Quality Assessment Methods for 3D Protein Structure Models Based on a Residue–Residue Distance Matrix Prediction

Genki Terashi, Yuuki Nakamura, Hiromitsu Shimoyama, and Mayuko Takeda-Shitaka*

School of Pharmacy, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan.
Received December 16, 2013; accepted April 10, 2014

In the absence of experimentally determined three dimensional (3D) structures of proteins, the prediction of protein structures using computational methods is a standard alternative approach in bioinformatics. When using the predicted protein models to compute the native structure of an unknown target protein, estimating the actual quality of the protein models is important for selecting the best or near-best model. Moreover, estimates of the differences between the protein models and the native protein structure are obviously useful to end users who can then decide on the utility of the models for their specific problems. This article describes two new single-model quality assessment (QA) programs, pure single-model QA method (psQA) and a template based QA method (tbQA), that we developed. psQA is a pure single-model QA program that uses a neural network method to predict residue–residue distance matrices of the native protein structures. tbQA is a quasi-single-model QA program that mainly uses target-template sequence alignments and template structures. The performance of these two model QA programs was analyzed in a data set of 24022 models for 94 targets from the 10th critical assessment of protein structure prediction (CASP10) experiment.

Key words model quality assessment; protein structure prediction; model accuracy; Critical Assessment of Protein Structure Prediction

When experimental three dimensional (3D) structures of proteins are not available, protein structure prediction using a computational method is widely used in biology and medical research, and has proven to be a useful tool. In the tenth community-wide Critical Assessment of Structure Prediction (CASP10), various protein structure prediction methods, including the methods of 69 individual servers, were tested and showed significant progress over the past CASP experiments. Currently, protein structure prediction methods are still being developed and a significant progress has been made to improve their accuracy.

In general, protein structure prediction methods consist of two steps: (1) generation of many candidate models from different alignments and templates, and (2) estimation of the quality of the candidate models to select the best or near best model. For protein structure prediction, the usefulness of a predicted protein model depends on its quality (i.e., similarity with the native protein structure); however, the quality of the model cannot be ascertained when the experimental protein structure is unknown. At present, many protein structure prediction methods can generate high quality models as candidates, but it remains difficult to identify the best or near best model. Therefore, computational biologists have developed various quality assessment (QA) methods to estimate the quality of predicted models when the experimental structure is not available. Accurate QA methods that are based on the coordinates of predicted models will contribute to the improvement of protein structure prediction strategies. Moreover, estimates of differences of the models from the native structure provide useful information to end users who can then decide on whether the models are useful for their specific problems. Therefore, an accurate QA method is an essential component of accurate protein structure prediction and a critical approach.

In principle, existing QA methods can be divided into three categories according to their algorithms: (1) the pure single-model QA; (2) the quasi-single model QA by the clustering QA method; and (3) the clustering QA method. The pure-single model QA method can estimate a quality score (Qscore) to a single model without relying on the similarity between other candidate models or homologous protein structures. Quasi-single model QA methods firstly find evolutionally related proteins (templates) or the best models by the pure single-model QA method, and then re-rank the set of models to be evaluated according to the similarity with the best models or templates. Clustering QA methods use a number of models and structurally compare each model to the other models to calculate the consensus value. Existing clustering QA methods use similar strategies based on the assumption that the model that has the greatest structural similarity to all other models (i.e., the model in the centroid of the cluster) will be the model that is structurally similar to the native structure. However, based on the correlation based assessment in earlier tests, the clustering QA method has been shown to outperform other methods, the top performed QA methods used very similar strategies and obtained very similar results. If almost all of the models to be evaluated are very high quality models (i.e., very close to the native structure), the clustering QA method performs very well. However, if the majority of the models are poor models, the clustering method may not perform well. Currently, no clustering QA can outperform other clustering QA methods consistently. However, single-model QA methods were noticeably behind the clustering QA methods and this indicates that there is a need for improving pure and quasi-single model QA methods. From the realistic viewpoint, the situation in which various servers produce large numbers of models at the same time, such as in the CASP experiments, is not the standard situation for most researchers who employ QA methods to obtain accurate models or to ascertain the usability of a model. Researchers often need to estimate the quality of a single model. Moreover, only single-model methods can be used as a guide for refinement.
and generation of models.

Here, we describe a new pure single-model QA method (psQA) and a template based QA method (tbQA) that is categorized as a quasi-single model QA method. The psQA can estimate the quality of a single model based on the contact prediction method, secondary structure prediction method and the neural network training-prediction method without requiring any template information and consensus based score. The tbQA simply combines the template information and the Qscore from psQA. psQA and tbQA have been trained on a set of predicted protein structure models from CASP9, and tested on a CASP10 data set.

Experimental

In psQA, our single model quality assessment method includes five components as shown in Fig. 1. The procedure consists of three phases: (1) neural network (NN) input generation which includes the generation of multiple sequence alignments (MSA) by HHblits, the secondary structure prediction by PSIPRED and the residue–residue contact prediction by PSICOV; (2) distance matrix prediction by the NN method; and (3) model assessment by comparing the predicted residue–residue distance matrix and the actual distance matrix of the model.

In this paper, the residue–residue distance is defined as the distance between the side-chain centers (SCs), as shown in Fig. 3. Here, we define that $D_{ij}$ denotes the residue–residue distance between $i$-th and $j$-th residues of the three-dimensional protein structure. The predicted distance matrix is denoted by $PD_{ij}(d)$, where $d$ is the particular residue–residue distance. Therefore, the $PD_{ij}(d)$ is a three dimensional matrix. For example, the high value of $PD_{ij}(d)$ corresponds to the high confidence for a particular distance $d$ to be the $D_{ij}$ of the native structure.

The main idea of psQA is that the high quality model which is closest to the native structure is in good agreement with the predicted residue–residue distance matrix. This idea is simply described by the following equation:

$$Q_{MODEL} = \sum_{i,j} PD_{ij}(D_{ij})$$

(1)

where $Q_{MODEL}$ is an estimated quality score of the MODEL to be evaluated, $D_{ij}$ is a residue–residue distance between $i$- and $j$-th residues in the MODEL.

In tbQA, the procedure consists of three steps as follows: (1) template selection by HHblits, (2) structural comparison to calculate a structural similarity score ($Q_{template}$) between the model and top three templates, and (3) model assessment by simply combining the $Q$score from psQA and $Q_{template}$ by linear combination. The basic idea of tbQA is that the high quality model, which is closest to the native structure, is also close to the genetically related proteins (homologous templates).

In this section, we describe the three phases of psQA and the procedure of tbQA in detail.

NN Input Generation

As shown in Fig. 1, the NN was used to predict a residue–residue distance matrix. The input features of NN are generated from: (1) the predicted residue–residue contact score by PSICOV, which employs a sparse inverse covariance estimation to predict residue contacts and (2) the predicted secondary structures (SS) by PSIPRED, which uses a two-stage neural network on the position specific scoring matrices. These two types of the input features are calculated from a MSA. Therefore, to initially build a MSA, HHblits is performed for the target sequence against Uniprot, which was generated by clustering the UniProt database into groups of similar sequences which have at least 20% pairwise sequence identities to each other. HHblits is based on the pairwise comparison of profile hidden Markov models and can build high-quality MSAs.

The predicted residue–residue contact score ($PC_{ij}$) is represented by a single element vector for a pair of residues. The predicted SS of the target sequence is represented by a three element vector for random coils, alpha helix and beta strands, respectively.

For prediction of the distance matrix for a pair of residues (such as between $i$-th and $j$-th residues), we considered that the residues which exist around the $i$-th and $j$-th residues also affect the residue–residue contact between the $i$-th and $j$-th residues. Thus, we used a size 13 window which consists of residues from $i-6$th to $i+6$th for a window of the $i$-th residue. The number of input features of the NN for a pair of residues is 247 as follows: (1) the predicted SS is represented by 78 features (a three element vector for a residue, $13 \times 2$ residues for a pair of windows, and therefore, $3 \times 13 \times 2$ features for a pair of windows). (2) The predicted residue–residue contact score from PSICOV for the pair of windows is represented by 169 features (all pair of residues between the windows of $i$-th and $j$-th residues, such as $PC_{i-6,j-6}, PC_{i-6,j-5}, ..., PC_{i,j}, PC_{i+6,j+6}$, therefore $13 \times 13 \times 13$ features).

Distance Matrix Prediction by NN

During the trial of various types of NN outputs, we noticed that the NN is weak in predicting the exact value of $D_{ij}$, such as using a single or multiple element of NN output to represent the distance.
directly. To solve this problem, we introduced the two-stage algorithm: (1) NN predicts two types of 17-element vectors, which represent the distribution of lower and higher values of $D_{i,j}$ (denoted by $L_{i,j}(d)$ and $H_{i,j}(d)$, respectively); (2) to generate the predicted distribution of the residue–residue distance ($PD_{i,j}(d)$), products of the two distributions ($L_{i,j}(d)$ and $H_{i,j}(d)$) are calculated.

Figure 2 shows an example of the two stage algorithm. The NN output is given as two sets of 17-element vectors ($L_{i,j}(d)$ and $H_{i,j}(d)$), which represent the distribution of distances from 3.5 to 20 Å. The first set of the 17-element vector denoted by $L_{i,j}(d)$ represents the lower value of the $D_{i,j}$, where larger values in the particular bin indicate higher confidence for the particular distance to be lower than the actual $D_{i,j}$. For example, in Fig. 2B the value in the bin of “4.5–5.5” is ca. 0.7 and it means that the high confidence for the value (4.5–5.5 Å) to be lower than the actual $D_{i,j}$. Same as the first set of the NN output $L_{i,j}(d)$, the second set of the 17-element vector denoted by $H_{i,j}(d)$ represents the higher value of the $D_{i,j}$. As shown in the Fig. 2C, the bin of “18.5–19.5” shows a very high value (about 1.0), and indicates that the value (18.5–19.5 Å) should be higher than the actual $D_{i,j}$ (i.e., the actual $D_{i,j}$ is strongly predicted to be lower than 18.5 Å). The predicted distance distribution for the pair of $i$ and $j$-th residues ($PD_{i,j}(d)$) are calculated from the products of $L_{i,j}(d)$ and $H_{i,j}(d)$. As seen in Fig. 2D, the distance 4.5–5.5 Å has the highest confidence to be the actual $D_{i,j}$.

**NN Training** The NN was trained on residue–residue distance data from the 3D protein structures in the Protein Data Bank (PDB$^{20}$). The training data set consists of 1298871 residue–residue pairs obtained from 5664 protein chains with less than 30% sequence identity. These protein chains each have at
least 50, but no more than 400 residues, and resolutions better than 2.0 Å. The actual $D_{ij}$ data were converted to $L_{ij}(d)$ and $H_{ij}(d)$. Figure 3 shows an example of the distribution of $L_{ij}(d)$ and $H_{ij}(d)$ for $D_{ij}=5.1$ Å.

The NN software which was used for training and prediction is from the Fast Artificial Neural Network Library (FANN\textsuperscript{20}). The training algorithm is a back propagation algorithm. According to our benchmark test, we choose a single hidden layer with 68 hidden neurons for the NN. To determine how much training is required to perform well without over-fitting, 10% of the training data was withheld to test the performance of the network during the first stage of training. The performance is defined by the mean square error (MSE), which represents the average of the difference between the actual distributions ($L_{ij}(d)$ and $H_{ij}(d)$), and the predicted distributions. The first stage of the NN training was terminated when the MSE of the excluded 10% of the training data was not improving and then the MSE threshold was set to the MSE of the remaining 90% of the training data. After the first stage of the training, the NN training was performed again for all of the training data until the MSE of the training data reached the MSE threshold. According to the amino acid types of the pairs of residues, a total of 400 NN profiles were generated by this NN training process.

**Model Assessment** In psQA, the estimated quality score ($Q_{psQA}$) of the model is described as follows:

$$Q_{psQA} = \sum_{i,j} w_1(SS_i, SS_j, PSS_i, PSS_j) \times w_2(i, j) \times PD_{ij}(D_{ij})$$  \hspace{1cm} (2)

where function $w_1$ is a weight function determined by the secondary structures of the model (SS and SS) and the predicted secondary structures ($PSS$ and $PSS$) of the target sequence for $i$- and $j$-th residues. The function $w_1$ was optimized by the 27926 models, which were predicted by the servers that participated in CASP9.\textsuperscript{21} $PD_{ij}(D_{ij})$ represents the confidence for the $D_{ij}$ of the model to be the actual $D_{ij}$ of the target structure. As mentioned above, the distribution of $PD_{ij}(D_{ij})$ was generated by the NN. The $w_2(i, j)$ is a step function defined as:

$$w_2(i, j) = \begin{cases} 1 & \text{if } |i-j| > 12 \text{ and } PC_{i,j} \geq 0.1 \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (3)

where $|i-j|$ is a sequence distance and $PC_{i,j}$ is the predicted residue–residue contact score from PSICOV. Moreover, we also considered whether $i$- and $j$-th residues are the members of the same secondary structure unit, and if so, the $w_2(i, j)$ is set to zero.

**Template Based Single Model Quality Assessment (tbQA)** For template based targets, it is well known that the target protein and known protein structures that are evolutionarily related have similar conformations. In other words, the high-quality models share similar conformations with the template protein structures. Therefore, the quasi-single model QA method, tbQA, has the advantage of the availability of known protein structures.

In tbQA, HHBlits was performed against PDB70, which consists of 27132 protein chains (release 28, April 2012), to find homologous templates. The top three alignments were selected to build coarse C-alpha only models and these models were used as references to compare with all models by sequence dependent structural superimposing. For each alignment, we generated ten C-alpha only models, according to the confidence value of PSIPRED. Here, $REF_{a,b}$ was defined as representing the C-alpha only model that was generated from $a$-th alignment of the HHBlits output, and coordinates of the position where the confidence value of PSIPRED was less than $b$ were removed. Thus, the confidence value was used for weighting the position in the template structure.

The template similarity score ($Q_{template}$) is calculated using the following equation:

$$Q_{template} = \frac{1}{30 Len_{target}} \times \sum_{Y=0}^{26} \sum_{X=50}^{100} GD(T(MODEL, REF_{a,b})) \times Len_{REF_{a,b}}$$  \hspace{1cm} (4)

where $Len_{target}$ and $Len_{REF_{a,b}}$ are the length of the target sequence and $REF_{a,b}$, respectively. $GD(T(MODEL, REF_{a,b}))$ is a value of GD-TS between the MODEL given to be evaluated and $REF_{a,b}$. The details of the GD-TS are described in the “Results and Discussion.”

The estimated quality score from tbQA ($Q_{tbQA}$) is a combined value of normalized $Q_{template}$ and $Q_{psQA}$:

$$Q_{tbQA} = nls(Q_{template}) + w_3 \times Q_{psQA}$$  \hspace{1cm} (5)

where the function $nls$ is a nonlinear weight function to normalize $Q_{template}$, $w_3$ is a weight and set to 0.5. Parameters of these function and weight were determined according to the optimization on CASP9 dataset which consists of 34101 models for 116 targets.

**Results and Discussion**

**Actual Quality of Model (Global Distance Test Total Score (GD-TS) and Contact Area Difference Based Score (CAD-Score))** In this article, we defined the actual quality by the GD-TS\textsuperscript{22} and CAD-score.\textsuperscript{23} For the official CASP assessment, the quality of a model was evaluated by assigning the model a GD-TS value. GD-TS is an averaged fraction percentage of residues in the predicted models that deviate from the native structure after four different sequence-dependent superpositions using C-alpha distance thresholds of 1, 2, 4 and 8 Å. In the CASP experiment, GD-TS is able to evaluate the actual quality of models effectively in not only easy targets but also hard targets when there are no high quality models, such as New Fold targets. Although GD-TS is a standard measure of model quality, it is known that there are two primary weaknesses. Firstly, GD-TS is not a robust evaluation on multi-domain protein models. GD-TS is based on superposition of Cartesian coordinates, therefore a small change of the domain orientation, which is not rare in nature, gives rise to large differences in the GD-TS. Secondly, GD-TS lacks side chain information, because it only considers the C-alpha atoms as the positions of model residues. Thus, GD-TS cannot evaluate the correctness of side chain conformations and unrealistic atom contacts (i.e., physical interatomic clashes).

Therefore, we additionally introduced the CAD-score based evaluation. The CAD-score, which uses residue–residue contact area at the atomic level for quantifying differences between physical contacts in a model and the native structure, has a high correlation with GD-TS, but displays a stronger
preference for physically more realistic models. Moreover, it provided a balanced assessment of multi-domain protein models.

**Evaluation of Estimated Quality Score** We evaluated the performance of psQA and tbQA from two points of view: (1) the correlations between the estimated quality and real quality (GDT-TS or CAD-score) of the models; and (2) the ability to pick out the best models from the sets of models that were being assessed. The former corresponds to the average and overall correlations of estimated quality score ($Q$, such as $Q_{\text{psQA}}$ or $Q_{\text{tbQA}}$) and the real quality ($RQ$, such as GDT-TS or CAD-score). The latter is the average loss of $RQ$ on the top ranked model. The average loss of $RQ$ is a measure of the difference between the $RQ$ of the model that was ranked top by the QA method and the $RQ$ of the actual best model.

The average correlations are the average per target correlations between $Q$ and $RQ$. The overall correlations are the correlations between $Q$ and $RQ$ of all models of the dataset. To evaluate performance by the linear correlation and the rank correlation, we calculated the Pearson’s $r$ linear correlation coefficient and the Kendall’s $\tau$ rank correlation coefficient. Pearson’s $r$ is a measure of the linear relationship between $Q$ of the model and its $RQ$ calculated as follows:

$$r = \frac{\sum_{i=1}^{N} (Q_i - \bar{Q})(RQ_i - \bar{RQ})}{\sqrt{\sum_{i=1}^{N} (Q_i - \bar{Q})^2 \cdot \sum_{i=1}^{N} (RQ_i - \bar{RQ})^2}}$$

where $N$ is the total number of models that were evaluated by QA methods, $Q_i$ is the estimated quality value of the $i$-th model, $\bar{Q}$ is the average value of the estimated quality values, $RQ_i$ represents the real quality of the $i$-th model, and $\bar{RQ}$ is the average value of $RQ$ for the evaluated models. Kendall’s $\tau$ is a measure of the probability that the model with the better estimated $Q$ is actually the better model in the given random pair of models to be assessed. Kendall’s $\tau$ is calculated as follows,

$$\tau = \frac{4 \sum_{i < j} P_{ij}}{N(N-1) - 1}$$

where $P_{ij}$ is 1 if the $i$-th or $j$-th model with the better estimated $Q$ is actually the better model, otherwise it is set to zero; and $N$ represents the number of models that were assessed.

**Performance on the 94 CASP10 Targets** We evaluated the performance of the developed two single model QA methods (psQA and tbQA) in the CASP10 dataset. In CASP10, three types of predicted protein structure datasets were released; STAGE1: 20 models of different quality per target, STAGE2: the best 150 server models of the target according to the clustering method, which were performed on all server models, and STAGE3: all models (over 250 models per target) that were generated from the protein structure prediction server, which participated in CASP10. Tables 1 and 2 show the summarized results of GDT-TS and CAD-score based evaluations, respectively. These two tables report the average correlations and overall correlations of Pearson’s $r$ and Kendall’s $\tau$, and the average difference between the models predicted to be the best and the actual best model in GDT-TS or CAD-score.

As shown in Tables 1 and 2, we additionally evaluated two extensive QA methods, “templateQA” and “Naïve Consensus.” We are interested in how much the performance of tbQA depends on the Qscore from psQA. Therefore we also analyzed the performance of the templateQA which uses the normalized $Q_{\text{template}}$, described in Eq. 5 as $\text{nls}(Q_{\text{template}})$. The templateQA uses only the structural similarity between model and homologous templates. To compare with the standard clustering QA method, we also evaluated the naïve consensus method, which

| Stage | Method       | Pearson $r$ | Kendall $\tau$ | dGDT | $\#$ |
|-------|--------------|-------------|----------------|------|-----|
| Stage 1 | **CAD-score**<sup>a</sup> | 0.80<sup>b</sup> | 0.61<sup>c</sup> | 2.0<sup>d</sup> | 94<sup>e</sup> |
|        | Naïve Consensus<sup>a</sup> | 0.71<sup>b</sup> | 0.54<sup>c</sup> | 5.5<sup>d</sup> | 94<sup>e</sup> |
|        | TemplateQA<sup>a</sup> | 0.69<sup>b</sup> | 0.54<sup>c</sup> | 5.0<sup>d</sup> | 94<sup>e</sup> |
|        | psQA         | 0.60<sup>b</sup> | 0.40<sup>c</sup> | 6.6<sup>d</sup> | 83<sup>e</sup> |
|        | tbQA         | 0.70<sup>b</sup> | 0.53<sup>c</sup> | 4.4<sup>d</sup> | 94<sup>e</sup> |
| Stage 2 | **CAD-score**<sup>a</sup> | 0.60<sup>b</sup> | 0.50<sup>c</sup> | 3.1<sup>d</sup> | 94<sup>e</sup> |
|        | Naïve Consensus<sup>a</sup> | 0.49<sup>b</sup> | 0.38<sup>c</sup> | 5.1<sup>d</sup> | 94<sup>e</sup> |
|        | TemplateQA<sup>a</sup> | 0.38<sup>b</sup> | 0.31<sup>c</sup> | 7.4<sup>d</sup> | 94<sup>e</sup> |
|        | psQA         | 0.34<sup>b</sup> | 0.26<sup>c</sup> | 5.5<sup>d</sup> | 83<sup>e</sup> |
|        | tbQA         | 0.41<sup>b</sup> | 0.32<sup>c</sup> | 6.5<sup>d</sup> | 94<sup>e</sup> |
| Stage 3 | **CAD-score**<sup>a</sup> | 0.78<sup>b</sup> | 0.65<sup>c</sup> | 3.1<sup>d</sup> | 94<sup>e</sup> |
|        | Naïve Consensus<sup>a</sup> | 0.89<sup>b</sup> | 0.65<sup>c</sup> | 5.1<sup>d</sup> | 94<sup>e</sup> |
|        | TemplateQA<sup>a</sup> | 0.80<sup>b</sup> | 0.54<sup>c</sup> | 9.5<sup>d</sup> | 94<sup>e</sup> |
|        | psQA         | 0.80<sup>b</sup> | 0.45<sup>c</sup> | 6.8<sup>d</sup> | 83<sup>e</sup> |
|        | tbQA         | 0.82<sup>b</sup> | 0.54<sup>c</sup> | 8.7<sup>d</sup> | 94<sup>e</sup> |

<sup>a</sup> Pearson linear correlation coefficient.  <sup>b</sup> Average of the correlation coefficient per target.  <sup>c</sup> Overall correlation coefficient for all quality estimations.  <sup>d</sup> Kendall rank correlation coefficient.  <sup>e</sup> Delta GDT-TS, average difference in GDT-TS between the models predicted to be the best and the actual best models based on the GDT-TS based evaluation.  <sup>f</sup> Number of targets assessed by the QA methods.  <sup>g</sup> Virtual QA method which uses the CAD-score of the model as the estimated quality score.  <sup>h</sup> Naïve Consensus method described by Eq. 5.  <sup>i</sup> Normalized $Q_{\text{template}}$, which is described by Eq. 5 as $\text{nls}(Q_{\text{template}})$.  <sup>j</sup> Quality estimation for 20 models of different quality per target released by the CASP organizers.  <sup>k</sup> Quality estimation for best 150 server models according to the clustering method, which was performed by the CASP organizers.  <sup>l</sup> Quality estimation for all server models submitted on the target.
is described as follows:

\[ Q_{\text{naive}} = \frac{\sum_{i=1}^{N} GDT(MODEL, M_i)}{N} \]  

(8)

where \( N \) is the total number of models that were evaluated by the Naïve Consensus method, and \( GDT(MODEL, M_i) \) is a value of GDT-TS between the MODEL and \( i \)-th model of the dataset \( (M_i) \).

Moreover, we compare the two quality definitions, GDT-TS and CAD-score by using these actual quality values as the estimated quality score (thus, these two methods are so-called “virtual QA methods”). For example, the “CAD-score” in the GDT-TS based evaluation (Table 1) represents the performance of the virtual QA method, which assigned the CAD-score to the estimated quality score. Of course, the CAD-score could be obtained by comparing the native protein structure and the model to be assessed. Therefore, the virtual QA methods, “CAD-score” and “GDT-TS” in Tables 1 and 2 are not available in reality and not the actual QA methods.

In the GDT-TS based evaluation (Table 1), the “CAD-score” outperforms the other QA methods due to the use of the native structure. The performance of “Naïve Consensus” is clearly better than other single-model QA methods, and it is a well-known result. Surprisingly, “Naïve Consensus” is slightly better than “CAD-score” in the correlations on Stage3. These results indicated that when a large number of diverse models are available, the consensus method correlates with GDT-TS beyond the performance of the CAD-score, even though the CAD-scores are obtained from the comparison between the model and native structure. When comparing three single-model QA methods to each other, tbQA showed the best or near best average and overall correlations among all stages. The reason for this result is due to the methodology of tbQA, which uses the combination of psQA and templateQA, and the optimization for correlations to GDT-TS. Interestingly in Stage3, although the average of the Pearson’s \( r \) of the three single model QA methods are comparable with the virtual QA method “CAD-score,” the overall linear correlation of psQA is worse than the other two single QA methods. This means the psQA is good for the relative quality of the models in each target, but not good for the absolute quality of the models that can be used among different targets. This tendency is also seen in the GDT-score based evaluation (Table 2). While the correlations of psQA are worse than tbQA and templateQA, the ability of psQA to select the high quality models from a set of predicted protein structures (dGDT-TS) is clearly better than the other single-model QA methods for Stage2 and Stage3.

The results of the CAD-score based evaluation is listed in Table 2. Similar to the results presented in Table 1, psQA showed better performance than the other single-model QA methods in the ability to select high quality models (dCAD-score) for Stage2 and Stage3. Despite the low performance of psQA in Table 1, the psQA average correlations are best or near best for Stage2 and Stage3. We attribute these performances of psQA to its scoring function (Eq. 2), which mainly uses the agreement between the predicted distance matrix and the distance matrix of the model. Instead of using the all positions based similarity of all C-alpha atoms (such as templateQA, tbQA), psQA considers the local residue–residue contacts denoted by the distances between the SCs. This point of view is more similar to the CAD-score than to the GDT-TS.

**Performance versus Difficulty of the Target** It is well known that QA methods show different performances, especially in the correlations, according to the distribution of the quality of the models to be assessed. Therefore, we analyzed the performance of our two single-model QA methods (psQA and tbQA) against the difficulties for each CASP10 target among all server models (Stage3 datasets). The difficulty of the target corresponds to the average of the actual quality value (GDT-TS or CAD-score) of the models, which were predicted for the target. Comparisons of the average of GDT-TS and correlation coefficients of Pearson’s \( r \) and Kendall’s \( \tau \) for
each CASP10 target are shown in Figs. 4A and C. The solid and dashed lines correspond to the smoothed data of psQA and tbQA by spline curves, respectively. For all targets, except for very hard targets where the average of GDT-TS is below 20, the Qscore estimated by tbQA had relatively higher correlations with GDT-TS of the model than that of psQA. As described above, in the scoring function, tbQA employs $Q_{\text{template}}$ (Eq. 4), which uses the same structural comparison method as the GDT-TS evaluation. Both methods use the coordinates of the C-alpha atoms and the superimposing method based on Cartesian coordinates. Therefore, tbQA can order the models according to the quality based on the position of the C-alpha atoms.

Conversely, the performance (correlation coefficients of Pearson’s $r$ and Kendall’s $\tau$) of psQA and tbQA based on the CAD-score are shown in Figs. 4B and D. The contact based evaluation (CAD-score) is more similar with the scoring function which uses the predicted residue–residue distance (Eq. 2) rather than the $Q_{\text{template}}$ (Eq. 4), which is used in tbQA. For this reason, we expected that the Qscore from psQA will have high correlations with the CAD-score among all targets. However, as shown in Figs. 4B and D, the Qscore estimated by psQA had higher correlations with the CAD-score for the medium and hard targets (the average CAD-score $<0.50$−$0.55$) than tbQA, while the tbQA had higher correlations in easy targets (the average CAD-score $>0.55$) when compared with the results of psQA.

From the point of view of examining the ability to select the best models, we compared the quality of the top ranked model by the QA methods and the difficulties. As shown in Figs. 5A,B, psQA and tbQA achieved clearly better performances than the average quality of each dataset for almost all targets. Figures 5C and D show that the comparison between the difficulties of each target and the ability to select the high quality models, denoted by $d_{\text{GDT-TS}}$ and $d_{\text{CAD-score}}$. For the correlation based evaluation (Figs. 4A–D), psQA performed better...
Fig. 5. Comparison between the Average of the Actual Quality (GDT-TS or CAD-Score) and That of the Selected Models by the QA Methods

The results of psQA are represented by the filled circles, whereas the results of tbQA are represented by the opened triangles. (A) Average GDT-TS versus GDT-TS of top ranked model. (B) Average CAD-score versus CAD-score of top ranked model. (C) Average GDT-TS versus dGDT-TS. (D) Average CAD-score versus dCAD-score.

Fig. 6. The Coverage of the Targets versus the Threshold of the (A) dGDT-TS and (B) dCAD-Score
than tbQA for medium and hard targets.

The performances of the three methods (psQA, tbQA and templateQA) were also compared using the coverage of targets against the threshold of dGDT-TS and the dCAD-score. Figure 6 plots the rate of targets versus the deviations of the quality (dGDT-TS and dCAD-score) of the top ranked models by the QA methods from those of the best model. For example, in Fig. 6A, where the threshold of dGDT-TS was set to 5.0, the target coverages of psQA and tbQA were 65.1 and 55.4%. This indicates that psQA can select the models whose GDT-TS are not 5.0 lower than that of the best model for 65.1% of CASP10 targets. Similarly in Fig. 6B, where the threshold of the dCAD-score was set to 0.05, the target coverages of psQA and tbQA were 57.8 and 47.0%. Figures 5C, D, 6A and B indicate that psQA frequently selected better models than tbQA and templateQA. These coverage based evaluations show that the performances of the three single model QA method can be ranked as psQA > tbQA > templateQA.

Performance of Estimating the Absolute Quality of the Single-Models

Figures 7A and B plot CAD-score (actual quality of model) against the Qscore estimated by psQA and tbQA on all models of CASP10 data set, respectively. As mentioned above, psQA is good for the relative quality of the models in each target, but not good for the absolute quality of the models that can be used among different targets. Additionally, Fig. 7C shows that, when the native structures were evaluated by psQA and tbQA, psQA can rank the native structures as good structures among over 250 models per target. The average rank of the native structures assigned by psQA and tbQA are 62nd and 140th, respectively. Moreover, we calculated the Z-score of the Qscore on native structures, by subtracting the mean of Qscores of data set from the Qscore of native structure and dividing by the standard deviation for each target. The average Z-score of Qscore from psQA and tbQA on the native structures are 0.79 and 0.29, respectively.

These data suggest that the Qscore estimated by psQA is a protein specific value which depends on the length and shape of target protein. Therefore, psQA can estimate a good relative score for model selection instead of an absolute model quality score. From Figs. 7A–C, we attributed the worse performance of estimating the absolute quality score to the lack of the normalization of the Qscore. The absolute quality score is critical for using predicted protein structure models and thus, it is an important future work to improve the overall correlations corresponding to the absolute quality score by developing the normalization method of the Qscore.

Conclusion

In this study, we have developed a single model quality assessment method (psQA) and quasi-single model quality assessment method (tbQA). The psQA uses the predicted residue–residue distance matrix that was generated from only the MSA of the target sequence. The tbQA combines the estimated quality score from psQA and the similarity between the model and homologous template structures which are delivered from the sequence search method.

In the CASP10 Stage3 dataset, the estimated quality score from tbQA has higher correlation coefficients with GDT-TS than that of psQA. However, psQA has better correlations with the CAD-score than tbQA. We have noted that these results are attributed to similarities between the evaluation methods and scoring functions of the QA methods. The results reveal that psQA is not good for the overall correlations that correspond to the ability for assigning absolute quality. This means that there is significant room to improve the psQA method. The developing of normalization method to estimate the absolute quality score which is not a target protein specific value should be an important task of future research.

Although the correlations are not good, psQA is better than tbQA for near best model selection. According to our analysis which considers the difficulties of each target, psQA performs better than tbQA for medium and hard targets in both correlation and ability to select the best models. These results indicate that psQA performed well whenever the homologous protein structure that can be easily identified by a sequence search method is not available.

One goal of the QA method is to develop a scoring function that improves the performance of the protein structure prediction method by selecting better models from the set of models generated as the candidate models. Our results revealed that the psQA, which uses only one metric, is useful for selecting the high quality models. Moreover, psQA can also be extended to improve protein structure modeling methods by assigning a local quality score for each residue from Eq. 2.

References

1) Kryshtafovych A., Monastyrskyy B., Fidelis K., Proteins, 82 (Suppl. 10), 1–5 (2011).
2) Moult J., Fidelis K., Kryshtafovych A., Tramontano A., Proteins, 79 (Suppl. 2), 7–13 (2014).
3) Moult J., Fidelis K., Kryshtafovych A., Rost B., Tramontano A., *Proteins*, 77 (Suppl. 9), 1–4 (2009).

4) Moult J., Fidelis K., Kryshtafovych A., Rost B., Hubbard T., Tramontano A., *Proteins*, 69 (Suppl. 8), 3–9 (2007).

5) Kryshtafovych A., Fidelis K., Tramontano A., *Proteins*, 79 (Suppl. 10), 91–106 (2011).

6) Kryshtafovych A., Barbato A., Fidelis K., Monastyrskyy B., Schwede T., Tramontano A., *Proteins*, 82 (Suppl. 2), 112–126 (2014).

7) Ray A., Lindahl E., Wallner B., *BMC Bioinformatics*, 13, 224 (2012).

8) Wang Z., Tegge A. N., Cheng J., *Proteins*, 75, 638–647 (2009).

9) Terashi G., Takeda-Shitaka M., Kanou K., Iwadate M., Takaya D., Hosoi A., Ohta K., Umezama H., *Proteins*, 69 (Suppl. 8), 98–107 (2007).

10) Larsson P., Skwark M. J., Wallner B., Elofsson A., *Proteins*, 77 (Suppl. 9), 167–172 (2009).

11) Wang Q., Vantasin K., Xu D., Shang Y., *Proteins*, 79 (Suppl. 10), 185–195 (2011).

12) Terashi G., Oosawa M., Nakamura Y., Kanou K., Takeda-Shitaka M., *Chem. Pharm. Bull.*, 60, 1359–1365 (2012).

13) Kryshtafovych A., Fidelis K., Moult J., *Proteins*, 79 (Suppl. 10), 196–207 (2011).

14) “10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. CASP10 Experiment.”: http://predictioncenter.org/casp10/, cited 1 August, 2012.

15) “Fast Artificial Neural Network Library.”: http://leenissen.dk/fann/wp/, cited 1 April, 2012.

16) Remmert M., Biegert A., Hauser A., Soding J., *Nat. Methods*, 9, 173–175 (2012).

17) Jones D. T., *J. Mol. Biol.*, 292, 195–202 (1999).

18) Jones D. T., Buchan D. W., Cozzetto D., Pontil M., *Bioinformatics*, 28, 184–190 (2012).

19) UniProt Consortium, *Nucleic Acids Res.*, 40 (D1), D71–D75 (2012).

20) Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E., *Nucleic Acids Res.*, 28, 235–242 (2000).

21) “9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. CASP9 Experiment.”: http://predictioncenter.org/casp9/, cited 14 April, 2010.

22) Zemla A., Venclovas C., Moult J., Fidelis K., *Proteins*, 37 (Suppl. 3), 22–29 (1999).

23) Olechnovič K., Kulberkyte E., Venclovas C., *Proteins*, 81, 149–162 (2013).