Published in final edited form as:

Int J Artif Intell Tools. 2013 October 25; 22(5): 1360006-. doi:10.1142/S0218213013600063.

NEW MDS AND CLUSTERING BASED ALGORITHMS FOR PROTEIN MODEL QUALITY ASSESSMENT AND SELECTION

QINGGUO WANG,

Bioinformatics and Systems Medicine Laboratory, Vanderbilt University Nashville, TN 37203, USA

CHARLES SHANG,

Computer Science Department, University of Illinois at Urbana-Champaign Urbana, IL 61801, USA

DONG XU, and

Computer Science Department, University of Missouri Columbia, MO 65211, USA

YI SHANG

Computer Science Department, University of Missouri Columbia, MO 65211, USA

QINGGUO WANG: qingguo.wang@vanderbilt.edu; CHARLES SHANG: cshang4@illinois.edu; DONG XU: xudong@missouri.edu; YI SHANG: shangy@missouri.edu

Abstract

In protein tertiary structure prediction, assessing the quality of predicted models is an essential task. Over the past years, many methods have been proposed for the protein model quality assessment (QA) and selection problem. Despite significant advances, the discerning power of current methods is still unsatisfactory. In this paper, we propose two new algorithms, CC-Select and MDS-QA, based on multidimensional scaling and k -means clustering. For the model selection problem, CC-Select combines consensus with clustering techniques to select the best models from a given pool. Given a set of predicted models, CC-Select first calculates a consensus score for each structure based on its average pairwise structural similarity to other models. Then, similar structures are grouped into clusters using multidimensional scaling and clustering algorithms. In each cluster, the one with the highest consensus score is selected as a candidate model. For the QA problem, MDS-QA combines single-model scoring functions with consensus to determine more accurate assessment score for every model in a given pool. Using extensive benchmark sets of a large collection of predicted models, we compare the two algorithms with existing state-of-the-art quality assessment methods and show significant improvement.

Keywords

Protein tertiary structure prediction; model quality assessment; consensus method; clustering; multidimensional scaling

1. Introduction

Protein structure prediction is one of the most challenging problems in bioinformatics. Knowing protein structures quickly and at low-cost will benefit many life science fields, such as medicine and biotechnology. In recent years, large-scale genome projects have produced massive amounts of protein sequences. To fully understand the functions of these proteins, the knowledge of their tertiary structures is indispensable. Although protein structure databases have been growing, they have been limited by the time-consuming and expensive experimental processes of X-ray crystallography and NMR spectroscopy. To complement experimental efforts in determining protein tertiary structures, computational methods to predict tertiary structures directly from one-dimensional amino acid sequences have been actively developed and achieved significant progress in the recent years. The availability of high-quality protein structures would not only enable us to understand the biological roles of proteins, but also enhance our ability of structure-based drug design.¹

Existing computational methods for protein structure prediction generally have two phases: (a) model generation—generating a large number candidate structures and (b) model assessment—assessing the quality of models and selecting the best ones. Models are constructed either by assembling different protein tertiary structure pieces aligned to the query sequence or by sampling and searching in different regions of the conformational space based on heuristic energy or scoring functions.² Oftentimes, a protein structure prediction software tool generates a large number of models and, as the result, the ability of picking out the good ones directly affects the final prediction result.

In model assessment, a few major approaches have been developed to distinguish high-quality models from poor ones. One approach is using physics or knowledge based scoring functions,^{3–6} which calculate a score as a quality estimate for a model. Although many scoring functions have been developed over the years, their abilities to identify near-native structures are still limited. Another approach is based on clustering.^{7–9} The clustering approach groups similar models together and picks a representative one, or centroid, from each cluster. Clustering-based methods are more robust than scoring functions, but require a sufficient number of good models to work well. The third approach uses consensus information among candidate structures.^{2,10–12} The naive consensus method computes the average similarity of each model to the other models as its quality score. Unlike scoring functions, the consensus and clustering approaches require a pool of models and are not applicable to the evaluation of a single protein model on its own. However, when a large set of candidate structures are provided, it has been shown that the consensus approach can perform significantly better than the other approaches.^{13–15}

Specifically, in this paper, the *model selection* problem refers to finding the best one or several models from a pool. *Model quality assessment (QA)* refers to a slightly different problem, defined according to the CASP (Critical Assessment of Techniques for Protein Structure Prediction). The CASP, organized by the Protein Structure Prediction Center and sponsored by NIH, has been spearheading the development of protein structure prediction methods for 20 years. CASP is a biennial competition for community-wide assessment of state-of-the-art prediction methods using unknown protein targets. To facilitate the

development of protein structure selection and quality assessment methods, a new category called model quality assessment (QA) was introduced into CASP in 2006. For each target, a QA method gives each predicted model a score. The Pearson correlation coefficient of the estimated scores against the true quality scores (i.e., the GDT_TS score of a model against the native structure) of the models is the performance metric of a QA method and is used to compare the performances of different QA methods. In 2010, more than 250 groups from the world participated in CASP9 and 46 groups were in the model QA category (<http://predictioncenter.org/>).

The accuracy of existing protein model quality assessment methods is not high enough for many real applications. The difficulty is illustrated by the example in Figure 1, which shows the distributions of estimate quality scores of predicted models vs. the ground truth for protein 3dcx, which is target T0446 in CASP8. For this target, 282 predicted models were submitted by CASP8 participants and were used for the evaluation of quality assessment methods. The true quality of each model is its GDTTS score against the native structure (see the next section for details). The GDTTS value has a range [0, 1] and measures the similarity between two 3-D structures. The larger the value, the more similar they are, with identical structures having value 1. In Fig. 1, the quality estimates of two state-of-the-art QA methods, the score function DFIRE⁵ and the naive consensus method, are plotted against the true quality values, respectively. Figure 1(a) shows that the model ranked top according to the DFIRE score (the smaller the value the better) is far away from the true best. Similarly, Figure 1(b) shows that the naive consensus method, which uses Eq. (5) to calculate a consensus score (the larger the value the better), fails to identify the best predicted models either.

Figure 1(b) suggests that the predicted models form clusters, which is true for most CASP targets, partly due to using some common techniques in the participating prediction software programs. Although the best models may not have highest overall consensus scores, they may have the highest values in their clusters. This hypothesis motivates this work to blend consensus approach with clustering.

In this paper, two new algorithms, CC-Select and MDS-QA, based on multi-dimensional scaling and *k*-means clustering are proposed. For the model selection problem, CC-Select combines consensus with clustering techniques to select the best models from a given pool. Given a set of predicted models, CC-Select first calculates a consensus score for each structure based on its average pairwise structural similarity to other models. Then, similar structures are grouped into clusters using multidimensional scaling and clustering algorithms. In each cluster, the one with the highest consensus score is selected as a candidate model. For the model QA problem, MDS-QA combines single-model scoring functions with consensus to determine more accurate assessment score for every model in a given pool. The experiments of MDS-QA focus on a new type of QA task appeared in the CASP10 competition in 2012. Instead of using the full set of submitted protein models as the QA test set, the organizer only released the best 150 models chosen by naive consensus as the QA test set (<http://www.predictioncenter.org/casp10>). This could severely reduce the effectiveness of consensus methods. Since this QA task is new, there is no performance comparison of QA methods on this kind of QA test sets available in the literature.

The rest of the paper is organized as follows. In the next section, we present key concepts and problem formulation. In Section 3, the new algorithm CC-Select is presented. In Section 4, the new algorithm MDS-QA is presented. Section 5 shows experimental results. Finally, we review related work in Section 6 and summarize the paper in Section 7.

2. Problem Definitions Model Quality Assessment and Model Selection

In this section, we first introduce some basic terminology and then define the model selection problem and the model QA problem, two different, yet closely related problems.

The similarity of two 3-D structures of a protein can be measured in various ways. Widely used metrics include pairwise root-mean-squared distance (RMSD), global distance test total score (GDTTS),¹⁶ and Q score.^{11,17} GDTTS is the main measurement used in CASP competitions. GDTTS score represents the percentage of aligned regions between two protein structures and is defined as follows:

$$GDTTS(s_i, s_j) = (P_1 + P_2 + P_4 + P_8) / 4 \quad (1)$$

where s_i and s_j are two 3-D models and P_d is the percentage of residues in s_i that is within distance d , $d \in \{1, 2, 4, 8\}$ angstrom (Å) from corresponding residues in s_j after s_i is superimposed on s_j .¹⁶ GDTTS values are in the range of 0 to 1. The larger the GDTTS score between two structures, the more similar they are. In this paper, GDTTS score is computed using software TM-score.¹⁸

The GDTTS value defined in Eq. (1) is asymmetric. We calculate a normalized symmetric GDTTS value, nGDTTS, by scaling the GDTTS value with the lengths of the compared structures as follows:

$$nGDTTS(s_i, s_j) = \frac{Len(s_j) \times GDTTS(s_i, s_j) + Len(s_i) \times GDTTS(s_j, s_i)}{2Len(seq)} \quad (2)$$

where $Len(seq)$ is the length of the query protein sequence, from which the 3-D predicted models s_i, s_j are generated, and $Len(seq) \geq Len(s_i), Len(seq) \geq Len(s_j)$.

The normalized score nGDTTS has several nice properties. First, it is between 0 and 1, because $Len(s_i) \leq Len(seq)$ holds for any predicted structure s_i . Second, it is symmetric, $nGDTTS(s_i, s_j) = nGDTTS(s_j, s_i)$. Finally, it is independent of model lengths and nGDTTS values of different pairs of models are directly comparable. In the rest of the paper, we only use the term GDTTS, which is actually nGDTTS for convenience purpose.

Next, we define the model selection problem. Let N be the native structure of a protein and $S = \{s_i, 1 \leq i \leq n\}$ be a set of predicted models for the protein. Each structure $s_i \in S$ has one or more estimated quality scores $e(s_i)$ (e.g., the naive consensus score, Opus ca, dDFIRE, and CalRW scores) and a truth quality score $t(s_i)$ (e.g., its GDTTS value to the native structure N). The objective is to select a subset of models $R = (R_1, R_2, \dots, R_k) \subseteq S$ based on $e(s_i)$ such that R contains the best true structures:

$$R = \operatorname{argmin}_{R \subseteq S} |\max t(R) - \max t(S)|. \quad (3)$$

In CASP competition, each prediction group can submit up to five models, which corresponds to $k = 5$ in the definition. When $k = 1$, it becomes finding the single best model.

Finally, we define the model QA problem. The goal of this problem is to maximize the correlation between the estimated quality for each of a set of models and their true quality. Specifically, given the model pool S , input quality estimations $e(S)$, and true quality score $t(S)$, compute a new quality score $m(S)$ based on S and $e(S)$, such that the Pearson correlation between $m(S)$ and $t(S)$ is maximized:

$$\max_m \operatorname{correlation}(m(S), t(S)). \quad (4)$$

The Pearson correlation coefficient r value is between -1 and 1 , with 1 for perfectly correlated, 0 for no correlation, and -1 for perfectly reversely correlated.

3. CC-Select: Combining Consensus and Clustering

In this section, a new consensus and clustering-based algorithm, CC-Select, is presented for model selection. CC-Select has five steps: consensus score calculation, filtering, dimension reduction, clustering, and final model determination, as shown in Fig. 2. Given a pool of models, their naive consensus scores are computed and bad models are dropped based on the scores. Then, the remaining models are mapped onto a Euclidean space based on their pairwise similarities using a multidimensional scaling algorithm, followed by the k -means clustering algorithm group them into k clusters. Finally, k models are selected, one from each cluster, as the final output.

3.1. Consensus score calculation and filtering

In the first step, the naive consensus method is applied to calculate a consensus score for each model. Analysis of models generated in CASP competitions suggests that a correctly-folded model is more likely to be similar to other predicted models than an incorrectly folded model.^{13,19} Therefore, the similarity between a model and other models is a good estimation of the quality of the model, which is the main idea of the naive consensus method. Let S be a set of predicted structures of a protein. For each structure s_i , $s_i \in S$, its naive consensus score q_i is

$$q_i = \frac{1}{|S|} \sum_{s_j \in S} GDTTS(s_i, s_j). \quad (5)$$

The example in Fig. 1(b) shows that the naive consensus scores are strongly correlated with the ground truth. Though simple, the naive consensus method is among the best for near-native model selection.^{15,20}

In the second step, the models are ranked based on their consensus scores and those with the lowest consensus scores are discarded. Bad models are filtered out to prevent them from exerting undue influence on the results of clustering. Based on a separate training set, we

determined a threshold, 20%, for the filtering step, i.e., 20% predicted structures with the lowest scores are discarded.

3.2. Dimension reduction using MDS

In the third step, models are mapped onto a Euclidean space based on their pairwise similarities using a multidimensional scaling algorithm. Pairwise GDTTS scores between each pair of models are computed to form a similarity matrix. Given the pairwise GDTTS matrix M , the models are mapped onto a Euclidean space using multidimensional scaling (MDS): MDS has its origins in psychometrics and psychophysics. It is a set of data analysis techniques that display the structure of dissimilarity-like data as a geometrical picture.²¹ MDS starts with one or more dissimilarity matrices that are presumed to have been derived from points in a multidimensional space, and it finds a placement of the points in a low-dimensional space, where the distances between points resemble the original dissimilarities. MDS is often used as part of exploratory data analysis or information visualization. It is also related to principal component analysis, factor analysis, and cluster analysis.

Classical metric MDS (CMDS) is used for its simplicity and efficiency. The goal of CMDS is to find a configuration of points in a multidimensional space such that the inter-point distances are related to the provided proximities by some transformation (e.g., a linear transformation). If the proximity data were measured without error in the Euclidean space, then CMDS would exactly recreate the configuration of points. Otherwise, it minimizes the sum of the least squared errors between the proximity and the computed distances between all pair of points. It is worth mentioning that the technique tolerates error gracefully, due to the over-determined nature of the solution. This is important for our method as dissimilarity restraints can be inaccurate, inconsistent, and over-restrained.

Specifically, given a dissimilarity matrix $D = (d_{ij}) = 1 - M$, $i, j = 1, \dots, n$, where M is the pairwise GDTTS matrix, CMDS computes the coordinates, $X_i \in R^n$, $i = 1, \dots, n$, through SVD on the double centered squared D . Double centering a matrix is subtracting the row and column means of the matrix from its elements, adding the grand mean and multiplying by $-1/2$. For an n by n D matrix, we have the following equation:

$$-\frac{1}{2} \left(d_{ij}^2 - \frac{1}{n} \sum_{i=1}^n d_{ij}^2 - \frac{1}{n} \sum_{j=1}^n d_{ij}^2 + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 \right) = X_i' X_j \text{ for } i, j = 1, \dots, n. \quad (6)$$

Let us call the double centered matrix on the left hand side B , and then perform singular value or Eigen value decomposition on B gives $B = UVU'$ which leads to the coordinate solution $X = UV^{1/2}$. However, with inaccurate d_{ij} values, the B matrix can be indefinite with negative as well as zero or positive roots. Let V_3 represent the 3 largest Eigen values of V , the corresponding Eigen vectors be U_3 , and $Y = U_3 V_3^{1/2}$, then $C = Y'Y$ is a least squares approximation to B .

As an example, Fig. 3 shows the 2-D mapping generated by CMDS for the models of the aforementioned target T0446. It shows that these models form three main clusters. In our algorithm, the models are mapped to an n -D space, where n is the number of models.

3.3. Clustering

In the fourth step, k -means clustering is used to group the models into k clusters.^{22,23} K -means partitions a set of points into clusters in a way that the distance between the members of each cluster is minimized while the distance between the members of different clusters maximized. Comparing with other clustering algorithms, e.g. hierarchical clustering⁸ and SPICKER,²⁴ k -means produces better results empirically in our experiments. An advantage of k -means is that the number of clusters can be controlled as an input parameter.

The cosine similarity is used in CC-Select to measure the distance between two points, say p_1 and p_2 , in n -D Euclidean space:

$$DIST(p_1, p_2) = 1 - \frac{p_1 \cdot p_2}{\|p_1\| \|p_2\|} \quad (7)$$

In CC-Select, a two-phase k -means clustering algorithm is used. In order to pick k models, k clusters are generated and one model from each cluster is chosen as the final output. Specifically, the clustering algorithm is as follows.

Algorithm CC-Select-Clustering (S, k)

1. Initial step: Select k models, i.e., k points, at random as the initial cluster centroids. Assign each model to the cluster with the closest centroid.
2. Batch updates: Reassign models to their nearest cluster all at once. Then, recompute cluster centroids. Repeat this steps iteratively to reduce the sum of distances.
3. Online updates: Reassign a model if doing so reduces the sum of distances. Recalculate cluster centroids immediately after moving each model. Repeat this step iteratively until the algorithm converges, i.e., reassigning any single model increases the total sum of distances.

Finally, in the last step of CC-Select, k models are selected as output, one from each cluster. In each cluster, the model with the highest consensus score (the original naive consensus score computed based on the whole pool of input models) is selected.

4. MDS-QA, Combining Consensus and Scoring Functions

MDS-QA is a new QA algorithm that combines the consensus idea with scoring functions, such as the publicly available Opus ca, dDFIRE, and CalRW scores. The main rational behind it is to correct naive consensus's tendency to assign larger scores to larger clusters of similar models even when there is a separate cluster with fewer but better models. The algorithm is as follows:

Algorithm MDS-QA

Given a set of 3-D models of a protein,

1. Compute Opus ca, dDFIRE, and CalRW scores of each protein. For each type of scores, normalize their values based on the whole model set to z-scores, i.e., a distribution with mean 0 and standard deviation 1. Let's call the three z-scores S_o , S_d , and S_c , respectively.
2. Compute GDT_TS values of all pairs of models.
3. Run CMDS on the pairwise GDT_TS matrix to map the models to points in 2-D space.
4. Run k -means clustering on the resulting points to find 2 clusters.
5. For each cluster:
 - a. Find the three models closest to the centroid (cluster representatives), m_i , $i = 1, 2, 3$.
 - b. For each of the 3 models, find the maximum of the Opus_ca, dFIRE, and CalRW z-scores, $S = \max(S_o, S_d, S_c)$.
 - c. Set the average of the two largest S values as the cluster's raw weight.
6. Normalize the two clusters' weights to make their sum to be 1, i.e., $w_1 + w_2 = 1$;
7. To compute the final quality score, S_f , of a model p :

$$S_f(p) = w_1 \left(\sum_{i=1}^3 GDT_TS(p, m_i^{(1)}) \right) / 3 + w_2 \left(\sum_{i=1}^3 GDT_TS(p, m_i^{(2)}) \right) / 3,$$

where $m_i^{(1)}$, $i = 1, 2, 3$, are the 3 representatives in the 1st cluster and $m_i^{(2)}$, $i = 1, 2, 3$, the 3 models in the 2nd.

As an example, Figure 4 shows the color map of the pairwise GDT_TS similarity matrix of 150 models for target T0623 from CASP9 and the CMDS mapping of the models onto 2-D space and the two clusters found by k -means clustering. One cluster roughly corresponds to the upper left cluster in the color map, while the other cluster roughly corresponds to the lower right cluster. When a model set has a large, but lower quality cluster and a small, but higher quality cluster, naive consensus will falsely assign higher scores to models in the large cluster and thus perform poorly. This happens for the T0623 models from CASP9, where models within the smaller cluster are generally better than those in the larger cluster.

In Step 5, the weight for each cluster is found based on the largest z-scores of three single-model scoring function values. Each cluster is represented by three models closest to its centroid. The average of the two largest z-scores of the three models is set as the raw weight for a cluster. Then, the raw weights of the two clusters are normalized to make their sum to be 1. In the T0623 example, the larger cluster 1 has a weight of 0.2831, while cluster two had a weight of 0.7169.

The final step uses the consensus idea as the score of each model is related to how similar it is to other models. The added weight prevents small, but good, clusters from being overwhelmed by large clusters. The benefit of MDS-QA and the weakness of naive consensus are shown in the T0623 example, where naive consensus has a QA correlation of -0.597 , very poor, while MDS-QA has a correlation of 0.958 , almost 1.

Experimental results show that MDS-QA works best on targets with multiple clusters of models. Naive consensus does better when there is a lack of clusters. To combine the strengths of naive consensus and MDS-QA, a hybrid algorithm, MDS-QA + consensus, is proposed as follows:

Algorithm MDS-QA+consensus

Given a set of 3-D models of a protein,

1. Compute GDT_TS values of all pairs of models.
2. Compute the average of all pairwise GDT_TS values, A .
3. If $A \geq 0.5$ and $A \leq 0.8$

Run MDS-QA

else

Run naive consensus

end

In the hybrid algorithm, MDS-QA is only used for a model set with an average pairwise GDT_TS value between 0.5 and 0.8. The thresholds are set similar to the classification used in CASP, where high accuracy models are those with $GDT_TS > 0.8$, medium accuracy $0.5 \leq GDT_TS \leq 0.8$, and low accuracy $GDT_TS < 0.5$. The average pairwise GDT_TS value of all models for a target is correlated to how hard the target is. For easy targets, all models are similar and are likely to form one cluster. On the other hand, for hard targets, the models are dissimilar to each other and likely to be spread out in the 2-D space. In both cases, naive consensus is generally better.

5. Experimental Results

In this section, the new algorithms, CC-Select and MDS-QA, are compared with several state-of-the-art methods in extensive experiments and shown to achieve improved results.

5.1. Results of CC-Select

In the experiments, CC-Select was compared mostly with the state-of-the-art QA method MUFOLD-QA,^{15,20} a top performer in CASP9 QA category, in experiments using data from CASP8, CASP9, Rosetta, and I-TASSER.

First, CC-Select was evaluated using 122 CASP8 targets and around 300 predicted models for each target. The lengths of their primary sequences range from 55 to 829 amino acids. Based on the average GDTTS scores of the best 10 models of each target, these targets are

partitioned in the order from hard to easy into five categories: (1) free modeling (FM), (2) fold recognition (FR), (3) comparative modeling — hard (CM H), (4) comparative modeling — medium (CM M), and (5) comparative modeling — easy (CM E). The detailed description of domain category is provided at <http://prodata.swmed.edu/CASP8/evaluation/Categories.htm>.

Table 1 and Figure 5 show the results of selecting k , $k = 5, 10, 15, 20$, models for each target by three algorithms, CC-Select, MUFOLD-QA, and hierarchical clustering (using hierarchical clustering instead of MDS+ k -means in CC-Select). Two MATLAB functions, clusterdata and linkage, were used to implement the hierarchical clustering method. When $k = 1$, CC-Select does not perform clustering any more. It simply picks the model having the highest consensus score in the pool as output. Hence, the three methods have the same result. For the rest k values, CC-Select is consistently better than the other two algorithms. In paired T test between the results of CC-Select and MUFOLD-QA, the P values are 0.0005, 0.0019, 0.0228 and 0.0064, respectively, indicating the improvement of CC-Select over MUFOLD-QA is statistically significant.

Figure 6 shows the results of CC-Select and MUFOLD-QA on the 117 CASP9 targets for picking k , $k = 5, 10, 15, 20$, models for each target. Again CC-Select is better.

Next, we compare the performance of CC-Select and MUFOLD-QA on models generated from a single prediction software, different from the previous experiments where models are generated from a large number of prediction software programs and more diverse. In this experiments, all models were generated by Rosetta, one of the best software for *ab initio* protein structure prediction.²⁵ The test models were generated using Rosetta version 2.2.0²⁶ in *ab initio* mode on 22 CASP8 targets. For each target, Rosetta generated 10 000 *ab initio* models and 300 models were randomly selected as the test models to evaluate CC-Select and MUFOLD-QA. The best model for each target all has a GDTTS score over 0.5. Although many existing prediction software can generate good models, they have difficulty to identify the best ones. Thus, the ability to select good models from a single server predictions is importance in practice. Figure 7 shows the results and CC-Select outperforms MUFOLD-QA.

Finally, we compare the performance of CC-Select and MUFOLD-QA on models generated by another state-of-the-art prediction software I-TASSER,²⁷ which participated in CASP as Zhang-Server, was ranked top in the server category of automated structure prediction in the past three CASP experiments. The Decoy Set II (<http://zhanglab.ccmb.med.umich.edu/decoys/>) generated by I-TASSER is a structurally non-redundant set of protein models. It consists of 56 protein targets and each target contains 300–500 decoys. Each decoy was first generated by I-TASSER Monte Carlo simulations and then refined by GROMACS4.0 MD simulation to remove steric clashes and improve hydrogen-bonding network. This decoy set has been used to test the ability of atomic potentials for near-native structure identification and RMSD-energy correlation. Figure 8 shows the result and CC-Select is slightly better than MUFOLD-QA.

5.2. Results of MDS-QA

MDS-QA's performance was evaluated for the new 150-model QA problem in CASP10. Because CASP10 native structures are not fully available at present, CASP9 targets and corresponding server prediction models were used in the experiments. Out of the 117 CASP9 targets, the Opus ca program could not run on 10 targets because they contained more than 400 amino acids. These ten targets were not used in our experiments. The remaining 107 targets were separated into three difficulty groups: a) easy: best model's $GDT_TS > 0.8$; b) medium: $0.8 \geq \text{best model's } GDT_TS > 0.5$; and c) hard: best model's $GDT_TS \leq 0.5$. For a target, given all models available from CASP9, which is usually more than 300, naive consensus was applied to them and only the top 150 models with the highest naive consensus scores were kept for the experiments. Thus all experimental results reported in this section are for 150 models per target.

The performance of a QA method is evaluated using the Pearson correlation r between the QA scores of all models and the true GDT_TS values of all models with respect to the native structure. Naive consensus was used in our comparison because it is among the best QA methods.

Figure 9 compares the performances of MDS-QA, naive consensus, and the MDS-QA + consensus hybrid algorithm on the medium difficulty group of 61 targets. The purple triangles are for naive consensus, red crosses for MDS-QA, and blue circles for the hybrid algorithm. For this group of 61 targets, MDS-QA has an average correlation 0.382, slightly better than naive consensus' 0.37. MDS-QA is better than naive consensus for about half of them and worse for the other half. The hybrid algorithm is better than either one on average, achieving an average correlation 0.437. Although the hybrid algorithm does not always choose the better result of the two, it is able to avoid most of the very bad scores.

For the full set of 107 targets, naive consensus had an average correlation of 0.434, while MDS-QA only had an average correlation of 0.352. Still, the hybrid algorithm is the best, with an average correlation of 0.469, an improvement of 8% over naive consensus. Experimental results show that MDS-QA performs better on medium difficulty targets than easy or hard targets.

6. Related Work

A major approach for evaluating protein structure uses energy or scoring functions, which fall into two main categories: physics-based energy functions³ and knowledge-based statistical potentials.⁴⁻⁶ Physics-based energy functions are derived from physical laws by evaluating directly the true energy of a protein tertiary structure. Knowledge-based statistical potentials are generally based on the inverse Boltzmann equation. These potentials assume protein structure in a thermodynamic equilibrium and exploit data of known protein structures to derive energy functions that reflect the distribution of observed structures. In comparison with physics-based energy functions, knowledge-based potentials are fast, simple to construct, and have been widely used for structure quality assessment. However, although a variety of scoring functions have been proposed over the years, with some recent

ones that are very complicated and incorporated information of dozens of sources, none of them can consistently and reliably identify near-native structures.

Another approach for protein structure selection is based on clustering.⁷⁻⁹ The idea of using clustering to discriminate protein structures suggests that near-native structures have more predicted structural neighbors than poor structures. Some clustering approach, e.g. SCAR,⁷ applied *k*-means clustering algorithm and used cluster centroids as representative models for clusters. A centroid is the average of all the structures in a cluster and is constructed by minimizing distance constraint between each pair of residues. Some clustering approaches, e.g. SPICKER²⁴ and its variant,⁹ found clusters by looking for one or more structures with the largest number of neighbors within a certain clustering radius. Surrounding a structure with most neighbors, a cluster is formed and the cluster center is constructed. SPICKER also used cluster centroids as the candidates of near-native conformations. Though centroid is more robust than an existing structure in a cluster, it often contains significant atomic clashes.

The third approach is based on consensus among predicted models.^{2,10,12} Like clustering methods, consensus approach uses structure similarity information. But instead of building cluster centroids, consensus approach employs the pairwise similarity between structures to rank predicted models. The idea behind the consensus approach is that a near-native model is more likely to be similar to other predicted models than an incorrect model. Consensus methods have been very successful in the recent CASP competition.¹³⁻¹⁵ In the model quality assessment (QA) category in CASP8 and CASP9, the accuracy of structure evaluation achieved by consensus-based approach was consistently better than other methods.

7. Summary

As the most effective approach for model quality evaluation in recent CASP competitions, consensus-based methods have attracted significant attention. With intensive research efforts since CASP7, QA methods have been steadily improved. In this paper, the two new methods, CC-Select and MDS-QA, improve the selection and QA performance by combining consensus with clustering or consensus with scoring functions. In extensive experiments, we demonstrated that CC-Select outperforms existing methods on both single-server model sets and multiple-server model sets. MDS-QA and a hybrid algorithm outperform naive consensus on the challenging 150-model QA problem.

Within MDS-QA, multidimensional scaling was used to map the data into two dimensions so that it would be easy to visualize and manipulate. However, it may be more effective to use more dimensions based on the properties of the models. At the same time, using too many dimensions may remove some characteristics of the method and become too similar to naive consensus. The number of clusters and the clustering algorithm could also be varied in order to best match each target. Due to the CASP competition rule of using only the top 150 models based on naive consensus, there were usually no more than a few major clusters. But even if there were not clear cluster divisions in a data set, clustering may be able to increase accuracy. In the future work, it may be more effective to dynamically decide certain

parameters in the method, such as the number of clusters and the ways to use single-model score functions to weight the different clusters, so that the method could tailor itself to match each target more precisely. The hybrid method of MDS-QA could also be improved. In this work, a simple average consensus score threshold was used to decide when to use MDS-QA and when not to. More sophisticated selection methods based on the structures of the clusters or the distribution of consensus or single model QA scores could be developed.

Acknowledgments

This work has been partially supported by NIH Grant R21/R33-GM078601 and R01-GM100701. We are thankful to Ioan Kosztin, Bogdan Barz, Zhiqian He, Jingfen Zhang, and Jianlin Cheng for valuable suggestions and feedbacks.

References

1. Baker D, Sali A. Protein structure prediction and structural genomics. *Science*. 2001; 294(5):93–96. [PubMed: 11588250]
2. Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. *Proteins: Structure, Function, and Bioinformatics*. 2008; 71(3):1175–1182.
3. Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy minimization and dynamic calculations. *Journal of Computational Chemistry*. 1983; 4(2):187–217.
4. Kalman M, Ben-Tal N. Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics*. 2010; 26(10):526–531.
5. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*. 2002; 11:2714–2726. [PubMed: 12381853]
6. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Science*. 2006; 15(11):2507–2524. [PubMed: 17075131]
7. Betancourt MR, Skolnick J. Finding the needle in a haystack: Educating protein native folds from ambiguous *ab initio* folding predictions. *Journal of Computational Chemistry*. 2001; 22(3):339–353.
8. Gront D, Hansmann UH, Kolinski A. Exploring protein energy landscapes with hierarchical clustering. *International Journal of Quantum Chemistry*. 2005; 105(6):826–830. [PubMed: 16479277]
9. Wang, Q.; Shang, Y.; Xu, D. A new clustering-based method for protein structure selection, in; *Proc. of the IEEE Int. Joint Conf. on Neural Networks*; 2008 Jun. p. 2891–2898.
10. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: A simple approach to improve protein structure predictions. *Bioinformatics*. 2003; 19(8):1015–1018. [PubMed: 12761065]
11. McGuffin LJ, Roche DB. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*. 2010; 26(2): 182–188. [PubMed: 19897565]
12. Wallner B, Elofsson A. Pcons5: Combining consensus, structural evaluation and fold recognition scores. *Bioinformatics*. 2005; 21(23):4248–4254. [PubMed: 16204344]
13. Cozzetto D, Kryshchuk A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. *Proteins: Structure, Function, and Bioinformatics*. 2007; 69(Suppl. 8):175–183.
14. Cozzetto D, Kryshchuk A, Tramontano A. Evaluation of casp8 model quality predictions. *Proteins: Structure, Function, and Bioinformatics*. 2009; 77(Suppl. 9):157–166.
15. Wang, Q.; Shang, Y.; Xu, D. Protein structure selection based on consensus; *Proc. of the IEEE Congress on Evolutionary Computation*; 2010 Jul. p. 1–7.
16. Zemla A. LGA: A method for finding 3d similarities in protein structures. *Nucleic Acids Res*. 2003; 31(13):3370–3374. [PubMed: 12824330]

17. Ben-David M, Noivirt-Birk O, Paz A, Prilusky J, Sussman JL, Levy Y. Assessment of casp8 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*. 2009; 77(9):50–65.
18. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*. 2004; 57(4):702–710.
19. Venclovas, eslovas; Margelevi ius, M. Comparative modeling in casp6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins: Structure, Function, and Bioinformatics*. 2005; 7(Suppl. 7):99–105.
20. Wang Q, Shang Y, Xu D. Improving consensus approach for protein structure selection by removing redundancy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2011; 8(6):1708–1715. [PubMed: 21519117]
21. Borg, I.; Groenen, P. *Modern Multidimensional Scaling, Theory and Applications*. New York: Springer-Verlag; 1997.
22. MacQueen, JB. Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. Vol. 1. University of California Press; 1967 Jun. Some methods for classification and analysis of multivariate observations; p. 281-297.
23. Spath, H. *The Cluster Dissection and Analysis Theory FORTRAN Programs Examples*. New York: Prentice-Hall, Inc.; 1985.
24. Zhang Y, Skolnick J. Spicker: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*. 2004; 25(6):865–871. [PubMed: 15011258]
25. Zhang J, Wang Q, Barz B, He Z, Kosztin I, Shang Y, Xu D. Mufold: A new solution for protein 3d structure prediction. *Proteins: Structure, Function, and Bioinformatics*. 2010; 78(5):1137–1152.
26. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the robetta server. *Nucleic Acids Res*. 2004; 32:526–531.
27. Zhang J, Zhang Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One*. 2010; 5(10):e15386. [PubMed: 21060880]

Fig. 1.

(Color online) The distributions of estimated quality scores against ground truth (GDTTS) for 282 predicted models of protein 3dcx (CASP8 target T0446). The truth quality of each model, i.e., its GDTTS score to the native model, is plotted against a) its DFIRE scoring and b) naive consensus score.

Fig. 2.

The flow chart of CC-Select. Here, $R_1, R_2, \dots, R_k$ denotes the final k models outputted, each from one cluster C_i .

Fig. 3.

(Color online) The distribution of models for the CASP8 target T0446 being mapping by CMDS onto a 2-D space.

Fig. 4.

(Color online) Top: The pairwise similarity (GDT_TS) of 150 models for target T0623 from CASP9. Bottom: The CMDS mapping of T0623 models and the centroids of the two clusters found by *k*-means clustering.

Fig. 5.

(Color online) Performance comparison of three model selection algorithms on the data set of 122 CASP8 targets.

Fig. 6.

(Color online) Performance comparison of CC-Select and MUFOLD-QA on the data set of 117 CASP9 targets.

Fig. 7.

(Color online) Performance comparison of CC-Select and MUFOLD-QA on the data set of Rosetta models of 22 CASP8 targets.

Fig. 8.

(Color online) Performance comparison of CC-Select and MUFOLD-QA on the decoy set II generated by I-TASSER for 56 proteins.

Fig. 9.

(Color online) Performance comparison of MDS-QA, naive consensus, and the MDS-QA+consensus hybrid algorithm on 61 CASP9 medium difficulty targets for the 150-model QA problem. MDS-QA has an average correlation 0.382, naive consensus 0.37, and the hybrid algorithm 0.437.

Comparison of the average GDT_TS of the best of k models selected by three methods, MUFOLD-QA, Hierarchical clustering, and CC-Select, on the data set of 122 CASP8 targets.

Table 1

Target Category	MUFOLD-QA				Hierarchical Clustering				CC-Select			
	$k = 1$	$k = 5$	10	20	5	10	20	5	10	20		
FM	0.232	0.249	0.261	0.265	0.237	0.243	0.249	0.247	0.247	0.256		
FR	0.374	0.395	0.400	0.409	0.387	0.398	0.402	0.398	0.407	0.412		
CM_H	0.586	0.594	0.600	0.606	0.592	0.606	0.615	0.616	0.619	0.625		
CM_M	0.717	0.728	0.730	0.732	0.724	0.726	0.737	0.731	0.734	0.740		
CM_E	0.854	0.864	0.869	0.875	0.866	0.873	0.875	0.871	0.875	0.879		
Overall	0.635	0.647	0.652	0.657	0.644	0.651	0.658	0.654	0.659	0.664		