BIOINFORMATICS

Vol. 20 no. 10 2004, pages 1612–1621
doi:10.1093/bioinformatics/bth136

Protein backbone angle prediction with machine learning approaches

Rui Kuang[1], Christina S. Leslie[1,4] and An-Suei Yang[2,3,4,*]

[1]Department of Computer Science, [2]Department of Pharmacology, [3]Columbia Genome Center and [4]Center for Computational Biology and Bioinformatics, Columbia University, West 168th Street, PH 7 W Room 318, New York, NY 10032, USA

Received on August 14, 2003; revised on December 19, 2003; accepted on January 5, 2004
Advance Access publication February 26, 2004

ABSTRACT

Motivation: Protein backbone torsion angle prediction provides useful local structural information that goes beyond conventional three-state (α, β and coil) secondary structure predictions. Accurate prediction of protein backbone torsion angles will substantially improve modeling procedures for local structures of protein sequence segments, especially in modeling loop conformations that do not form regular structures as in α-helices or β-strands.

Results: We have devised two novel automated methods in protein backbone conformational state prediction: one method is based on support vector machines (SVMs); the other method combines a standard feed-forward back-propagation artificial neural network (NN) with a local structure-based sequence profile database (LSBSP1). Extensive benchmark experiments demonstrate that both methods have improved the prediction accuracy rate over the previously published methods for conformation state prediction when using an alphabet of three or four states.

Availability: LSBSP1 and the NN algorithm have been implemented in PrISM.1, which is available from www.columbia.edu/~ay1/.

Contact: ay1@columbia.edu; cleslie@cs.cloumbia.edu

Supplementary information: Supplementary data for the SVM method can be downloaded from the Website www.cs.columbia.edu/compbio/backbone.

INTRODUCTION

Protein backbone torsion (Φ, Ψ) angles are highly correlated to protein secondary structures. The distribution of the Φ–Ψ angles in protein structures is mostly clustered around the alpha (centered at $\Phi = -60°$, $\Psi = -40°$), beta (centered at $\Phi = -120°$, $\Psi = 120°$) and L-alpha (centered at $\Phi = 60°$, $\Psi = 0°$) regions of the Ramachandran plot. α-helices and β-sheets consist of residues with backbone torsion angles distributed mostly in the alpha and beta Φ–Ψ angle regions, respectively. Backbone structures in the loop regions are not as regular as in α-helices and β-sheets and can have Φ–Ψ angles in any region of the Ramachandran plot. Figure 1a summarizes the relationship between backbone torsion angles and secondary structure by plotting the distributions of Φ–Ψ angles in the alpha, beta and loop regions.

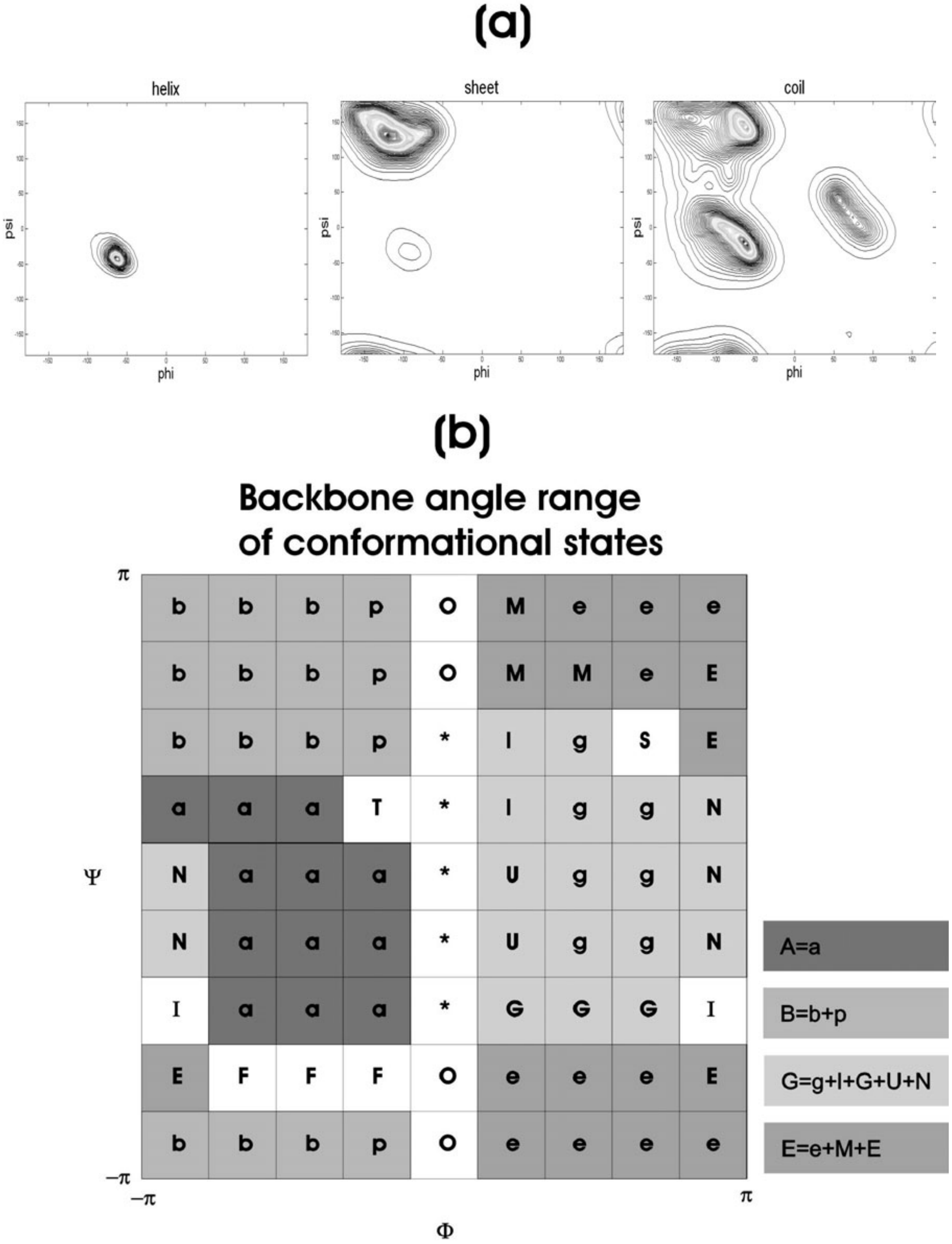
Both α-helices and β-strands are relatively straight in structure; the turning points in protein chains are made up of residues in loop regions. The loop residues in a protein chain play important roles as structural determinants in connecting regular secondary structure elements, leading to a specific protein folding topology for the protein structure (Richardson, 1981). Moreover, many loop residues involve enzymatic activities and protein–protein interactions, such as in antibody–antigen interactions. Local structural information provided by predictive algorithms will facilitate significantly the analysis of protein sequence–structure–function relationships.

Although many loop regions contain recurrent local structural motifs (see recent reviews, de Brevern *et al.*, 2002; Wojcik *et al.*, 1999), the large conformational variability makes the characterization and prediction of loop conformations one of the most challenging molecular modeling problems (see e.g. de Bakker *et al.*, 2003; Fiser *et al.*, 2000; Galaktionov *et al.*, 2001; Wojcik *et al.*, 1999; Xiang *et al.*, 2002). Accurate predictions of protein backbone torsion angles will improve further the prediction capacities of loop modeling procedures.

Three-state (α, β and coil) secondary structure prediction methods have reached ~80% accuracy (Petersen *et al.*, 2000; Pollastri *et al.*, 2002; Rost, 2001). Although these methods are powerful tools in protein structure prediction from amino acid sequences, three-state secondary structure predictions do not distinguish one loop conformation from the other. Backbone torsion angle predictions, on the other hand, provide local structural information that is useful in defining local structures for highly variable loop regions in amino acid sequences.

While three-state secondary structure prediction methods have been developed with increasing accuracy, the procedure for prediction of protein backbone torsion angles has received relatively little attention. The hidden Markov model HMMSTR, based on local sequence–structure correlations in

[*]To whom correspondence should be addressed.



b	b	b	p	O	M	e	e	e
b	b	b	p	O	M	M	e	E
b	b	b	p	*	l	g	S	E
a	a	a	T	*	l	g	g	N
N	a	a	a	*	U	g	g	N
N	a	a	a	*	U	g	g	N
I	a	a	a	*	G	G	G	I
E	F	F	F	O	e	e	e	E
b	b	b	p	O	e	e	e	e

A=a

B=b+p

G=g+l+G+U+N

E=e+M+E

Fig. 1. (**a**) Density plot of the joint distribution of Φ−Ψ angles in α-helix, β-sheet and loop (coil) region from left to right. The density at a point in the plot is estimated by the area of the disk that is centered at the point and contains exactly 100 observations. (**b**) Protein backbone conformational states. The backbone torsion angle ranges of the backbone conformational states (A, B, G and E) are defined in the right-hand side of the Ramachandran plot. The definitions of the conformational states shown in the Ramachandran plot were obtained from Oliva *et al*. (1997).

proteins, has been demonstrated to make backbone torsion angle predictions with significant accuracy (Bystroff *et al.*, 2000). This was the first protein backbone torsion angle prediction method benchmarked with a large set of test proteins (Bystroff *et al.*, 2000). HMMSTR uses an alphabet of 11 conformation states, 10 corresponding to Φ–Ψ angle regions and one for *cis*-peptide bonds.

Other authors have presented an extensive study of the predictability of different definitions and alphabet sizes of local structural states (de Brevern *et al.*, 2000). One of the goals of using a predicted local structure is improving the performance of profile HMMs for fold recognition (Karchin *et al.*, 2003). The focus of this work, however, is on the fold recognition problem rather than optimizing conformation state prediction.

Recently, a local structure prediction method based on a local structure-based sequence profile database (LSBSP1) has been devised and tested for prediction accuracy (Yang and Wang, 2003). Although the LSBSP1 local structure prediction procedure has been demonstrated to predict reasonably accurate local structures for sequence segments of nine consecutive residues based on the root mean square deviation (RMSD) measure, the backbone torsion angle prediction accuracy of the LSBSP1-based procedure has made only marginal progress in comparison with HMMSTR (Yang and Wang, 2003) (see also Table 2 for a comparison of the two published prediction results).

In this work, we report two novel protein backbone torsion angle prediction procedures. One extends the previous LSBSP1 prediction procedure by using an artificial neural network (NN) algorithm to process and summarize the prediction results. The other uses support vector machines (SVMs) to make protein backbone torsion angle predictions based on the protein sequence profile produced with PSI-BLAST (Altschul *et al.*, 1997) and the three-state secondary structure prediction from PSI-PRED (Jones, 1999). The goal of the prediction procedures is to predict the backbone conformational state of each residue in protein chains. Instead of using fine conformational states as in, e.g. HMMSTR, our prediction procedures focused on prediction accuracy based on four (A, B, G and E; Fig. 1b for definition) or three (A, B and G/E) conformational states. This goal reflects the general observation that, as shown in Figure 1a, there are only three major backbone conformational states for residues in proteins. Bystroff and Baker (1998) have demonstrated that the backbones of two eight-residue segments can be superimposed with RMSD less than 1.4 Å if none of the backbone torsion angles in one segment deviates from the corresponding torsion angles in the other segment by more than 120°. This indicates that accurate coarse-grained backbone conformational state prediction can be extremely useful in local structure prediction. This is particularly true for highly variable local structures as in the coil regions, which constitute slightly less than half the residues in proteins, and those residues for which the conventional three-state (α, β and coil) secondary structure prediction methods have provided essentially no local structural information.

The two prediction methods have been benchmarked against extensive testing cases. The results show that these two methods improve backbone torsion angle prediction accuracy over those for which results have been published previously.

METHODS

Protein backbone torsion angle conformational states

Following previous work, protein backbone torsion angles are mapped onto the Φ–Ψ plot (Oliva *et al.*, 1997; Yang and Wang, 2003). We divided the Φ–Ψ map into four major conformational states: A, B, G and E. Figure 1b shows the Φ–Ψ angle ranges of these conformational states. Almost all the residues in our training/testing proteins (see below for more details on the training/testing sets) have backbone torsion angles distributed in one of the four major conformational states. Only 0.38% of the residues have a backbone conformation outside the four major conformational states. In addition, 1.8% of the residues, most of which are the N- or C-terminal residues, are not assigned to any of the conformational states because they lack well-defined backbone atoms to calculate the Φ–Ψ angles. These residues were removed from our training/testing set.

Local structure-based sequence profile database LSBSP1

We only briefly describe the construction of the LSBSP1 database; more details can be found in a recently published work (Yang and Wang, 2003). The procedure has also been summarized in a flow chart available from our ftp server (ftp://ps7ayang.cpmc.columbia.edu/pub/LSBSP1flow1.pdf). The LSBSP1 database contains a total of 138 604 position-specific score matrices (PSSMs). Each PSSM has dimensions of 9 × 20. Each of the PSSMs was calculated from a structure-based multiple alignment constructed with a seed nine-residue segment from a protein structure. The seed segments in LSBSP1 are nine consecutive residue sequence segments from the non-redundant protein structures in PDB_SELECT_25 [PDB_SELECT_25; Hobohm *et al.* (1992) version Feb/2001, with no pairwise sequence identify >25%]. To construct a PSSM based on a seed segment, we first used the seed segment as a probe to search through the non-redundant proteins. Sequence segments from the non-redundant protein set that are identical in backbone conformational state (Fig. 1) and have the amino acid replacement scores above a threshold in comparison with the seed sequence were aligned to construct a preliminary local structure-based sequence profile for the seed segment. The sequence similarity was calculated with the structure-specific amino acid substitution matrices that we have developed to align distantly related protein pairs (Yang, 2002). This preliminary local structure-based sequence profile was then converted into a pre-PSSM in half-bit units with the Bayesian prediction pseudo-count method (Tatusov *et al.*, 1994):

$$W_{Ji} = 2\log_2\left(\frac{q_{Ji}}{p_i}\right), \tag{1}$$

where p_i is the background probability (Tatusov *et al.*, 1994) for amino acid type i and

$$q_{Ji} = \frac{C_{Ji} + (B + M - \sum_{k=1}^{20} C_{Jk})p_i}{M + B}, \tag{2}$$

where C_{Ji} is the number of amino acid type i that appear in the column J of the sequence profile. M is the number of rows in the sequence profile. The term $(B + M - \sum_{k=1,20} C_{Jk})$ in the numerator is the pseudo-count, where $B = M^{0.5}$ is considered adequate (Tatusov *et al.*, 1994). This pre-PSSM ($[W_{Ji}]$, a 20 × 9 matrix) was further refined by removing from the preliminary structure-based multiple alignment the sequence segments that did not score higher than a threshold (>15) with the pre-PSSM. The remaining set of segments form a refined local structure-based sequence profile, and the PSSM was recalculated and saved along with the sequence and structural information of the seed nine-residue segment in the LSBSP1 database. The procedure described above was applied to all the nine-residue sequence segments in the non-redundant protein structures to construct the LSBSP1 database in the PrISM.1 system.

Protein backbone torsion angle prediction with LSBSP1 database and artificial NN algorithm: the LSBSP1 + NN method

The goal of the LSBSP1+NN prediction procedure is to predict the backbone conformational state of the central residue in a nine-residue segment from a query protein sequence. A standard feed-forward back-propagation artificial NN (Rumelhart *et al.*, 1986) with single hidden layer is used in the torsion angle prediction procedure. The input layer has 216 input units, representing a window of nine consecutive residues in a protein chain. The hidden layer has 50 units. Architectures with more hidden layer units did not improve the performance of the prediction capacities. The output layer has three units, representing are A, B or G/E backbone conformational state of the central residue in the input nine-residue segment. We group the G and E states into one class in the prediction output because the E conformational state has only 1.7% of the total training cases. The scarcity of the training cases made the prediction for E conformational state by itself extremely difficult with the artificial NN and the SVM algorithm (see e.g. the results shown in Table 2). By grouping together the G and E training cases (6.4% of the training cases), we were able to train the NN algorithm to predict residues in the G/E state with reasonable accuracy.

The 216 input units are divided into nine groups, representing a window of nine consecutive residues. We use a nine-residue window for prediction input because the LSBSP1 database was constructed with nine-residue segments (see above). An orthogonal representation of an amino acid type requires 21 input units. The input unit that specifies the amino acid type is set to 1, while all other 19 input units are set to zero. The 21st input unit is set to 1 for residue positions in the nine-residue window outside the N- or C-terminus of the protein chain. The last three (the 22nd to the 24th) input units in each group are encoded with values summarized from the backbone torsion angle predictions with the LSBSP1 database. To make the LSBSP1-based backbone torsion angle prediction, we use each window of nine-residue segments in the query protein sequence as a probe sequence segment to match for nine-residue structure-based sequence profiles in the LSBSP1 database. All the LSBSP1 profiles for which the matching scores are more than a threshold of 20 and for which the secondary structure assignments are consistent with the PSI-PRED secondary structure prediction of the query sequence segment by more than 50% (Yang and Wang, 2003) are aligned to the query sequence to form a multiple alignment. Positions in the multiple alignment represent the predictions of the backbone conformational states, which can be A, B or G/E, for the corresponding residue in the query sequence. As the nine-residue window slides through the query protein chain one residue at a time, the backbone conformational state predictions accumulate in the multiple alignment. Each column in the multiple alignment shows all the backbone conformational state predictions for the corresponding residue in the query protein. After the predictions for all nine-residue windows in the query protein, the multiple alignment for all the predicted backbone conformational states is then converted into a PSSM with Equation (1) to calculate W_{Ji}, where i can be A, B or G/E. The W_{Ji} is a log-odds ratio in half-bit units for the backbone conformational state predictions versus random predictions based on background probabilities of the conformational states at position J in the query sequence; a large positive W_{Ji} value indicates that the residue position J is consistently predicted to be the i conformational state, and a negative W_{Ji} value indicates that the J position is consistently predicted to be the non-i conformational state. Three values within the range between 0 and 1 are calculated from W_{Ji} for each position J in the query protein sequence using the standard logistic function (Jones, 1999):

$$a_{Ji} = \frac{1}{1 + \mathrm{e}^{-W_{Ji}}}, \tag{3}$$

where i can be A, B or G/E. These a_{Ji} values are used in the 22nd to the 24th input units for each group to encode the information of predicted backbone conformational states.

An on-line back-propagation training procedure was used to update the weights connecting the nodes after each training pattern presentation (Rumelhart *et al.*, 1986). Each training pattern was randomly selected from a pool of nine-residue structural segments from training proteins. The input units are encoded based on the sequence of the nine-residue segment and the a_{Ji} values derived from Equation (3) (see above). The three output target values are set to 0 or 1 based on the backbone conformational state of the central residue of the nine-residue segment. The momentum value of 0.9 is used to prevent oscillation. The learning rate of 0.001 was found to be adequate in all the training procedures.

A total of 97 365 nine-residue segments derived from a non-redundant protein set were used as training and testing cases. These non-redundant proteins are from the PDB_SELECT_25 list (version Dec/2002) and are not related to any of the proteins in LSBSP1 (from PDB_SELECT_25 version Feb/2001) with a p-value threshold of 10^{-6} (or average sequence ID < 18%). The proteins in the non-redundant protein set are not related to each other by more than 25% sequence identity. It is important to have the training/testing proteins unrelated to the proteins in LSBSP1 because close homologs to the proteins in LSBSP1 tend to have a high accuracy in backbone torsion angle prediction with the LSBSP1-based method. The training and testing processes were carried out with 10-fold jackknife cross-validation: 10% of the segments were used as testing cases, while the remaining 90% segments were used in training; the processes repeat 10 times, each with a different 10% of the nine-residue segments. For each training–testing process, the training iteration was terminated when the prediction capacities of the network started to degrade on the 10% testing cases. The prediction accuracy was calculated by averaging the prediction accuracy rate for the testing cases over the 10-fold cross-validation processes. Prediction accuracies were calculated by comparing the true backbone conformational state with the predicted conformational state. The predicted conformational state was indicated by the output node with the largest output value.

Finally, the NN was again trained with all the 97 365 nine-residue segments and the trained network was tested with a set of recently released proteins that are not related to any of the training proteins and the proteins in LSBSP1. The same p-value threshold described in the previous paragraph was used to identify the test proteins in a recently released PDB_SELECT_25 list (version Apr/2003). The prediction accuracy for the new test proteins was then compared with the average accuracy from the 10-fold cross-validation. The comparison is to ensure that the LSBSP1+NN method has not been over-trained and the benchmarked accuracy is generally applicable to protein sequences of unknown structure.

The NN input nodes combine two types of information: the amino acid sequence of the query sequence segment (the first 21 input units) and the a_{Ji} values, the information on the predicted backbone conformational states (the 22nd to the 24th input units). The combination of these two types of information gives the optimum overall prediction accuracy of 78.2% in the 10-fold cross-validation results (Table 4 for details). Two tests have been performed to isolate the prediction effect of the two types of information: first, we used only the sequence information and removed the 22nd to the 24th input units from the input nodes and re-ran the 10-fold cross-validation (see above). The results showed an overall prediction accuracy rate of 61.5%. Second, we used only the a_{Ji} values, the information of predicted backbone conformational states (the 22nd to the 24th input units) and removed the first 21 input units from the input nodes and re-ran the 10-fold cross-validation. The results showed an overall prediction accuracy rate of 67.8%. These tests indicate that the latter information contributed more to the prediction accuracy and that the artificial NN algorithm does indeed combine these two types of information to make optimum prediction.

Protein backbone torsion angle prediction with SVM and PSI-PRED

We also developed an SVM method to predict the backbone conformation of the middle amino acid in a nine-residue sequence segment. Here again, we either classify four types of conformational states (A, B, E and G) or combine the two smallest states into a single class (E/G) for three-state classification. Our main effort is to design the feature representation of nine-residue amino acid segments. The three kinds of information we use for features are amino acid sequences, PSI-BLAST (Altschul *et al.*, 1997) profiles and secondary structures that are known (for training data) and that are predicted by PSI-PRED (Jones, 1999) (for test data).

Support vector machines Support vector machines are a family of algorithms for classification problems (Vapnik, 1998). Given a training dataset with m labeled training samples $(x_i,\ y_i)$ $(1 \leq i \leq m, x_i \in \Re^n$ and $y_i \in \{1, -1\})$, the goal is to learn a 'large margin' linear classifier f to discriminate between the two classes. Here, a linear classifier can be represented as a function:

$$f(x) = \langle w, x\rangle + b(w \in \Re^n, x \in \Re^n, b \in \Re). \qquad (4)$$

The decision boundary is a hyperplane $\langle w, x\rangle + b = 0$, and the margin for a training example x_i is the value $y_i f(x_i)$ (>0 if the example is correctly classified). A test example will be classified as positive if $f(x) > 0$, negative otherwise. The linear classification function can be learned with a soft margin SVM (Cristianini and Shawe-Talor, 2000), which incorporates a trade-off between maximizing the geometric margin and minimizing margin violations on the training set. An important property of the SVM optimization problem is that we can replace the inner product $\langle x_i, x_j\rangle$ by a kernel function $K(x, y)$; here, the kernel implicitly represents the inner product between feature vectors for pairs of input examples, $K(x, y) = \langle\Phi(x), \Phi(y)\rangle$, for some feature mapping Φ from the original input vector space to a feature space $\Re^N$ (or a Hilbert space). Typical kernels include polynomial kernels,

$K(x, y) = (\langle x, y\rangle + 1)^d$, or radial basis kernels,

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$

A one-versus-all classification is used to make multiclass predictions from trained binary SVM classifiers. We simply choose our prediction to be the class that gives the maximum margin for each test example x:

$$y = \arg_{\max}[f_j(x)],$$

where f_j is the SVM classifier for the j-th class.

We use the publicly available SVM$^{\text{light}}$ package to learn the binary classifiers (Joachims, 1999) in our experiments.

Binary encoding feature map A simple way of representing an amino acid sequence is through binary encoding. Here, each amino acid is represented as a 21-dimensional vector, where each dimension corresponds to one type of amino acid or to a special null character that is used to fill in the blank positions in window segments containing entries before the beginning or after the end of a protein sequence. The introduction of a null character helps to make predictions for boundary positions without affecting the overall accuracy. To encode an amino acid or blank at a particular position in the sequence, we put a positive constant, δ, in the corresponding entry in the feature vector and 0 in all other entries. A length 9-segment S is mapped to a 189 (9 × 21)-dimension vector by the feature map Φ_{binary}.

PSI-BLAST profile feature map Instead of using a binary encoding of the amino acid sequence, we can represent the sequence segment with its PSI-BLAST log-odds score profile. These scores are calculated as $[\ln(Q_i/P_i)]/\lambda_i$, where Q_i is the estimated probability for residue i to be found in that column, P_i is the background frequency of i and λ_i is a scale parameter. The way we construct the profile kernel is as follows: the PSSM is constructed for all protein sequences with the PSI-BLAST program running under the standard setup of PSI-PRED, and then these PSSMs are cut along with the sequence into 9 × 20 matrices (each of the nine positions encodes a probability distribution over 20 amino acids). For the blank positions at the beginning and the end of a protein sequence, 20 zeros are filled in to represent the background distribution. The feature map, Φ_{profile}, assigns to a length nine segment S the concatenation of nine 20-dimension vectors in the PSSM.

Predicted secondary structure feature mapping In addition to the protein sequence, another useful source of information we can use is the secondary structure. The secondary structure of training sequences is derived from the DSSP program (Kabsch and Sander, 1983). For testing sequences, secondary structures are predicted with the PSI-PRED program (Jones, 1999). As with binary encoding for amino acid sequences, we can represent a length nine segment as a binary encoding of nine positions, each of which has a 4-dimensional vector (three kinds of secondary structures plus the blank position). We denote the secondary structure feature map as Φ_{sec}.

Different feature representations can be combined by concatenation of feature vectors (direct product of vectors); we write, e.g. $\Phi_{\text{binary}} \times \Phi_{\text{sec}}$ for the direct product of binary and secondary structure feature maps.

Two datasets are used for evaluation of the SVM method: PDB_SELECT_25 and a modified version of the Dunbrack-culled PDB (Karchin *et al.*, 2003). This version of Dunbrack-culled PDB has a sequence identity cutoff of 20%, a resolution cutoff of 3.0 Å and a R-factor cutoff of 1.0 with fragments shorter that 20 residues removed. The Dunbrack-culled PDB dataset provides a more accurate non-redundant benchmark and allows us to compare the SVM performance with other results from the literature.

For the PDB_SELECT_25 dataset, we performed two sets of benchmark experiments for evaluation of the SVM classifiers. First we used proteins in the LSBSP1 database (PDB_SELECT_25 version Feb/2001) as the training set, and we tested on PDB_SELECT_25 (version Dec/2002) proteins that are not related to any proteins in the training set. For the second set of experiments, we performed 10-fold cross validation on PDB_SELECT_25 as described in the neural net approach above. The results produced from the two training/testing procedures were essentially identical, and we report only the 10-fold cross validation experiments below. Finally, we performed 3-fold cross validation experiments on the dunbrack-in-scop dataset, where the number of folds was chosen for consistency with previous published results (Karchin *et al.*, 2003) in order to allow comparison.

RESULTS AND DISCUSSION

Feature representations and kernel selection for SVM

We first report results on the PDB_SELECT_25 dataset. Table 1 compares the prediction accuracy between four different types of feature maps for SVM classification. The profile feature mapping outperforms the binary mapping by around 6% and is outperformed by the secondary structure feature mapping by 3%. A breakdown of results by conformation state suggests that both the profile and secondary structure feature maps have good results in the alpha and beta regions but are less helpful in the loop regions; in particular, the binary encoding is more successful than profile or secondary structure for prediction of the E/G state, which almost always occurs in loops. These results are understandable, given the strong correlation between secondary structure and local conformation for α-helices and β-strands and given that profiles help predict these regular secondary structures; for loop regions, secondary structure information is complementary to local conformation and not directly useful for prediction. By integrating the secondary structure with the profile or binary mapping, both are improved significantly to approximately

Table 1. Prediction results using 10-fold cross-validation by SVM classification with various feature maps (PDB_SELECT_25 dataset)

	Test case	Φ_{binary} (%)	Φ_{profile} (%)	Φ_{sec} (%)	$\Phi_{\text{binary}} \times \Phi_{\text{sec}}$ (%)	$\Phi_{\text{profile}} \times \Phi_{\text{sec}}$[a] (%)	$\Phi_{\text{profile}} \times \Phi_{\text{true\_sec}}$[b] (%)	$\Phi_{\text{profile}} \times \Phi_{\text{pred\_sec}}$[c] (%)
A	50 689	77.10	77.30	71.30	80.50	82.00	83.41	68.31
B	40 268	51.00	65.40	87.22	78.80	79.00	86.78	89.51
G/E	4760	54.20	42.60	7.76	48.50	51.10	52.68	54.28
Total	97 365	64.80	70.10	73.70	77.70	78.70	82.78	76.15

The combination of Φ_{binary} and Φ_{profile} is not included since Φ_{profile} is a richer representation of Φ_{binary}. The definitions of these feature mappings are described in the Methods section.

[a] $\Phi_{\text{profile}} \times \Phi_{\text{sec}}$: uses true secondary structure for training and predicted secondary structure for testing.

[b] $\Phi_{\text{profile}} \times \Phi_{\text{true\_sec}}$: uses true secondary structure for both training and testing.

[c] $\Phi_{\text{profile}} \times \Phi_{\text{pred\_sec}}$: uses predicted secondary structure for both training and testing.

Table 2. Comparison of the prediction accuracies from the SVM method against the two previously published results

	SVM $\Phi_{\text{profile}} \times \Phi_{\text{sec}}$		LSBSP1+consensus prediction (consensus level=1)		HMMSTR	
	Test cases	Accuracy (%)	Test cases	Accuracy (%)	Test cases	Accuracy (%)
A	50 689	82.5	17 466	82.7	$A' = 9625$	82.0
B	40 268	79.6	12 732	71.2	$B' = 7749$	71.6
G	4760	32.9	1491	32.8	$G' = 837$	15.5
E	1648	0.3	461	6.5	$E' = 199$	22.6
Total	97 365	77.3	32 150	74.6	18 410	74.0

The training and testing of the SVM method are described in the Methods section. The LSBSP1+consensus data are reproduced from Table 1 of Yang and Wang (2003). The HMMSTR data are reproduced from Table 5 of Bystroff *et al.* (2000). The definitions of the conformational states in HMMSTR predictions are not completely identical to the definitions of the A, B, G and E conformational states shown in Figure 1. For comparison, $A' = H + G$, $B' = B + E + d + b + e$, $G' = L + l$ and $E' = x$; the backbone conformational states on the right-hand side of the equations were defined by Bystroff *et al.* (2000). The A′, B′, G′ and E′ states are approximately equivalent to the A, B, G and E states defined in Figure 1. The test cases listed under the HMMSTR predictions are the residues in the A′, B′, G′ and E′ backbone conformational state, respectively.

the same accuracy. The similarity in performance could be explained by the fact that the predicted secondary structures used for testing segments are also derived from profiles in PSI-PRED. Finally, in the last two columns of Table 1, we use true secondary structure and predicted secondary structure for both training and testing. We find that using true secondary structure improves results dramatically for conformation state A but slightly degrades the performance for E/G, again showing that secondary structure information is not predictive of conformation states in loop regions.

All the results shown in the table are produced with linear kernels. We also performed experiments using polynomial kernels and RBF kernels on the combined profile feature map and secondary structure feature map and obtained a slight improvement of about 1% in each case (see supplementary Website for results).

Comparison with other methods

Table 2 compares the SVM results of linear kernels with profile and secondary structure prediction with results reproduced from our previous work (LSBSP1+consensus prediction) (Yang and Wang, 2003) and the published HMMSTR prediction accuracy (Bystroff *et al.*, 2000). Since HMMSTR uses a larger alphabet of 11 conformation states, we facilitate the comparison by grouping the 10 states that correspond to Φ–Ψ angle ranges into four (A, B, G and E) states. (The final state corresponds to *cis*-peptide bonds rather than the backbone angle state, and we omit this small set of residues in the comparison.) One might consider this comparison unfair with HMMSTR, which is trying to perform a more difficult multi-class prediction problem; however, if the predictions, when grouped into this coarser four-state setting, are less accurate than four-state prediction methods, one could argue that a smaller alphabet is better justified. We present the SVM prediction accuracy in Table 2 by making explicit G and E state predictions. The separation of the two states slightly decreases the overall prediction accuracy (Tables 1 and 2). Even so, the comparisons are not straightforward because the three methods were benchmarked with different sets of test cases. One interesting negative result is that the SVM method performs poorly on the smallest class (E) compared with both LSBSP1+consensus and HMMSTR, indicating perhaps that the simple feature representation is not expressive enough to detect this class. However, overall, the SVM prediction clearly outperforms the previous methods by a few percent. More importantly, the SVM predictions were benchmarked with a much larger set of test cases: the SVM test set is 3-fold larger than the test set used in benchmarking the LSBSP1+consensus method and is 5-fold larger than the test set used in benchmarking the HMMSTR method. Based on the

Table 3. Position-wise predicted conformation states are tabulated according to true values

	dunbrack-in-scop					PDB_SELECT_25			
	A^{pred}	B^{pred}	G/E^{pred}	Total		A^{pred}	B^{pred}	G/E^{pred}	Total
A^{obs}	125 244	25 008	2835	153 087	A^{obs}	41 571	8165	953	50 689
B^{obs}	21 921	102 542	3344	127 807	B^{obs}	7385	31 803	1080	40 268
G/E^{obs}	4019	4366	11 004	19 389	G/E^{obs}	1570	1565	3273	6408
Total	151 184	131 916	17 183	300 283	Total	50 526	41 533	5306	97 365

Table 4. Comparison of the prediction results from SVM with linear kernel and LSBSP1+NN methods

	All residues			Loop residues only		
	Test cases	SVM (%)	LSBSP1+ NN (%)	Test cases	SVM (%)	LSBSP1+ NN (%)
A	50 689	81.4	81.9	1 4262	61.4	61.3
B	40 268	79.5	78.0	1 7109	71.0	70.1
G/E	6408	52.2	50.1	5 082	55.2	47.1
Total	97 365	78.7	78.2	36 453	65.1	63.5

Here, the loop residues are the residues in the coil regions that connect two flanking regular secondary structure elements in the test proteins. Coil residues in the N- and C-termini are not included. Regular secondary structure elements were defined by the DSSP program: α-helices are regions with at least four consecutive H residues characterized by the DSSP program, and β-strands are regions with at least two consecutive E residues characterized by the DSSP program. The test cases are obtained from the PDB_SELECT_25 dataset.

large benchmark, we expect that the SVM prediction accuracy will generalize to backbone torsion angle predictions for protein sequences of unknown structure.

Table 3 further shows the SVM prediction (Table 2) details in a number of predicted residues. Prediction errors are shown as the off-diagonal numbers. The SVM prediction has been validated with the same procedure and parameters but with a different protein set: the dunbrack-in-scop dataset. The results are compared side-by-side in Table 3. The differences are comparable with <1% (Tables 4 and 5), indicating that the benchmark results shown in this work are relatively insensitive to the choice of the test and/or the training datasets.

Prediction accuracy in loop region

The trained LSBSP1+NN was tested with proteins in the most recent PDB_SELECT_25 list (April 2003). All new non-redundant proteins that are not related to the training proteins for the LSBSP1+NN method are used to test the prediction method. The results are summarized as follows: 14 898 residues are in A backbone conformational state, and 81.5% are correctly predicted; 11 462 residues are in B backbone conformational state, and 76.6% are correctly predicted; 2065 residues are in G/E backbone conformational state, and 45.6% are correctly predicted. Overall, 77.0% of the residues are correctly predicted. The test results are similar to the prediction accuracy derived from the 10-fold cross-validation (Table 4), indicating that the prediction capacities of the LSBSP1+NN method shown in Table 4 have not been over-trained.

Table 5. SVM predictions for test cases from the dunbrack-in-scop dataset

	All residues		Loop residues only	
	Test cases	SVM (%)	Test cases	SVM (%)
A	153 087	81.8	47 497	60.9
B	127 807	80.2	58 981	71.4
G/E	19 389	56.8	16 756	58.2
Total	300 283	79.5	123 234	65.6

Table 4 compares the prediction performance of the SVM method and the LSBSP1+NN method. The comparison shows that the SVM prediction is slightly better than the LSBSP1+NN method but by <1%. To compare further the two methods on more level ground, we trained the SVM methods with the proteins used in constructing the LSBSP1 database and then tested on the 97 365 test cases. The prediction accuracies are almost identical to the results shown in Tables 1 and 2. We conclude that the SVM method is more accurate than the LSBSP1+NN method by a small margin.

Finally, to verify our results, one additional SVM experiment is done on the dunbrack-in-scop dataset with the profile and secondary structure feature map using 3-fold cross-validation. The results shown in Table 5 are slightly better than those on PDB_SELECT_25, probably due to cleaner structural data in the second dataset. [We can compare the overall accuracy of 79.5% on this dataset for three-state prediction and 78.4% for four-state prediction with previous results obtained using neural nets (Karchin *et al.*, 2003) that found an accuracy of 58.8% on a 10-state conformation alphabet and 64.9% on a four-state alphabet.]

The usefulness of the backbone torsion angle prediction resides in prediction of local structures in protein sequences of an unknown structure. Our prediction assessments have shown that the backbone torsion angle predictions for α-helix and β-strand residues are highly accurate (89% on average and 59% baseline; baseline is evaluated with a random prediction based on the conformational state population in the training proteins). However, local structures in α-helices and β-strands

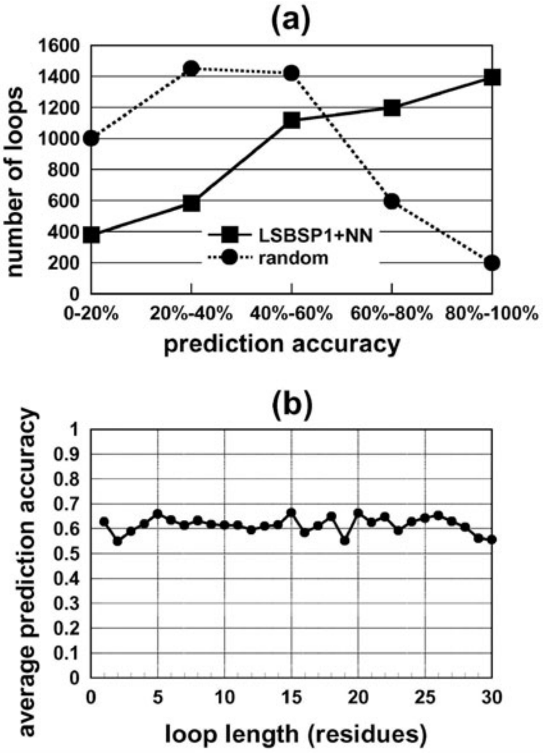


Fig. 2. (**a**) Distribution of protein loops against backbone torsion angle prediction accuracy rate. A total of 4673 loops with 2–10 residues were used as test loops, which have not been used in the training of the LSBSP1+NN prediction methods. The prediction accuracy for each loop was calculated as the ratio of the correctly predicted residues over the residues in the loop. The distribution of the loops against the prediction accuracy rate is shown in the solid line. The dotted line shows the distribution of the same set of loops against a random prediction accuracy rate. The random prediction accuracy rate for each loop was calculated with random predictions for the loop residues based on the background probabilities for the conformational states. (**b**) Average prediction accuracy rate plotted against loop length. The dataset shown in this plot is the same as the dataset shown in Table 3. In this figure, the dataset was subdivided into groups based on the loop length, and the average accuracy rate for each subgroup was calculated by averaging over residues in the subgroup.

are relatively regular, and hence torsion angle predictions are not particularly informative in providing additional structural information in comparison with three-state secondary structure predictions. In contrast, local structures in loop regions are highly variable. Tables 4 and 5 show that it is more difficult to make accurate backbone torsion angle predictions for loop residues. Still, the prediction accuracies with both the prediction methods are far greater than the 39.0% baseline calculated with random assignment of backbone conformational states to residues in the test set with background probabilities.

Figure 2 analyzes the backbone torsion angle prediction capacities on loop residues using LSBSP1+NN. The solid line in Figure 2a shows that 30% of the loops (1394 out of 4673 loops) with 2–10 residues can be predicted with high accuracy (more than 80% residues in the loop are predicted correctly). In contrast, the dotted line in Figure 2a indicates that random prediction can only produce high accuracy prediction for 4% of the loops. Figure 2b shows that the average backbone torsion angle predictions are relatively insensitive to the loop length. Together, Figure 2a and b indicate that some of the local conformations of residues in protein loop regions are recognizable from their local amino acid sequences and that these local structural motifs are not restricted to short segments of residues connecting two secondary structural elements.

CONCLUSION

Optimally combining available information is one of the difficult challenges in knowledge-based protein structure prediction procedures. The SVM and the LSBSP1+NN methods are fundamentally different in using structural and sequence information derived from the database of known protein structures. A similar prediction accuracy rate suggests that we are approaching the prediction limit for the prediction methods with the current knowledge in the protein structural database. The results show that backbone torsion angles in regular secondary structure elements can be predicted with high accuracy, and backbone torsion angles in loop residues are more difficult to predict. However, the current prediction accuracy for loop backbone torsion angles suggests that some of the structural motifs in loop regions can be recognized with a high accuracy from local sequence information alone. The predictions will provide useful information for modeling protein structures from protein sequences. With the upcoming expansion of protein structural databases, we expect to improve further our prediction capacities in recognizing protein local conformations from local sequence segments.

ACKNOWLEDGEMENTS

This work was partially supported by the NIH (PAR-00-102). Planning Grant: National Programs of Excellence in Biomedical Computing for Computational Biology and Bioinformatics. C.L. was also supported by an Award in Informatics from PhRMA Foundation.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bystroff,C., Thorsson,V. and Baker,D. (2000) HMMSTR: a hidden Markov model for local sequence–structure correlation in proteins. *J. Mol. Biol*., **301**, 173–190.

Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence–structure motif. *J. Mol. Biol*., **281**, 565–577.

Cristianini,N. and Shawe-Talor,J. (2000) *A Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

de Bakker,P.I., DePristo,M.A., Burke,D.F. and Blundell,T.L. (2003) *Ab initio* construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins*, **51**, 21–40.

de Brevern,A.G., Valadie,H., Hazout,S. and Etchebest,C. (2002) Extension of a local backbone description using a structural alphabet: a new approach to the sequence–structure relationship. *Protein Sci*., **11**, 2871–2886.

de Brevern,A.G., Etchebest,C. and Hazout,S. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, **41**, 271–287.

Fiser,A., Do,R.K. and Sali,A. (2000) Modeling of loops in protein structures. *Protein Sci*., **9**, 1753–1773.

Galaktionov,S., Nikiforovich,G.V. and Marshall,G.R. (2001) *Ab initio* modeling of small, medium, and large loops in proteins. *Biopolymers*, **60**, 153–168.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein datasets. *Protein Sci*., **1**, 409–417.

Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press.

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

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karchin,R., Cline,M., Mandel-Gutfreund,Y. and Karplus,K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins: Struct. Func. Genet*., **51**, 504–514.

Oliva,B., Bates,P.A., Querol,E., Aviles,F.X. and Sternberg,M.J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol*., **266**, 814–830.

Petersen,T.N., Lundegaard,C., Nielsen,M., Bohr,H., Bohr,J., Brunak,S., Gippert,G.P. and Lund,O. (2000) Prediction of protein secondary structure at 80% accuracy. *Proteins*, **41**, 17–20.

Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.

Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem*., **34**, 167–339.

Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol*., **134**, 204–218.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci., USA*, **91**, 12091–12095.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Springer.

Wojcik,J., Mornon,J.P. and Chomilier,J. (1999) New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J. Mol. Biol*., **289**, 1469–1490.

Xiang,Z., Soto,C.S. and Honig,B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl Acad. Sci., USA*, **99**, 7432–7437.

Yang,A.S. (2002) Structure-dependent sequence alignment for remotely related proteins. *Bioinformatics*, **18**, 1658–1665.

Yang,A.S. and Wang,L. (2003) Local structure prediction with local structure-based sequence profiles. *Bioinformatics*, **19**, 1267–1274.