Position-specific residue preference features around the ends of helices and strands and a novel strategy for the prediction of secondary structures

MOJIE DUAN, MIN HUANG, CHUANG MA, LUN LI, AND YANHONG ZHOU

Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China

(RECEIVED April 7, 2008; FINAL REVISION May 22, 2008; ACCEPTED May 23, 2008)

Abstract

It has been many years since position-specific residue preference around the ends of a helix was revealed. However, all the existing secondary structure prediction methods did not exploit this preference feature, resulting in low accuracy in predicting the ends of secondary structures. In this study, we collected a relatively large data set consisting of 1860 high-resolution, non-homology proteins from the PDB, and further analyzed the residue distributions around the ends of regular secondary structures. It was found that there exist position-specific residue preferences (PSRP) around the ends of not only helices but also strands. Based on the unique features, we proposed a novel strategy and developed a tool named E-SSpred that treats the secondary structure as a whole and builds models to predict entire secondary structure segments directly by integrating relevant features. In E-SSpred, the support vector machine (SVM) method is adopted to model and predict the ends of helices and strands according to the unique residue distributions around them. A simple linear discriminate analysis method is applied to model and predict entire secondary structure segments by integrating end-prediction results, tri-peptide composition, and length distribution features of secondary structures, as well as the prediction results of the most famous program PSIPRED. The results of fivefold cross-validation on a widely used data set demonstrate that the accuracy of E-SSpred in predicting ends of secondary structures is about 10% higher than PSIPRED, and the overall prediction accuracy (Q_3 value) of E-SSpred (82.2%) is also better than PSIPRED (80.3%). The E-SSpred web server is available at http://bioinfo.hust.edu.cn/bio/tools/E-SSpred/index.html.

Keywords: secondary structure prediction; position-specific residue preference; ends of secondary structures; protein structure prediction

The knowledge of protein structures plays an important role in understanding protein functions (Watson et al. 2005), reconstructing protein structures (Dwyer et al. 2004), studying protein–protein interactions (Russell et al. 2004), and rationally designing drugs (Thiel 2004). Recently, the gap between available protein sequences and the experimental determination of their structures increased rapidly, making the prediction of protein structures more and more important (Koehl and Levitt 1999; Dunbrack 1999; Baker and Sali 2001). Accurate prediction of protein secondary structures can provide constraints for or be part of a tertiary structure prediction (Russell et al. 1996; Rost 1997; Jones 1999a). Furthermore, knowledge of secondary structures alone can

Reprint requests to: Yanhong Zhou, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China; e-mail: yhzhou@hust.edu.cn; fax: 86-27-87792170.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.035691.108.

also help the design of site-directed mutants that will not destroy the native protein structures (Chasman and Adams 2001; Bao and Cui 2005).

Secondary structure predictions methods have been developing for many years. The early methods were based on simple statistics (Chou and Fasman 1974; Lim 1974) or stereochemistry principles (Garnier et al. 1978). Thereafter, Qian and his coworkers used the neural network to take the influence of local interactions on secondary structure formation into account, which effectively improved the prediction accuracy (Qian and Sejnowski 1988). In the early 1990s, Rost and Sander (1993) proposed the method of using a sequence profile constructed by a similar sequence search and multiple sequence alignment to predict secondary structures, which exploited the evolution information and improved the prediction accuracy significantly. Later, based on Rost's method (Rost and Sander 1993), Jones (1999b) used PSI-BLAST to improve the homology sequence search, and developed a famous tool named PSIPRED that can get better results.

Today, almost all secondary structure prediction methods follow the Rost's idea (Rost and Sander 1993). These methods build models to predict the secondary structure class of a single residue position according to the information of its neighboring residues (Hua and Sun 2001; Kim and Park 2003; Guo et al. 2004; Qin et al. 2005). Apparently, these methods treat different positions on a protein sequence equally since they predict the secondary structure class of each residue position with the same models. That is, these methods assume that the residue distributions are distinctive for different secondary structure classes, but nondistinctive for different positions of a specific secondary structure class. In fact, the residue distributions on some positions of regular secondary structures are of specificity, which is especially obvious for positions around the ends of regular secondary structures and can be proved by the concept of helix capping (Presta and Rose 1988; Richardson and Richardson 1988; Padmanabhan et al. 1990; Blader et al. 1993; Aurora et al. 1994). Some researchers even declared that the helix ends are determined by the residues around them (Baldwin and Rose 1999). Unfortunately, this position-specific residue preference feature has not been exploited to predict secondary structures. As a result, their prediction performance around the ends of regular secondary structures are quite unsatisfactory, which remarkably limits the application of secondary structure prediction results (Russell and Barton 1993; Rost et al. 1994).

In this study, we collected a relatively large data set consisting of 1860 high-resolution, non-homology proteins from PDB, and further analyzed the residue distributions around the ends of regular secondary structures (i.e., α-helix and β-strands). It was found that there exist position-specific residue preferences around the ends of not only helices but also strands. On this basis, we proposed a novel strategy and developed a tool named E-SSpred to predict the secondary structures. This strategy treats the secondary structure as a whole, and builds models to predict entire secondary structure segments, instead of the class of a single residue, by integrating such information as the residue distribution features around ends, the composition and length distribution features of secondary structure segments, and so on. The results of fivefold cross-validation on the widely used data set CB513 (Cuff and Barton 1999) demonstrate that the accuracy of E-SSpred in predicting ends of secondary structures is about 10% higher than PSIPRED, and the whole prediction accuracy (Q_3 value) of E-SSpred (82.2%) is also better than PSIPRED (80.3%).

Results

Position-specific residue preference around the ends of helices and strands

Based on the DB1860 data set, we analyzed the residue distribution on positions around the ends of helices and strands. Similar to Aurora and coworkers (Aurora et al. 1994), the nomenclatures for these positions are labeled as follows:

$$\cdots N''_\alpha - N'_\alpha - N^{end}_\alpha - N^1_\alpha - N^2_\alpha - N^3_\alpha - \cdots - C^3_\alpha - C^2_\alpha - C^1_\alpha - C^{end}_\alpha - C'_\alpha - C''_\alpha \cdots,$$

$$\cdots N''_\beta - N'_\beta - N^{end}_\beta - N^1_\beta - N^2_\beta - \cdots - C^2_\beta - C^1_\beta - C^{end}_\beta - C'_\beta - C''_\beta \cdots,$$

where N^{end}_α, C^{end}_α represent the N-terminal and C-terminal of the α-helices, respectively, and N^{end}_β and C^{end}_β represent the N-terminal and C-terminal of the β-strands.

We calculated the residue preference scores (see Equation 1 in Materials and Methods) for each of theses positions. The results for partial positions are given in Table 1. For the purpose of comparison, Richardson's position-specific residue preference results around helix ends (Richardson and Richardson 1988) are also listed in Table 1 (R^+ denotes that a residue appears on a position with high frequency, and R^- means that the frequency is low). From Table 1, it can be found that many positions exhibit strong position-specific residue preference. For example, on the N-terminal of the helices, the hydrophobic residues such as *Val*, *Leu*, and *Ile* appear infrequently, and the electronegative, polar residues like *Asp* and *Glu*, are more likely to present.

Table 1. *Residue preference scores for partial positions around the ends of helices and strands*

	Positions around helix ends							Positions around strand ends			
a.a.	N''_α	N'_α	N^{end}_α	N^1_α	N^2_α	N^3_α	C^{end}_α	N'_β	N^{end}_β	C^{end}_β	C'_β
Gly	**1.41**	**1.66**	**1.76R$^+$**	1.31R$^-$	0.77	0.72	**2.36R$^+$**	**1.77**	1.23	0.94	**1.72**
Ala	0.42	0.80	0.90	0.85	1.05	1.06	1.18	0.73	0.93	0.77	0.87
Val	0.26	1.06	0.66R$^-$	0.99	**1.70**	0.68	**0.46R$^-$**	0.61	0.87	0.61	0.77
Leu	0.36	0.88	0.51R$^-$	0.68	**1.45**	1.17	1.05	0.51	0.79	0.68	0.78
Ile	0.29	0.89	0.53	0.71	**1.78**	0.73	0.48	0.50	0.85	0.62	0.60
Pro	—	—	—	—	—	—	—	**1.84**	0.63	**2.15**	1.05
Ser	**2.56**	1.13	**1.49R$^+$**	1.10	0.65	**1.43**	1.19	0.98	1.03	1.23	1.28
Thr	**2.18**	1.11	1.14	**1.45**	0.93	1.20	0.70	1.02	1.14	0.91	1.29
Cys	0.95	0.68	0.60	0.74	1.21	1.06	1.11	0.82	0.80	0.92	1.20
Met	0.42	0.82	0.52	0.72	1.35	1.11	1.07	0.65	0.95	0.67	0.77
Asn	**2.24**	0.78	1.16R$^+$	0.94	0.63	**1.52**	**1.68**	**1.58**	0.98	**1.55**	**1.71**
Gln	0.60	0.78	1.04	1.38	0.83	1.00	1.21	0.95	1.27	0.93	0.67
Phe	0.52	1.07	0.67	0.91	**1.52**	1.07	0.90	0.66	0.90	0.74	0.72
Tyr	0.63	1.09	0.82	0.89	1.30	1.20	0.98	0.73	1.00	0.73	0.77
Trp	0.42	1.34	0.90	0.93	1.38	0.73	0.55	0.87	0.90	0.78	0.70
Lys	*0.56*	0.93	0.95	0.73	0.69	1.23	1.14	1.15	**1.52**	0.86	0.88
Arg	0.61	0.87	0.78	0.65	1.05	1.08	1.06	1.09	1.27	0.86	0.80
His	1.18	0.88	1.05	1.17	0.85	1.28	1.32	1.06	1.10	1.07	1.08
Asp	**2.56**	1.07	**2.06R$^+$**	**1.95**	0.47R$^+$	0.80R$^+$	0.78	**1.65**	0.73	**2.06**	**1.59**
Glu	0.58	0.93	**1.94**	**1.63R$^+$**	0.40R$^+$	0.93R$^+$	0.80	0.83	1.12	0.93	0.74

The scores larger than 1.4 are in bold type, smaller than 0.6 are underlined. Data of proline on positions of the helices are omitted.

In order to find out whether the residue distributions are influenced by the length of the secondary structures, we further calculated and compared the position-specific residue preference scores for secondary structures of different lengths. Shown in Figure 1 are the results for four selected positions.

From Figure 1 it can be seen that the residue preference scores for the position N^1_α (Fig. 1A) and C^{end}_β (Fig. 1B) are almost not varying with the length of the secondary structures. It implies that the residue distributions on positions close to the ends of the helices and strands are scarcely influenced by the length of the secondary structures. On the contrary, the residue preference scores for the sixth position of helices (Fig. 1C) and the third position of stands (Fig. 1D), both of them are relatively far away from the ends of secondary structures, greatly vary with the structure length. The results suggest that it is feasible to build a unified model to predict the ends of helices and strands of different lengths.

Accuracy of secondary structure prediction

Fivefold cross-validation has been used on RS126 and CB513 to test the performance of E-SSpred, and the results are given in Tables 2 and 3. For the purpose of comparison, the prediction performance of the PMSVM (Guo et al. 2004), SVMpsi (Kim and Park 2003), and PSIPRED (Jones 1999b) on the same data sets, are also given in these tables. In Table 2, three kinds of widely used measures, the per-residue accuracy for overall proteins (Q_3 value) and for each class of secondary structure (Q_H, Q_E, Q_C, Q_H^{pre}, Q_E^{pre}, Q_c^{pre}), Matthew correlation coefficient for each class of secondary structure (C_H, C_E, C_C) (Matthews 1975), and segment overlap measure score (*SOV*) (Zemla et al. 1999) are used to evaluate the prediction results. The details for calculating per-residue accuracy Q_3, Q_I, and Q_I^{pre}, Matthew correlation coefficient C_I (here, I = H, E, and C), and the segment overlap measure score *SOV* are given in a previous paper (Kim and Park 2003). In Table 3, three measures, the sensitivity *Sn*, specificity *Sp*, and Matthew correlation coefficient *CC*, are used to evaluate the performance for predicting the ends of helices and strands. The sensitivity is defined as $Sn = TP/(TP + FN)$, the specificity is $Sp = TP/(TP + FP)$, and the Matthews correlation coefficient *CC* is

$$CC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}}$$

where the symbols *TP*, *TN*, *FP*, and *FN* denote the number of true positives, true negatives, false positives, and false negatives, respectively.

It can be seen from Table 2 that the results from the E-SSpred method are very good. On both data sets of RS126 and CB513, the Q_3 value of E-SSpred is improved >5% compared with the recently developed tools PMSVM and SVMpsi, and compared with PSIPRED, one of the most popular secondary structure prediction tools, E-SSpred can also get better performance in terms of the Q_3 (increased 2%), correlation coefficient, and *SOV* value.

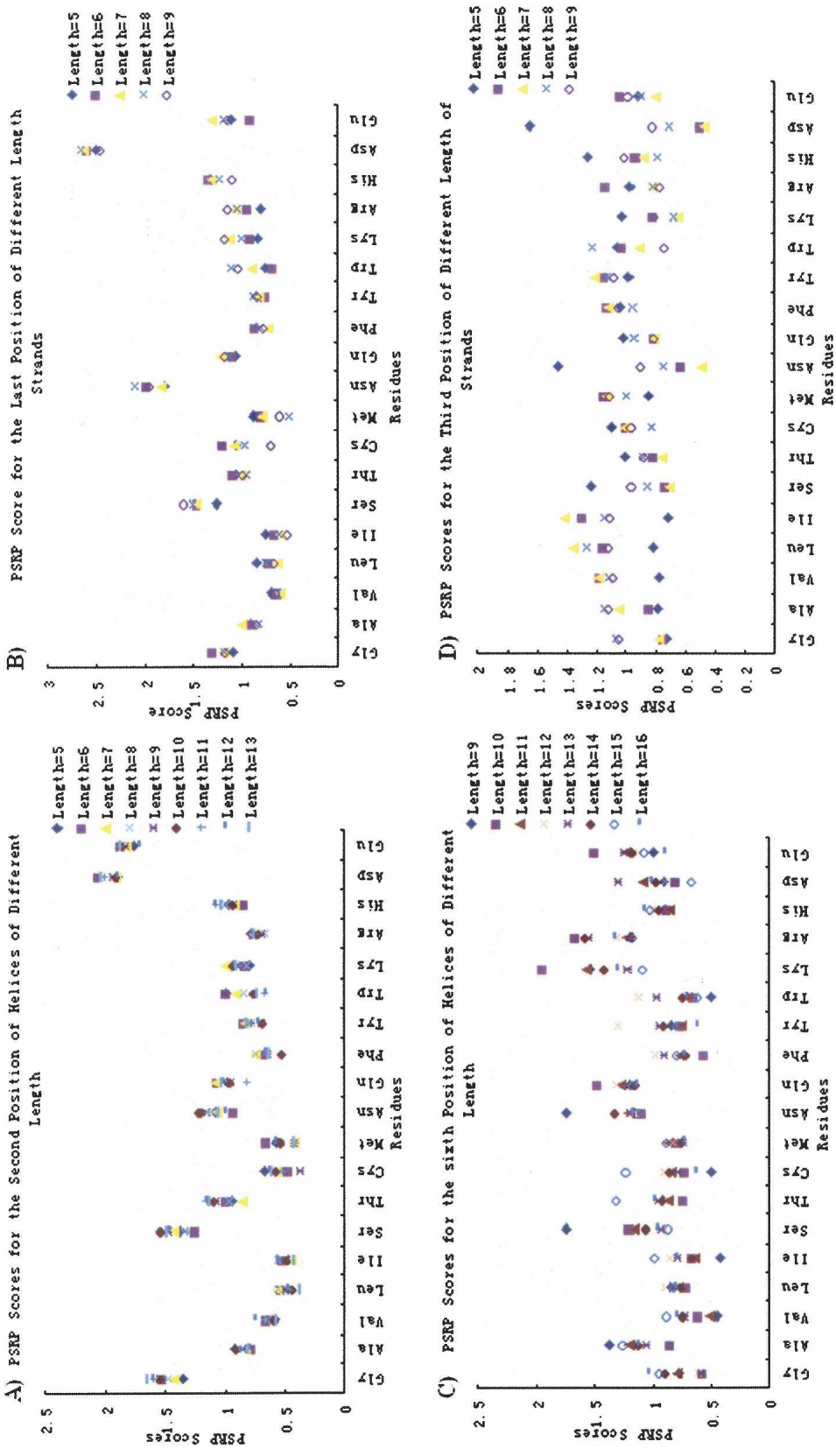

Figure 1. Position-specific residue preference scores for some positions of secondary structures of different lengths. (*A*) The second position of helices, helix lengths from 5 to 13; (*B*) the last position of β-strands, strand lengths from 5 to 9; (*C*) the sixth position of helices, helix lengths from 9 to 16; (*D*) the third position of strands, stand lengths from 5 to 9. Data of proline are omitted.

Table 2. *The prediction performance of E-SSpred and the comparison with PMSVM, SVMpsi, and PSIPRED*

	Q_3 (%)	Q_H (%)	Q_H^{pre} (%)	Q_E (%)	Q_E^{pre} (%)	Q_C (%)	Q_C^{pre} (%)	C_H	C_E	C_C	SOV[a]
PMSVM[b]	75.81	79.41	79.5	69.33	66.52	72.11	73.65	0.71	0.61	0.61	74.21
SVMpsi[c]	76.1	77.2	—	63.9	—	81.5	—	—	—	—	72.0
SVMpsi[b]	76.6	78.1	—	65.6	—	81.1	—	—	—	—	73.5
PSIPRED[c]	79.69	84.01	81.41	72.71	71.59	79.31	78.15	0.75	0.69	0.63	76.0
PSIPRED[b]	80.30	84.76	83.01	74.30	75.76	79.84	80.37	0.76	0.69	0.63	76.20
E-SSpred[c]	81.63	84.43	82.31	73.59	71.43	80.39	81.85	0.76	0.69	0.65	76.39
E-SSpred[b]	82.15	84.91	83.74	75.78	75.28	80.67	82.84	0.78	0.70	0.65	76.72

[a] *SOV*: this criterion is following the definition of Zemla et al. (1999).
[b] SVMpsi, PSIPRED, E-SSpred: result obtained on RS126 set. SVMpsi results are from Kim and Park (2003). E-SSpred is the new method proposed in this paper.
[c] PMSVM, SVMpsi, PSIPRED, E-SSpred: results obtained on CB513 set.

The data in Table 3 demonstrate that the performances of both PMSVM and PSIPRED for the prediction of secondary structure ends are quite low. For example, the prediction sensitivity and specificity for helix C-terminal, strand N-terminal, and C-terminal are all under 40%, implying that these tools are incompetent to locate secondary structures exactly. Compared with PMSVM and PSIPRED, the performance of E-SSpred for locating the ends of secondary structures is significantly better. It can be seen from Table 3 that the overall performance (*CC*) of E-SSpred for predicting helix N-terminal, helix C-terminal, strand N-terminal, and strand C-terminal are higher than PSIPRED, respectively.

Discussion

Position-specific residue preference around the ends of helices and strands

The residue distribution around the ends of helices has been analyzed by Richardson and Richardson (1988) using a small data set containing 45 protein sequences. It was found that, on certain positions, some residues present with preference (e.g., *Asp* on N_α^{end}, *Gly* on C_α^{end} etc.), and some others are unlikely to occur (e.g., *Leu* on N_α^{end}, *Val* on C_α^{end} etc.). In this study, we collected a large data set containing 1860 high-resolution, non-homology proteins to further analyze the residue distributions around the ends of regular secondary structures. It was found that there exist more position-specific residue preferences around the ends of not only helices but also strands (see Table 1). For example, *Glu* also presents high frequently on position N_α^{end} as *Asp*, and *Asp* also prefers to occur on position N_α^1 as *Glu*.

From our results, some interesting conclusions about the residue distribution around the ends of helices and strands can be obtained. For instance, polar and electronegative residues *Glu* and *Asp* prefer to present on the first (N_α^{end}) and second positions (N_α^1) of α-helices, but hydrophobic residues such as *Val*, *Leu*, *Ile*, and *Met* are unlikely to appear on these positions. On the third position of the α-helices (N_α^2), however, hydrophobic residues are of preference but electronegative residues are unlikely to occur. This residue preference on positions next to the helix starts may be one of the requirements to form the helix structure.

Our results show that the position-specific residue preference around the ends of helices and strands is more obvious than the inner positions, which is consistent with the results of Richardson and Richardson (1988). In addition, we also analyzed the residue distributions of secondary structures with different length, and found that the influences of structure length on residue distributions for positions around the ends of helices and strands are much less than those inner positions. These results imply that it is necessary to build specific models to predict the end positions of regular secondary structures.

There are also some conflicts between Richardson and Richardson's (1988) results and ours. For example, their research showed that both *Asp* and *Glu* prefer to present

Table 3. *The prediction performance of E-SSpred for the ends of helices and strands and the comparison with PMSVM and PSIPRED*

	Helix N-terminal			Helix C-terminal			Strand N-terminal			Strand C-terminal		
	Sn (%)	*Sp* (%)	*CC*	*Sn* (%)	*Sp* (%)	*CC*	*Sn* (%)	*Sp* (%)	*CC*	*Sn* (%)	*Sp* (%)	*CC*
PMSVM	38.34	36.69	0.366	27.02	36.09	0.304	32.36	34.33	0.325	29.35	31.14	0.293
PSIPRED	51.96	50.14	0.503	34.69	32.14	0.325	39.50	38.11	0.38	36.19	34.92	0.347
E-SSpred	61.73	60.56	0.606	45.06	45.73	0.447	50.15	47.61	0.482	46.04	45.35	0.45

on N_α^2, but our results indicate that these two residues are unlikely to present on N_α^2 (Richardson and Richardson 1988). The main cause for these conflicts is possibly that too few protein sequences were used in Richardson and Richardson's (1988) research to estimate the residue distributions, which might lead to some statistical biases.

The performance of secondary structure prediction

From Table 3 it can be seen that the performance of E-SSpred for predicting the ends of α-helices and β-strands is significantly better than PMSVM, SVMpsi, and PSIPRED, indicating that the position-specific residue preference around the ends of helices and strands is a very useful feature to help predict the ends of secondary structures and locate secondary structures more accurately. Moreover, it means that E-SSpred can locate the secondary structures on protein sequences more accurately, and therefore its prediction results can be applied to solve related problems such as protein tertiary structure prediction, protein function analysis, and so on, more effectively. However, from Table 2, it can also be seen that, with the help of this feature, the improvement of secondary structure prediction performance in terms of the measures Q_3, Matthew correlation coefficient, and *SOV* is not very significant. The possible causes include: (1) The number of ends is very small relative to the number of residues in helices and strands; thus, the direct contribution of improving the prediction of ends is limited to the improvement of secondary structure prediction accuracy measured by Q_3, etc.; (2) the main cause that greatly influences the performance of existing secondary structure prediction tools is that some helices and strands are easily predicted as loops completely. The novel strategy proposed in this study, to predict entire secondary structure segments directly by integrating relevant features in such aspects as the residue distribution around ends, tripeptide composition, and so on, has the potential to change this situation. However, the algorithms currently used in E-SSpred are still too simple to adequately bring into play the potential of this novel strategy. We expect to develop, in the near future, more advanced algorithms that can significantly improve the prediction performance.

Materials and Methods

Data sets

Three data sets were used in this study. One is the data set we collected to analyze the statistical features of secondary structures and to train models for secondary structure ends prediction. This data set contains 1860 non-homology proteins and is called DB1860. The proteins in this data set were picked from the PDB database using the tool, PISCES, developed by Dunbrack (Wang and Dunbrack 2003). These proteins meet the following criterions: (1) they were detected by an X-ray diffraction method; (2) the sequence identity between any two of them is <30%; (3) the experiment resolution is <2.0 angstroms; (4) there are no homology proteins between DB1860 with RS126 and CB513 data sets. The list of proteins in DB1860 can be downloaded from the website: http://bioinfo.hust.edu.cn/bio/tools/E-SSpred/.

The other two data sets are RS126 (constructed by Rost and Sander [1993]) and CB513 (constructed by Cuff and Barton [1999]); these two data sets contains 126 and 513 non-homology proteins, respectively, and have been widely used to test secondary structure prediction methods (Hua and Sun 2001; Kim and Park 2003), and they also are used to compare the prediction performance of our method with that of other methods.

The secondary structure of proteins in these data sets is assigned from the experimentally determined tertiary structure by DSSP (Kabsch and Sander 1983), which has been the most widely used secondary structure definition. It has eight secondary structure classes: H(α-helix), G(3_{10}-helix), I(π-helix), E(β-strand), B(isolated β-bridge), T(turn), S(bend), and –(rest). We reduced the eight classes to three states, helix(H), sheet(E), and coil(C) using the following strategy: H, G to H; E, B to E; all other states to C. This strategy is now widely used, and considered to be the strictest definition in secondary structure prediction methods (Hua and Sun 2001; Kim and Park 2003; Guo et al. 2004).

Assessment of position-specific residue preference around ends of secondary structures

The position-specific residue preference is defined as the statistical frequencies where residues occur on a certain position around the ends of secondary structures. The preference score for residue a on position i of secondary structure class ss is denoted as $f_{ss}(a, i)$, and is determined by:

$$f_{ss}(a,i) = p_{ss}(a,i)/p_{ss}^0(a), \quad (1)$$

where $p_{ss}(a, i)$ is the frequency of residue a occurring on position i of secondary structure class ss, and $p_{ss}^0(a)$ is the average frequency of residue a on all positions of ss.

The position-specific residue preference for secondary structures of different lengths is denoted as $f_{ss}^l(a, i)$, and is determined by

$$f_{ss}^l(a,i) = p_{ss}^l(a,i)/p_{ss}^{l0}(a), \quad (2)$$

where l is the length of secondary structures, $p_{ss}^l(a, i)$ is the frequency of residues a occurring on position i of structure ss whose length is l, and $p_{ss}^{l0}(a)$ is the average frequency.

Prediction of ends of secondary structures

We first predict the probabilities of each position in the protein belonging to the end positions of regular secondary structures. The SVM method is adopted to do this job according to the residue distributions around each position. For each kind of ends, the helix N-terminal, helix C-terminal, strand N-terminal, and strand C-terminal, a binary SVM classifier, is built to predict them, respectively. By the analysis of the

position-specific residue preference scores of different positions in secondary structures, we find that the position-specific residue preference features on some positions, such as nine positions around the helix N-terminal (i.e., upstream three residues, downstream five residues, and the helix N-terminal itself), and seven positions around other terminals (i.e., upstream three residues, downstream three residues, and the end itself), is more intense than on other positions. Based on this, to predict the helix N-terminal, nine residues are encoded with PSSM (position-specific scores matrix) scores to construct feature vectors, and for the prediction of the other three ends, seven residues are selected to construct feature vectors.

In this study, the LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) program is used as the implementation of SVM, in which the radial basis function kernel is adopted and the two parameters, C and γ, are empirically set to 10 and 0.0015, respectively.

In the course of prediction, for each position of a protein sequence, these SVMs can output four scores reflecting the probabilities of this position being a helix N-terminal, helix C-terminal, strand N-terminal, and strand C-terminal, respectively, and these scores will be used as features in the secondary structure prediction of the whole protein.

Tri-peptide composition in secondary structures

Similar to the idea of using codon usage to help distinguish exons from introns in the field of predicting gene structures in eukaryotic DNA sequences, in this study, the tri-peptide composition is used as an additional feature to help distinguish different secondary structures. For a tri-peptide, its probability score, appearing in secondary structure *ss*, is defined as:

$$p(a_i a_j a_k \mid ss) = n(a_i a_j a_k|ss) \Big/ \sum_i \sum_j \sum_k n(a_i a_j a_k|ss), \qquad (3)$$

where a_i, a_j, a_k denote a residue type, respectively, $n(a_i a_j a_k \mid ss)$ is the number that the tri-peptide $a_i a_j a_k$ appears as in *ss*.

Length distribution of secondary structures

As shown in Figure 2, the length distribution of helices is different from that of strands. Thus, the length can be used as an additional feature to help distinguish different secondary structures. In this study, the length score for predicting a segment of length l as a helix or strand is determined by:

$$S_{len}(l, ss) = \ln(n_l^{ss} / \overline{n^{ss}}), \qquad (4)$$

where n_l^{ss} is the number of secondary structures of length l of secondary structures of class *ss*, and $\overline{n^{ss}}$ is the average number of all lengths of *ss*. Both n_l^{ss} and $\overline{n^{ss}}$ are determined by the training data set DB1860.

Linear discriminate analysis of secondary structures

Similar to the strategy widely used in predicting exons and introns of genes from DNA sequences, in this study, the linear discriminate analysis method is used to integrate the end-prediction results, tri-peptide composition score, length distribution score, and the PSIPRED prediction results to predict entire secondary structure segments.

For a protein sequence, let $Seg[i, j]$ be a segment of this sequence (i and j are the start and end position, respectively), then the potential score where this segment belongs to helices or strands is determined by:

$$S_{ij}^{ss} = p_1 \cdot s_i^{ss,n} + p_2 \cdot s_j^{ss,c} + p_3 \cdot s_{ij}^{ss,psipred} + p_4 \cdot s_{ij}^{ss,tri-res} + p_5 \cdot s_{ij}^{ss,len} + p_6, \qquad (5)$$

where, $s_i^{ss,n}$ is the probability score for position i to be the N-terminal of class *ss*, $s_j^{ss,c}$ is the score for position j to be the C-terminal of *ss*, $s_{ij}^{ss,len}$ and $s_{ij}^{ss,tri-res}$ are the length distribution score and average tri-peptide composition score of this segment, respectively. $s_{ij}^{ss,psipred}$ is the PSIPRED prediction score determined by

$$s_{ij}^{ss,psipred} = \frac{1}{l} \sum_{k=i}^{j} s_k^{ss,psipred},$$

and the $s_k^{ss,psipred}$ is the score of residues k predicted to be *ss* by PSIPRED. The parameters $p_1 - p_6$ are weight coefficients that are determined by the least-square approach used for the REGRESS function of Matlab7.0.

A segment whose potential score S_{ij}^{ss} is bigger than a threshold $S_{threshold}^{ss}$ is predicted as an *ss* candidate, and all the candidates in a protein sequence can be obtained after scanning the whole sequence. For overlapped candidates of the same class, only the one with the biggest S_{ij}^{ss} value is kept. Then, for overlapped candidates of different classes (for example, a helix candidate from position i to j, a strand candidate from k to l, and $i < k < j < l$), the following rules are used to make the decision:

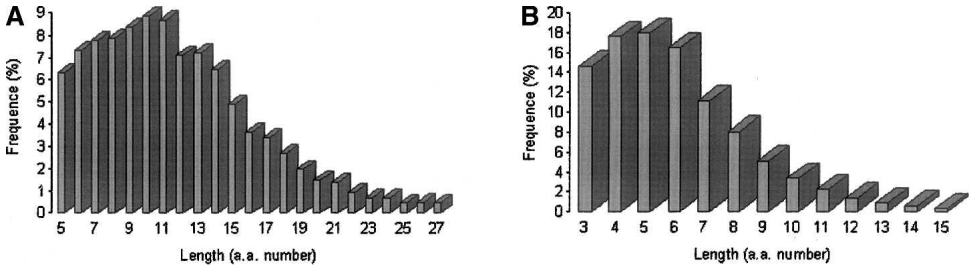

Figure 2. The length distribution of (*A*) helices and (*B*) strands in DB1860.

$$\begin{cases} struct(i \to j) = H, struct(j \to l) = E \cdots\cdots if(S_{ij}^{H} > S_{kl}^{E}) \\ struct(i \to k) = H, struct(k \to l) = E \cdots\cdots others \end{cases}, \quad (6)$$

where H and E represent the helix and the strand, respectively, and $struct(i \to j) = H$ means the segment from position i to j is predicted as a helix and so do the others.

Acknowledgments

We thank J.A. Cuff and G.J. Barton for providing the CB513 data set. This work was supported by the National Natural Science Foundation of China (Grant Nos. 90608020, 30370354, and 90203011), NCET-060651, the National Platform Project of China (Grant No. 2005DKA64001), and the Ministry of Education of China (Grant Nos. 20050487037 and 505010).

References

Aurora, R., Srinivasan, R., and Rose, G.D. 1994. Rules for α-helix termination by glycine. *Science* **264:** 1126–1130.

Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93–96.

Baldwin, R.L. and Rose, G.D. 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24:** 26–33.

Bao, L. and Cui, Y. 2005. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* **21:** 2185–2190.

Blader, M., Zhang, X.J., and Matthews, B.W. 1993. Structural basis of amino acid α-helix propensity. *Science* **260:** 1637–1640.

Chasman, D. and Adams, R.M. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation. *J. Mol. Biol.* **307:** 683–706.

Chou, P.Y. and Fasman, G.D. 1974. Conformational parameters for amino acids in helical, -sheet, and random coil regions calculated from proteins. *Biochemistry* **13:** 211–222.

Cuff, J.A. and Barton, G.J. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34:** 508–519.

Dunbrack, R.L. 1999. Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.* **16:** 374–384.

Dwyer, M.A., Looger, L.L., and Hellinga, H.W. 2004. Computational design of a biologically active enzyme. *Science* **304:** 1967–1971.

Garnier, J., Osguthorpe, D.J., and Robson, B. 1978. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120:** 97–120.

Guo, J., Chen, H., Sun, Z.R., and Lin, Y.L. 2004. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins* **54:** 738–743.

Hua, S.J. and Sun, Z.R. 2001. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J. Mol. Biol.* **308:** 397–407.

Jones, D.T. 1999a. GenTHREADER: An efficient and reliable protein folding recognition method for genomic sequences. *J. Mol. Biol.* **287:** 797–815.

Jones, D.T. 1999b. Protein secondary structure prediction based on position-specific score matrix. *J. Mol. Biol.* **292:** 195–202.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hudrogen bonded and geometrical features. *Biopolymers* **22:** 2577–2637.

Kim, H. and Park, H. 2003. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng.* **16:** 553–560.

Koehl, P. and Levitt, M. 1999. A bright future for protein structure prediction. *Nat. Struct. Biol.* **6:** 108–111.

Lim, V.I. 1974. Algorithms for prediction of α-helices and structural regions in globular proteins. *J. Mol. Biol.* **88:** 872–894.

Matthews, B.W. 1975. Comparison of the prediction and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405:** 442–451.

Padmanabhan, S., Marquesee, S., Ridgeway, T., Laue, T.M., and Baldwin, R.L. 1990. Relative helix-forming tendencies of nonpolar amino acids. *Nature* **344:** 268–270.

Presta, L.G. and Rose, G.D. 1988. Helix signals in proteins. *Science* **240:** 1632–1641.

Qian, N. and Sejnowski, T.J. 1988. Predicting the neural network models. *J. Mol. Biol.* **202:** 865–884.

Qin, S.B., He, Y., and Pan, X.M. 2005. Predicting protein secondary structure and solvent accessibility with an improved multiple linear regression method. *Proteins* **61:** 473–480.

Richardson, J.S. and Richardson, D.C. 1988. Amino acid preference for specific locations at the ends of helices. *Science* **240:** 1648–1652.

Rost, B. 1997. Protein fold recognition by prediction based threading. *J. Mol. Biol.* **270:** 1–10.

Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232:** 584–599.

Rost, B., Sander, C., and Schneider, R. 1994. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **235:** 13–26.

Russell, R.B. and Barton, G.J. 1993. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.* **234:** 951–957.

Russell, R.B., Copley, R.R., and Barton, G.J. 1996. Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259:** 349–365.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pochaud, M., Topf, M., and Sali, A. 2004. A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol.* **14:** 313–324.

Thiel, K.A. 2004. Structure-aided drug design's next generation. *Nat. Biotechnol.* **22:** 513–519.

Wang, G. and Dunbrack, J.R. 2003. PISCES: A protein sequence-culling serve. *Bioinformatics* **19:** 1589–1591.

Watson, J.D., Laskowski, R.A., and Thornton, J.M. 2005. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **15:** 275–284.

Zemla, A., Venclovas, C., Fidelis, K., and Rost, B. 1999. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* **34:** 220–223.