ModFOLDclust method was updated to incorporate per-residue model quality scores. The local model quality was evaluated by using a score similar to the average *S*-score,[13] which has previously been used for model evaluation in both the 3D-SHOTGUN method[14] and the Pcons server.[5,15] For a residue in a pairwise superposition the *S*-score was defined as:

$$S_i = \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2}$$

where S_i was the *S*-score for residue i in a model, d_i was the distance between aligned residues according to the TM-score superposition and d_0 was the distance threshold (3.9). An S_i score of 0 was given if $d_i > 3.9$ Å. The *S*-scores for each residue were summed and the mean score was calculated:

$$S_r = \frac{1}{N-1}\sum_{a \in A} S_{ia}$$

where S_r was the predicted residue accuracy for the model, N was the number of models for the target, A was the set of alignments and S_{ia} was the S_i score for a residue in a structural alignment (a). The size of set A was equal to $N-1$. The mean *S*-score for each residue was then converted to the predicted distance from the equivalent residue in the native structure (d_r), by simply rearranging the equation for the *S*-score:

$$d_r = d_0\sqrt{\frac{1}{S_r} - 1}$$

An upper limit of 15 Å was set for d_r. Missing residues in the model were represented by an "X" in the prediction.

ModFOLD version 2.0—a combined approach to model QA (Group 379)

A novel version of ModFOLD (v2.0) was also developed, which attempted to combine the single-model based Mod-

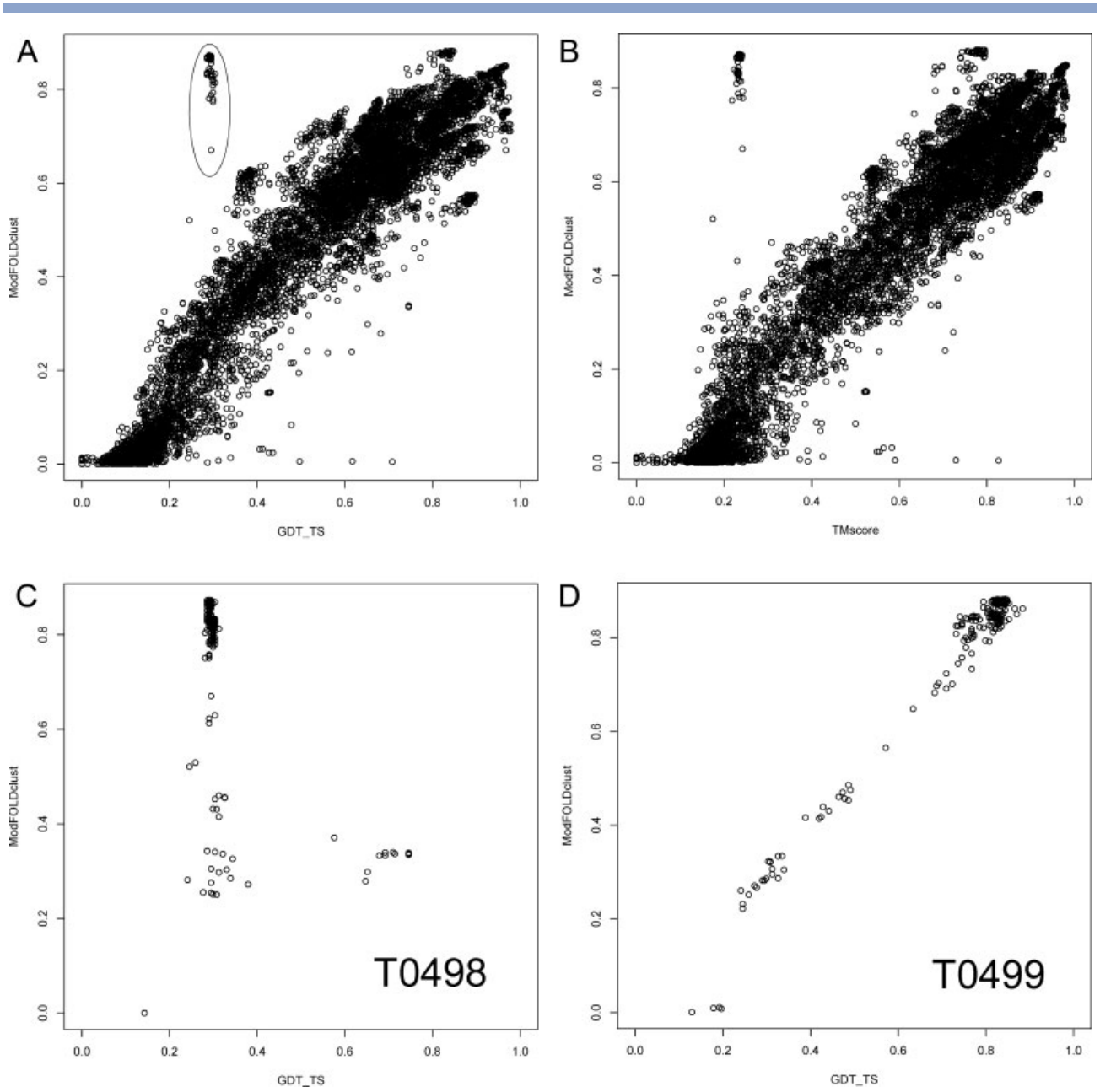

Figure 1

Scatter plots of the predicted global model quality scores versus the observed global model quality scores. (**A**) The ModFOLDclust scores versus the GDT-TS scores for all CASP8 TS1 server models. The ellipse encompasses outliers for target T0498. (**B**) The ModFOLDclust scores versus the TM-scores for all CASP8 TS1 server models. (**C**) The ModFOLDclust scores versus the GDT-TS scores for all models submitted for T0498. (**D**) The ModFOLDclust scores versus the GDT-TS scores for all models submitted for T0499, which differs from T0498 by only three residues.

FOLD v1.1 method with the clustering based ModFOLDclust method. While the ModFOLDclust method is more accurate for ranking multiple models than ModFOLD v1.1, it is not able to produce scores for single models as it relies on comparisons of several alternative models. Conversely, while ModFOLD v1.1 can produce a score for a single model, it is generally less accurate when many models are available and it does not produce per-residue accuracy predictions. Version 2.0 of the ModFOLD method attempts to overcome these issues by using an artificial neural network to combine the six scores from the ModFOLD v1.1 method, the global clustering score from ModFOLDclust and a score based on the original server ranking, to produce a single global score for each model.

The QA predictions from ModFOLD v2.0 were submitted manually in QMODE1 format. However, Mod-

Table I
Global Measures of the Correlations Between Predicted and Observed Model Quality Scores

QA method	Pearson's *r*	Spearman's ρ	Kendall's τ
ModFOLDclust	0.915	0.887	0.735
ModFOLD v2.0	0.898	0.870	0.708
ModFOLD v1.1	0.714	0.697	0.505

Observed model quality was measured using the GDT-TS score. Results based on the TM-score and MaxSub scores are available at http://www.reading.ac.uk/bioinf/CASP8/.

FOLD v2.0 has now been developed as a server that provides both global and per-residue predictions for either single or multiple models. In single-model mode, an individual model is firstly analysed using ModFOLD version 1.1 and then compared with models obtained from the nFOLD3 server using the ModFOLDclust protocol. The combined global quality and local quality scores are then returned to the user in QMODE2 format. However, in multiple-model mode, the ModFOLDclust scores are calculated using all submitted models, as well as those obtained from the nFOLD3 server. For CASP8, ModFOLD v2.0 was run in multiple-model mode.

The manual TS predictions from the McGuffin group were made by purely using ModFOLD v2.0 for model selection. The top five server models, according the ModFOLD 2.0 global score ranking, were selected and submitted. The only modifications made to the models were in cases where the full backbone did not exist, in which case the program BBQ[16] was used to reconstruct the chain. In addition, for each model the predicted per-residue error was calculated using ModFOLDclust and added into the B-factor column for each set of ATOM records.

RESULTS AND DISCUSSION

In Figure 1(A), the ModFOLDclust global predicted model quality scores are plotted against the observed global quality (GDT-TS[17]) scores, for all TS1 or AL1 server models for the 123 CASP8 targets with released structures. A strong positive linear relationship between predicted and observed quality scores is shown. The ModFOLDclust method is also shown to be the best performing of the three methods according to the three different correlation coefficients, both in terms of the global correlations (Table I), where the quality scores for all targets are pooled, and in terms of the per-target correlations (Table II), where correlations are taken for each target and the mean correlation is determined.

A similar linear relationship between the ModFOLDclust global predicted model quality scores and the TM-scores is also observed in Figure 1(B). However, according to the TM-scores the predicted model quality scores appear to be slightly more consistent between targets. The slight inconsistency of the ModFOLDclust scores from target to target according to the GDT-TS score is manifested by the different trajectories of points with varying gradients, as seen in Figure 1(A). This behavior was also previously observed when Pcons scores were compared with GDT-TS scores for the CASP7 models.[15] These inconsistencies are shown to occur to a lesser extent when the ModFOLDclust scores are compared with the TM-scores [Fig. 1(B)]. If ModFOLDclust scores are compared with TM-scores instead of GDT-scores, then the mean per-target correlations are slightly improved ($r = 0.919$, $\rho = 0.827$, $\tau = 0.684$) when compared with the results shown in Table II.

Outliers for Target T0498 are shown in both Figure 1(A,B). Several models for T0498 have high predicted quality scores (>0.8), however, the observed scores are relatively low (<0.3). These data points are shown more clearly in Figure 1(C) where it appears that the ModFOLDclust method has completely failed. However, target T0498 is a special case as it differs by only three residues from T0499, yet it adopts a different fold. Many automated servers produced very similar models for T0489 and T0499 as a consequence of the high sequence homology. As ModFOLDclust works by clustering models based on their similarity, it will fail if most models contain the same errors. The ModFOLD v2.0 neural network output score is also heavily weighted by the ModFOLDclust input score and is therefore also affected by such errors. However, the inaccurate QA results from T0498 are contrasted by the near 1:1 relationship between the ModFOLDclust output scores and the GDT-TS scores shown for the T0499 models [Fig. 1(D)].

An alternative method for benchmarking the performance of MQAP methods is to compare the observed quality scores of the top ranked models for each target. This allows the performance of MQAP methods to be compared with the TS prediction severs. Table III shows the *P*-values for the Wilcoxon signed rank sum tests, which indicate whether any significant differences occur between in the top models selected by ModFOLDclust, ModFOLD v2.0 and two of the top server methods (Zhang-Server[18] and RAPTOR[19]). The results in Table III indicate that there is no significant difference between the top models selected by ModFOLDclust, ModFOLD

Table II
Mean Per-Target Measures of the Correlations Between Predicted and Observed Model Quality Scores

QA method	Pearson's *r*	Spearman's ρ	Kendall's τ
ModFOLDclust	0.917	0.821	0.679
ModFOLD v2.0	0.890	0.772	0.609
ModFOLD v1.1	0.621	0.587	0.440

Observed model quality was measured using the GDT-TS score. Results for individual targets are available at http://www.reading.ac.uk/bioinf/CASP8/.

Table III
Calculated *P*-values for Wilcoxon Signed Rank Tests

	ModFOLDclust	Zhang-Server	ModFOLDv2_0	RAPTOR
ModFOLDclust	1.000	0.129	0.122	0.000
Zhang-Server	0.871	1.000	0.170	0.000
ModFOLDv2_0	0.879	0.831	1.000	0.000
RAPTOR	1.000	1.000	1.000	1.000

QA methods are compared with servers in terms of the observed model quality scores for the top ranked models for each target. The null hypothesis is that the method in the row selects models that are equal or lower in quality than those produced by the method in the column, according to the GDT-TS scores. The alternative hypothesis is that the method in the row produces higher quality models than the method in the column, according to the GDT-TS scores. Gray cells highlight values of $P < 0.01$, indicating that the method in the row significantly outperforms the method in the column. The full table of results is available at http://www.reading.ac.uk/bioinf/CASP8/.

v2.0 and those from the Zhang-Server, yet it is clear that both MQAP methods select models that are significantly better than the next best server—RAPTOR. However, server methods, such as RAPTOR and the Zhang-Sever, selected their best models from their own pools, which were different from the CASP8 pool of models used by ModFOLDclust and MoFOLD 2.0.

Figure 2 shows an example of the predicted and observed per-residue errors occurring within a server model that was built for target T0388. The image of the model on the left is coloured by the predicted per-residue errors according to the ModFOLDclust method. The colouring by predicted error closely matches the colouring by the observed residue errors shown on the right. Overall, the ModFOLDclust method was one of the better performing methods at predicting the per-residue errors in models from all servers. In addition, the predicted per-residue errors featured in B-factor columns of the TS submissions from McGuffin group (Group 379) were the most accurate compared with those from other groups.

CONCLUSIONS

Despite the obvious success of simple clustering methods such as ModFOLDclust, it is clear that there is still room for the improvement of model quality prediction strategies. Given all server models, no MQAP method is able to consistently recognise the best model for each target. Clustering methods will fail in situations where most models contain the same errors; an extreme example of this occurred in the case of T0498. If the ModFOLDclust method is used to select the best model for each target, then the method can achieve a higher cumulative GDT-TS score than the best server (Zhang-Server), however this improvement is not significant. In addition, whilst the ModFOLDclust method is one of the most simple and most effective methods for QA, it is also one of the most CPU intensive.

Because of these shortcomings, developers should be encouraged to explore new methods that will add value to simple clustering, in terms of both accuracy and speed. Attempts were made to add value to ModFOLDclust predictions, through the integration of scores from ModFOLD, but no significant difference in performance was observed.

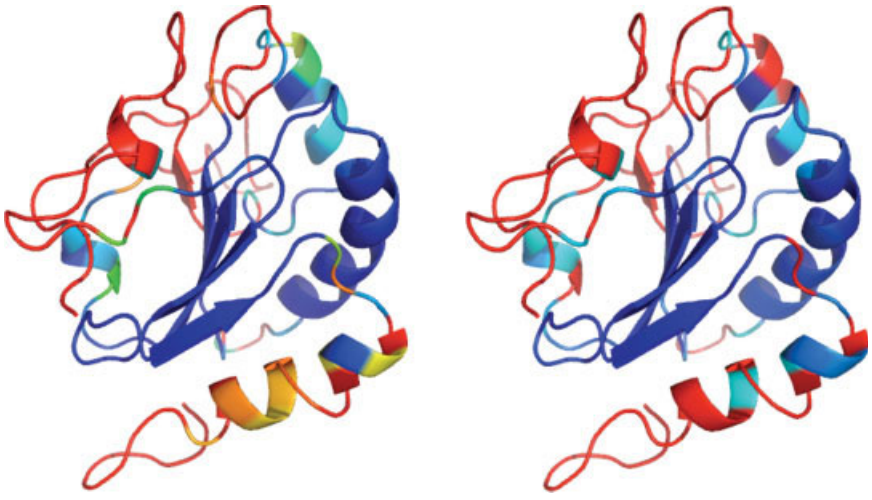

Figure 2
The predicted per-residue accuracy (left) is shown against the observed per-residue accuracy (right) for a model of target T0388. In each case the model is coloured from red to blue according to the predicted or observed distance of each residue in the model to its equivalent residue in the native structure. Blue indicates residues closest to those in the native structure. Residues furthest from those in the native structure (>3.9 Å) are coloured red.

REFERENCES

1. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins 2007;69(Suppl 8):175–183.
2. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci 2006;15:1653–1666.
3. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12:1073–1086.
4. McGuffin LJ. Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics 2007;8:345.
5. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 2006;15:900–913.
6. Benkert P, Tosatto SC, Schomburg D. QMEAN: a comprehensive scoring function for model quality assessment. Proteins 2008; 71:261–277.
7. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics 2008;9:403.
8. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
9. McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 2008;24: 586–587.
10. Pettitt CS, McGuffin LJ, Jones DT. Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 2005;21:3509–3515.
11. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
12. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
13. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
14. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
15. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 2007;69(Suppl 8):184–193.
16. Gront D, Kmiecik S, Kolinski A. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J Comput Chem 2007; 28:1593–1597.
17. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;(Suppl 3):22–29.
18. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007;69(Suppl 8):108–117.
19. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol 2003;1:95–117.