FragQA: predicting local fragment quality of a sequence-structure alignment

Xin Gao[1] Dongbo Bu[1,3]
x4gao@cs.uwaterloo.ca dbu@cs.uwaterloo.ca

Shuai Cheng Li[1] Jinbo Xu[2] Ming Li[1]
scli@cs.uwaterloo.ca j3xu@tti-c.org * mli@cs.uwaterloo.ca

[1] *David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada, N2L 3G1*
[2] *Toyota Technological Institute at Chicago, Chicago, IL, USA, 60637*
[3] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100080*

Motivation.

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 structure biology community can benefit from automated structure prediction. It is also important to identify badly-predicted regions in a model so that some refinement measurements can be applied to.

Results.

We present a novel technique FragQA to accurately predict local quality of a sequence-structure (i.e., sequence-template) alignment generated by comparative modeling (i.e., homology modeling and threading). Different from previous local quality assessment methods, FragQA directly predicts cRMSD between a continuously aligned fragment determined by an alignment and the corresponding fragment in the native structure. FragQA uses an SVM (Support Vector Machines) regression method to perform prediction using information extracted from a single given alignment. Experimental results demonstrate that FragQA performs well on predicting local quality. More specifically, FragQA has prediction accuracy better than a top performer ProQres [18]. 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) structure information such as solvent accessibility and secondary structure helps improving prediction performance.

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

1. Introduction

The biennial CASP (Critical Assessment of Structure Prediction) [12–15] events have demonstrated that the three-dimensional structures of many new target pro-

[*] To whom correspondence should be addressed.

teins 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 FM (free modeling) targets, a structural model generated by protein threading usually contains some good local regions, although the overall conformation of the model is incorrect [21].

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 (as referred to global quality assessment) as well as correctly-predicted residues from badly-predicted ones (as referred to local quality assessment). To make automated structure prediction really useful for the structure biology community, a reliable model quality evaluation program is indispensable when hundreds of models are predicted for a single target protein. There are a variety of global quality prediction methods [3, 5, 10, 17, 19]. This kind of programs can be used to pick up the best few from a bunch of models generated by different structure prediction programs, which enables structure biologists to focus on the most possible models. In addition, a common practice taken by some human predictors or consensus-based automatic predictors to further improve the accuracy of 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, 3D-SHOTGUN [4] and TASSER [21] 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 [1, 16].

Local quality assessment methods are either structure-based or alignment-based. ERRAT [2] is a program that uses only structure information. This program employs a Gaussian error function based on the statistics of non-bonded interactions to predict incorrect regions in a protein model. These 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 [16]. The results indicate that an alignment position with high profile-derived alignment score often has good quality. Wallner *et al* developed four neural network-based methods [18] to identify correct regions in a protein model, using either structure information or alignment information: ProQres, ProQprof, ProQlocal and Pcons-local. ProQres uses structure 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.

Our contribution. In this paper, we present a novel method FragQA to accurately predict local quality of a sequence-structure alignment. Distinguishing itself from its peers, FragQA predicts the quality of an ungapped region (referred to as fragment) 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 is the predicted structure. Furthermore, statistical significance is introduced to improve FragQA's performance. As opposed to cRMSD, statistical significance can cancel out the impact of region length. FragQA utilizes only information in a single alignment. Structure information in the alignment-derived protein model is not directly used. However, in calculating features from an alignment, we use structure information in the template.

2. Methods

2.1. *Problem description*

This paper studies the following problem: 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 conformations of the ungapped region, denoted as "cRMSD of an ungapped region", after they are optimally superimposed. Please note that the two conformations are superimposed without taking into consideration other parts of the alignment. The reason to do this local superimposition is to eliminate the impact by some badly predicted regions of the model, and evaluate how truly similar a region in a model is to the native one. The alignment is cut into ungapped regions at gap positions.

2.2. *Development of FragQA*

Our SVM regression model uses only features extracted from a single sequence-template alignment, generated by any threading program. To exploit the evolutionary information of proteins, we utilize sequence profile of both target protein and template protein in calculating features. The sequence profile of the template, denoted by $PSSM_{template}$ (position specific mutation 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. We also apply PSI-BLAST 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 later we will discuss their relative importance. In summary, we tested the following features in FragQA:

(1) *Mutation score*: Mutation score measures the sequence similarity between two segments of an aligned region: one corresponds to the target protein and the

other to the template. The mutation score (S_m) of a region is calculated as:

$$S_m = \sum_{i \in T_{temp}} \sum_a PSFM_{target}(A(i), a) \times PSSM_{template}(i, a) \tag{1}$$

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

- Three types of secondary structure are used: α-helix, β-strand, and loop.
- 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. Let $F(env, a)$ denote the environment fitness potential for amino acid a and environment combination *env*, which is taken from PROSPECT-II [9]. The environment fitness score (S_e) for an aligned region is calculated as:

$$S_e = \sum_{i \in T_{temp}} \sum_a PSFM_{target}(A(i), a) \times F(env_i, a) \tag{2}$$

(3) *Secondary structure score*: In addition to 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. We use PSIPRED [7] 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:

$$S_{ss} = \sum_{i \in T_{temp}} SS(i, A(i)) \tag{3}$$

(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 8Å. 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:

$$CC(a, k) = -log \frac{N(a, k)N}{N(k)N'(a)} \tag{4}$$

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:

$$S_c = \sum_{i \in T_{temp}} \sum_a PSFM_{target}(A(i), a) \times CC(a, C(i)) \tag{5}$$

(5) *Aligned region length*: The cRMSD between 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 predicted alignment accuracy normalized by target protein size, and calculated by Xu's SVM module [19].

(7) *Alignment topology*: We test 3 separate topology features: template protein size, target protein size, alignment length (i.e., the number of aligned positions).

(8) *Sequence identity*: We use the fraction of identical residues in the whole alignment to measure the sequence identity.

Meanwhile, feature (1)-(5) are specific to the ungapped region; while feature (6)-(8) are for the whole sequence-structure alignment.

3. Results

3.1. *FragQA Training*

Training and Test Data. Choosing good training and test sets is one of the key steps in objectively evaluating the performance of a machine learning method. We test our method on several threading methods, such as RAPTOR [20] (with three different threading algorithms), PROSPECT-II [9], and GenTHREADER [8]. The results are similar. In this paper, we only show the results on alignments generated by RAPTOR default threading algorithm (with NoCore option). Our training and test data is from recent CASP7 event. There are 104 target proteins in CASP7 while only 92 of them have native structures published after the event. Ninety-one target proteins are left after we removed redundancy at 40% sequence identity level using CD-HIT [11]. 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. Here, we took top 10 alignments generated by RAPTOR 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 5 residues are not considered in our experiments. Table 1 shows the statistics on the four sets. It is clear that the four data sets are very similar.

Training. We used the software SVM-light [6] with RBF (radial basis function) kernel 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

Table 1. Statistics on the four data sets. Column 2-5 show the number of target proteins, the number of fragments, the average cRMSD of the fragments, and the standard deviation of cRMSD of each set, respectively.

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

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.

We executed a 4-fold cross validation. Each time we used three of the four data sets as the training set, and the other one for testing.

3.2. *Performance of FragQA*

After studying the relative importance of the 8 features, which will be discussed later, we encoded following features 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.

3.2.1. *Comparing to ProQres*

As far as we know, FragQA is the first method to directly predict the local fragment quality. Thus, there is no existing method for us to compare with. However, there are some well-known methods that predict local quality for each residue. So it is possible to convert the prediction on residues by such methods to a prediction of a fragment. Since the objective function of FragQA is cRMSD, to fairly evaluate FragQA, we compared FragQA to a top-notch method ProQres [18], which uses a residue-based cRMSD-related objective function. We tested all three available methods by ProQ-group in terms of the ability to predict fragment quality : ProQlocal, ProQres, and ProQprof. ProQres yielded the best results (slightly better than ProQlocal and ProQprof in terms of fragment cRMSD prediction). Thus, in this paper, we will compare FragQA to ProQres. The objective function of ProQres is $D_i = 1/(1+\frac{(d_i)^2}{(d_0)^2})$ [18], where d_i denotes the cRMSD at position i, and d_0 is set to $\sqrt{5}$. From the prediction of ProQres, we can calculate d_i from D_i for each residue of a fragment, then use $cRMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {d_i}^2}$ to compute the predicted cRMSD by ProQres for the fragment, where n is the length of the fragment. Note cRMSD calculated by this way has a slightly different meaning to the one used by FragQA, because this cRMSD is based on the optimal superposition between the whole target and the template on all similar regions, while FragQA's cRMSD is based on the optimal superposition between two fixed regions. However, the superposition between two aligned regions determined by the optimal superposition of the whole target and template is usually very similar to the optimal one between the two regions, because aligned regions are usually very similar. Thus, FragQA and ProQres are comparable from this point of view.

3.2.2. *Prediction Error and Correlation Coefficient of FragQA*

The prediction error is defined as the difference between the predicted cRMSD values and the real ones. Table 2 lists the average prediction errors of FragQA and ProQres, under different cRMSD thresholds on the four test sets, together with average fraction of fragments with real cRMSD under such thresholds, and the correlation coefficient between the predicted and real cRMSD by FragQA and ProQres on the four test sets. As shown in this table, the prediction error of FragQA ranges from 0.9Å to 1.6Å, while the error of ProQres ranges from 0.9Å to 2.4Å. In most cases, the prediction error of FragQA is much smaller than that of ProQres. In fact, when there is no restriction on cRMSD, the error of FragQA is on average 0.5Å smaller than that of ProQres. The smallest error of FragQA happens when cRMSD threshold is set to 3Å, which means FragQA is most accurate when dealing with fragments with cRMSD to native smaller than 3Å. However, when the real cRMSD is very small ($\leq 1\AA$), the prediction error tends to be big. In other word, it is hard to obtain an accurate prediction when cRMSD is very small. As indicated in Table 2, the correlation coefficient between predicted cRMSD by FragQA and the real cRMSD is about 0.5 for each test set, while that of ProQres is at most 0.22.

Table 2. The prediction error of FragQA (denoted as FQA) and ProQres (denoted as PQr), under different cRMSD thresholds on the four test sets, average fraction of fragments with real cRMSD under such thresholds, and the correlation coefficient of FragQA and ProQres.

cRMSD	Test Set 1		Test Set 2		Test Set 3		Test Set 4		Ave. Fraction
	FQA	PQr	FQA	PQr	FQA	PQr	FQA	PQr	
≤1Å	1.36	1.50	1.57	1.10	1.41	1.35	1.54	1.30	14%
≤2Å	1.11	1.06	1.28	0.90	1.08	1.01	1.18	1.01	42%
≤3Å	1.00	1.84	1.16	1.12	0.94	0.98	1.04	1.01	69%
≤4Å	1.03	1.79	1.12	1.23	0.97	1.21	1.04	1.13	85%
≤5Å	1.12	1.88	1.14	1.34	1.06	1.37	1.09	1.34	92%
≤6Å	1.20	1.98	1.19	1.46	1.16	1.50	1.20	1.51	95%
≤7Å	1.33	2.13	1.26	1.57	1.22	1.58	1.25	1.62	97%
≤8Å	1.41	2.20	1.32	1.68	1.29	1.67	1.31	1.72	98%
≤9Å	1.48	2.27	1.36	1.73	1.37	1.78	1.36	1.77	99%
≤10Å	1.57	2.37	1.39	1.77	1.41	1.84	1.41	1.83	99%
Correlation Coefficient	0.51	0.07	0.46	0.22	0.50	0.22	0.48	0.16	-

3.2.3. *Sensitivity and Specificity*

Given a cRMSD threshold, sensitivity is calculated as the fraction of ungapped regions with real cRMSD smaller than the threshold, that are also predicted to be smaller than the threshold. Specificity measures the fraction of ungapped regions with predicted cRMSD under a given threshold, that indeed have cRMSD smaller than the threshold. Figure 1 illustrates the sensitivity and specificity of FragQA and ProQres under various cRMSD thresholds on the four test sets.

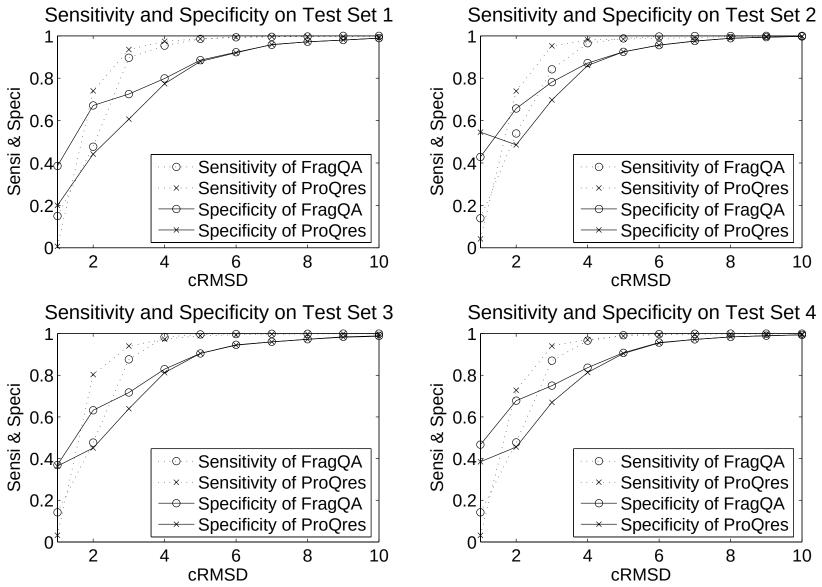


Fig. 1. Comparison of sensitivity and specificity between FragQA and ProQres under different cRMSD thresholds on the four test sets. Circle dotted line: sensitivity of FragQA, cross dotted line: sensitivity of ProQres, circle solid line: specificity of FragQA, cross solid line: specificity of ProQres. Please see Section 3.2.3 for the definition of sensitivity and specificity.

As shown in Figure 1, there is no obvious difference between sensitivity or specificity of FragQA and ProQres when cRMSD is larger than 4Å. When cRMSD is smaller than 4Å, the sensitivity of ProQres is higher than that of FragQA for most cases, while the specificity of FragQA is higher than that of ProQres. In particular, when cRMSD threshold is 2.5Å, approximately 70% of ungapped regions with predicted cRMSD by FragQA under 2.5Å indeed have cRMSD less than 2.5Å, while 70% of ungapped regions with real cRMSD under this threshold are predicted by FragQA correctly. However, the sensitivity of ProQres on cRMSD 2.5Å is about 80%, while the specificity is only 50%. This implies that ProQres has a strong trend to predict the cRMSD of a fragment to be smaller than the real value. This makes ProQres to have a high sensitivity while having a low specificity. As shown in Figure 1, the specificity curve of ProQres is not smooth sometime, nor it is monotonous. It is clear that for both FragQA and ProQres, the sensitivity curve increases much more quickly than the specificity curve. However, all curves are quite low when cRMSD is small. A possible explanation is that when the ungapped region is short, a small cRMSD does not necessarily mean that this region has good quality. Therefore, it is hard for FragQA to predict cRMSD accurately under such cases. Later in this paper, we will replace cRMSD with its statistical significance and show that when statistical significance is high, even as high as 1, FragQA can still yield a good prediction.

3.2.4. *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. We studied the importance of each feature by excluding it from the feature set, training a new FragQA, and then testing the performance of this new predictor. Thus, we can compare the performance resulting from different sets of features and then detect the important features.

Table 3. Sensitivity of FragQA with different feature sets. The 2^{nd} column lists the sensitivity of FragQA with all features. Starting from the 3^{rd} column, each column lists the sensitivity 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, *Topo*: topology features, *SeqId*: sequence identity.

cRMSD	All	No *Len*	No S_z	No S_m	No S_e	No S_c	No S_{ss}	No *Topo*	No *SeqId*
$\leq$1Å	0.12	0	0.04	0.09	0.11	0.13	0.13	0.12	0.12
$\leq$1.25Å	0.16	0.01	0.08	0.15	0.14	0.22	0.18	0.15	0.16
$\leq$1.5Å	0.25	0.04	0.16	0.19	0.22	0.27	0.26	0.25	0.25
$\leq$1.75Å	0.35	0.12	0.27	0.27	0.29	0.34	0.36	0.35	0.34
$\leq$2Å	0.42	0.21	0.38	0.35	0.39	0.48	0.42	0.42	0.43
$\leq$2.25Å	0.50	0.42	0.52	0.46	0.48	0.58	0.51	0.51	0.51
$\leq$2.5Å	0.62	0.61	0.64	0.55	0.56	0.65	0.63	0.62	0.62
$\leq$2.75Å	0.70	0.74	0.73	0.65	0.67	0.74	0.71	0.69	0.69
$\leq$3Å	0.76	0.82	0.79	0.74	0.75	0.81	0.77	0.76	0.76
$\leq$3.25Å	0.83	0.90	0.86	0.82	0.80	0.85	0.84	0.83	0.83
$\leq$3.5Å	0.88	0.94	0.90	0.88	0.84	0.89	0.89	0.88	0.88

Table 4. Specificity of FragQA with different feature sets. The 2^{nd} column lists the specificity of FragQA with all features. Starting from the 3^{rd} column, each column lists the 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, *Topo*: topology features, *SeqId*: sequence identity.

cRMSD	All	No *Len*	No S_z	No S_m	No S_e	No S_c	No S_{ss}	No *Topo*	No *SeqId*
$\leq$1Å	0.19	0	0.10	0.17	0.16	0.32	0.17	0.18	0.18
$\leq$1.25Å	0.28	0.22	0.20	0.27	0.22	0.43	0.27	0.28	0.28
$\leq$1.5Å	0.42	0.23	0.37	0.35	0.36	0.49	0.41	0.41	0.42
$\leq$1.75Å	0.52	0.41	0.51	0.46	0.47	0.57	0.51	0.52	0.52
$\leq$2Å	0.59	0.48	0.58	0.53	0.57	0.65	0.56	0.59	0.60
$\leq$2.25Å	0.64	0.56	0.64	0.60	0.62	0.68	0.63	0.64	0.64
$\leq$2.5Å	0.72	0.63	0.70	0.66	0.69	0.73	0.70	0.72	0.72
$\leq$2.75Å	0.78	0.67	0.75	0.73	0.76	0.78	0.77	0.77	0.78
$\leq$3Å	0.79	0.70	0.77	0.77	0.80	0.79	0.79	0.79	0.79
$\leq$3.25Å	0.82	0.75	0.80	0.81	0.83	0.82	0.80	0.82	0.82
$\leq$3.5Å	0.86	0.79	0.84	0.83	0.85	0.86	0.84	0.86	0.86

Tables 3 and 4 list the sensitivity and specificity of FragQA with different sets of features under different cRMSD thresholds on test set 1. The results are similar on other test sets. There is no obvious difference among different sets of features when cRMSD threshold is larger than 3.75Å. As shown in these two tables, if we remove the aligned region length, the performance of FragQA will drop obviously, except for cRMSD threshold larger than 2.75Å, the sensitivity of FragQA without fragment length is a little higher than that with all 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 2.25Å, removing Z-score will increase sensitivity slightly and have no obvious influence on 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 specificity. Surprisingly, removing contact capacity score will increase both sensitivity and specificity. This implies contact score is a noise feature. On the other hand, removing secondary structure score will decrease the specificity but increase the sensitivity slightly. Removing any other features, such as alignment topology features and sequence identity feature, does not obviously deteriorate either sensitivity or 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.

3.2.5. *Statistical Significance*

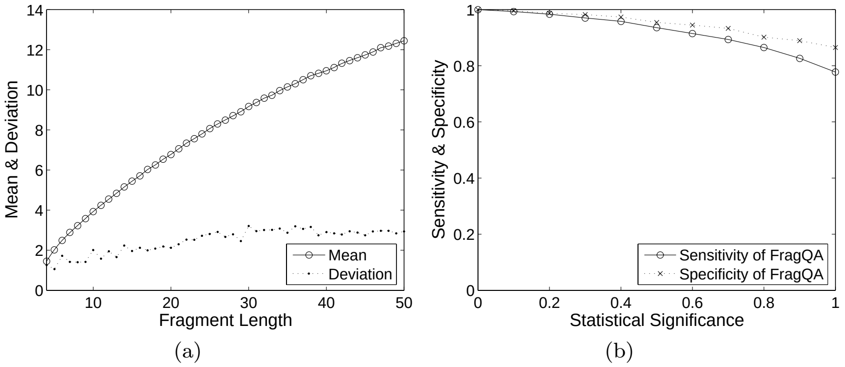
The cRMSD between the predicted structure of an ungapped region and its native is closely relevant to the length of the region. Thus, a 5-residue ungapped region with 3Å cRMSD may not be better than a 15-residue region with 4Å cRMSD. To better evaluate the quality of a region, we calculate the statistical significance of its cRMSD 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, we randomly sampled 10,000 pairs of fragments of this length from PDB30 and then calculated their cRMSDs. PDB30 is a subset of PDB, in which any two proteins share no more than 30% sequence identity. As shown in Figure 2(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 not shown). For a given ungapped region with length l and (real or predicted) cRMSD r, its statistical significance (denoted as $StatSig$) is calculated as follows:

$$StatSig = \frac{\#random\ pairs\ of\ length\ l\ with\ cRMSD \geq r}{10,000} \tag{6}$$

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

We calculated the sensitivity and specificity of FragQA in terms of statistical significance in a way similar to that calculates them 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 or equal than the threshold, that also have predicted values larger or equal than the threshold. The specificity is defined as the percentage of ungapped regions with predicted significance larger or equal than the threshold, that have real statistical significance better or equal than the threshold. Figure 2(b) illustrates the sensitivity and specificity of FragQA in terms of statistical significance on test set 1. Results are similar on other three sets. As shown in this figure, when statistical significance is 0.8 (about 81% fragments in our test sets have such values), both the sensitivity and specificity is around 90%. Even when statistical significance threshold is 1 (about 48% fragments in our test sets have this value), the sensitivity is 78%, and the specificity is 88%.

We also studied the prediction error and correlation coefficient of FragQA in terms of statistical significance. The prediction error increases from 0.02 to 0.16 when the statistical significance threshold decreases from 1 to 0. Recall that it is hard for FragQA to predict cRMSD accurately when cRMSD is small, this result implies that FragQA is able to predict statistical significance well on high-quality fragments. On the other hand, the correlation coefficient of FragQA on each set in terms of statistical significance is higher than 0.60. This means statistical significance is probably a better way to measure quality of a fragment. We further compared FragQA to ProQres in terms of statistical significance. FragQA also outperforms



(a) (b)

Fig. 2. (a) Mean (circle solid line) and standard deviation (point dotted line) of cRMSD for random region sets with length from 5 residues to 50 residues. (b) FragQA's sensitivity (circle solid line) and specificity (cross dotted line) in terms of statistical significance on test set 1. Please see Section 3.2.5 for the definitions of sensitivity and specificity.

ProQres on prediction error, correlation coefficient, and specificity, but is not as good as ProQres on sensitivity (data not shown here).

4. Discussion

To the best of our knowledge, FragQA is the first program that directly predicts the quality of an ungapped region in an alignment. Currently FragQA utilizes only alignment information in a single alignment, although some structure information from the template is also taken into consideration. We plan to further improve the method along the following avenues: 1) Combine structure 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 prediction performance.

Although our experiments used alignments generated by RAPTOR as data source, FragQA can take as input alignments generated by any comparative modeling methods, since FragQA is totally independent of threading methods. We benchmarked FragQA using RAPTOR's results in CASP7 because CASP7 target protein structures were published only recently. Most CASP7 target proteins have low sequence similarity with proteins in PDB. The template database used by RAPTOR for CASP7 was generated before any CASP7 target structures were deposited into PDB. This can reduce bias introduced by template database to its minimum level.

A potential application of FragQA is to identify high-quality regions in an alignment. These regions can often cover a large portion of the target protein even if it is a hard target and thus, they can be re-assembled to obtain a better overall structural model. For example, Zhang-server [21] is the best automated server in CASP7. It first cuts a threading-generated alignment into some ungapped regions and then rearranges the physical orientations of these regions. Zhang-server uses all the ungapped regions without considering their quality. A further improvement over Zhang's method is to first predict the quality of each region and then refold only those high-quality regions to obtain a better structural model.

5. Conclusion

This research develops a novel local quality predictor, FragQA. Experimental results on the CASP7 data set demonstrate that FragQA performs well, better than a top-notch predictor ProQres. Our experimental results indicate that local sequence evolutionary information is the major factor in predicting local quality. Other information such as secondary structure and solvent accessibility also helps improving prediction accuracy.

Acknowledgment

This work is supported by NSERC Grant OGP0046506, and NSF of China Grant 60496324.

References

[1] Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., Casadio, R., and Ben-Tal, N., ConSeq: the identification of functionally and structurally important residues in protein sequences, *Bioinformatics*, 8:1322–1324, 2004.

[2] Colovos, C., and Yeates, T.O., Verification of protein structures: patterns of non-bonded atomic interactions, *Proteins Science* 2:1511-1519, 1993.

[3] Eisenberg, D., Lthy, R., and Bowie, J.U., VERIFY3D: assessment of protein models with three-dimensional profiles, *Methods Enzymol*, 277:396–404, 1997.

[4] Fischer, D., 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor, *Proteins: Structure, Function and Genetics*, 51:434–441, 2003.

[5] Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L., 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics*, 19:1015–1018, 2003.

[6] Joachims, T., *Making large-scale SVM learning practical. Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 1999.

[7] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology*, 292(2):195–202, 1999.

[8] Jones, D.T., GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences, *Journal of Molecular Biology*, 287:797–815, 1999.

[9] Kim, D., Xu, D., Guo, J., Ellrott, K., and Xu, Y., PROSPECT II: protein structure prediction method for genome-scale applications, *Protein Engineering* 16(9):641–650, 2003.

[10] Laskowski, R.A., Macarthur, M.W., Moss, D.S., and Thornton, J.M., PROCHECK: a program to check the stereochemical quality of protein structures, *Journal of Applied Crystallography*, 26(2):283–291, 1993.

[11] Li, W. and Godzik, A., CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, 22:1658–1659, 2006.

[12] Moult, J., Hubbard, T., Fidelis, K., and Pedersen, J., Critical assessment of methods of protein structure prediction (CASP):round III, *Proteins*, 37:2–6, 1999.

[13] Moult, J., Fidelis, K., Zemla, A., and Hubbard, T., Critical assessment of methods of protein structure prediction (CASP):round IV, *Proteins*, 45:2–7, 2001.

[14] Moult, J., Fidelis, K., Zemla, A., and Hubbard, T., Critical assessment of methods of protein structure prediction (CASP):round V, *Proteins*, 53:334–339, 2003.

[15] Moult, J., Fidelis, K., Rost, B., Hubbard, T. and Tramontano, A., Critical assessment of methods of protein structure prediction (CASP):round 6, *Proteins*, 61:3–7, 2005.

[16] Tress, M., Jones, D., and Valencia, A., Predicting reliable regions in protein alignments from sequence profiles, *Journal of Molecular Biology*, 330:705–718, 2003.

[17] Wallner, B., and Elofsson, A., Can correct protein models be identified? *Protein Science*, 12(5):1073–1086, 2003.

[18] Wallner, B., and Elofsson, A., Identification of correct regions in protein models using structural, alignment, and consensus information, *Protein Science*, 15:900–913, 2005.

[19] Xu, J., Protein fold recognition by predicted alignment accuracy, *ACM/IEEE Transactions on Computational Biology and Bioinformatics*, 2(2):157–165, 2005.

[20] Xu, J., Li, M., Kim, D., and Xu, Y., RAPTOR: optimal protein threading by linear programming, *JBCB*, 1(1):95–117, 2003.

[21] Zhang, Y., and Skolnick, J., Automated structure prediction of weakly homologous proteins on a genomic scale, *Proceedings of National Academy of Sciences*, 101(20):7594–7599, 2004.