Journal of Bioinformatics and Computational Biology
Vol. 7, No. 5 (2009) 1–22

ICP Imperial College Press
www.icpress.co.uk

PREDICTING LOCAL QUALITY OF A SEQUENCE–STRUCTURE ALIGNMENT

XIN GAO[*,‡], JINBO XU[†,§], SHUAI CHENG LI[*,¶] and MING LI[*,‖]

[*]*David R. Cheriton School of Computer Science*
University of Waterloo, 200 University Avenue West
Waterloo, Ontario, N2L 3G1 Canada

[†]*Toyota Technological Institute at Chicago*
1427 East 60th Street, Chicago, IL, 60637, USA
[‡]*x4gao@cs.uwaterloo.ca*
[§]*j3xu@tti-c.org*
[¶]*scli@cs.uwaterloo.ca*
[‖]*mli@cs.uwaterloo.ca*

Received 16 January 2009
Revised 6 April 2009
Accepted 7 April 2009

Although protein structure prediction has made great progress in recent years, a protein model derived from automated prediction methods is subject to various errors. As methods for structure prediction develop, a continuing problem is how to evaluate the quality of a protein model, especially to identify some well-predicted regions of the model, so that the structural biology community can benefit from the automated structure prediction. It is also important to identify badly-predicted regions in a model so that some refinement measurements can be applied to it. We present two complementary techniques, FragQA and PosQA, to accurately predict local quality of a sequence–structure (i.e. sequence–template) alignment generated by comparative modeling (i.e. homology modeling and threading). FragQA and PosQA predict local quality from two different perspectives. Different from existing methods, FragQA directly predicts cRMSD between a continuously aligned fragment determined by an alignment and the corresponding fragment in the native structure, while PosQA predicts the quality of an individual aligned position. Both FragQA and PosQA use an SVM (Support Vector Machine) regression method to perform prediction using similar information extracted from a single given alignment. Experimental results demonstrate that FragQA performs well on predicting local fragment quality, and PosQA outperforms two top-notch methods, ProQres and ProQprof. Our results indicate that (1) local quality can be predicted well; (2) local sequence evolutionary information (i.e. sequence similarity) is the major factor in predicting local quality; and (3) structural information such as solvent accessibility and secondary structure helps to improve the prediction performance.

Keywords: Local quality assessment; protein structure prediction; SVM regression; sequence-structure alignment; CASP7.

[‡]Corresponding author.

1. Introduction

The biennial CASP (Critical Assessment of Structure Prediction)[1–4] events have demonstrated that the three-dimensional structures of many new target proteins can be predicted at a reasonable resolution, although in most cases, the predicted models are still not accurate enough for functional study. In particular, comparative modeling methods can generate reasonably good models for approximately 70% of target proteins in recent CASP events. Even for those free modeling (FM) targets, a structural model generated by protein threading usually contains some good local regions, although the overall conformation of the model is incorrect.[5]

As methods for structure prediction develop, a continuing problem is how to evaluate the quality of a protein model in details. The challenge is to distinguish a good model from a bad one (referred to as global quality assessment), as well as correctly-predicted residues from badly-predicted ones (referred to as local quality assessment). To make automated structure prediction really useful for the structural biology community, a reliable model quality evaluation program is indispensable when hundreds of models are predicted for a single target protein.

Global quality prediction has been an active research topic for two decades.[6–35] This kind of programs can be used to pick up the best few models from a bunch of models generated by different structure prediction programs, which enables structure biologists to focus on the most native-like models. However, a structural model is not able to provide enough information for functional study if it has a bad quality.[36]

A common practice taken by some human predictors or consensus-based automatic predictors to further improve the accuracy of the structure prediction is to identify correctly-predicted regions from each structural model and then assemble them together to obtain a better overall model for the target protein; for example, TASSER[5] and 3D-SHOTGUN[37] are two such top-scoring methods. This kind of refinement methods often perform better than the classical threading-based protein structure prediction methods. The key factor underlying the success of these refinement methods is identifying the correctly-predicted regions in a structural model. Besides being used to examine and improve the accuracy of a protein model, local quality prediction methods can also be used to recognize functional residues in a protein model.[38,39]

Local quality assessment methods are either structure-based[32,34,40–44] or alignment-based.[36,38,45–47] ERRAT[42] is a program that uses only structural information. This program employs a Gaussian error function based on the statistics of non-bonded interactions to predict incorrect regions in a protein model. Such methods can recognize incorrect structural regions which obviously deviate from their natives. There are also some programs using alignment information to predict local quality. Tress *et al.* developed a method to evaluate local quality of a given alignment and tested the method on alignments generated by five comparative-modeling methods.[38] The results indicate that an alignment position with a high

profile-derived alignment score often has good quality. Wallner *et al.* developed four neural network-based methods, i.e. ProQres, ProQprof, ProQlocal and Pcons-local, to identify correct regions in a protein model, using either structural information or alignment information.[36] ProQres uses only structural information in a protein model, while ProQprof uses alignment information such as profile-profile scores, information scores, and gap penalty. ProQlocal combines ProQres and ProQprof together to achieve a better performance. Pcons-local is a consensus-based local quality predictor, taking as input protein models generated by different structure-prediction programs. These four methods evaluate local quality by comparing the sequence alignments used to build the models with the optimal structure alignments. However, to make local quality assessment methods really useful for structure prediction and refinement approaches, it is crucial to assess the real quality of regions of the structural models. Meanwhile, it is also important to evaluate the single position quality, so that local refinement strategies can be applied to.

In this paper, we present two complementary methods, FragQA and PosQA, to accurately predict local quality of a sequence-structure alignment. Distinguishing itself from previous methods, FragQA directly predicts the quality of an ungapped region in the alignment. The quality is measured using the cRMSD (i.e. C_α-based RMSD) between two fragments corresponding to the ungapped region: one is the native structure of the region and the other one is the predicted model. Note that the quality measurement used here is "absolute" quality, which is independent of the optimal structure alignment. Furthermore, statistical significance is introduced to improve FragQA's performance. As opposed to cRMSD, statistical significance can cancel out the impact of region length. Some preliminary results of FragQA have been discussed in Ref. 46. Complementary to FragQA, our recently developed PosQA predicts the quality of an individual aligned position in a given alignment. The single position quality is measured using a normalized cRMSD described in Ref. 36. FragQA and PosQA utilize only information in a single alignment. Structural information in the alignment-derived protein model is not directly used. However, in calculating features from an alignment, we use structural information in the template.

2. Results

2.1. *Problem description*

This paper studies the following two problems:

(1) Given a sequence-structure alignment, what is the quality of an ungapped region in this alignment? The quality is defined as the cRMSD between the native and the predicted local structures of the ungapped region, denoted as "cRMSD of an ungapped region", after they are optimally superimposed. Note that the two local structures are superimposed without taking into consideration other parts of the alignment. The alignment is cut into ungapped regions

at gap positions. Thus, the fragments studied here are different from the fixed-length fragments studied in Refs. 45 and 47. *FragQA is developed to solve this problem.*

(2) Given a sequence-structure alignment, what is the quality of a single aligned position in this alignment? To measure the quality of a single position, we optimally superimpose the predicted structural model, derived from this alignment, and the native structure, and then calculate cRMSD at each position to measure its quality. The final quality measure is normalized cRMSD as described in Ref. 36. More specifically, let D_i and d_i denote the normalized cRMSD and the original cRMSD at position i, respectively. Then D_i is defined as $1/(1 + \frac{(d_i)^2}{(d_0)^2})$ where d_0 is set to $\sqrt{5}$ according to Ref. 36. Different from the quality measure of an ungapped region, the single-point quality depends on the superimposition between the whole predicted model and its native structure. *PosQA is developed to solve this problem.*

2.2. *FragQA training*

Training and test data. Choosing good training and test data is one of the key steps in objectively evaluating the performance of a machine-learning method. FragQA and PosQA are tested on several threading methods, such as RAPTOR,[48] PROSPECT-II,[49] and GenTHREADER.[50] The results are similar. In this paper, we only show the results on alignments generated by RAPTOR default threading algorithm. The training and test data are from the CASP7 event. As suggested by Fasnacht *et al.*,[45] CASP dataset is the most practical and challenging set, which covers a very broad range of types of target proteins and local errors. There are 104 target proteins in CASP7 while 92 of them were considered as valid targets and were used for final assessment by CASP7 assessors. Ninety-one target proteins are left after we removed redundancy at 40% sequence identity level using CD-HIT.[51] Only T0346 is removed because it shares 71% sequence identity with T0290. To do a cross validation, the 91 target proteins are randomly divided into four sets. Top 10 alignments generated by RAPTOR are considered for each target protein. If one target protein belongs to a set, then all of its 10 alignments belong to this set. Each alignment is cut into a set of ungapped regions with cutting points being at the gap positions. The ungapped regions containing less than five residues are not considered in our experiments. Table 1 shows the statistics on the four sets. It is clear that the four datasets are very similar.

Training. SVM-light[52] with RBF (radial basis function) kernel is used to train FragQA. The parameter gamma in the RBF kernel function is trained using the leave-one-out error estimation method. Other parameters are set to their default values or calculated automatically by SVM-light. Experimental results indicate that the RBF kernel with its gamma parameter set to 0.2 can yield the best training performance. Other kernel functions such as linear kernel and polynomial kernel are also tested, but they cannot yield as good performance as the RBF kernel.

Table 1. Statistics of ungapped regions on the four datasets.

Set Name	# of proteins	# of fragments	Average cRMSD	Deviation
1	23	1347	2.93 Å	1.50 Å
2	22	1108	2.57 Å	1.46 Å
3	23	1519	2.86 Å	1.47 Å
4	23	1461	2.73 Å	1.49 Å

Columns 2–5 show the number of target proteins, the number of fragments, the average quality in terms of cRMSD of the fragments, and the standard deviation of cRMSD of each set, respectively.

A four-fold cross validation is applied. Each time three of the four datasets are used as the training set, and the other one is used for testing.

2.3. *Performance of FragQA*

After studying the relative importance of eight features (see Sec. 4 for the description of the features), which will be discussed later, the following features are encoded into FragQA: (1) length of the ungapped region, (2) Z-score of the whole alignment, (3) mutation score of the region, (4) environmental fitness score of the region, and (5) secondary structure score of the region.

2.3.1. *Prediction error and correlation coefficient of FragQA*

To the best of our knowledge, FragQA is the first method to directly predict the quality of fragments that are automatically determined by the sequence–structure alignments rather than fragments with fixed length. Thus, there is no existing method for us to compare with. The prediction error is defined as the absolute difference between the predicted cRMSD value and the real one. Table 2 lists the

Table 2. Prediction accuracy and correlation coefficient of FragQA.

cRMSD	Test Set 1	Test Set 2	Test Set 3	Test Set 4	Ave. Fraction (%)
≤ 1 Å	1.36	1.57	1.41	1.54	14
≤ 2 Å	1.11	1.28	1.08	1.18	42
≤ 3 Å	1.00	1.16	0.94	1.04	69
≤ 4 Å	1.03	1.12	0.97	1.04	85
≤ 5 Å	1.12	1.14	1.06	1.09	92
≤ 6 Å	1.20	1.19	1.16	1.20	95
≤ 7 Å	1.33	1.26	1.22	1.25	97
≤ 8 Å	1.41	1.32	1.29	1.31	98
≤ 9 Å	1.48	1.36	1.37	1.36	99
≤ 10 Å	1.57	1.39	1.41	1.41	99
Correlation coefficient	0.51	0.46	0.50	0.48	—

Column 1 lists different cRMSD thresholds. Columns 2–5 list prediction errors of FragQA under different cRMSD thresholds on the four test sets. Column 6 lists average fraction of fragments with real cRMSD under such thresholds.

average prediction errors of FragQA, under different cRMSD thresholds on the four test sets, together with the average fraction of fragments with real cRMSD under such thresholds, and the correlation coefficient between the predicted and real cRMSD by FragQA on the four test sets. As shown in this table, the prediction error of FragQA ranges from 0.9 Å to 1.6 Å. The smallest error of FragQA happens when cRMSD threshold is set to 3 Å, which means FragQA is most accurate when dealing with fragments with cRMSD smaller than 3 Å to the native. However, when the real cRMSD is very small (≤ 1 Å), the prediction error tends to be big. In other words, it is hard to obtain an accurate prediction when cRMSD is very small. As indicated in Table 2, the correlation coefficient between the predicted cRMSD by FragQA and the real cRMSD is about 0.5 for each test set.

2.3.2. *Feature selection for FragQA*

It is important to detect which features are closely relevant to the prediction capability of FragQA since unrelated features may introduce extra noise. The importance of each feature is investigated by excluding it from the entire feature set, training a new FragQA, and then testing the performance of this new predictor. Thus, the performance resulting from different sets of features can be compared, and the important features can be detected.

Table 3 lists the sensitivity and specificity of FragQA with different sets of features under different cRMSD thresholds on test set 1. The results are similar on the other test sets. There is no obvious difference among different sets of features when cRMSD threshold is larger than 3.75 Å. As shown in this table, if the aligned region length is removed, the performance of FragQA will drop obviously, except for cRMSD threshold larger than 2.75 Å, the sensitivity of FragQA without fragment

Table 3. Sensitivity and specificity of FragQA with different feature sets.

cRMSD	All	No Len	No S_z	No S_m	No S_e	No S_c	No S_{ss}	No $SeqId$	No Seq
≤ 1 Å	12/19	0/0	4/10	9/17	11/16	13/32	13/17	12/18	12/18
≤ 1.25 Å	16/28	1/22	8/20	15/27	14/22	22/43	18/27	16/28	15/28
≤ 1.5 Å	25/42	4/23	16/37	19/35	22/36	27/49	26/41	25/42	25/41
≤ 1.75 Å	35/52	12/41	27/51	27/46	29/47	34/57	36/51	34/52	35/52
≤ 2 Å	42/59	21/48	38/58	35/53	39/57	48/65	42/56	43/60	42/59
≤ 2.25 Å	50/64	42/56	52/64	46/60	48/62	58/68	51/63	51/64	51/64
≤ 2.5 Å	62/72	61/63	64/70	55/66	56/69	65/73	63/70	62/72	62/72
≤ 2.75 Å	70/78	74/67	73/75	65/73	67/76	74/78	71/77	69/78	69/77
≤ 3 Å	76/79	82/70	79/77	74/77	75/80	81/79	77/79	76/79	76/79
≤ 3.25 Å	83/82	90/75	86/80	82/81	80/83	85/82	84/80	83/82	83/82
≤ 3.5 Å	88/86	94/79	90/84	88/83	84/85	89/86	89/84	88/86	88/86

Column 1 lists different thresholds. Column 2 lists the sensitivity/specificity of FragQA with all features. Starting from column 3, each column lists the sensitivity/specificity when one feature is removed. Len: region length, S_z: Z-score, S_m: mutation score, S_e: environmental fitness score, S_c: contact capacity score, S_{ss}: secondary structure score, $SeqId$: sequence identity, and Seq: other sequential features. All values are percentiles.

length is a little higher than that with all the features. This complies with a fact that cRMSD itself is closely related to the length of an ungapped region. Removing mutation score or the overall Z-score will also have an obvious reduction on the performance of FragQA, except for cRMSD larger than 2.25 Å, where removing Z-score will increase the sensitivity slightly and have no obvious influence on the specificity. This also makes sense: mutation score measures the sequence similarity in the aligned region, and Z-score evaluates the overall quality of the alignment. An alignment with good overall quality often contains good aligned regions. However, when the overall quality of an alignment is poor (Z-score is low), the fragments can be either good or bad. In such case, Z-score will not be an influential factor any more. Removing environmental fitness score will decrease both the sensitivity and the specificity. Surprisingly, removing contact capacity score will increase both the sensitivity and the specificity. This implies that contact score is a noisy feature. On the other hand, removing secondary structure score will decrease the specificity but increase the sensitivity slightly. Removing any other features, such as sequence identity feature and other sequential features, does not obviously deteriorate either the sensitivity or the specificity. Thus, the final version of FragQA uses the following features: (1) aligned region length, (2) overall alignment Z-score, (3) mutation score, (4) environmental fitness score, and (5) secondary structure score. Meanwhile, mutation score, Z-score, and the region length are the most important factors in quality prediction.

2.3.3. *Statistical significance*

The cRMSD between the predicted structure of an ungapped region and its native is closely relevant to the length of the region. Thus, a five-residue ungapped region with 3 Å cRMSD may not be better than a 15-residue region with 4 Å cRMSD. To better evaluate the quality of a region, the statistical significance of its cRMSD is calculated to reduce the bias introduced by region length. To calculate statistical significance, statistical distribution of cRMSD for a given region length is empirically calculated as follows. For a given region length, 10,000 pairs of fragments of this length are randomly sampled from PDB30, and their pairwise cRMSDs are calculated. PDB30 is a subset of PDB (the Protein Data Bank),[53] in which any two proteins share no more than 30% sequence identity. As shown in Fig. 1(a), the mean of cRMSD increases clearly with respect to the length, but the standard deviation increases much more slowly. The cRMSD distribution looks like a normal distribution. Figure 1(b) shows the statistical distribution of cRMSD calculated from 10,000 randomly sampled pairs of fragments with length 10. Fragments with different length give similar distributions. For a given ungapped region with length l and (real or predicted) cRMSD r, its statistical significance (denoted as $StatSig$) is calculated as follows:

$$StatSig = \frac{\#random\ pairs\ of\ length\ l\ with\ cRMSD \geq r}{10{,}000}. \quad (1)$$

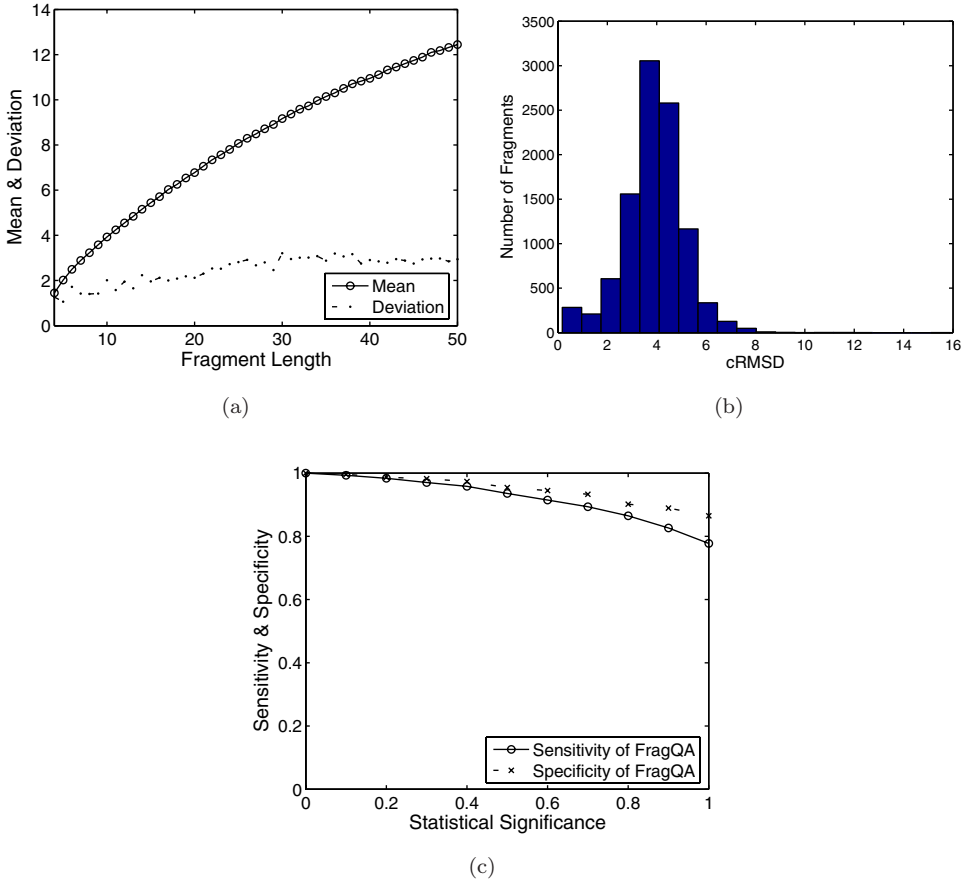


Fig. 1. (a) Mean (circled solid line) and standard deviation (point dotted line) of cRMSD for random region sets with length from 5 residues to 50 residues. (b) The statistical distribution of cRMSD calculated from 10,000 randomly sampled pairs of fragments with length 10. (c) FragQA's sensitivity (circle solid line) and specificity (cross dotted line) in terms of statistical significance on test set 1. Please see Sec. 2.3.3 for the definitions of sensitivity and specificity.

Thus, the smaller the cRMSD is, the larger its statistical significance is.

The sensitivity and specificity of FragQA in terms of statistical significance are calculated in a way similar to that calculated in terms of cRMSD. For each statistical significance threshold varying from 0 to 1, the sensitivity is defined as the percentage of ungapped regions with real statistical significance larger than or equal to the threshold, that also have predicted values larger than or equal to the threshold. The specificity is defined as the percentage of ungapped regions with predicted significance larger than or equal to the threshold, that have real statistical significance better than or equal to the threshold. Figure 1(c) illustrates the sensitivity and specificity of FragQA in terms of statistical significance on test set 1. Results are similar on the other three sets. As shown in this figure, when

statistical significance is 0.8 (about 81% of fragments in our test sets have such values), both the sensitivity and specificity are around 90%. Even when statistical significance threshold is 1 (about 48% of fragments in our test sets have this value), the sensitivity is 78%, and the specificity is 88%.

We also studied the prediction error of FragQA in terms of statistical significance. As shown in Table 4, the prediction error decreases quickly from 0.26 to 0.05 when the statistical significance threshold increases from 0 to 1. When the threshold is 0.9, the prediction error is approximately 0.12. This indicates that FragQA is able to predict the statistical significance well when the ungapped region has a good quality. By contrast, FragQA is not able to accurately predict cRMSD when it is small because a small cRMSD does not imply a high-quality region. This result also shows that statistical significance is a better measure than cRMSD. All the test alignments are further divided into three classes, "high-quality" alignments, "medium-quality" alignments, and "low-quality" alignments, based on their Z-scores (calculated by RAPTOR) at cutting points 0.33 and 0.66. A "high-quality", "medium-quality", and "low-quality" alignment has Z-score at least 0.66, between 0.33 and 0.66, and less than 0.33, respectively. Table 4 indicates that different sets have different prediction errors. The underlying reason may be that different sets have different distributions of ungapped regions under a given threshold.

On the other hand, the correlation coefficient of FragQA on each set in terms of statistical significance is higher than 0.60. This means that statistical significance is probably a better way to measure the quality of a fragment.

2.4. *PosQA training*

PosQA uses the same data source as FragQA to train and test the SVM model. The only difference is that a data entry in FragQA is an ungapped region while a data entry in PosQA is a single aligned position. If a residue in the target protein

Table 4. Prediction errors of FragQA in terms of statistical significance.

StatSig	Whole	High-quality	Medium-quality	Low-quality
≥ 0	0.26	0.21	0.25	0.28
≥ 0.1	0.23	0.20	0.23	0.25
≥ 0.2	0.21	0.19	0.21	0.22
≥ 0.3	0.19	0.16	0.18	0.20
≥ 0.4	0.17	0.14	0.17	0.18
≥ 0.5	0.15	0.12	0.16	0.16
≥ 0.6	0.14	0.10	0.15	0.14
≥ 0.7	0.13	0.08	0.14	0.14
≥ 0.8	0.12	0.08	0.14	0.13
≥ 0.9	0.12	0.08	0.14	0.13
$= 1.0$	0.05	0.03	0.04	0.08

Column 1 lists different significance thresholds. Column 2 lists the overall prediction errors of FragQA. Columns 3–5 are the prediction errors on the three classes of alignments: "high-quality", "medium-quality", and "low-quality". Please see the text for the definition of these three classes.

is aligned to a gap, the quality of this position is set to zero, and this residue is not used for training or test. The whole CASP7 dataset is also divided into four sets as in FragQA. In summary, there are 26 432, 27 018, 26 982, and 26 831 entries in the four sets, respectively. Their average normalized cRMSD values, D_i's, are 0.57, 0.51, 0.52 and 0.54, respectively.

The SVM-light software[52] is also applied to train PosQA with the RBF kernel, following almost the same procedure that trains FragQA. The objective values in the SVM regression training are D_i values. Experimental results indicate that PosQA yields the best performance when the RBF kernel function is used with gamma being 0.3. After selecting features by adopting the similar approach used by FragQA, PosQA encodes the following features: (1) overall alignment Z-score, (2) mutation score, (3) environmental fitness score, and (4) secondary structure score. Again, contact capacity score has no contribution to the performance of PosQA, and is thus not encoded in PosQA.

2.5. *Performance of PosQA*

2.5.1. *Prediction error of PosQA*

We compared the prediction error of PosQA, ProQres, and ProQprof, which is defined as the average absolute difference between the predicted D_i and its real value. Table 5 shows the prediction errors above different D_i thresholds. As shown in this table, the overall prediction errors for PosQA, ProQres, and ProQprof range from 0.13 to 0.29, 0.14 to 0.41, and 0.15 to 0.40, respectively. This implies that the overall prediction accuracy of PosQA is better than that of ProQres and ProQprof. When D_i increases, the overall prediction errors of PosQA decrease clearly, while the lowest errors of ProQres and ProQprof happen when D_i threshold is 0.6. Recall

Table 5. Comparison of prediction errors of PosQA, ProQres, and ProQprof.

	Whole			High-quality			Medium-quality			Low-quality		
D_i	PQA	PQr	PQp	PQA	PQr	PQp	PQA	PQr	PQp	PQA	PQr	PQp
≥ 0	0.29	0.41	0.40	0.27	0.36	0.44	0.29	0.47	0.54	0.29	0.41	0.20
≥ 0.1	0.28	0.31	0.35	0.27	0.26	0.32	0.29	0.31	0.36	0.29	0.39	0.37
≥ 0.2	0.26	0.26	0.29	0.25	0.22	0.27	0.26	0.26	0.30	0.29	0.31	0.30
≥ 0.3	0.23	0.22	0.24	0.21	0.19	0.23	0.22	0.22	0.26	0.27	0.25	0.24
≥ 0.4	0.22	0.18	0.20	0.20	0.16	0.19	0.21	0.18	0.22	0.25	0.22	0.20
≥ 0.5	0.21	0.16	0.17	0.18	0.14	0.15	0.20	0.15	0.18	0.23	0.19	0.18
≥ 0.6	0.19	0.14	0.15	0.16	0.13	0.12	0.19	0.13	0.15	0.20	0.18	0.19
≥ 0.7	0.17	0.15	0.15	0.15	0.12	0.10	0.15	0.12	0.14	0.21	0.21	0.24
≥ 0.8	0.15	0.16	0.17	0.14	0.14	0.10	0.10	0.14	0.13	0.20	0.22	0.29
≥ 0.9	0.13	0.19	0.19	0.13	0.17	0.13	0.12	0.17	0.13	0.24	0.25	0.33

Column 1 lists different D_i thresholds. Columns 2–13 list the prediction errors of PosQA (denoted as PQA), ProQres (denoted as PQr), and ProQprof (denoted as PQp) on the whole set, "high-quality" alignments, "medium-quality" alignments, and "low-quality" alignments, respectively.

that a large D_i indicates a high-quality position. This means that PosQA predicts the well-aligned positions better than ProQres and ProQprof.

All the test alignments are also divided into three classes: "high-quality" alignments, "medium-quality" alignments, and "low-quality" alignments, based on their Z-scores (calculated by RAPTOR) at cutting points 0.33 and 0.66. Table 5 shows the prediction errors of PosQA, ProQres, and ProQprof on the three classes of alignments. It is clear that different sets have different prediction errors, which means Z-score is an informative factor for local quality. For all the three classes, the overall errors, which correspond to $D_i \geq 0$, and the errors on high-quality residues, which correspond to $D_i \geq 0.9$, of PosQA are better than those of ProQres and ProQprof. However, ProQres outperforms the other two methods on both "high-quality" and "medium-quality" alignments, whereas PosQA is the best method on "low-quality" alignments. This makes sense because ProQres and ProQprof are both trained on high-quality models and alignments, while PosQA is trained on the comprehensive set of CASP7 targets, which contains high-quality (HA) targets, template-based modeling (TBM) targets, as well as free modeling (FM) targets.

2.5.2. *Sensitivity and specificity*

Receiver Operating Characteristic (ROC) plots are used to evaluate the trade-off between the ability of PosQA, ProQres, and ProQprof to correctly identify positive cases and the number of negative cases that are incorrectly classified. Figure 2 shows the ROC curves for PosQA, ProQres, and ProQprof on the four cross-validation test sets. The discrimination threshold for differentiate positive cases and negative cases is set to 4 Å in this figure. PosQA clearly outperforms the other two methods on all the four test sets. Meanwhile, the ROC curves also show that the performance for a method on test sets 1 and 3 is higher than that on test sets 2 and 4, which reveals that test sets 1 and 3 are easier than test sets 2 and 4 in terms of single position quality assessment.

We further evaluated the performance of PosQA, ProQres, and ProQprof on "high-quality", "medium-quality", and "low-quality" alignment sets. As shown in Figs. 3(a)–3(c), ProQres outperforms PosQA and ProQprof on "high-quality" alignments, whereas PosQA is the best method on both "medium-quality" and "low-quality" alignments. It is noteworthy that PosQA performs significantly better than both ProQres and ProQprof on "low-quality" alignments. One may argue that the difference on the performance is the result of the settings of ROC discrimination thresholds. Thus, we drew the ROC curves of PosQA with different discrimination thresholds on test set 1 in Fig. 3(d). Since there is almost no difference between different curves when false positive rate is higher than 0.4, only the ROC curves with false positive rate lower than 0.4 are shown. Again, the difference is not obvious when different discrimination thresholds are used. Similar observations are found on

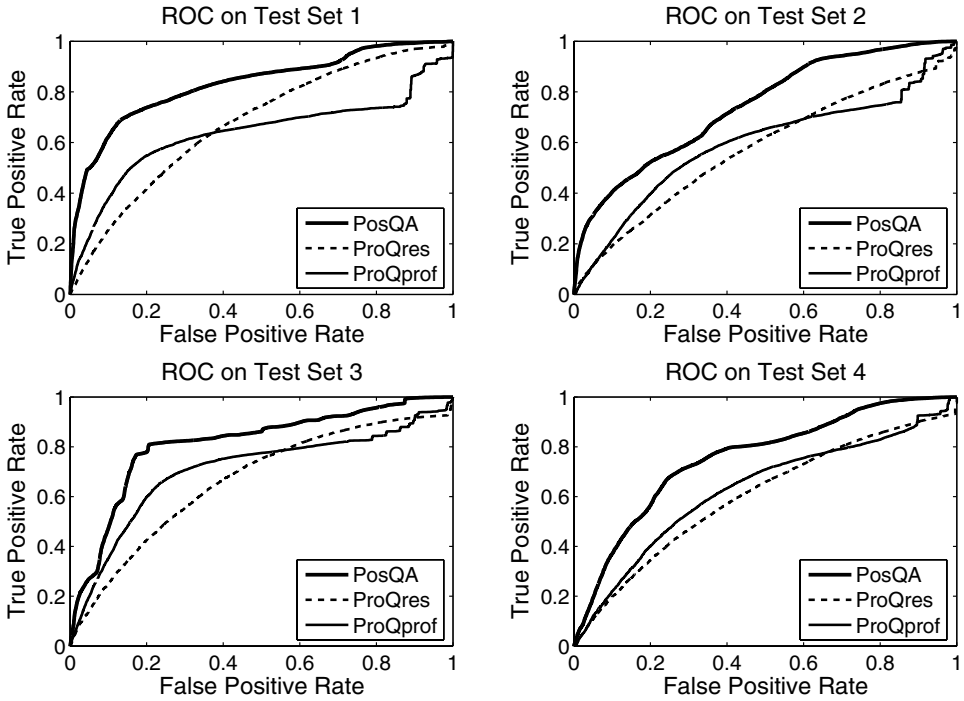


Fig. 2. ROC curves for PosQA, ProQres, and ProQprof on the four test sets. Discrimination threshold is 4 Å.

the other test sets and on the other two methods. Thus, all ROC curves shown here reveal the actual comparisons of the three methods regardless of the discrimination thresholds.

2.5.3. *Prediction examples of PosQA*

In this section, three representative alignments generated by RAPTOR in CASP7 are shown, and the performance of PosQA and ProQres on them is carefully studied. ProQres has been used for protein structure prediction by its developer, a top-ranked group in the CASP events.[36] These three alignments are T0346 (target) versus 1a33 (template), T0323 versus 1dizA, and T0372 versus 1sqhA; the structural models derived from these alignments have very different GDT_TS[54] scores 97.67, 53.69 and 24.75, respectively. For the sake of clearness, only the results of PosQA and ProQres are compared here, because ProQprof performs worse than ProQres on these three alignments. Since PosQA does not predict the quality of an unaligned position, to do a fair comparison between PosQA and ProQres, the average prediction errors for both PosQA and ProQres are calculated on only the aligned positions. As shown in Fig. 4, the prediction errors of both PosQA and ProQres are related to the overall alignment quality. The better the overall quality

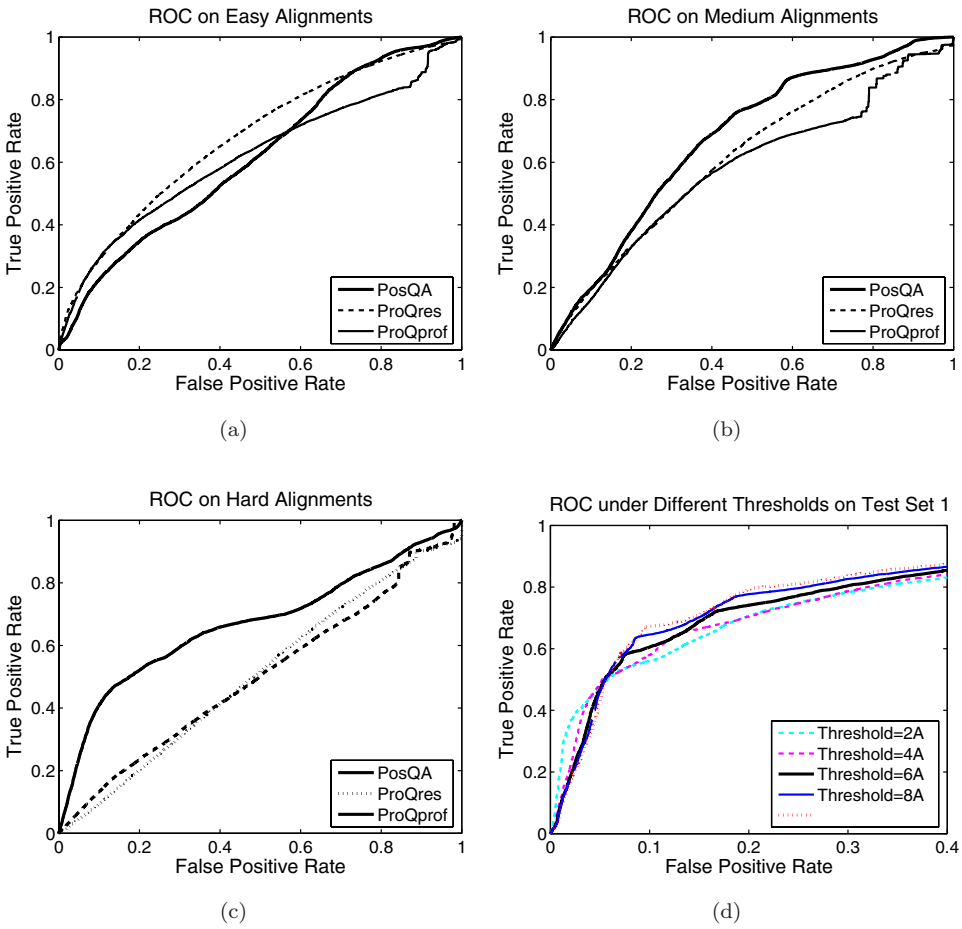
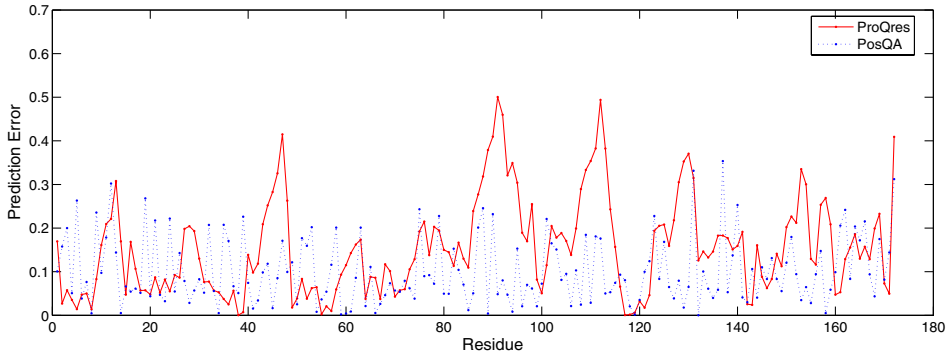
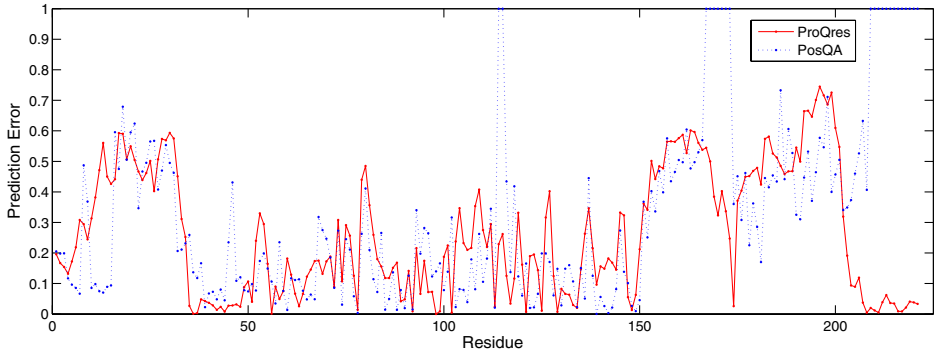


Fig. 3. (a) ROC curves for PosQA, ProQres, and ProQprof on "high-quality" alignments (Z-score ≤ 0.33). Discrimination threshold 2 Å; (b) ROC curves for PosQA, ProQres, and ProQprof on "medium-quality" alignments ($0.33 <$ Z-score ≤ 0.66). Discrimination threshold 4 Å; (c) ROC curves for PosQA, ProQres, and ProQprof on "low-quality" alignments ($0.66 <$ Z-score ≤ 1.0). Discrimination threshold 6 Å; (d) ROC curves for PosQA with different discrimination threshold values on test set 1.

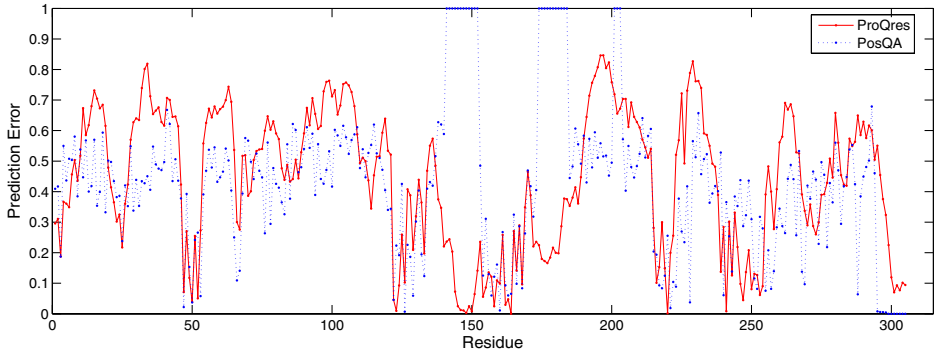
is, the smaller the prediction error is. PosQA performs better than ProQres on all these three test cases. The difference between the prediction errors of PosQA and ProQres is large on "high-quality" and "low-quality" alignments, i.e. T0346 versus 1a33 and T0372 versus 1sqhA, but relatively small on "medium-quality" alignment, T0323 versus 1dizA. The average prediction errors of PosQA and ProQres are 0.10 and 0.15 for T0346 versus 1a33, respectively, 0.24 and 0.27 for T0323 versus 1dizA, respectively, and 0.39 and 0.47 for T0372 versus 1sqhA, respectively. It is clear that for most residues of these alignments, the prediction errors of PosQA are smaller than that of ProQres. In particular, ProQres has obviously large prediction errors at



(a)



(b)



(c)

Fig. 4. Prediction errors of PosQA and ProQres on three typical alignments generated by RAPTOR in the CASP7 event. Since PosQA does not predict the quality at unaligned positions, the prediction errors at these positions are set to 1. (a) Prediction errors on T0346 versus 1a33 (GDT_TS score 97.67). The average errors of PosQA and ProQres are 0.10 and 0.15, respectively. (b) Prediction errors on T0323 versus 1dizA (GDT_TS score 53.69). The average errors of PosQA and ProQres are 0.24 and 0.27, respectively. (c) Prediction errors on T0372 versus 1sqhA (GDT_TS score 24.75). The average errors of PosQA and ProQres are 0.39 and 0.47, respectively.

some positions on the "high-quality" alignment between T0346 and 1a33, whereas PosQA's prediction errors are mostly contained within 0.3.

3. Discussion

FragQA and PosQA predict two different aspects of the local quality of an alignment. The cRMSD used in FragQA is calculated without considering other parts of an alignment, while the cRMSD used in PosQA depends on the overall alignment between the structural model and its native. To the best of our knowledge, FragQA is the first program that directly predicts the quality of an ungapped region in an alignment. A potential application of local quality predictors such as FragQA and PosQA is that they can be used to identify those high-quality regions in an alignment. These high-quality regions can often cover a large portion of the target protein even if it is a hard target and thus, they can be refolded to obtain a better structural model for the target protein. For example, Zhang-server[55,56] achieved an impressive performance in CASP7 and CASP8 by first cutting a threading-generated alignment into some ungapped regions, and then rearranging the physical orientations of these regions. Zhang-server uses all the ungapped regions without considering their quality. A further improvement over Zhang-server is to first predict the "absolute" quality of each region, and then refold only those high-quality regions to obtain a better structural model. FragQA provides such a powerful tool to directly evaluate the fragment quality cut from the alignments, which is independent of the optimal superimposition of the two whole structures. Currently, both FragQA and PosQA utilize only alignment information in a single alignment, although some structural information from the template is also taken into consideration. We plan to further develop these two programs along the following avenues: (1) combine structural information in a protein model with alignment information; and (2) utilize various alignments generated by independent threading programs so that consensus information can be used to boost the prediction performance. As demonstrated in recent CASP events, consensus information from independent prediction programs can help to improve prediction accuracy.

Although our experiments use alignments generated by RAPTOR as data source, both FragQA and PosQA can take alignments generated by other comparative modeling methods as inputs, since these two predictors are totally independent of threading methods. Thus, researchers can use these two programs to predict the local quality of an alignment generated by their own threading methods. On the other hand, as demonstrated by feature selection and the experiments, the local quality is also related to the overall quality of an alignment. We benchmarked our predictors using RAPTOR's results in CASP7, because most CASP7 target proteins have low sequence similarity with proteins in RAPTOR's template database. The template database used by RAPTOR for CASP7 was generated before any CASP7 target structures were deposited into the PDB database. This can reduce the bias introduced by template database to its minimum level. Moreover, as suggested in

Ref. 45, CASP dataset is the most comprehensive set, which is suitable to evaluate the broad range of the performance of our methods.

4. Methods

4.1. *Development of FragQA*

Our SVM regression model uses only features extracted from a single sequence-template alignment, generated by any comparative modeling program (i.e. homology modeling and threading). To exploit the evolutionary information of proteins, sequence profiles of both the target protein and the template protein are utilized in calculating features. The sequence profile of the template, denoted by $PSSM_{template}$ (position-specific scoring matrix), is generated by PSI-BLAST with five iterations; $PSSM_{template}(i, a)$ encodes mutation information for amino acid a at position i of the template. PSI-BLAST is also applied with five iterations to generate position-specific frequency matrix, $PSFM_{target}$, for each target protein; $PSFM_{target}(j, b)$ encodes occurring frequency of amino acid b at position j of the target. Let $A(i)$ denote the aligned sequence position of template position i, and T_{temp} denote the set of template positions belonging to an aligned region. We studied a variety of features extracted from the alignment, and their relative importance is studied in Sec. 2.3.2. In summary, the following features are tested in FragQA:

(1) **Mutation score.** Mutation score measures the sequence similarity between the two segments corresponding to an aligned region: one corresponds to the target protein and the other one corresponds to the template. The mutation score (S_m) of a region is calculated as:

$$S_m = \sum_{i \in T_{temp}} \sum_{a} PSFM_{target}(A(i), a) \times PSSM_{template}(i, a). \quad (2)$$

(2) **Environmental fitness score.** This score measures how well one target protein region aligns to the environment where the corresponding template region lies in. The environment consists of two types of local structure features.

   (a) Three types of secondary structure are used: α-helix, β-strand, and loop.
   (b) Solvent accessibility: There are three levels: buried (inaccessible), intermediate, and accessible. The Equal-Frequency discretization method is used to determine boundaries between these three levels. The calculated boundaries are 7% and 37%.

   Thus, there are nine environment combinations (denoted as env) in total. Define $F(env, a)$ to be the environment fitness potential for amino acid a and environment combination env. $F(env, a)$ is calculated and taken from PROSPECT-II.[49] For more details about $F(env, a)$, please see to Ref. 49. The environment fitness score (S_e) for an aligned region is then calculated as:

$$S_e = \sum_{i \in T_{temp}} \sum_{a} PSFM_{target}(A(i), a) \times F(env_i, a). \quad (3)$$

(3) **Secondary structure score.** In addition to the secondary structure information encoded in environmental fitness score, we also use $SS(i, A(i))$, the secondary structure difference between position i in template and position $A(i)$ in target, to measure the quality of an ungapped region from another aspect. PSIPRED[57] is called to predict the secondary structure of the target protein. Let $\alpha(j)$, $\beta(j)$, and $loop(j)$ denote the predicted confidence levels of α-helix, β-sheet, and loop at sequence position j, respectively. If the secondary structure type at template position i is α-helix, then $SS(i, A(i)) = \alpha(A(i)) - loop(A(i))$. If the secondary structure type at template position i is β-sheet, then $SS(i, A(i)) = \beta(A(i)) - loop(A(i))$. Otherwise, we set $SS(i, A(i))$ to be 0. The secondary structure score (S_{ss}) of an ungapped region is calculated as:

$$S_{ss} = \sum_{i \in T_{temp}} SS(i, A(i)). \tag{4}$$

(4) **Contact capacity score.** Contact capacity potentials describe the hydrophobic contribution of free energy, measured by the capability of a residue making a certain number of contacts with other residues in the protein. Two residues are in physical contact if the spatial distance between their C_β atoms (C_α for glycine) is smaller than 8 Å. Let $CC(a, k)$ denote the contact potential of amino acid a having k contacts. $CC(a, k)$ is calculated by statistics on PDB as:

$$CC(a, k) = -log \frac{N(a, k)N}{N(k)N'(a)}, \tag{5}$$

where $N(a, k)$ is the number of amino acid a with k contacts; $N(k)$ is the number of residues with k contacts; $N'(a)$ is the number of amino acid a; and N is the total number of residues in PDB. Let $C(i)$ denote the number of contacts at template position i. The contact capacity score (S_c) is calculated as:

$$S_c = \sum_{i \in T_{temp}} \sum_{a} PSFM_{target}(A(i), a) \times CC(a, C(i)). \tag{6}$$

(5) **Aligned region length.** The cRMSD between the two fragments of an ungapped region is relevant to its length. The longer the ungapped region is, the more likely larger the cRMSD is.

(6) **Z-score.** Z-score measures the overall quality of a sequence-structure alignment. An alignment with a good Z-score likely contains more good ungapped regions. In this paper, Z-score is the predicted alignment accuracy normalized by the target protein size, and calculated by Xu's SVM module.[28] Z-score ranges from 0 to 1: Z-score equals to 0 means the alignment is likely random, while 1 means it is probably a perfect alignment.

(7) **Sequence identity.** The fraction of identical residues in the whole alignment is used to measure the sequence identity.

(8) **Other sequential features.** Three other features are tested: template protein size, target protein size, and alignment length (i.e. the number of aligned positions).

Meanwhile, mutation score, environmental fitness score, secondary structure score, contact capacity score, and aligned region length are specific to the ungapped region; while Z-score, sequence identity, and other sequential features are for the whole sequence-structure alignment.

4.2. *Development of PosQA*

Instead of directly using cRMSD between the native C_α position and the predicted position of a residue, a normalized cRMSD is used as the objective function of PosQA. Let D_i and d_i denote the normalized cRMSD and cRMSD at position i, respectively. Then D_i is defined as $1/(1+\frac{(d_i)^2}{(d_0)^2})$ where d_0 is set to $\sqrt{5}$ according to Ref. 36. Thus, the larger the D_i is, the higher the quality of this position is.

PosQA uses almost the same set of features as FragQA. In particular, PosQA tests the following information: (1) mutation score, (2) environmental fitness score, (3) secondary structure score, (4) contact capacity score, and (5) Z-score. The only difference between PosQA and FragQA is that the values of the first four features are calculated at a single position.

5. Conclusions

This research develops two local quality predictors: FragQA and PosQA, which can be used to evaluate the local quality of a given sequence-template alignment from two different aspects: FragQA directly predicts the "absolute" quality of ungapped aligned regions, while PosQA predicts the quality for single aligned positions. Experimental results on the CASP7 dataset demonstrate that both FragQA and PosQA can predict the local quality well, especially when the local quality is good. Meanwhile, we conclude that local sequence evolutionary information is the major factor in predicting local quality. Other information such as secondary structure and solvent accessibility also helps to improve the prediction accuracy.

Acknowledgment

We thank Björn Wallner and Arne Elofsson for providing us ProQres and ProQprof programs, and Dongbo Bu, Xuefeng Cui, and William Wong for their thought-provoking discussions. We are grateful to Gloria Rose for proofreading the manuscript. This work is supported by the NSERC grant OGP0046506, the Canada Research Chair Program, an NSERC collaborative grant, CFI, MITACS, and an 863 Grant from the Ministry of Science and Technology of China.

References

1. Moult J, Hubbard T, Fidelis K, Pedersen J, Critical assessment of methods of protein structure prediction (CASP): Round III, *Proteins* **37**:2–6, 1999.
2. Moult J, Fidelis K, Zemla A, Hubbard T, Critical assessment of methods of protein structure prediction (CASP): Round IV, *Proteins* **45**:2–7, 2001.

3. Moult J, Fidelis K, Zemla A, Hubbard T, Critical assessment of methods of protein structure prediction (CASP): Round V, *Proteins* **53**:334–339, 2003.
4. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A, Critical assessment of methods of protein structure prediction (CASP): Round VI, *Proteins* **61**:3–7, 2005.
5. Zhang Y, Skolnick J, Automated structure prediction of weakly homologous proteins on a genomic scale, *Proc Natl Acad Sci USA* **101**(20):7594–7599, 2004.
6. Laskowski RA, Macarthur MW, Moss DS, Thornton JM, PROCHECK: A program to check the stereochemical quality of protein structures, *J Appl Crystallogr* **26**:283–291, 1993.
7. Abagyan RA, Totrov MM, Contact area difference (CAD): A robust measure to evaluate accuracy of protein models, *J Mol Biol* **268**:678–685, 1997.
8. Park BH, Huang ES, Levitt M, Factors affecting the ability of energy functions to discriminate correct from incorrect folds, *J Mol Biol* **266**:831–846, 1997.
9. Melo F, Feytmans E, Assessing protein structures with a non-local atomic interaction energy, *J Mol Biol* **277**:1141–1152, 1998.
10. Samudrala R, Moult J, An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, *J Mol Biol* **275**:895–916, 1998.
11. Lazaridis T, Karplus M, Discrimination of the native from misfolded protein models with an energy function including implicit solvation, *J Mol Biol* **288**:477–487, 1999.
12. Petrey D, Honig B, Free energy determinats of tertiary structure and the evaluation of protein models, *Protein Sci* **9**:2181–2191, 2000.
13. Siew N, Elofsson A, Rychlewski L, Fischer D, MaxSub: An automated measure for the assessment of protein structure prediction quality, *Bioinformatics* **16**(9):776–785, 2000.
14. Lu H, Skolnick J, A distance-dependent atomic knowledge-based potential for improved protein structure selection, *Proteins* **44**:223–232, 2001.
15. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A, Reliability of assessment of protein structure prediction methods, *Structure* **10**:435–440, 2002.
16. Feig M, Brooks CL, Evaluating CASP4 predictions with physical energy functions, *Proteins* **49**:232–245, 2002.
17. Felts AK, Gallicchio E, Wallqvist A, Levy RM, Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized born solvent model, *Proteins* **48**:404–422, 2002.
18. Zhou HY, Zhou YQ, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Sci* **11**:2714–2726, 2002.
19. Ginalski K, Elofsson A, Fischer D, Rychlewski L, 3D-Jury: A simple approach to improve protein structure predictions, *Bioinformatics* **19**:1015–1018, 2003.
20. McConkey B, Sobolev V, Edelman M, Discrimination of native protein structures using atom-atom contact scoring, *Proc Natl Acad Sci USA* **100**(9):3215–3220, 2003.
21. Wallner B, Elofsson A, Can correct protein models be identified? *Protein Sci* **12**(5):1073–1086, 2003.
22. Berglund A, Head RD, Welsh EA, Marshall GR, ProVal: A protein-scroing function for the selection of native and near-native folds, *Proteins* **54**:289–302, 2004.
23. Lee MC, Duan Y, Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model, *Proteins* **55**:620–634, 2004.
24. Buchete NV, Straub JE, Thirumalai D, Orientational potentials extracted from protein structures inprove native fold recognition, *Protein Sci* **13**:862–874, 2004.

25. Zhang Y, Skolnick J, Scoring function for automated assessment of protein structure template quality, *Proteins* **57**:702–710, 2004.
26. Zhou HY, Zhou YQ, Single body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, *Proteins* **55**:1005–1013, 2004.
27. Zhou HY, Zhou YQ, Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments, *Proteins* **58**:321–328, 2005.
28. Xu J, Protein fold recognition by predicted alignment accuracy, *ACM/IEEE Trans Comput Biol Bioinform* **2**(2):157–165, 2005.
29. Fang QJ, Shortle D, A consistent set of statistical potentials for quantifying local side-chain and backbone interactions, *Proteins* **60**:90–96, 2005.
30. Summa CM, Levitt M, DeGrado WF, An atomic environment potential for use in protein structure prediction, *J Mol Biol* **352**:986–1001, 2005.
31. Tosatto SC, The victor/FRST function for model quality estimation, *J Comput Biol* **12**:1316–1327, 2005.
32. Wallner B, Elofsson A, Prediction of global and local model quality in CASP7 using Pcons and ProQ, *Proteins* **69**(Suppl 8):184–193, 2007.
33. Sadowski MI, Jones DT, Benchmarking template selection and model quality assessment for high-resolution comparative modeling, *Proteins* **69**:476–485, 2007.
34. McGuffin LJ, The ModFOLD server for the quality assessment of protein structural models, *Bioinformatics* **24**:586–587, 2008.
35. Wang Z, Tegge AN, Cheng J, Evaluating the absolute quality of a single protein model using support vector machines and structural features, *Proteins*, in press.
36. Wallner B, Elofsson A, Identification of correct regions in protein models using structural, alignment, and consensus information, *Protein Sci* **15**:900–913, 2005.
37. Fischer D, 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor, *Proteins* **51**:434–441, 2003.
38. Tress M, Jones DT, Valencia A, Predicting reliable regions in protein alignments from sequence profiles, *J Mol Biol* **330**:705–718, 2003.
39. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N, ConSeq: The identification of functionally and structurally important residues in protein sequences, *Bioinformatics* **20**(8):1322–1324, 2004.
40. Luthy R, Bowie JU, Eisenberg D, Assessment of protein models with 3-dimensional profiles, *Nature* **356**:83–85, 1992.
41. Sippl M, Recognition of errors in three-dimensional structures of proteins, *Proteins* **17**:355–362, 1993.
42. Colovos C, Yeates TO, Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Sci* **2**:1511–1519, 1993.
43. Eisenberg D, Lüthy R, Bowie JU, VERIFY3D: Assessment of protein models with three-dimensional profiles, *Meth Enzymol* **277**:396–404, 1997.
44. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM, MetaMQAP: A meta-server for the quality assessment of protein models, *BMC Bioinformatics* **9**:403, 2008.
45. Fasnacht M, Zhu J, Honig B, Local quality assessment in homology models using statistical potentials and support vector machines, *Protein Sci* **16**:1557–1568, 2007.
46. Gao X, Bu D, Li SC, Xu J, Li M, FragQA: Predicting local fragment quality of a sequence-structure alignment, *Genome Informatics* **19**:27–39, 2007.
47. Rangwala H, Karypis G, fRMSDPred: Predicting local RMSD between structural fragments using sequence information, *Proteins* **72**:1005–1018, 2007.

48. Xu J, Li M, Kim D, Xu Y, RAPTOR: Optimal protein threading by linear programming, *J Bioinform Comput Biol* **1**(1):95–117, 2003.
49. Kim D, Xu D, Guo J, Ellrott K, Xu Y, PROSPECT II: Protein structure prediction method for genome-scale applications, *Protein Eng* **16**(9):641–650, 2003.
50. Jones DT, GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences, *J Mol Biol* **287**:797–815, 1999.
51. Li W, Godzik A, CD-HIT: A fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics* **22**:1658–1659, 2006.
52. Joachims T, Making large-scale support vector machine learning practical, in Smola A, Schölkopf B, Burges C (eds.), *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA, 1998.
53. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The protein data bank, *Nucleic Acids Res* **28**:235–242, 2000.
54. Zemla A, Venclovas C, Moult J, Fidelis K, Processing and evaluation of predictions in CASP4, *Proteins* **45**:13–21, 2001.
55. Zhang Y, Template-based modeling and free modeling by I-TASSER in CASP7, *Proteins* **8**:108–117, 2007.
56. Zhang Y, I-TASSER server for protein 3D structure prediction, *BMC Bioinformatics* **9**:40, 2008.
57. Jones DT, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol* **292**(2):195–202, 1999.

Xin Gao received his B.S. degree from the Computer Science and Technology Department at Tsinghua University, China, in 2004. He then applied to David R. Cheriton School of Computer Science at the University of Waterloo, Canada, where he began working on his doctoral thesis in the area of bioinformatics and algorithms. His doctoral work mainly focuses on fully automated NMR protein structure determination and protein structure prediction. His research interests include computational methods and machine learning techniques in structural biology, sequence analysis, and systems biology.

Jinbo Xu is an Assistant Professor at the Toyota Technological Institute at Chicago. He received his B.S. in Computer Science from the University of Science and Technology of China in 1996, his M.Sc. from Chinese Academy of Sciences in 1999, and Ph.D. from the University of Waterloo, Canada, in 2003. He was also a postdoctoral fellow in the Department of Mathematics and Computer Science and AI Laboratory at MIT. Dr. Xu's primary research interest is computational biology and bioinformatics including protein structure prediction, biological network analysis, and biological sequence analysis. He has developed several protein structure prediction tools, such as RAPTOR, ACE, and TreePack.

Shuai Cheng Li received his B.S. (Hons) and his M.S. from National University of Singapore in 2001 and 2002, respectively. Between 2002 and 2004, he was a full-time research associate in the database research group at National University of Singapore. Since 2004, he is a doctoral student at University of Waterloo. His current research interests include bioinformatics and algorithms. He now focuses on protein structure-related problems.

Ming Li is the Canada Research Chair in Bioinformatics and a Professor at the University of Waterloo. He is a fellow of Royal Society of Canada, ACM, and IEEE. He was a recipient of Canada's E.W.R. Steacie Fellowship Award in 1996, and the 2001 Killam Fellowship. Together with Paul Vitanyi, they have pioneered the applications of Kolmogorov complexity and co-authored the book "*An Introduction to Kolmogorov Complexity and Its Applications*. His research interests recently include protein-structure determination and next-generation internet search engines.