TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 02/10 pp559-567
Volume 19, Number 6, December 2014

A New Hidden Markov Model for Protein Quality Assessment Using Compatibility Between Protein Sequence and Structure

Zhiquan He, Wenji Ma, Jingfen Zhang, and Dong Xu*

Abstract: Protein structure Quality Assessment (QA) is an essential component in protein structure prediction and analysis. The relationship between protein sequence and structure often serves as a basis for protein structure QA. In this work, we developed a new Hidden Markov Model (HMM) to assess the compatibility of protein sequence and structure for capturing their complex relationship. More specifically, the emission of the HMM consists of protein local structures in angular space, secondary structures, and sequence profiles. This model has two capabilities: (1) encoding local structure of each position by jointly considering sequence and structure information, and (2) assigning a global score to estimate the overall quality of a predicted structure, as well as local scores to assess the quality of specific regions of a structure, which provides useful guidance for targeted structure refinement. We compared the HMM model to state-of-art single structure quality assessment methods OPUSCA, DFIRE, GOAP, and RW in protein structure selection. Computational results showed our new score HMM.Z can achieve better overall selection performance on the benchmark datasets.

Key words: protein structure prediction; structure quality assessment; Hidden Markov Model (HMM)

1 Introduction

Proteins are large biological molecules consisting of one or more chains of amino acids performing a vast array of functions within living organisms. Proteins with various sequences of amino acids fold into different and unique three-dimensional (3-D) structures. The functions of proteins are determined by their structures. Massive amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. Protein structure prediction is an effective and efficient way to bridge the growing gap between the number of protein sequences and the number of experimental tertiary structures. Although numerous efforts for more than three decades have been made, the prediction of protein 3-D structures from their amino acid sequences still has a large room to improve[1,2].

Sequence-structure compatibility plays a critical role in protein structure prediction, such as fold recognition, threading alignment (or sequence-structure alignment), and protein structure Quality Assessment (QA).

To measure the sequence-structure compatibility, the structure environment of a protein residue is specified by a number of variables. Then a score is assigned for the observation of an amino acid type occurring in a structure environment. Some simple measures have been widely used in threading alignment methods[3-6]. For example, the secondary structure matching score is the match ratio between the predicted secondary structure from an amino acid sequence and the actual

• Zhiquan He, Jingfen Zhang, and Dong Xu are with both Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, MO 65211, USA. E-mail: zhy78@mizzou.edu; zhangjingf@gmail.com; xudong@missouri.edu.
• Wenji Ma is with both Christopher S. Bond Life Sciences Center, University of Missouri, MO 65211, USA and Department of Computer Science, City University of Hong Kong, Hong Kong, China. E-mail: wenjima821@gmail.com.
* To whom correspondence should be addressed.
Manuscript received: 2014-06-18; accepted: 2014-06-25

ones calculated from the template structure. Another one is the environmental fitness score, which measures the propensity of an amino acid type to appear in the structure environment specified by three types of secondary structures (Helix, Sheet, and Coil) or three types of solvent accessibilities (Buried, Intermediate, and Exposed). More advanced studies have been done to address this problem, which mainly differ in the definition of structure environment and the method to calculate the compatibility score (probability or pseudo-energy)[7-11]. For example, in Ref. [8], three-dimensional profiles were derived from native structures to measure the compatibility in which the structural environment was defined by parameters such as the area of the side chain that is buried and the secondary structure type; in Ref. [11], more complex structural environment was defined in which side chain packing and hydrogen bonding were used as one of its four measurement functions; a neural network was trained to predict the probability of observing an amino acid type given the structural environment[7].

It is commonly observed that proteins have recurrent local sequences and structure patterns. The sequence-structure dependency at local levels leads researchers to use the Hidden Markov Model (HMM) approach to describe the proteins. In Refs. [12, 13], an HMM was used to compress protein three-dimensional conformations into a one-dimensional series of letters of a structural alphabet, where the emission of the HMM at each state is a multi-dimensional Gaussian distribution for the distance configuration of four consecutive neighboring C_α atoms. In Ref. [14], a more complex HMM-based method, HMMSTR, was proposed to capture local sequence-structure correlations, in which four types of emissions were defined, i.e., amino acid types, secondary structure types, backbone angle region (i.e., using the (ϕ, ψ) Ramachandran plot to partition the protein chain into several non-overlapping regions), and structural context descriptor (for example, distinguishing a hairpin turn from a diverging turn). However, the structure information contained in this HMM is not informative enough as it only contains discretized backbone angle region types and secondary structures.

A number of knowledge-based scoring functions such as OPUSCA[15], DFIRE[16], GOAP[17], and RW[18] for protein structure quality assessment can also be considered as sequence-structure compatibility measures at global levels. Most of these scores are weighted sums of several energy terms obtained through statistics over native structures. For example, OPUSCA uses the distance distributions of residue pairs and DFIRE constructs residue-specific all-atom potential of mean force from a database of native structures. GOAP is the extension of DFIRE score.

More advanced descriptions of local structures are important for improving HMM's capability of capturing sequence-structure relationship. For this purposes, we defined new emission functions for the HMM to describe the local sequence-structure relationship. The structural emission contains information for every four consecutive C_α atoms, which is represented as three-dimensional Gaussian distributions in the angular space. Another important emission is about the sequence profile, which contains the distribution of 20 types of amino acids, and the insertion and deletion during the evolution process. The HMM model has two capabilities: (1) encoding local structure of each position by considering the local sequence-structure relationship, and (2) assigning a global score to estimate the overall quality of a predicted structure, as well as local scores to assess the quality of a specific structural segment.

Our new model was tested and compared with the state-of-art single structure QA methods. Test results demonstrated that our model can achieve better overall selection performance than the other QA methods that were compared.

2 Methods

Our goal is to construct a new Hidden Markov Model to encode the compatibility between protein sequence and structure, and capture their complex relationships. First, the emission of the HMM is defined based on protein local structures in the angular space, secondary structures, and sequence profile. Second, with a training data set, the proposed HMM was trained using the Expectation Maximization (EM) algorithm.

2.1 Sequence and structure representation

For each protein, we calculated the sequence profile matrix **PSFM**[] and **SEQ**[] from the output alignments of PSI-BLAST[19] running against the Non-Redundant (NR) sequence database (ftp://ftp.ncbi.nih.gov/blast/db/) three rounds with an *E*-value cutoff of 0.001. Each row of the matrix **PSFM**[] is a vector of 21 dimensions containing frequencies for 20 types of amino acids and indels (insertions and deletions) in

the Multiple Sequence Alignment (MSA), while **SEQ**[] only contains amino acids distribution information with each row being 20 dimensions.

The local structure of a protein is represented in the angular space according to the work of Ref. [20]. Specifically, for each residue x_k in a protein structure, we calculated an angle triplet $(\theta_k, \tau_k, \theta_{k+1})$ for four consecutive C_α atoms $(x_{k-2}, x_{k-1}, x_k, x_{k+1})$, where θ_k is the bend angle of (x_{k-2}, x_{k-1}, x_k), τ_k is the dihedral angle of $(x_{k-2}, x_{k-1}, x_k, x_{k+1})$, and θ_{k+1} is the bend angle of (x_{k-1}, x_k, x_{k+1}), as shown in Fig. 1. Let $x_k \equiv (\theta_k, \tau_k, \theta_{k+1})$, a protein of length L is represented by a list of x_k, where k goes from 3 to $L-1$. The probability distribution of an angle triplet x for the entire structure space was approximated by a Gaussian mixture model of 17 components[20], i.e.,

$$P(\boldsymbol{x}) = \sum_{i=1}^{17} \pi_i N_i(\boldsymbol{x}; \boldsymbol{u}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$N_i(\boldsymbol{x}; \boldsymbol{u}_i, \boldsymbol{\Sigma}_i) = (2\pi)^{-3/2} |\boldsymbol{\Sigma}_i|^{-1/2} \mathrm{e}^{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u}_i)\cdot\boldsymbol{\Sigma}_i^{-1}\cdot(\boldsymbol{x}-\boldsymbol{u}_i)} \tag{2}$$

where $N_i(\boldsymbol{x}; \boldsymbol{u}_i, \boldsymbol{\Sigma}_i)$ is the i-th normal distribution, π_i is the corresponding weight, and $\boldsymbol{u}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix, respectively.

2.2 HMM definition

Let $Y = [y_1, y_2, .., y_T]$ and $O = [o_1, .., o_T]$ be the state sequence and observation sequence of length T, respectively. The basic form of HMM, denoted by $\boldsymbol{\Lambda}$, can be written as the joint probability of Y and O,

$$p(Y, O|\boldsymbol{\Lambda}) = \pi_{y_0} \prod_{t=1}^{T} a_{y_{t-1}y_t} \prod_{t=1}^{T} b_{y_t}(o_t) \tag{3}$$

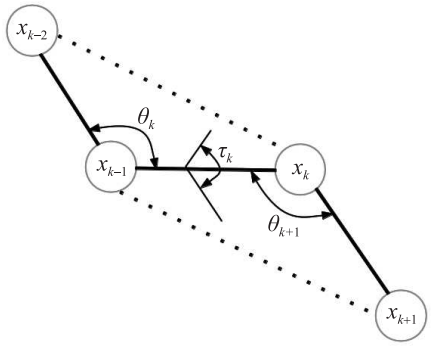


Fig. 1 Angles of four consecutive C_α atoms. For a residue x_k in a protein structure, the associated angle triplet for four consecutive C_α atoms $(x_{k-2}, x_{k-1}, x_k, x_{k+1})$ is represented as $(\theta_k, \tau_k, \theta_{k+1})$. θ_k is the bend angle of (x_{k-2}, x_{k-1}, x_k), τ_k is the dihedral angle of $(x_{k-2}, x_{k-1}, x_k, x_{k+1})$, and θ_{k+1} is the bend angle of (x_{k-1}, x_k, x_{k+1}).

where y_t is the state of position t, $a_{y_{t-1}y_t}$ is the transition probability from state y_{t-1} to y_t, and $b_{y_t}(o_t)$ is the emission probability for state y_t. In this work, the emission probability is defined as the multiplication of the following three terms,

$$b_{y_t}(o_t) = \left[\sum_{k=1}^{17} w_{y_t,k} \cdot N_k(x_t; \boldsymbol{u}_k, \boldsymbol{\Sigma}_k)\right] \cdot \left[\sum_{b=1}^{20} s_{y_t,b} \cdot \mathbf{SEQ}[t, b] \cdot \mathrm{env}(b, \mathrm{SS}_d, \mathrm{SA}_d)\right] \cdot \left[\sum_{a=1}^{21} f_{y_t,a} \cdot \mathbf{PSFM}[t, a]\right] \tag{4}$$

The first part of Eq. (4) describes the structure information, where $N_k(x_t; \boldsymbol{u}_k, \boldsymbol{\Sigma}_k)$ is the k-th Gaussian function, whose parameters were taken from Eq. (2) and x_t is the angle triplet defined above. The second part of Eq. (4) is the sequence profile distribution, where **PSFM**[] is the sequence profile matrix. The third part of Eq. (4) describes the sequence-structure distribution, where $\mathrm{evn}(b, \mathrm{SS}_d, \mathrm{SA}_d)$ is the probability score of amino acid type b appearing in the structure environment specified by three types of secondary structures SS_d and three types of solvent accessibilities SA_d[3,4]. For the simplicity of implementation, currently only parameters $w_{y_t,k}$, $f_{y_t,a}$, and $s_{y_t,b}$ in the emission function need to be trained by the learning procedure. Therefore the number of states is set to 17 by default[20], which can be optimized by Bayes Information Criteria (BIC) or other model selection techniques such as cross validation. We have tried different numbers of states, and the test results did not show any significant improvement.

2.3 Scoring structures by HMM

Once the HMM is given, we can assign a score to measure the global sequence structure compatibility of a protein by

$$V = \arg\max_{Y*} P(Y, O|\Lambda) \tag{5}$$

or

$$Z = \sum_{Y} P(Y, O|\Lambda) \tag{6}$$

where Λ is the model, O is the observation, and Y is the state sequence. Practically, the probability given by Eq. (6) is more robust than that of Eq. (5). Throughout this paper, we use HMM.Z to denote the score defined by Eq. (6).

2.4 Training data set

Considering the diversity of structural space, each test protein will have its own training dataset. First, for each protein in the testing data, we use PSI-BLAST to search the sequence against the PDB[21] database to get statistically significant (E-value less than 0.001) templates, and remove those templates having more than 70% sequence identity to test sequence. If too few or no template remains, we add a random subset from the following default data set to constitute the training data set of about 200 chains for this protein. The default data set for training is extracted from PDB according to the following steps:

(1) Filtering the entire PDB database with the following setting:

- X-Ray structure with resolution less than 0.2 nm;
- All residues have 3-D coordinates, at least for backbone atoms;
- Sequence length $L \in [50\ 300]$.

(2) Remove all chains that have sequence similarity higher than 70% to any test sequence using BLAST[19].

(3) Remove redundant proteins within the training data set by decreasing the mutual sequence similarity to 40% using CD-Hit[22].

With this data set, the proposed HMM is trained using the EM algorithm.

2.5 Test data set

We tested the method in protein structure selection scenario using Global Distance Test score (GDT)[23] as structure similarity measure. GDT is defined as $\frac{N_1+N_2+N_4+N_8}{4L}$, where $N_i, i = 1, 2, 4, 8$, is the number of positions with distance less than $0.1i$ nm after optimal structural superimposition and L is the protein length. Therefore, GDT value being 1 means two structures are the most similar. We applied the method to four benchmark datasets from different protein structure prediction methods. The first dataset, I-TASSER-DATA, contains 56 targets (proteins) with decoys generated by I-TASSER *ab initio* method[24-26] (http:// zhanglab.ccmb.med.umich.edu/decoys/). The second one, Modeller-DATA, has 55 targets, with decoys generated by Modeller[27]. In both datasets, each target has about 500 decoys, and the best decoy for each target has a GDT score greater than 0.4, which ensures that the pool contains at least some good-quality decoys. Figures 2a and 2b show the GDT distribution information, i.e., the maximum, average, and minimum GDT of I-TASSER-DATA and Modeller-DATA, respectively. The third benchmark data has 20 targets, containing FISA, LMDS_V2, and SEMFOLD from the Decoys 'R' Us decoy set[28]. The fourth one is HG_STRUCTAL from Decoys 'R' Us containing 29 targets.

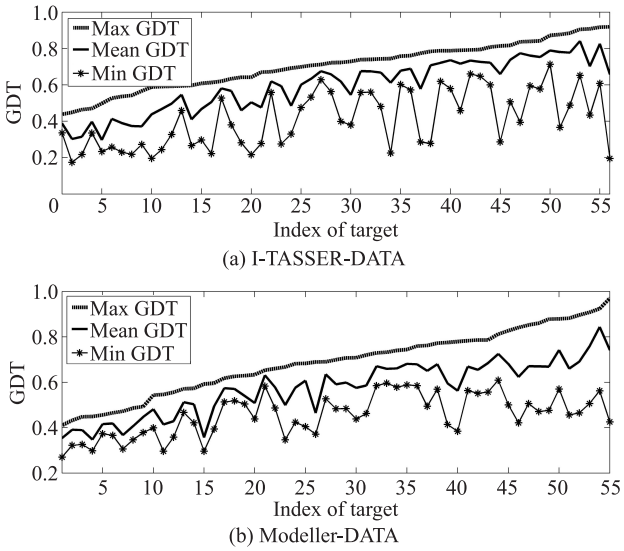


Fig. 2 Decoy distributions of I-TASSER-DATA (a) and Modeller-DATA (b). The horizontal axis indicates the index of each target and the vertical axis shows the GDT score. The dashed curve shows the maximum GDT score, the solid curve without stars shows the mean GDT score, and the curve with stars shows the minimum GDT score in the pool for each target.

3 Results

We compared the score HMM.Z with the state-of-art QA tools, OPUSCA, DFIRE, GOAP, and RW, all of which make use of global contact information in protein structures. We also compared the score HMM.Z with the secondary structure matching score (SSMatch) and environmental fitness (Fitness) which is the summation of compatibility score of all positions in a protein structure[3,4]. In the test, scores were used to rank the decoys of a given protein. In the following tables, we used the criteria below to study the selection and ranking performance:

- Top1: the GDT score of the top-1 selected model;
- Top5: the best GDT of selected top 5 models;
- Mean5: the average GDT score of the top 5 models;
- Pearson: Pearson correlation coefficient between the QA score and the true GDT score;
- Spearman: Spearman correlation coefficient between

the QA score and true GDT score.

Tables 1 and 2 showed the global QA performance of score HMM.Z, compared with OPUSCA, DFIRE, GOAP, and RW on I-TASSER-DATA and Modeller-DATA, respectively. We can see that HMM.Z achieved the best average top-1 selection performance on both datasets and the best correlation (Pearson and Spearman) to GDT score on Modeller-DATA. In particular, in Table 1, HMM.Z has comparable performance to the four QA methods in which OPUSCA is the best. But in Table 2, HMM.Z achieved the best top-1 selection performance (GDT: 0.594), which is significantly better than that of OPUSCA (0.579) and RW (0.569). Figure 3 compared the top-1 selection performance of HMM.Z to that of OPUSCA on I-TASSER-DATA and Fig. 4 compared HMM.Z to DFIRE on Modeller-DATA. We can find that for many targets, the decoys selected by our method were significantly better than those from OPUSCA or DFIRE. In Fig. 3, although the average performance was similar to that of OPUSCA, for a significant number of targets HMM.Z selected almost the best model in the pool. The result in Fig. 4 showed that HMM.Z outperformed DFIRE or GOAP, which ranked the best in the four QA methods on Modeller-DATA. Table 3 compared the global QA performances on FISA, LMDS_V2, and SEMFOLD together. As we can see, HMM.Z achieved the best selection performance with top-1 selection performance of 0.485 which was

Table 1 Global QA performance on I-TASSER-DATA.

	Top1	Top5	Mean5	Pearson	Spearman
GDT	0.705	0.705	0.693	1.000	1.000
OPUSCA	0.614	0.646	0.613	**0.322**	**0.237**
DFIRE	0.609	0.641	0.608	0.312	0.231
RW	0.610	0.636	0.609	0.278	0.196
GOAP	0.603	0.643	0.610	0.285	0.230
Fitness	0.607	0.641	0.606	0.176	0.119
SSMatch	**0.617**	0.651	0.616	0.216	0.166
HMM.Z	0.615	**0.651**	**0.616**	0.265	0.192

Table 2 Global QA performance on Modeller-DATA.

	Top1	Top5	Mean5	Pearson	Spearman
GDT	0.688	0.688	0.675	1.000	1.000
OPUSCA	0.579	0.627	0.584	0.192	0.175
DFIRE	0.587	0.623	0.585	0.175	0.157
RW	0.569	0.613	0.574	0.104	0.093
GOAP	0.588	0.628	0.589	0.192	0.179
Fitness	0.558	0.621	0.567	0.018	0.020
SSMatch	0.578	0.624	0.580	0.075	0.067
HMM.Z	**0.594**	**0.631**	**0.593**	**0.227**	**0.205**

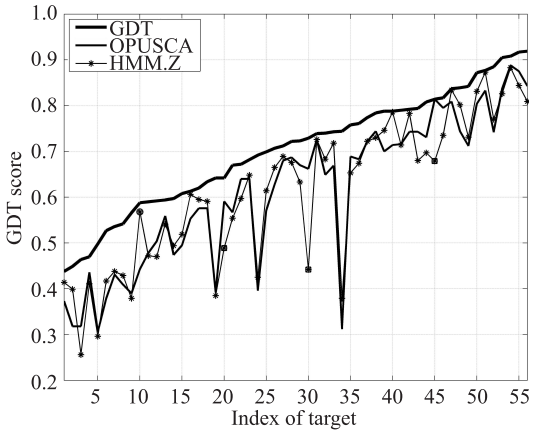


Fig. 3 Detailed comparison of global QA on the I-TASSER-DATA. The thickest curve represents the best true GDT score of the decoy for each target. The middle curve shows the performance of OPUSCA. The thinnest one represents the GDT score achieved by our method HMM.Z. The circled stars indicate the corresponding targets where our method HMM.Z performs significantly better than OPUSCA. The boxed stars show that HMM.Z significantly underperforms over OPUSCA on the corresponding targets.

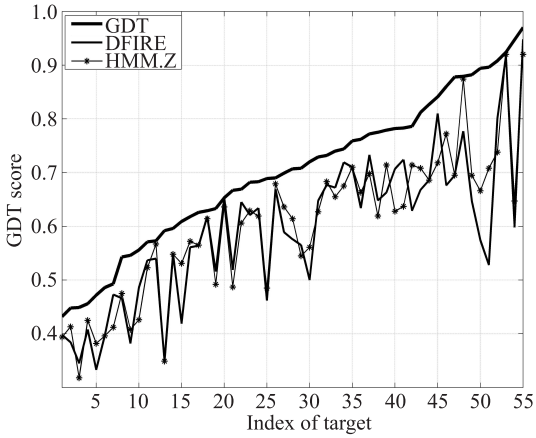


Fig. 4 Detailed comparison of global QA on the Modeller-DATA. The thickest curve represents the best true GDT score of the decoy for each target. The middle curve shows the performance of DFIRE. The thinnest one represents the GDT score achieved by our method HMM.Z.

0.016 higher than the second best method, DFIRE, and 0.021 higher than OPUSCA and RW, although HMM.Z does not stand out in other metrics. Figure 5 showed the detailed comparison between HMM.Z and DFIRE. For this dataset, we do not have GOAP result as several targets have too large number of decoys for GOAP to calculate all the scores. Table 4 showed the average performance on the HG_STRUCTRAL data set. HMM.Z had nearly the same average performance as OPUSCA, DFIRE, GOAP, and RW, all of which were close to the limit.

Table 3 Global QA performance on data of FISA + LMDS_V2 + SEMFOLD.

	Top1	Top5	Mean5	Pearson	Spearman
GDT	0.623	0.623	0.598	1.000	1.000
OPUSCA	0.464	0.517	0.450	0.274	0.274
DFIRE	0.469	0.525	**0.468**	**0.288**	**0.282**
RW	0.463	0.524	0.465	0.268	0.268
Fitness	0.470	**0.542**	0.465	0.190	0.186
SSMatch	0.467	0.519	0.451	0.172	0.166
HMM.Z	**0.485**	0.525	0.464	0.236	0.218

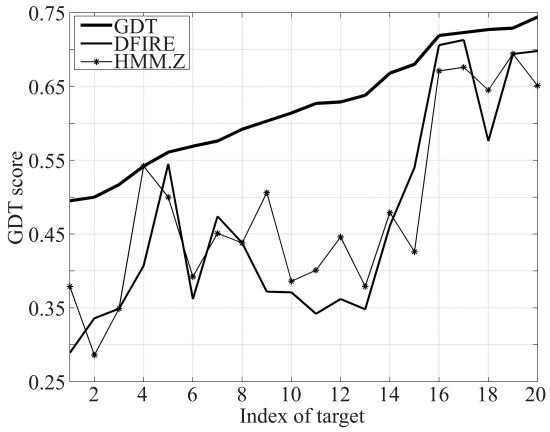


Fig. 5 Detailed comparison of global QA on FISA+LMDS_V2+SEMFOLD data. The thickest curve represents the best true GDT score of the decoy for each target. The middle curve shows the performance of DFIRE. The thinnest represents the GDT score achieved by our method HMM.Z.

Table 4 Global QA performance on HG_STRUCTAL data.

	Top1	Top5	Mean5	Pearson	Spearman
GDT	0.860	0.860	0.836	1.000	1.000
OPUSCA	0.840	0.858	0.823	0.779	0.739
DFIRE	0.844	0.856	0.824	0.806	0.756
RW	**0.847**	**0.858**	**0.824**	**0.812**	**0.759**
GOAP	0.844	0.859	0.824	0.842	0.754
Fitness	0.826	0.854	0.803	0.740	0.592
SSMatch	0.789	0.845	0.795	0.680	0.625
HMM.Z	0.839	0.857	0.813	0.780	0.721

From Fig. 6, we see that except for one case, HMM.Z can select almost the best decoy from the decoy pool. Tables 1-4 also compared the global QA performance of HMM.Z with Fitness and SSMatch. Overall, HMM.Z was mostly better than Fitness and SSMatch in terms of selection and correlation performance on four benchmark datasets.

4 Discussion and Conclusions

Our Hidden Markov Model is modeled in the sequence-structure space, in which the emission contains

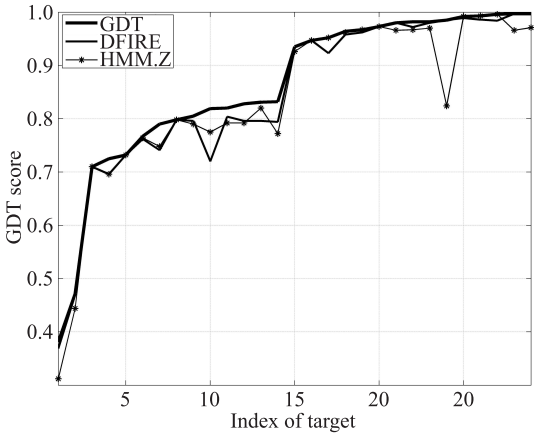


Fig. 6 Detailed comparison of global QA on the HG_STRUCTAL data. The thickest curve represents the best true GDT score of the decoy for each target. The middle curve shows the performance of RW. The thinnest one represents the GDT score achieved by our method HMM.Z.

sequence profile information and continuous (instead of discrete) structural content. As one of its advantages, HMM considers the dependency between adjacent local sequences and structures. The emission of HMM contains rich information about the sequence profile, secondary structures, and solvent accessibilities as well as local conformation represented in the angular space. The model for each test protein is trained on its homologous structures (if available) obtained by template search, which enhances the discerning power of the model and greatly reduces the noise in the training procedure, helping better capture the underlying relationship between the sequence and the native structure. From the test results, comparing to the four single model QA methods OPUSCA, DFIRE, GOAP, and RW, our test results have shown clear improvement of score HMM.Z in selection performance on the second (Modeller-DATA) and third (FISA+LMDS_V2+SEMFOLD) datasets and comparable performance on the first one (I-TASSER-DATA) and the fourth one (HG_STRUCTAL). From the detailed comparisons, we can conclude that for a significant number of cases HMM.Z is able to select almost the best model from the pool and achieve significant better selection performance than other scores, which means our HMM method is more sensitive in selecting near-native structures.

Another advantage of our HMM method is the less computation time. The most computation intensive step in our method is to generate the sequence profile using

PSI-BLAST. But this only needs to be calculated once. For a large number of decoys, our method is much faster than the four comparing methods, among which DFIRE and GOAP are the slowest.

However, our HMM method has room for improvement. In a few cases HMM.Z are significantly worse than the corresponding best method. One example is the 30th target in Fig. 3, which is 2CR7 from I-TASSER-DATA. And another example is the one in Fig. 6 from the FISA+LMDS_V2+SEMFOLD data set. We manually checked the case of 2CR7. Our HMM mis-selected a protein decoy whose local structures are very similar to the native one, but having a different packing, as shown in Fig. 7. Table 5 shows the pairwise GDT score between the native structure and the top-1 models selected by all the methods. RW and DFIRE also selected an incorrect decoy similar to the one selected by HMM.Z, while OPUSCA chose a decoy with correct packing. This indicates that adding global pairwise contact information into our method for HMM.Z might lead to further improvement. We are investigating those cases that HMM.Z loses more than 10 GDT points (in the 100 scale) to the best decoy for further possible improvement. As one of the future studies, we will derive informative scores from this method for local structure assessment and compare with existing local QA methods.

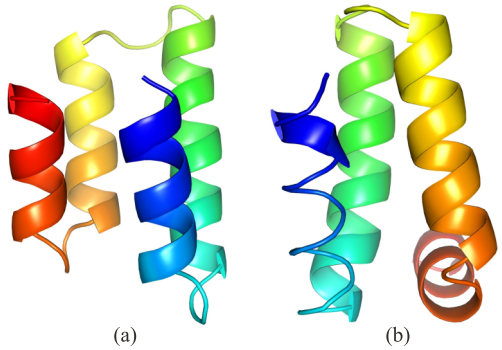


Fig. 7 Native structure (a) and top-1 model selected by HMM.Z (b) for 2CR7 from I-TASSER-DATA.

Table 5 Pairwise GDT of selected top-1 models for protein 2CR7 from I-TASSER-DATA.

	Native	OPUSCA	DFIRE	RW	HMM.Z
Native	1.000	0.662	0.525	0.525	0.442
OPUSCA		1.000	0.521	0.521	0.463
DFIRE			1.000	1.000	0.762
RW				1.000	0.762
HMM.Z					1.000

Note: Native means the native structure of protein 2CR7, OPUSCA means its selected top-1 model, and similarly for DFIRE, RW, and HMM.Z.

In summary, our HMM method can be used as a component tool for protein structure prediction to evaluate the global structure quality of predicted decoys. It will be released to the public when the stand-alone tool is fully tested.

Acknowledgements

This work was supported by National Institutes of Health grants R21/R33-GM078601 and R01-GM100701. We thank the authors of OPUS-Ca, DFIRE, and RW for making their scoring functions available to the community. We like to thank Prof. Yi Shang and Prof. Ioan Kosztin for helpful discussions.

References

[1] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, Critical assessment of methods of protein structure prediction (CASP) round x, *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. S2, pp. 1-6, 2014.

[2] Y. Zhang, Progress and challenges in protein structure prediction, *Current Opinion in Structural Biology*, vol. 18, no. 3, pp. 342-348, 2008.

[3] Y. Xu, D. Xu, and E. C. Uberbacher, An efficient computational method for globally optimal threading, *J. Comput. Biol.*, vol. 5, no. 3, pp. 597-614, 1998.

[4] Y. Xu and D. Xu, Protein threading using PROSPECT: Design and evaluation, *Proteins*, vol. 40, no. 3, pp. 343-354, 2000.

[5] J. Peng and J. Xu, Boosting protein threading accuracy, *Res. Comput. Mol. Biol.*, vol. 5541, pp. 31-45, 2009.

[6] J. Soding, Protein homology detection by HMM-HMM comparison, *Bioinformatics*, vol. 21, no. 7, pp. 951-960, 2005.

[7] K. Lin, A. C. May, and W. R. Taylor, Threading using neural network (TUNE): The measure of protein sequence-structure compatibility, *Bioinformatics*, vol. 18, no. 10, pp. 1350-1357, 2002.

[8] J. U. Bowie, K. Zhang, M. Wilmanns, and D. Eisenberg, Three-dimensional profiles for measuring compatibility of amino acid sequence with three-dimensional structure, *Methods Enzymol.*, vol. 266, pp. 598-616, 1996.

[9] S. Sunyaev, E. Kuznetsov, I. Rodchenkov, and V. Tumanyan, Protein sequence-structure compatibility criteria in terms of statistical hypothesis testing, *Protein Eng.*, vol. 10, no. 6, pp. 635-646, 1997.

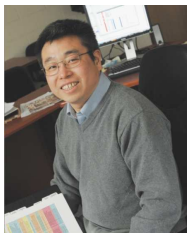
[10] H. Sumikawa, K. Fukuhara, E. Suzuki, Y. Matsuo, and K. Nishikawa, Tertiary structural models for human interleukin-6 and evaluation by a sequence-structure compatibility method and NMR experimental information, *FEBS Lett.*, vol. 404, nos. 2-3, pp. 234-240, 1997.

[11] Y. Matsuo and K. Nishikawa, Protein structural similarities predicted by a sequence-structure compatibility method, *Protein Sci.*, vol. 3, no. 11, pp. 2055-2063, 1994.
[12] A. C. Camproux and P. Tuffery, Hidden Markov model-derived structural alphabet for proteins: The learning of protein local shapes captures sequence specificity, *Biochim. Biophys. Acta*, vol. 1724, no. 3, pp. 394-403, 2005.
[13] A. C. Camproux, P. Tuffery, J. P. Chevrolat, J. F. Boisvieux, and S. Hazout, Hidden Markov model approach for identifying the modular framework of the protein backbone, *Protein Eng.*, vol. 12, no. 12, pp. 1063-1073, 1999.
[14] C. Bystroff, V. Thorsson, and D. Baker, HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins, *J. Mol. Biol.*, vol. 301, no. 1, pp. 173-190, 2000.
[15] Y. Wu, M. Lu, M. Chen, J. Li, and J. Ma, OPUS-Ca: A knowledge-based potential function requiring only Calpha positions, *Protein Sci.*, vol. 16, no. 7, pp. 1449-1463, 2007.
[16] H. Zhou and Y. Zhou, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Sci.*, vol. 11, no. 11, pp. 2714-2726, 2002.
[17] H. Zhou and J. Skolnick, GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction, *Biophysical Journal*, vol. 101, no. 8, pp. 2043-2052, 2011.
[18] J. Zhang and Y. Zhang, A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction, *PLoS One*, vol. 5, no. 10, p. e15386, 2010.
[19] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389-3402, 1997.
[20] W. Zheng and X. Liu, A protein structural alphabet and its substitution matrix CLESUM, *Transactions on Computational Systems Biology II*, vol. 3680, pp. 59-67, 2005.
[21] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola, Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules, *Acta Crystallogr. D Biol. Crystallogr.*, vol. 54, pp. 1078-1084, 1998.
[22] W. Li and A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.
[23] A. Zemla, LGA: A method for finding 3D similarities in protein structures, *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3370-3374, 2003.
[24] Y. Zhang, A. K. Arakaki, and J. Skolnick, TASSER: An automated method for the prediction of protein tertiary structures in CASP 6, *Proteins*, vol. 61, no. Supp. l7, pp. 91-98, 2005.
[25] S. Wu, J. Skolnick, and Y. Zhang, Ab initio modeling of small proteins by iterative TASSER simulations, *BMC Biol.*, vol. 5, p. 17, 2007.
[26] A. Roy, A. Kucukural, and Y. Zhang, I-TASSER: A unified platform for automated protein structure and function prediction, *Nat. Protoc.*, vol. 5, no. 4, pp. 725-738, 2010.
[27] A. Sali and T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.*, vol. 234, no. 3, pp. 779-815, 1993.
[28] R. Samudrala and M. Levitt, Decoys R Us: A database of incorrect conformations to improve protein structure prediction, *Protein Science*, vol. 9, no. 7, pp. 1399-1401, 2000.

Zhiquan He received his bachelor's degree from Xiangtan University, Hunan, China in 2000 and master's degree in Institute of Electronics, Chinese Academy of Sciences, Beijing, China in 2004. He obtained his PhD degree in computer science at University of Missouri-Columbia, USA in 2014. His PhD research was focused on applying machine learning and data mining methods in computational protein structural analysis and prediction.

Wenji Ma received her bachelor's degree in software engineering from Shandong University, Jinan, China in 2009 and is now a PhD candidate in the Department of Computer Science, City University of Hong Kong. She was a visiting scholar at University of Missouri-Columbia from Dec. 2012 to June 2013. Her research interests include algorithms, bioinformatics, and computational biology.

Jingfen Zhang obtained her PhD degree from Institute of Computing Technology, Chinese Academy of Sciences in 2006. She was a Postdoc Fellow and later a Research Scientist at University of Missouri-Columbia from 2006 to 2013. Her research interests are algorithms design such as combinatorial optimization, machine learning and pattern recognition, etc. Her research includes DNA sequencing, high-throughput mass-spec data analysis, and protein structure prediction.



Dong Xu is James C. Dowell Professor and Chair of Computer Science Department, with appointments in the Christopher S. Bond Life Sciences Center and the Informatics Institute at the University of Missouri-Columbia. He obtained his PhD degree from the University of Illinois, Urbana-Champaign in 1995 and did two years of postdoctoral work at the US National Cancer Institute. He was a Staff Scientist at Oak Ridge National Laboratory until 2003 before joining the University of Missouri. His research includes protein structure prediction, high-throughput biological data analysis, and *in silico* studies of plants, microbes, and cancers. He has published more than 240 papers. He is a recipient of 2001 R&D 100 Award, 2003 Federal Laboratory Consortium's Award of Excellence in Technology Transfer, and 2010 Outstanding Achievement Award from International Society of Intelligent Biological Medicine. He is Editor-in-Chief of *International Journal of Functional Informatics and Personalized Medicine* and Associate Editor-in-Chief of *IEEE/ACM Transactions on Computational Biology and Bioinformatics.*