A KNOWLEDGE-BASED APPROACH TO PROTEIN LOCAL STRUCTURE PREDICTION[*]

CHING-TAI CHEN, HSIN-NAN LIN, KUN-PIN WU, TING-YI SUNG[†] AND WEN-LIAN HSU

Institute of Information Science, Academia Sinica, Taipei, Taiwan
{caster, arith, kpw, tsung, hsu}@iis.sinica.edu.tw

Abstract

Local structure prediction can facilitate ab initio structure prediction, protein threading, and remote homology detection. However, previous approaches to local structure prediction suffer from poor accuracy. In this paper, we propose a knowledge-based prediction method that assigns a measure called the local match rate to each position of an amino acid sequence to estimate the confidence of our approach. To remedy prediction results with low local match rates, we use a neural network prediction method. Then, we have a hybrid prediction method, HYPLOSP (HYbrid method to Protein LOcal Structure Prediction) that combines our knowledge-based method with a neural network method. We test the method on two different structural alphabets and evaluate it by Q_N, which is similar to Q_3 in secondary structure prediction. The experimental results show that our method yields a significant improvement over previous studies.

1. Introduction

Protein local structure is a set of protein peptides that share common physiochemical and structural properties. Researchers usually cluster protein fragments by different local criteria, such as solvent accessibility, residue burial [8], and backbone geometry [9], and represent these fragment clusters by an *alphabet*, called a *local structure alphabet* (also known as a *structural alphabet* or *structural motifs*) [9]. Local structure prediction predicts the local structure of a protein fragment expressed by a *letter* of the structural alphabet from its amino acid sequence. Local structure prediction helps improve the performance of both profile and threading/fold-recognition methods for tertiary structure prediction [3, 6].

Various local structure libraries have been constructed, some of which focus on the reconstruction of protein tertiary structures. In such libraries, the number of letters in each structural alphabet is large, e.g., 100 in Unger et al. [16], 40 and 100 in Micheletti et al. [12], 100 in Schuchhardt et al. [15], and 25-300 with fragment lengths from 5 to 7 in Kolodny et al. [10]. Though large alphabet sets can better approximate protein tertiary

[*] This work is partially supported by the Thematic program of Academia Sinica under Grant 94B003 and by the National Science Council, Taiwan under Grant NSC94-2213-E-001-008.

[†] Correspondence to: Ting-Yi Sung, Institute of Information Science, 128 Sec. 2, Academia Rd, Nankang, Taipei, 115 Taiwan. E-mail: tsung@iis.sinica.edu.tw

structures, predicting protein local structures from amino acid sequences is much more challenging.

Thus, smaller structural alphabet sets have been proposed, and their associated local structure libraries have been constructed. Moreover, local structure prediction algorithms using these libraries have been developed. Bystroff et al. [2] generated a library called *I-site*, which contains 13 structural motifs of different length. Prediction is based on profile-profile alignment between each structural motif and the PSI-BLAST [1] result of the input sequence. They further proposed a new model, HMMSTR, to improve prediction accuracy. The *structural alphabet of HMMSTR,* denoted by *SAH*, is used in this paper to test our method. In [5], de Brevern et al. built their library, called *Protein Blocks* (*PB*), by clustering 5-mer protein fragments into a structural alphabet of 16 letters according to a torsion angle space. They then used a Bayesian probabilistic approach for prediction. Karchin et al. [9] constructed an *STR* library, in which the structural alphabet consists of 13 letters obtained from eight secondary structure states by dividing β-sheet into 6 types. They used a hidden Markov model (HMM) for local structure prediction.

The performance of local structure prediction depends on the definition of the underlying structural alphabet and the prediction algorithm. However, there is no unifying performance measure for evaluation. Bystroff et al. regard a local structure correctly predicted if the *MDA* (*Maximum Deviation of backbone torsion Angle*) of an eight-residue window is less than 120 degrees to their native structure [2, 4]. However, a straightforward evaluation measure, Q_N, is used in [5], which is similar to Q_3 used in secondary structure prediction. Q_N compares the predicted results with the encoded structural letter sequence, where N is the alphabet size, for example, N= 10 for SAH. Specifically, Q_N of a protein, *p,* is calculated as follows:

$$Q_N = \frac{\text{the number of residues of } p \text{ correctly predicted}}{\text{the number of all residues of } p} \times 100 \text{ .}$$

In [5], the accuracy of Q_N is 40.7%. Q_N is also used by Karchin et al. in [8, 9]. Thus in this paper, we use Q_N to evaluate different prediction methods, as discussed in Section 3.

Previous studies indicate that accuracy is the main difficulty in local structure prediction. In this study, we propose a local structure prediction algorithm to improve the current accuracy. The proposed method is alphabet-independent, i.e., it is not designed for a specific structural alphabet. Furthermore, we use Q_N to evaluate the method and demonstrate its capability.

2. Methods

We propose a knowledge-based prediction method and use a measure called the *local match rate* to estimate the prediction confidence. The *local match rate* represents the amount of information at each position of an amino acid sequence acquired from the knowledge base. Empirically, by this method, a high match rate results in high prediction accuracy. To improve the low prediction accuracy of low-match-rate positions, we propose a neural network prediction method that also provides confidence from its output. We propose a hybrid method, called HYPLOSP (HYbrid method to Protein LOcal Structure Prediction), which combines the results of these two methods according to the local match rate and neural network confidence.

2.1 Knowledge-based approach

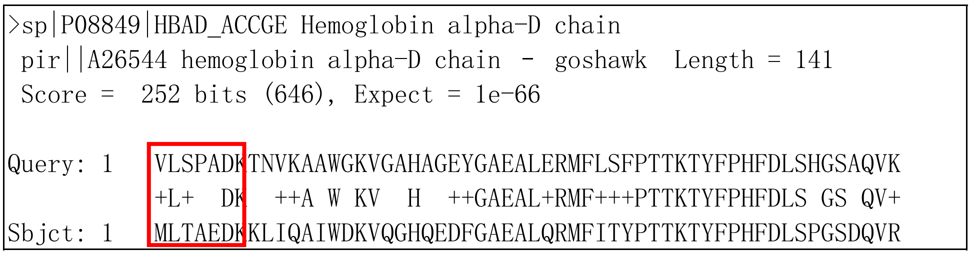
2.1.1 Construction of a sequence-structure knowledge base (SSKB)

Our knowledge base contains both local structure information and secondary structure information about peptides. The former is expressed by a structural alphabet (discussed in Section 3.1), and the latter is obtained from the DSSP database. For ease of exposition, we assume that we use a protein dataset with a known secondary structure and local structure based on a given structural alphabet.

The strength of a knowledge base depends on its size. Since the number of proteins with known secondary structures is relatively small, we amplify our knowledge base by finding homologous proteins to inherit the structural information of the given dataset. To this end, we utilize PSI-BLAST [1] to find proteins remotely homologous to a protein with a known structure, referred to as a *Query protein* in the PSI-BLAST output. While using PSI-BLAST, we set the parameter j to 3 (3 iterations), e to 10 (E-value < 10), and use the NCBI nr database as the sequence database. For each Query protein, PSI-BLAST generates a large number of homologous protein segments as well as their pairwise alignment called *high-scoring segment pairs*, *HSPs*. In each HSP, the counterpart sequence aligned with the Query protein is denoted by *Sbjct* in the PSI-BLAST output.

Performing PSI-BLAST on a Query protein, we obtain a large set of HSPs. Now we need to find the peptides in the Sbjct protein of each HSP that are similar to those of the Query protein so that similar peptides can inherit the structural information of the Query protein. We use a sliding window of length w to determine the peptides. In our experiments, we choose $w = 7$, which yields the best results among other lengths. Let p and q denote a pair of peptides in Query and Sbjct, respectively. We define the *similarity score*, S, between p and q as the number of positions that are identical or have a "+" sign in the sliding window. We call p and q *similar* if $S \geqq 5$. For the peptide q, which is similar to p, we define the *voting score* of q with respect to p as $(S \times A) / w$ to measure the confidence level for q to inherit the structural information of p, where A denotes the alignment score of the HSP reported in PSI-BLAST output. If p and q do not contain any gap, we add the record (q, the secondary structure of p, local structure of p, and voting score of q) to the knowledge base, in addition to the record (p, the secondary structure of p, local structure of p, and voting score of p). Otherwise, we discard this pair of similar peptides.

Figure 1 shows part of an HSP. The pair of peptides marked by a box have a similarity score of 5 and are thus considered similar. The voting score of the peptide in Sbjct with respect to that in Query is 180 (= 5×252 / 7). Suppose the structural alphabet is a set of {A, B, C, D, E, F}, and the secondary structure and local structure of peptide VLSPADK are CCHHHHC and BBEEECD, respectively. Since this peptide pair does not contain any gap, the record (MLTAEDK, CCHHHHC, BBEEECD, 180) is added to the knowledge base as shown in Table 1(a). Note that a peptide may inherit structural information from multiple peptides; if this is the case, we simply add new records to the existing record. For example, suppose the peptide MLTAEDK also inherits the structural information from another similar peptide with a voting score of 65. Then, the record of MLTAEDK in the knowledge base is updated, as shown in Table 1(b).



```
>sp|P08849|HBAD_ACCGE Hemoglobin alpha-D chain
 pir||A26544 hemoglobin alpha-D chain  -  goshawk  Length = 141
 Score =  252 bits (646), Expect = 1e-66

Query: 1    VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVK
            +L+  DK  ++A W KV  H  ++GAEAL+RMF+++PTTKTYFPHFDLS GS QV+
Sbjct: 1    MLTAEDKKLIQAIWDKVQGHQEDFGAEALQRMFITYPTTKTYFPHFDLSPGSDQVR
```

Figure 1. An example of HSPs found by PSI-BLAST

Table1. An example of knowledge base entries

Peptide fragment		M	L	T	A	E	D	K
Secondary Structure	H	0	0	180	180	180	180	0
	E	0	0	0	0	0	0	0
	C	180	180	0	0	0	0	180
Structural Alphabet	A	0	0	0	0	0	0	0
	B	180	180	0	0	0	0	0
	C	0	0	0	0	0	180	0
	D	0	0	0	0	0	0	180
	E	0	0	180	180	180	0	0
	F	0	0	0	0	0	0	0

(a)

Peptide fragment		M	L	T	A	E	D	K
Secondary Structure	H	0	0	180	180	245	245	65
	E	0	0	0	0	0	0	0
	C	245	245	65	65	0	0	180
Structural Alphabet	A	0	0	0	0	0	0	0
	B	245	245	0	0	0	0	0
	C	0	0	0	0	0	180	0
	D	0	0	0	0	0	0	180
	E	0	0	180	180	180	65	65
	F	0	0	65	65	65	0	0

(b)

2.1.2 Local structure prediction based on SSKB

Using the constructed knowledge base, SSKB, our knowledge-based local structure prediction method is comprised of the following steps:

Step 1: Use PSI-BLAST to find all HSPs with respect to a target protein (i.e., a protein whose secondary and local structures are unknown and to be predicted).

Step 2: Use similar peptides found in SSKB to vote for the local structure of each amino acid in the target protein.

In Step 1, the parameters and the sequence database used in PSI-BLAST are the same as those used in knowledge base construction. To define similar peptides stated in Step 2, we use the same sliding window length of 7, same voting score, and the same similarity score of 5 with no gap to define similar peptides as before. We match all peptides of the target protein and their similar peptides against SSKB. We then use the local structure information of the matched peptides in SSKB to vote for the local structure of the target protein. Let *p* be a peptide of the target protein. Throughout this section, we assume the structural alphabet is a set of $\{A_1, A_2, \ldots, A_n\}$. We associate each position, *x*, in *p* with *n* variables given by $\mathrm{V}^{\mathrm{p}}_{\mathrm{i}}$, where $i = 1, \ldots, n$. Let *q* be *p*'s counterpart peptide with similarity score *S* in an HSP with an alignment score *A*. If *q* is similar to *p* and can be found in SSKB, the voting score of *q* is added to that of *p*, which is updated as follows: For each position, *x*, compute

$$V^p_i(x) \leftarrow V^p_i(x) + V^q_i(x) \times (S \times A) / 7,\ i=1,\ldots n,$$

and repeat the above calculation for all similar peptides. The local structure of *x* in *p* is given by the letter corresponding to $Max\ \{V^p_1(x), V^p_2(x), \ldots, V^p_n(x)\}$.

2.2 Neural network method

2.2.1 Neural network architecture

We use a standard feed-forward back-propagation neural network [14] with a single hidden layer. The number of hidden units in the hidden layer is 35, which has been found to be the most effective number in our training stage.

Taking each protein in the training set or testing set, we partition it into peptides by a sliding window of length 7. We also perform PSI-BLAST query to obtain the profile of the sequence, which is the Position-Specific Scoring Matrix (PSSM). Our neural network takes each peptide as input. Specifically, the input vector consists of the peptide's corresponding segment of PSSM as well as its secondary structure. So, the length of each input vector is 161, i.e., 7×20 for PSSM and 7×3 for the secondary structure. The output reports the results corresponding to the amino acid located at the center of the peptide (called the "peptide center" for short). Specifically, the output is a vector of size *n*, i.e., the size of the underlying structural alphabet, and each entry represents the confidence score of the peptide center to be assigned a specific alphabet letter.

2.2.2 Training procedure

An online back-propagation training procedure is used to optimize the weights of the network, whereby the weights are randomly initialized and updated with each input vector. The learning parameters of the hidden layer and the output layer are 0.075 and 0.05, respectively. In addition, the sum of square errors is used during back propagation.

In the training stage, the secondary structure information in the input vector is given by the *true* secondary structure from the DSSP database. The desired output is a vector with 1 at the entry corresponding to the *true* alphabet letter of the peptide center, and 0 elsewhere.

2.2.3 Local structure prediction based on a neural network

Our neural network prediction method consists of two steps:

Step 1: Perform secondary structure prediction on a target protein.

Step 2: Use the neural network method to predict the local structure of each amino acid in the target protein.

Unlike proteins in the training set, target proteins do not have secondary structure information. Thus, in Step 1 we use HYPROSP II [7] to predict the secondary structure. The predicted secondary structure and PSSM, extracted by a sliding window of length 7, constitute the input to the trained neural network. The letter with the highest confidence score in the output vector is then considered to be the local structure of the peptide center. Step 2 is repeated to predict all amino acids in the target protein.

2.3 Hybrid mechanism

Our knowledge-based method and the neural network method have different strengths. To better utilize their respective strengths, we propose a hybrid mechanism that uses the local match rate, to combine the two methods. At each position, *x*, of the target protein, we obtain from HSPs a set of similar peptides, *Q*(*x*), that contains the position *x*. The local match rate is defined as follows:

$$\text{Local Match Rate(x)} = \frac{|\mathrm{Q(x)} \cap \mathrm{SSKB}|}{|\mathrm{Q(x)}|} \times 100\% \,.$$

The local match rate represents the amount of information for each position *x* that can be extracted from the knowledge base. It is possible for the target protein to have a high local match rate in some positions and a low local match rate in others. Intuitively, a higher local match rate implies higher confidence in the result of the knowledge-based prediction method.

2.4 HYPLOSP: a hybrid method for protein local structure prediction

Our hybrid prediction method, HYPLOSP, combines the prediction results of the knowledge-based method and the neural network method at each position of the target protein. The neural network returns a confidence score for each output letter. In order to output these values to a text file, we normalize them into a range of 0 to 94, since there are only 95 printable ASCII codes. Then the neural network generates a set of normalized confidence scores {$Conf\_NN_1$, $Conf\_NN_2$,…, $Conf\_NN_n$} associated with each letter.

The knowledge-based method generates a set of voting scores, denoted by {V_1, V_2,…, V_n}, associated with each position. We define the confidence score of letter A_i as

$$Conf\_KB_i = \text{Min}\{ \frac{V_i}{\sum_j V_j} \times LocalMatchRate(x), 94\}.$$

Using $Conf\_NN_i$ and $Conf\_KB_i$, we determine the final predicted structure at position x to be A_k if

$$Conf\_NN_k + Conf\_KB_k = \text{Max} \bigcup_m (Conf\_NN_m + Conf\_KB_m) .$$

3. Experimental Results

3.1 Datasets

We downloaded 25,288 proteins from the DSSP database (dated 9/22/2004) as our first dataset. These proteins were divided into 46,745 protein chains. In our method, we use PSI-BLAST and pairwise sequence alignment to filter out protein chains with a pair-wise sequence identity over 25%. Moreover, protein chains of length less than 80 are removed. Finally, we have a non-redundant DSSP dataset, called *nrDSSP*, containing 3,925 unique protein chains along with their secondary structures. To evaluate our prediction methods, we transform nrDSSP into structural alphabets of our choice.

Furthermore, we use another dataset, containing new proteins for the period of Oct. 2004 to May 2005, to compare HYPLOSP with other prediction methods. Fifty-six protein chains remain after we filter out those with a sequence identity over 25% in this dataset and in nrDSSP.

We test our methods on two structural alphabets: SAH and PB. There are originally 11 alphabet letters in SAH, including 10 Φ–Ψ plane partitions for *trans* peptide and one for *cis* peptide. We follow Karchin's approach [9] and assign the *cis* residues among the other 10 partitions according to their Φ–Ψ values. We encode each amino acid with a SAH letter by assigning the letter of the Φ–Ψ plane that is the nearest to the amino acid. The PB alphabet is composed of 16 letters, each of which is 5-residue in length and represented by 8 dihedral angles. We use a sliding window of length 5 to extract peptides from amino acid sequences. The Root Mean Square Deviation on Angular values (RMSDA) between the peptide and each of the 16 PB letters is calculated, and the letter with the smallest RMSDA is assigned to the peptide center.

3.2 Cross-validation test of our methods

We perform 10-fold cross-validation experiments on each chosen structural alphabet to evaluate our knowledge-based (KB) method, neural network (NN) method, and the hybrid method, HYPLOSP. In each experiment, the dataset is randomly divided into ten sets. A set is selected as the testing set (containing *predicted* secondary structure information) and the other nine sets are integrated as the training set (containing *true* secondary structure information) for neural network training and the construction of SSKB. This process must be repeated for each set to be a testing set. In addition, we modify our methods that do not use secondary structure information as follows. For the knowledge-based method, we do not record secondary structure element (SSE) information while constructing SSKB, or while finding similar peptides in SSKB. For the neural network method, we do not take the SSE of a peptide as input for either training or testing (prediction); thus, the input of the network becomes a vector of size 140.

The performance results using SSE information are shown in Table 2. For the SAH alphabet, HYPLOSP reports a Q_N of 61.51% and outperforms our KB and NN methods (which report a Q_N of 56.7% and 59.53% on average) by approximately 5% and 2%, respectively. For the PB alphabet, our KB and NN methods achieve on average a Q_N of 57.79 % and 59.54%, respectively. Our hybrid method, HYPLOSP with an overall Q_N of 63.24% outperforms the KB and NN methods by 3.7% and 5.5%, respectively. In summary, HYPLOSP reports a Q_N over 60%, whether on the 10-letter SAH alphabet or the 16-letter PB alphabet.

The results not using SSE information are also shown in Table 2. Both the KB and NN methods suffer a considerable decrease in Q_N (between 3% and 5%). Therefore, the SSE information plays a role in these two methods. However, the Q_N of HYPLOSP is reduced by at most 1.37%, which is comparatively lower than the KB and NN methods. This implies that HYPLOSP is less sensitive to the absence of SSE and better utilizes both the neural network and knowledge-based methods.

Table 2. The performance of our methods on SAH and PB

	Using SSE		Not using SSE	
	Q_N on SAH	Q_N on PB	Q_N on SAH	Q_N on PB
NN	59.53%	59.54%	55.72%	54.65%
KB	56.70%	57.79%	53.14%	53.79%
HYPLOSP	61.51%	63.24%	60.14%	61.91%

3.3 Comparison with the previous studies

To compare HYPLOSP with the prediction methods used by the authors of SAH and PB, we use the second dataset (56 new proteins) for evaluation. The HYPLOSP model is trained on nrDSSP and tested on the testing dataset. We compare our methods with the HMMSTR server developed by Bystroff et al. [4] for the SAH alphabet, and with the LocPred server developed by de Brevern et al. [5] for the PB alphabet. Note that there are three models in LocPred server: Bayesian prediction, sequence families, and a new version of sequence families. We only compare HYPLOSP with the last one, since it is the best of the three.

The experimental results are shown in Table 3. HYPLOSP outperforms HMMSTR on the SAH alphabet by 4.4% and achieves a 13.24% improvement over LocPred on the PB alphabet. Furthermore, HYPLOSP demonstrates an alphabet-independent prediction capability and a relatively stable performance irrespective of the alphabet size. To be specific, HYPLOSP has a Q_N of 57.44% for the 10-letter SAH alphabet, and 55.17% for the 16-letter PB alphabet. Although the alphabet size grows by 60% ($(16-10)\div 10\times 100\%$), Q_N only decreases by 2.27%.

Table 3. Comparison of HYPLOSP with other prediction methods

		Q_N
SAH	HMMSTR	53.04%
	HYPLOSP	57.44%
	Improvement	4.40%
PB	LocPred	41.93%
	HYPLOSP	55.17%
	Improvement	13.24%

5. Concluding Remarks

Existing local structure prediction methods show that prediction accuracy is a very challenging issue. We use two different prediction methods: one is knowledge-based and the other is neural network-based. To better utilize the advantage of these two methods, we propose a hybrid method called HYPLOSP, which is alphabet-independent. We select two current structural alphabets, SAH and PB, to evaluate HYPLOSP. We have performed a 10-fold cross-validation test on the nrDSSP dataset of nearly 4,000 protein chains to evaluate our KB, NN methods in comparison with HYPLOSP. In addition, we have also performed a test on 56 protein chains to compare HYPLOSP with the prediction methods used the authors of SAH and PB. The experimental results not only show better performance of HYPLOSP in terms of Q_N accuracy, but also demonstrate its capability to be alphabet-independent. We further analyze the relation between our prediction accuracy rate and the secondary structure. The analysis shows that improving current secondary structure prediction accuracy leads to a substantial improvement in local structure prediction.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25(17):3389-3402, 1997.

2. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, 281: 565-577, 1998.
3. Bystroff C, Shao Y. Fully automated ab initio protein structure prediction using I-Sites, HMMSTR and Rosetta. *Bioinformatics*, 18: 54-61, 2002.
4. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.*, 301: 173-190, 2000.
5. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41:271-287, 2000.
6. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, 53:491-496, 2003.
7. Lin HN, Chang JM, Wu KP, Sung TY, Hsu WL. A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics*, 21:3227-3233, 2005.
8. Karchin R, Cline M, Karplus K. Evaluation of local structure alphabets based on residue burial. *Proteins*, 55:508-518, 2004.
9. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*, 51:504-514, 2003.
10. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.*, 323:297-307, 2002.
11. Kuang R, Leslie CS, Yang AS. Protein backbone angle prediction with machine learning approaches. *Bioinformatics*, 20: 1612-1621, 2004.
12. Micheletti C, Seno F, Maritan A. Recurrent Oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins*, 40: 662-674, 2000.
13. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. *Proteins*, 5:192-199, 2001.
14. Rumelhart, D., G. Hinton, and R. Williams. Learning internal representations by error propagation. In *Neurocomputing: Foundations of Research*, 675-695. Cambridge, MA: MIT Press, 1988.
15. Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng.*, 9: 833-842, 1996.
16. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5:355-373, 1989.