Analysis and prediction of protein local structure based on structure alphabets

Qiwen Dong, Xiaolong Wang, Lei Lin*

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

E-mail: qwdong, wangxl, linl @insun.hit.edu.cn

ABSTRACT

In recent years, protein structure prediction using local structure information has made great progress. Many fragment libraries or structure alphabets have been developed. In this study, the entropies and correlations of local structures are first calculated. The results show that neighboring local structures are strongly correlated. Then, a dual-layer model has been designed for protein local structure prediction. The position specific score matrix, generated by PSI-BLAST, is inputted to the first-layer classifier, whose output is further enhanced by a second-layer classifier. The neural network is selected as the classifier. Two structure alphabets are explored, which are represented in Cartesian coordinate space and in torsion angles space respectively. Testing on the non-redundant dataset shows that the dual-layer model is an efficient method for protein local structure prediction. The Q-scores are 0.456 and 0.585 for the two structure alphabets, which is a significant improvement in comparison with related works.

Keywords: local structure prediction, correlation, dual-layer model, neural network

Introduction

Protein structure prediction from amino acid sequences is an important problem in computational biology. The sequences in public database increase dramatically with the completion of various genome projects, while the proteins with known structures are very limited. Theoretical prediction methods are one possible way to fill in this gap [1].

The earliest work on prediction of protein tertiary structures can be traced back to the 1970s [2]. Significant advancement has been made in the capability of protein structure prediction since then [3]. Many methods have been presented such as homology modeling [4,5], fold recognition [6,7] and *ab initio* prediction methods [8,9]. Recently, the methods based on three-dimensional fragment assembly have yielded great progress [10-12]. The strategy used in these studies is to divide known protein structures into short overlapping fragments, which are then collected and clustered by geometric similarity. Each cluster is represented by its centroid, and these centroids are then used to construct or analyze protein three-dimensional structures.

The secondary structure provides a three-state description: α-helices, β-strands and coils. This description of protein structures is very crude [13]. Many research groups have designed fragment libraries or Structure Alphabets (SA) to try to describe the local structural features of known protein structures more accurately. One of the most interesting structural alphabets is the I-Sites library [14,15], which, together with a sophisticated procedure for three-dimensional reconstruction, has been used with great efficiency to improve *ab initio* prediction methods. Camproux et al. first derived a 12-letter alphabet of fragment by Hidden Markov Model [16] and then extended it to 27 letters by Bayesian information criterion [17]. This

approach learns simultaneously the geometry and connections of the alphabets and has been used for protein structure mining [18]. De Brevern et al. [19] proposed a 16-letter alphabet generated by a self-organizing map based on a dihedral angle similarity measure. The prediction accuracy of local three-dimensional structure has been steadily increased by taking sequence information and secondary structure information into consideration [20]. Recently, the prediction performance on this alphabet was improved by Benros et al. [12]. A comprehensive evaluation of different structural alphabets was done by Karchin et al. [21]. Besides being used for the structure prediction, such structure alphabets have also been used for constructing substitution matrix [22], structure similarity searching [18,23], developing structure analysis web-server [24] etc.

Structure prediction based on fragment assembly has received increasing attention in recent years, since small libraries of protein fragments can model native protein structures accurately [25]. Addressing the problem of local protein structure prediction from sequence may therefore constitute a first step toward global structure prediction. Considerable work is focused on analyzing the local conformations of available protein structures and predicting them from their sequences [12,26,27]. However, the prediction accuracies are too low to be used in practice. An alternative way is to search the conformation space extended by the structure alphabets using an energy function [28,29]. However, this method is time consuming and the performances are strongly dependent on the energy function [30].

Crooks and Brenner [31] has shown that the neighboring secondary structure elements are strongly correlated. Such correlations result in the dual-layer methods for protein secondary structure prediction's success [32]. The PHD [33] method trained a two-layered feed-forward

neural network making the three-state accuracy above 70%. The PSIPRED method [34] adopted the same architecture with Position Specific Scoring Matrices (PSSM) as input, which is a winning method at the third round of Critical Assessment of techniques for protein Structure Prediction experiment (CASP3) [35]. The YASPIN method [36] utilized a single neural network for predicting the secondary structure in 7-state and then optimized the output with a hidden Markov model. Guo et al. [37] applied dual-layer support vector machine and PSSM to protein secondary structure prediction. Other methods try to predict beta-strand [38,39] or loop [40] accurately. Since both of structure alphabet and secondary structure are derived from the backbone structure of proteins, there must be some common features between them. The successful secondary structure prediction method can be borrowed to predict the local structure.

In this study, the entropy and correlations of local structures are analyzed. The results reveal that the neighboring local structures are also strongly correlated. A dual-layer model is designed for local protein structure prediction. The first layer uses the PSI-BLAST profiles [41] as input and outputs the local structures. The second layer uses the output of the first layer as input and outputs the final prediction. Two structure alphabets are explored: the DW structure alphabet developed by ourselves (**D**ong Qi**W**en *et al*.) [42] and the PB (**P**rotein **B**locks) structure alphabet [20]. They are represented in Cartesian coordinate space and torsion angles space respectively. The results show that the dual-layer model is an effective method for protein local structure prediction.

Materials and methods

In this section, the dataset used in this study is first introduced, and then the calculation formulae of entropy and correlation for local structures are given. Lastly, the dual-layer model is introduced to predict the local structures of proteins.

Dataset

The dataset of proteins used in this study is a subset of PDB database [43] obtained from the PISCES [44] web-server. There is less than 25% sequence identity between any two proteins and any protein has a resolution better than 2.5 Å. The structures with missing atoms and chain breaks are also excluded. The resulting database contains 1400 chains. Each protein is decomposed into overlapping fragments with a fixed length (7 for the DW structure alphabet and 5 for the PB structure alphabet) and each fragment is assigned a structure alphabet that has the best local structure similarity with it. Totally, there are about 271, 818 fragments in the dataset.

Structure alphabets

Two structure alphabets are investigated in this study: the DW structure alphabet and the PB structure alphabet. They are represented in Cartesian coordinate space and in torsion angles space respectively.

The DW structure alphabet is developed in our recent study [42], which is represented in Cartesian coordinates space. This structure alphabet contains 28 letters with lengths of 7 residues. It is superior to other structure alphabets with a variety of comparisons [42]. The quantization error of DW structure alphabet is 0.82 Å. While Sander et al. [27] get a quantization error of 1.19 Å for 27 clusters and Hunter and Subramaniam [45] specify the quantization error of their method for 28 clusters as 1.71 Å. The quantization error is defined

as the average Root Mean Square Deviation (RMSD) distance of all fragments in the dataset to their centroids. The DW structure alphabet can model native protein structures with 3.32 Å global RMSD. The supplement material contains the detail process of generating the DW structure alphabet and the evaluating results.

The PB alphabet [20] is composed of 16 prototypes, each of which is 5-residue in length and represented by 8 dihedral angles. This structure alphabet remains valid although the size of the databank becomes large [46]. The prediction accuracy of local three-dimensional structure has been steadily increased by taking sequence information and secondary structure information into consideration [12,20].

Calculation of entropy and correlation

Crooks and Brenner [31] has shown that the neighboring secondary structure elements are strongly correlated. Here, we perform a similar analysis on protein local structures, since both the local structures and secondary structures are derived from the backbone structures of proteins.

Entropy is a measure of the information needed to describe a random variable [47]. The entropy *H(X)* of a discrete random variable *X* can be calculated as:

$$H(X) = -\sum_{x \in \Omega} p(x) \log_2 p(x) \qquad (1)$$

where *Ω* is the set of allowed states, *x* is an element of *Ω*, *p(x)* is the probability of the state *x*.

The mutual information *I(X; Y)* is a measure of correlation between two discrete random variables, *X* and *Y*, which is defined as:

$$I(X;Y) = H(X) - H(X \mid Y)$$
$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (2)$$

where *p(x ,y)* is the joint probability of observing states *x* and *y*. If the random variables are independent then the mutual information achieves its lowest bound of zero. Mutual information cannot exceed the entropy of either variable.

In this study, all the probabilities are calculated from the training set by the maximal likelihood estimation. The probability *p(x)* of state *x* is defined as:

$$p(x) = n_x / N \quad (3)$$

where n_x is the number of observations of state *x*, *N* is the total number of samples. No smoothing is used since the training set is sufficiently large.

Dual-layer prediction

The correlations between neighboring secondary structure elements have resulted in the dual-layer methods for protein secondary structure prediction successfully. Here a high correlation of the neighboring local structures is also discovered. Such correlation can be taken into account, at least in part, by using a second layer. A dual-layer model has been designed for local structure prediction. The first layer is called sequence-to-structure layer, which classifies the sequence information (PSSM) into local structures. The second layer is called structure-to-structure layer, which classifies the outputs of the first layer into final prediction.

Generation of sequence profiles

The input of the first layer is the Position Specific Scoring Matrices (PSSM) generated by PSI-BLAST [41]. Each protein sequence is queried with the default PSI-BLAST parameter values except for the number of iterations being set to 10. The search is performed against the NRDB90 database, which is derived from the NR database of NCBI (ftp://ftp.ncbi.nih.gov/blast/db) by removing the redundancy sequence with sequence identity larger than 90%, using the Perl script from EBI [48].

Dual-layer model

The input of the first layer is the PSSM with window length of *w1*. The target output is the local structure likelihood of the center residue, whose length is *N_LA*, the number of prototypes in the structure alphabet (28 for the DW structure alphabet and 16 for the PB structure alphabet). The window is shifted residue by residue through the protein chain, thus yielding *N-L_SA+1* prediction for a chain with *N* residues, where *L_SA* is the length of the prototypes in the structure alphabet (7 for DW structure alphabet and 5 for PB structure alphabet). In order to allow the window to extend over the N terminus and the C terminus, a further unit has to be added for both ends. So the length of the input vector is 21**w1* for the first layer.

The outputs of the first layer are further filtered by the second layer. Again an extra input is used to indicate that the window spans a chain terminus. For this network, a window with length *w2* is used. So the length of the input vector is (*N_LA*+1)**w2*. Fig. 1 gives the outline of the dual-layer model.

At each layer, a variant of classifiers can be used, such as Support vector machine (SVM) [27], Neural Network (NN) [49], Hidden Markov Models (HMM). In this study, standard

feed-forward back-propagation network architecture with a single hidden layer is used as classifier [50], since neural network has been widely used in protein structure prediction [51]. All the parameters of the network are optimized at the validation set (10% of the training set) including the number of hidden units, the momentum term and learning rate. The number of hidden units is dependent on the number of input layers and the structure alphabets used. For the DW structure alphabet, the numbers of hidden units for the two networks (one network at each layer) are 100 and 120 respectively. For the PB structure alphabet, they are 100 and 80 respectively. A variant of momentum term and learning rate are investigated and the results are varied by no more than 1 percent. A momentum term of 0.9 and a learning rate of 0.005 are found to be effective. To prevent over-training of the network, a validation set that contains 10% of the training set is kept aside to evaluate the performance of the network during training. This validation set is not used to update the weight in the network. The training process is halted when the performance of the network on the validation set begins to decrease. All results are obtained by 5-fold cross validation protocol. The dataset is randomly divided into 5 subsets. Four subsets are used for training and the other one is used for test. The above process is repeated until each subset is tested.

Reliability index

The prediction reliability index (*RI*) is used to assess the effectiveness of the approaches for the prediction of the secondary structure of a new sequence. The *RI* offers an excellent tool for focusing on key regions with high prediction accuracy. There are many definitions of the *RI*. Here, we use the definition proposed by Rost and Sander [33]:

$$RI = \text{INTERGER}(10 \times (out_{\max} - out_{next})) \quad (4)$$

where out_{max} is the output of the unit with highest value and out_{next} is that of the unit with the next highest value. The factor 10 normalizes *RI* to integer values from 0 to 9.

Results and Discussions

Entropy and correlation

Each protein has two sequences: one is the amino acid sequence and the other is the structure alphabet sequence. The structure alphabet sequence is obtained by decomposing the protein structure into overlapping fragments and assigning a structure alphabet letter that has the best local structure similarity with the fragment. The criteria of evaluating the similarity between the fragment and the structure alphabet letter are the Root Mean Squared Deviation on Ca coordinates (CaRMSD) for DW structure alphabet and the Root Mean Squared Deviation on angular values (RMSDA) for PB structure alphabet [20]. We use R_i to denote one kind of amino acids and S_i to represent one kind of local structures in the structure alphabet. Table 1 gives the entropies and correlations of primary sequences and local structures for DW and PB structure alphabets. Previous study shows that neighboring amino acids are almost independent [31]. Here, we get the same conclusion. The entropy of primary residues is 4.1822 bits, which is very large (The maximum entropy for 20 states is $\log_2{}^{20}$=4.3219 bits). The nearest neighbor mutual information, $I(R_i,R_{i+1})$=0.0068 bits, is so small, indicating that the neighboring amino acids are only weakly correlated. Moreover, the mutual information between neighboring amino acids, conditioned upon the corresponding local structures ($I(R_iR_{i+1}|S_iS_{i+1})$) is also insignificant: 0.0235 bits for DW structure alphabet, 0.0158 bits for PB structure alphabet. So the neighboring amino acids are approximately independent, irrespective of the local structure.

Similar to the primary amino acids, the local structure entropy $H(S_i)$ is also large, 4.3112 bits for the DW structure alphabet and 3.2902 bits for the PB structure alphabet. But neighboring local structures are strongly correlated, resulting in a large nearest neighbor mutual information $I(S_i,S_{i+1})$, 2.3388 bits for DW structure alphabet and 1.8426 for PB structure alphabet. Such correlation also exists in the neighboring secondary structure [31] and makes the dual-layer methods and Hidden Markov Model (HMM) for protein secondary structure prediction successfully [31,34,37]. In this study, we show that the dual-layer model can also improve the performance of the local structure prediction.

The correlation between the primary amino acids and local structures $I(R_i,S_i)$ is relatively small, 0.1566 bits for DW structure alphabet and 0.1778 bits for PB structure alphabet. However, the dipeptide mutual information $I(R_iR_{i+1};S_iS_{i+1})$ becomes large. So prediction of local structure from amino acids is feasible. Moreover, the correlation between amino acids and local structures, obtained from the PB structure alphabet, is larger than that obtained from the DW structure alphabet, so local structure prediction for PB structure alphabet is easier than that for DW structure alphabet. The experimental results support this viewpoint.

The influences of window length on the local structure entropy and mutual information are not addressed here, because the current dataset cannot provide sufficient parametric statistics. For example, the dataset used in this study contains 308, 359 residues. When the window length is 3, the total parameters for calculating the mutual information $I(R, S_{i-1}S_iS_{i+1})$ is 439, 040 (20*28^3) using the DW structure alphabet.

Double-layer method improves the performance

The *Q*-score is used to assess the prediction, that is, the proportion of structure alphabet prototypes correctly predicted. This score is equivalent to the Q_3 value for secondary structure prediction.

The standard feed-forward back-propagation network is selected as the classifier since it can quickly train on a large dataset. A number of parameters can influence the performance of network, e.g., window size, number of hidden units, criterion for stopping the optimization procedure, the momentum term and learning rate. However, the influence of window size on the performance is larger than those of other parameters. So we optimally select the window size to get the best results. The influence of other parameters can be negligible. We first scan the window lengths *w1* of the first layer and select the one with the best performance. Based on the optimal *w1*, the window length *w2* of the second layer is scanned and the one with the best performance is also selected. The results are shown in Fig. 2.

The two structure alphabet have the same optimal window lengths, i.e. *w1*=15, *w2*=11. The performance with these window lengths is given in table 2. The Q-scores of single layer classifiers are 43.2% for DW structure alphabet and 56.4% for PB structure alphabet. Adding the second layer classifiers can improve the performance by 2 percent, which is 45.6% and 58.5% for DW and PB structure alphabet respectively.

The prediction is well distributed among all the structural alphabet prototypes

Fig. 3 gives the fractions of different structure alphabet prototypes in the natural sequences and the predicted sequences. The predicted accuracies of different structure alphabet prototypes are shown in Fig. 4. Although the distribution of the different prototypes is not balanced, the fractions of different prototypes in predicted sequence are very close to those in

the natural sequences. In the DW structure alphabet, one of the prototypes (index 18) holds a large fraction, which is nearly 25%. It is a typical helix structure (see supplement materials). The prediction accuracy (Q-score) on this prototype is more than 80%, indicating that the helix structures are easier to predict than other structures. The prototypes (index 2, 6, 10) get low prediction accuracies, which are basic coil structures. In the PB structure alphabet, two prototypes (index 4 and 13) dominate nearly 50% of all fragments, which are beta strand and alpha helix structures respectively. The prediction accuracies of these two prototypes are relatively high. The other prototypes are coils or mixed structures and achieve low prediction accuracies.

Reliability index helps to evaluate the prediction

The results reported above use the final winner-take-all projection of the real output onto one local structure prototype. The overall performance is lower in comparison with that obtained from the secondary structure prediction (about 80%). This may be caused by the large number of states in the structure alphabet. However, the RI can help to evaluate the prediction. The distribution of Q-scores with different RIs is illustrated in Fig. 5. The percent of residues that are predicted with a given reliability index are also given. As expected, the prediction accuracy of residues with higher RI values is much better than those with lower RI values. The RI is used to measure the reliability of the prediction. Such measure is very useful in practice, especially for the globally three-dimensional structure prediction of proteins. For example, when we try to predict the structure of proteins by assembling the structure alphabet letters, the search space can be reduced by only considering the low RI

regions, since high RI value means high accuracy. However, the distribution of residues with different RIs is imbalanced. Much of residues are predicted with low RI value.

Comparison with related works

The direct comparison with others' work is difficult because of the different definitions of the structure alphabets and the datasets [52,53]. However, the Q-scores of our dual-layer model are significantly higher than those of other closely related works.

Sander et al. derived a 27-letter structure alphabet in Cartesian coordinate space [27]. They used the Support Vector Machine (SVM) and the profiles from HSSP database [54] for local protein structure prediction and got the Q-score of 0.3038. The results are further improved to 0.3426 by using the property profiles. Changing the size of the training set from 5,000 to 20,000, the Q-score rises from 0.3426 to 0.3615. However, the running time is increased significantly. While our DW structure alphabet is also developed in Cartesian coordinate space with the same fragment length and similar number of letters (28). Our dual-layer model gets Q-score of 0.456, which is higher than the best results of Sander et al. by nearly ten percent. Such huge difference may be caused by a variant of factors. One important factor is the size of the training set. The best results of Sander et al. are achieved using only 20,000 fragments while our method are cross-validated on the whole training set (217,455 fragments). The neural network can quickly train on a large dataset and get better results. So we use neural network as classifiers in the dual-layer model.

The PB structure alphabet is proposed by Brevern et al. [19]. The classical Bayesian method gets Q-score of 0.408. The results are improved up to 0.498 by taking sequence information and secondary structures information into consideration [12,20]. The performance is still lower

than that of our dual-layer model, which is 0.585 using PB structure alphabet. Overall, the methods with PB structure alphabet outperform those with DW structure alphabet since decreasing the number of states can get better results, for example, the secondary structure prediction in three states can get accuracy of 80%.

Tang et al. discovered 357 motifs by clustering protein segments with length of 8 [55]. They presented a novel segment distance measure that considered both structural similarity and sequential similarity together. Such measure may be suitable for protein structure prediction from sequence. The method for local structure prediction simply compared the frequency profile between the sequence segments and the clusters, while our method uses advanced machine-learning technique for local structure prediction. The evaluation scheme is different from this study, which takes two parameters, a window size *w* and a RMSD threshold *t*. Given a true structure and its prediction, the scheme computes the percentage of residues found in length-w segments whose predicted structures are within *t* from the true structure after superposition. To give a direct comparison, the experiment in this study is re-run and re-evaluated by the new scheme. The window size *w* of DW and PB structure alphabet are 7 and 5 respectively. For DW structure alphabet, the accuracy is 0.52 with t = 1.4 Å. For PB structure alphabet, the accuracy is 0.60 with t= 1.2 Å and 0.58 with t=1.1 Å. The highest accuracy reported by Tang et al. is 0.58 with w=8 and t=1.4 Å [55]. Such difference may be caused by the different datasets or more importantly by the different numbers of clusters and the evaluation parameters. The DW structure alphabet has only 28 letters, which may not provide an exact description of diverse structures of proteins. The DW and PB structure alphabet only consider the structural similarity and omit the sequential similarity between

segments. Incorporating the sequential similarity into the clustering process will be our future directions.

Further analysis

To give an unbiased comparison with other methods and explore what leads to the improvement of our method, we re-implement the SVM method and compare our method with the SVM method and Brevern's method on the same test data. The LIBSVM package [56] is used as the SVM implementation with radial basis function as the kernel. The values of γ and regularization parameter C are set to be 0.005 and 1, respectively. For Brevern's method, the web-server (http://condor.ebgm.jussieu.fr/~debrevern/LOCPRED/) with option 'sequence families (new)' is used to get results.

400 proteins are randomly selected for testing and the other 1000 proteins for training. Two different profiles are used here. The first is the profiles generated on the NRDB90 database, the second is the HSSP profiles [54] used by Sander et al. [27]. The size of training set increases from small to large. We also compare the running time of SVM and NN method. All the experiments are performed on a personal computer with CPU of Intel Pentium 3.0G and memory of 1G. The results are given at Table 3 for DW structure alphabet and Table 4 for PB structure alphabet.

It is obvious that the increase in the size of training set can significantly improve the performance. This conclusion has also been derived by Sander et al. [27]. The performance seems to get saturation when the training samples are 100,000. The classifiers with NRDB90 profiles as input outperform those with HSSP profile as input. The results of SVM are slightly better than those of NN when the training set is small. However, they get similar

performance when the whole dataset is used. Such phenomenon is not surprising since SVM has the foundation of statistical learning theory and shows better performance than other classifiers in many tasks such as protein classification [57], fold recognition [58], location of sub-cellular [59], etc. But the SVM method is time-consuming especially when the size of the training set is large. The Brevern's method only gets a Q-score of 0.451, which is far lower than the NN and SVM methods.

Overall our method is superior to other methods by two factors. One is the large training set; the other is the NRDB90 profile. Additional, the neural network can quickly train on a large dataset and get good results.

Conclusions

Success in resolving local structure prediction will be a major milestone toward understanding the folding process of proteins. In this study, we first calculate the entropy and correlations of protein local structures and find that neighboring local structures are strongly correlated. This conclusion is the same as that of protein secondary structures. Then, a dual-layer model has been designed for protein local structure prediction. This model, coupled with the PSI-BLAST profiles as input, provides an efficient method for protein local structure prediction. The performance is significantly better than those of other related works. These results are helpful for protein structure prediction, especially for the method based on fragment assembly. Future works will aim at exploring the long-range interactions of the local structures and the prediction of globally three-dimensional structures of proteins.

Acknowledgements

The authors would like to thank Xuan Liu for her comments on this work that significantly improve the presentation of the paper. Financial support is provided by the National Natural Science Foundation of China (60435020 and 60673019).

References

1. **Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294(5540):93-96.**

2. **Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253(5494):694-698.**

3. **Zintzaras E, Brown NP, Kowald A. Non-parametric classification of protein secondary structures. Comput Biol Med 2006;36(2):145-156.**

4. **Reddy Ch S, Vijayasarathy K, Srinivas E, Sastry GM, Sastry GN. Homology modeling of membrane proteins: A critical assessment. Comput Biol Chem 2006;30(2):120-126.**

5. **Yu W, Li X, Liu J, Wu B, Williams KR, Zhao H. Multiple Peak Alignment in Sequential Data Analysis: A Scale-Space-Based Approach. IEEE/ACM transactions on computational biology and bioinformatics 2006;3(3):208- 219.**

6. **Solis AD, Rackovsky S. Improvement of statistical potentials and threading score functions using information maximization. Proteins 2006;62(4):892-908.**

7. **Xu J. Fold recognition by predicted alignment accuracy. IEEE/ACM transactions on computational biology and bioinformatics 2005;2(2):157- 165.**

8. **Fujitsuka Y, Chikenji G, Takada S. SimFold energy function for de novo protein structure prediction: consensus with Rosetta. Proteins 2006;62(2):381-398.**

9. **Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins 2005;61 Suppl 7:67-83.**

10. **Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, Ward JJ. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 2005;61 Suppl 7:143-151.**

11. **Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D. Prediction of CASP6 structures using automated Robetta protocols. Proteins 2005;61 Suppl 7:157-166.**

12. **Benros C, de Brevern AG, Etchebest C, Hazout S. Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 2006;62(4):865-880.**

13. **Chu W, Ghahramani Z, Podtelezhnikov A, Wild DL. Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. IEEE/ACM transactions on computational biology and bioinformatics 2006;3(2):98- 113.**

14. **Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281(3):565-577.**

15. **Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301(1):173-190.**

16. Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S. Hidden Markov model approach for identifying the modular framework of the protein backbone. Protein Eng 1999;12(12):1063-1073.

17. Camproux AC, Gautier R, Tuffery P. A hidden markov model derived structural alphabet for proteins. J Mol Biol 2004;339(3):591-605.

18. Guyon F, Camproux AC, Hochez J, Tuffery P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. Nucleic Acids Res 2004;32(Web Server issue):W545-548.

19. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41(3):271-287.

20. Etchebest C, Benros C, Hazout S, de Brevern AG. A structural alphabet for local protein structures: Improved prediction methods. Proteins 2005;59(4):810-827.

21. Karchin R, Cline M, Karplus K. Evaluation of local structure alphabets based on residue burial. Proteins 2004;55(3):508-518.

22. Tyagi M, Gowri VS, Srinivasan N, de Brevern AG, Offmann B. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 2006;65(1):32-39.

23. Yang JM, Tung CH. Protein structure database search and evolutionary classification. Nucleic Acids Res 2006;34(13):3646-3659.

24. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, de Brevern AG, Offmann B. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res 2006;34(Web Server issue):W119-123.

25. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. J Mol Biol 2002;323(2):297-307.

26. Wang LY. Covariation analysis of local amino acid sequences in recurrent protein local structures. J Bioinform Comput Biol 2005;3(6):1391-1409.

27. Sander O, Sommer I, Lengauer T. Local protein structure prediction using discriminative models. BMC Bioinformatics 2006;7(1):14.

28. Lee J, Kim SY, Lee J. Protein structure prediction based on fragment assembly and parameter optimization. Biophys Chem 2005;115(2-3):209-214.

29. Grishaev A, Steren CA, Wu B, Pineda-Lucena A, Arrowsmith C, Llinas M. ABACUS, a direct method for protein NMR structure computation via assembly of fragments. Proteins 2005;61(1):36-43.

30. Lee J, Kim SY, Joo K, Kim I, Lee J. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins 2004;56(4):704-714.

31. Crooks GE, Brenner SE. Protein secondary structure: entropy, correlations and prediction. Bioinformatics 2004;20(10):1603-1611.

32. Karypis G. YASSPP: Better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins 2006.

33. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. JMol Biol 1993;232(2):584-599.

34. **Jones DT. Protein secondary structure based on position-specific scoring matrices. JMolBiol 1999;292(2):195-202.**

35. **Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins 1999;Suppl 3:2-6.**

36. **Lin K, Simossis VA, Taylor WR, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. BioInformatics 2005;21(2):152-159.**

37. **Guo J, Chen H, Sun Z, Lin YL. A Novel method for protein secondary structure prediction using dual-layer SVM and profiles. PROTEINS: Structure, Function, and Bioinformatics 2004;54(4):738-743.**

38. **Fooks HM, Martin AC, Woolfson DN, Sessions RB, Hutchinson EG. Amino acid pairing preferences in parallel beta-sheets in proteins. J Mol Biol 2006;356(1):32-44.**

39. **Bradley P, Baker D. Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins 2006;65(4):922-929.**

40. **Zhu K, Pincus DL, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. Proteins 2006;65(2):438-452.**

41. **Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped Blast and Psi-blast: a new generation of protein database search programs. Nucleic Acids Research 1997;25(17):3389-3402.**

42. **Dong QW, Wang XL, Lin L. Methods for optimizing the structure alphabet sequences of proteins. Computers in biology and medicine 2007;37(11):1610-1616.**

43. **Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. The RCSB PDB information portal for structural genomics. Nucleic Acids Res 2006;34(Database issue):D302-305.**

44. **Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589-1591.**

45. **Hunter CG, Subramaniam S. Protein fragment clustering and canonical local shapes. Proteins 2003;50(4):580-588.**

46. **de Brevern AG. New assessment of a structural alphabet. In Silico Biol 2005;5(3):283-289.**

47. **Cover TM, Thomas JA. Elements of Information Theory. 1991.**

48. **Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14(5):423-429.**

49. **Hawkins J, Boden M. The applicability of recurrent neural networks for biological sequence analysis. IEEE/ACM transactions on computational biology and bioinformatics 2005;2(3):243-253.**

50. **Fiesler E, Beale R. Handbook of Neural Computation. New York: Oxford Univ. Press; 1996.**

51. **Rost B. Neural networks predict protein structure: hype or hit? in 'Artificial intelligence and heuristic methods for bioinformatics' Paolo Frasconi and Ron Shamir (eds). Amsterdam: IOS Press; 2003. p 34-50.**

52. **Hunter CG, Subramaniam S. Protein local structure prediction from sequence. Proteins 2003;50(4):572-579.**

53. de Brevern AG, Benros C, Gautier R, Valadie H, Hazout S, Etchebest C. Local backbone structure prediction of proteins. In Silico Biol 2004;4(3):381-386.

54. Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res 1998;26(1):313-315.

55. Tang T, Xu J, Li M. Discovering sequence-structure motifs from protein segments and two applications. Pac Symp Biocomput 2005:370-381.

56. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001;Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

57. Lingner T, Meinicke P. Remote homology detection based on oligomer distances. Bioinformatics 2006;22(18):2224-2231.

58. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22(12):1456-1463.

59. Lei Z, Dai Y. An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics 2005;6:291.

Tables

Table 1. Summary of entropies and correlations of primary sequences and local structures for DW and PB structure alphabets.

	DW	PB
Primary sequences		
Residue entropy $H(R_i)$:	4.1822	4.1822
Neighbor mutual information $I(R_i,R_{i+1})$	0.0068	0.0068
Conditional neighbor MI $I(R_iR_{i+1}\|S_iS_{i+1})$	0.0235	0.0158
Local structures		
Local structure entropy $H(S_i)$:	4.3112	3.2902
Neighbor mutual information $I(S_i,S_{i+1})$	2.3388	1.8426
Inter-sequences		
Monomer MI $I(R_i,S_i)$	0.1566	0.1778
Dipeptide MI $I(R_iR_{i+1};S_iS_{i+1})$	0.4751	0.4916

Table 2. The average Q-scores for the two structure alphabets.

	DW	PB
Single-layer model	0.432	0.564
Dual-layer model	0.456	0.585

The Q-scores are derived based on the optimal window length.

Table 3. The influence of profiles and training size on the DW structure alphabet.

Profile \ Training size		5000		20000		100000		All	
	Method[a]	Time (s)[b]	Q-scores c	Time (s)	Q-scores	Time (s)	Q-scores	Time (s)	Q-scores
NRDB90	NN	193	0.266 (0.259)	430	0.379 (0.361)	8148	0.427 (0.414)	15692	0.456 (0.432)
	SVM	251	0.275 (0.267)	3870	0.388 (0.378)	92035	0.434 (0.425)	2349768	0.452 (0.440)
HSSP	NN	130	0.256 (0.249)	370	0.313 (0.304)	7089	0.378 (0.367)	13476	0.392 (0.378)
	SVM	179	0.272 (0.241)	3603	0.328 (0.319)	77852	0.397 (0.387)	2203892	0.389 (0.372)

[a]Method: NN, neural network, SVM, support vector machine

[b]The running time (second)

[c]The results of dual-layer model, given is the bracket are the results using only single-layer classifier.

Table 4. The influence of profiles and training size on the PB structure alphabet.

Profile \ Training size		5000		20000		100000		All	
	Method	Time (s)	Q-scores	Time (s)	Q-scores	Time (s)	Q-scores	Time (s)	Q-scores
NRDB90	NN	88	0.426 (0.399)	467	0.504 (0.489)	3536	0.552 (0.546)	10548	0.585 (0.564)
	SVM	241	0.427 (0.405)	2805	0.526 (0.501)	73760	0.567 (0.554)	1374215	0.582 (0.568)
HSSP	NN	55	0.391 (0.376)	2751	0.483 (0.469)	3735	0.542 (0.518)	12563	0.578 (0.556)
	SVM	156	0.405 (0.388)	2738	0.484 (0.473)	66852	0.553 (0.532)	1109243	0.571 (0.562)
LocPred[a]									0.451

[a]The results are obtained by the webserver (http://condor.ebgm.jussieu.fr/~debrevern/LOCPRED/) with the option 'sequence families (new)', no training set is used.

Figure Captions

Fig. 1. The architecture of the dual-layer model for protein local structure prediction. The PSI-BLAST profiles are transformed into one-dimensional vectors by a sliding window, which are used as the inputs of the first-layer classifier. The outputs of the first-layer classifier are a number of vectors representing the likelihood that the center residues belong to a particular structure alphabet prototype. Using the same method, the outputs of the first layer classifier are inputted to the second classifier. The final decisions are based on the outputs of the second layer classifier.

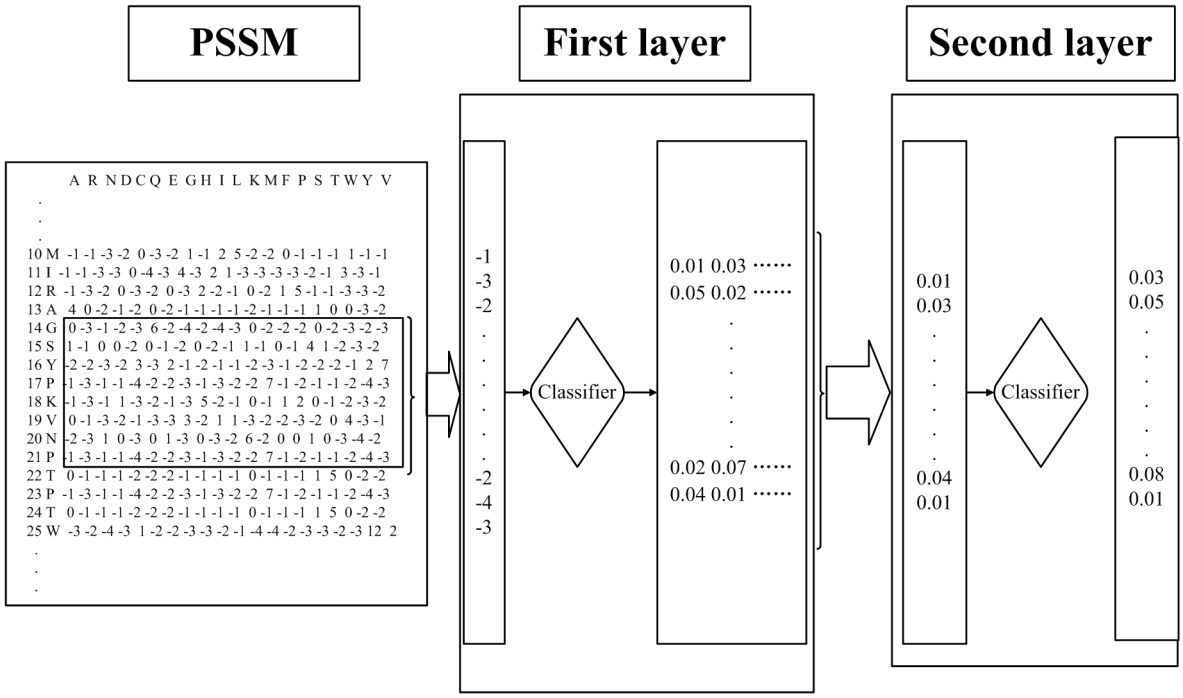

Fig. 2. The influence of window length on the performance. Figure (A1) and (A2) use the DW structure alphabet. Figure (B1) and (B2) use the PB structure alphabet.

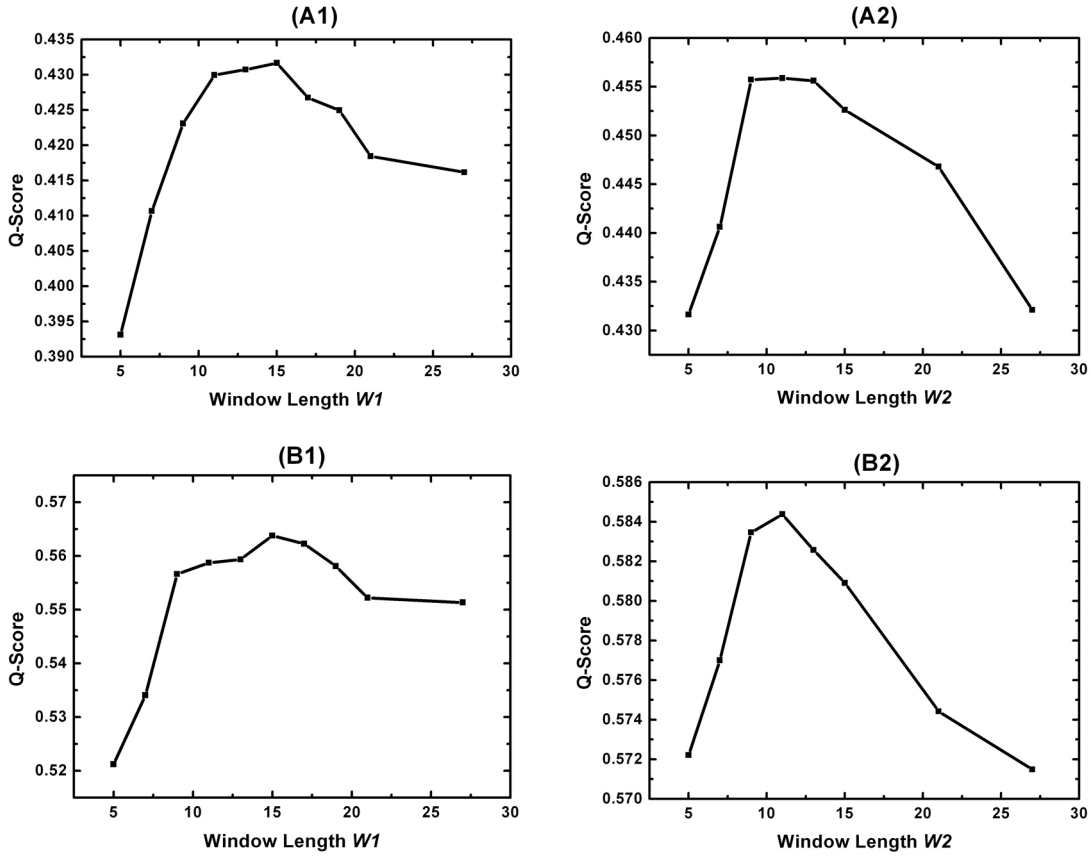

Fig. 3. The fractions of different structure alphabet prototypes in the natural sequences and the predicted sequences. The subfigure (A) and (B) are based on the DW and PB structure alphabet respectively.

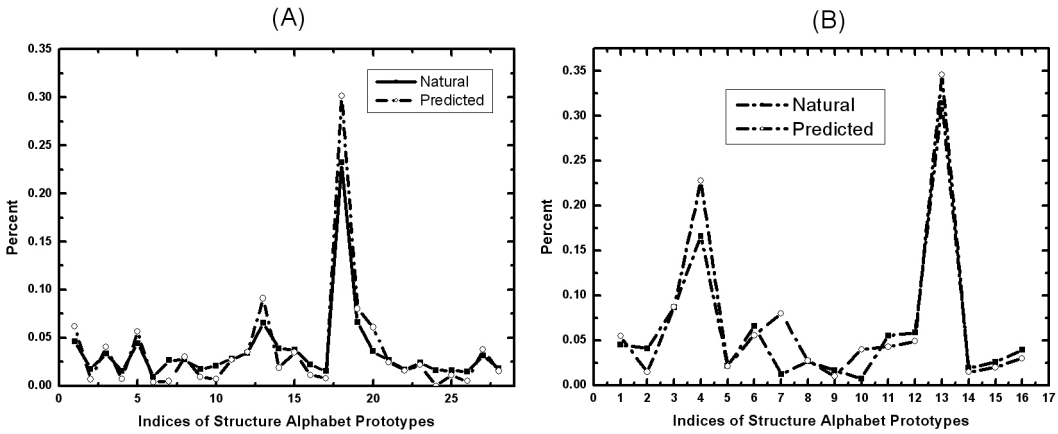

Fig. 4. The predicted accuracy of different structure alphabet prototypes. The subfigure (A) and (B) are derived using the DW and PB structure alphabet respectively.

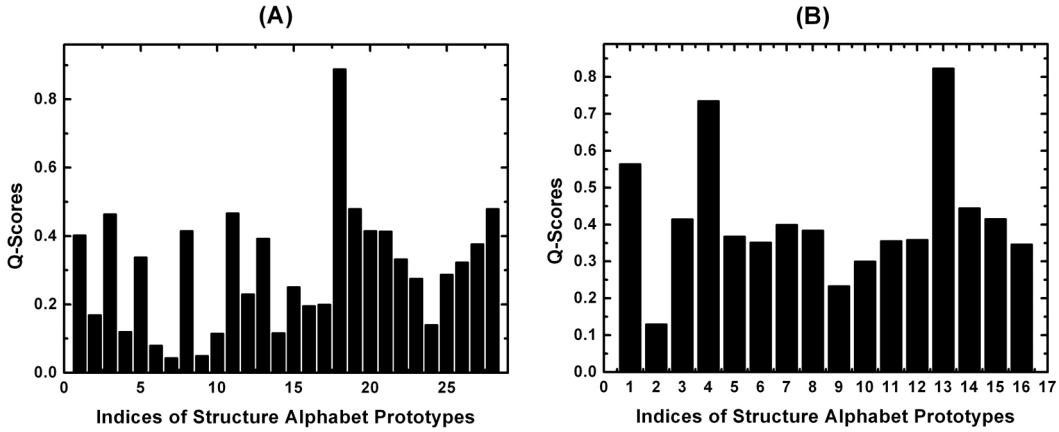

Fig. 5. The Q-scores for residues with a reliability index equal to a given value. The fraction of residues that are predicted with a given reliability index are also given. Figure (A) and (B) are based on the DW and PB structure alphabet respectively.