Journal of Bioinformatics and Computational Biology

HYPLOSP: A KNOWLEDGE-BASED APPROACH TO PROTEIN LOCAL STRUCTURE PREDICTION

Ching-Tai Chen

Institute of Information Science, Academia Sinica,
128 Sec. 2, Academia Rd,Taipei, Taiwan
caster@iis.sinica.edu.tw

Hsin-Nan Lin

Institute of Information Science, Academia Sinica,
128 Sec. 2, Academia Rd,Taipei, Taiwan
arith@iis.sinica.edu.tw

Ting-Yi Sung*

Institute of Information Science, Academia Sinica,
128 Sec. 2, Academia Rd,Taipei, Taiwan
tsung@iis.sinica.edu.tw

Wen-Lian Hsu*

Institute of Information Science, Academia Sinica,
128 Sec. 2, Academia Rd,Taipei, Taiwan
hsu@iis.sinica.edu.tw

Local structure prediction can facilitate *ab initio* structure prediction, protein threading, and remote homology detection. However, the accuracy of existing methods is limited. 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 method. Empirically, the accuracy of the method correlates positively with the local match rate; therefore, we employ it to predict the local structures of positions with a high local match rate. For positions with a low local match rate, we propose a neural network prediction method. To better utilize the knowledge-based and neural network methods, we design a hybrid prediction method, HYPLOSP (HYbrid method to Protein LOcal Structure Prediction) that combines both methods. To evaluate the performance of the proposed methods, we first perform cross-validation experiments by applying our knowledge-based method, a neural network method, and HYPLOSP to a large dataset of 3,925 protein chains. We test our methods extensively on three different structural alphabets and evaluate their performance by two widely used criteria, MDA (Maximum Deviation of backbone torsion Angle) and Q_N, which is similar to Q_3 in secondary structure prediction. We then compare HYPLOSP with three previous studies using a dataset of 56 new protein chains. HYPLOSP shows promising results in terms of MDA and Q_N accuracy and demonstrates its alphabet-independent capability.

Keywords: knowledge-based prediction method, local structure prediction, neural networks, secondary structure, structural alphabet.

* Corresponding authors.

1. Introduction

A protein's 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[1], and backbone geometry[2], and represent these fragment clusters by an *alphabet*, called a *structural alphabet*. By using the structural alphabet we can encode a native protein into a set of discrete representations. Local structure prediction predicts the local structure expressed by a *letter* of the structural alphabet from the amino acid sequence, which improves the performance of both *ab initio* and fold recognition methods of tertiary structure prediction[3-5].

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.[6], 40 and 100 in Micheletti et al.[7], 100 in Schuchhardt et al.[8], and 25-300 with a fragment length from 5 to 7 in Kolodny et al[9]. Although large alphabets can better approximate protein tertiary structures, predicting protein local structures from amino acid sequences is much more challenging.

As a result, smaller structural alphabets have been proposed, associated local structure libraries have been constructed, and local structure prediction algorithms have been developed for use with these libraries. Bystroff et al. generated a library called *I-site*[10], which contains 13 structural motifs of different length. Prediction is based on profile-profile alignment between each structural motif and the PSI-BLAST[11] result of the input sequence. To improve prediction accuracy, the authors also proposed a new model called HMMSTR[12]. In this paper, we use the *structural alphabet of HMMSTR*, denoted by *SAH*, to test our method. A.G. de Brevern et al[13] built their library, called *Protein Blocks* (*PB*), by clustering 5-mer protein fragments into a structural alphabet of 16 letters according to the torsion angle space. They used a Bayesian probabilistic approach for prediction. Karchin et al.[2] constructed an *STR* library, in which the structural alphabet consists of 13 letters obtained from eight secondary structure states by dividing β-sheets into 6 types. They then used a hidden Markov model (HMM)[14,15] for local structure prediction. Yang et. al.[16] used a local structure-based sequence profile database (LSBSP1) to predict the local structure of 4 states. Kuang et al.[17] incorporated a neural network which takes the result of LSBSP1 as input to enhance the model.

The accuracy of local structure prediction depends on the definition of the underlying structural alphabet and the prediction algorithm. Although there is no unifying measure for performance evaluation, one can use a straightforward measure called Q_N[13] that is similar to Q_3 in secondary structure prediction. Given an alphabet of size N, the Q_N of a protein, p, compares the predicted results with the encoded structural letter sequence. It is calculated as follows:

$$Q_N = \frac{\text{the number of residues of } p \text{ correctly predicted}}{\text{the total number of residues of } p} \times 100 \cdot \quad (1)$$

The accuracy of Q_N is 40.7% for PB[13], and 56.1% for STR[2]. Bystroff et al. introduced another measure, called the *MDA* (*Maximum Deviation of backbone torsion Angle*) score[10,12]. They regard a local structure as correctly predicted if the *MDA* of an eight-residue window is less than 120 degrees to the native structure, since the residues can be superimposed with an RMSD (Root-Mean-Square Deviation) of less than 1.4 Å. The MDA score is defined as the fraction of residues found in correctly predicted eight-residue segments[12]. Cross validation tests using I-site and HMMSTR yield MDA scores of approximately 48%[10] and 59.1%[12], respectively. The MDA score has been used extensively by different research groups[10, 12, 16-18].

To improve the accuracy of existing local structure prediction algorithms, which is no greater than 60%, is our main goal in this study. We construct a knowledge base that stores local structures of peptides, instead of profiles as in LSBSP1. Prediction results of knowledge base are combined with results of another method based on neural network by a sophisticated hybrid mechanism. We apply our method to three different structural alphabets (SAH, PB, and STR), and use Q_N and MDA to evaluate the prediction results on a non-redundant dataset of 3,925 protein chains. We also use a dataset of new submissions to PDB to conduct a benchmark comparison between HYPLOSP and prediction methods in previous studies. The results not only show some improvements over the previous methods but also demonstrate our method is alphabet-independent, i.e., its performance remains stable for different underlying structural alphabets.

2. Methods and materials

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, a high match rate of the protein results in high prediction accuracy. To improve the low prediction accuracy of low-match-rate positions, we use a neural network prediction method that provides confidence scores in its output. By combining the results of the above methods, based on the local match rate and the neural network prediction confidence, we propose a hybrid method called *HYPLOSP* (*HYbrid method to Protein LOcal Structure Prediction*).

2.1. Knowledge-based approach

2.1.1. *Construction of the sequence-structure knowledge base (SSKB)*

Our knowledge base contains both local structure information and secondary structure information of peptides. The former is expressed by a structural alphabet (discussed in Section 2.4), while the latter is obtained from the DSSP database[19]. For ease of exposition, we assume that we are given a protein dataset with known secondary structures and local structures based on a specific 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 to find proteins remotely homologous to a protein with a known structure, i.e. proteins in the DSSP database, referred to as a *Query protein* in the PSI-BLAST output. When using PSI-BLAST, we set the parameter *j* to 3 (3 iterations), *e* to 0.001 (E-value < 0.001), and use the NCBI nr[20] database as the sequence database. For each Query protein, PSI-BLAST generates a large number of homologous protein segments as well as their pairwise alignments, 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.

Since applying PSI-BLAST to a Query protein generates a large set of HSPs, we need to find the peptides in the Sbjct protein of each HSP that are similar to those of the Query protein so that they 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 various lengths. Let p and q denote a pair of peptides in the Query protein and Sbjct protein, respectively. We define the *similarity score, S,* of 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 $S \geqq 5$. We distinguish two cases of similarity between p and q for constructing the knowledge base. Case 1: p and q are similar. We define the *confidence score* of q with respect to p as $(S \times A) / w$. This enables us to measure the confidence level of q inheriting the structural information of p, where A denotes the alignment score of the HSP reported in PSI-BLAST output. If there is no gap between p and q, we create a new record corresponding to peptide q, (q, secondary structure of q, local structure of q, confidence score of q), given by (q, secondary structure of p, local structure of p, confidence score of q with respect to p), to the knowledge base. When q is later found in another HSP and is similar to the counterpart peptide r that already exists in the knowledge base, i.e., q can also inherit the structural information from r, we simply update the record of q by including the structural information of r and adding the confidence score of q with respect to r. Case 2: p and q are dissimilar. Then we simply ignore this pair of peptides for the construction of knowledge base.

Fig. 1 shows part of an HSP. The pair of peptides inside the box has a similarity score of 5, so the peptides are considered similar. The confidence score of the peptide in the Sbjct with respect to that in the 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 a gap, the record (MLTAEDK, CCHHHHC, BBEEECD, 180) is added to the knowledge base, as shown in Table I (a). Note that a peptide may inherit structural information from multiple peptides, in which case we simply add the structural information and confidence score to the existing record. For example, suppose the peptide MLTAEDK also inherits structural information from another similar peptide with a confidence score of 65. Then, the record of MLTAEDK in the knowledge base is updated, as shown in Table I (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
 Identities = 85/141 (60%), Positives = 108/141 (76%)
Query: 1    VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVK 55
             +L+  DK  ++A W KV  H  ++GAEAL+RMF+++PTTKTYFPHFDLS GS QV+
Sbjct: 1    MLTAEDKKLIQAIWDKVQGHQEDFGAEALQRMFITYPTTKTYFPHFDLSPGSDQVR 55
```

Fig.1. An example of an HSP generated by PSI-BLAST

Table I. 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 the construction of the knowledge base. To define the similar peptides in Step 2, we use the same sliding window length of 7 and the same similarity score of 5 with no gap to define similar peptides as before. We then match all the peptides of the target protein with similar peptides in SSKB, and use the local structure information of the matched peptides in SSKB to vote for the local structure of the target protein. Hereafter, we assume that the structural alphabet is a set of $\{A_1, A_2, \ldots, A_n\}$. Let p be a peptide of the target protein. We associate each position, x, in p with n variables, called the *voting score*, denoted by $V^p_i(x)$, where $i = 1, \ldots, n$; initially, $V^p_i(x)$ is set to be 0. In an HSP, let q be p's counterpart peptide with similarity score S and alignment score A. If q is similar to p and can be found in SSKB, the voting score of p is updated as follows. For each position, x, compute

$$V^p_i(x) \leftarrow V^p_i(x) + C^q_i(x) \times (S \times A) / 7,\ i=1,\ldots n, \tag{2}$$

where $C^q_i(x)$ is the *normalized* confidence score of q with the structural letter i given in SSKB. It is calculated as the confidence score of letter i divided by the total confidence score of all letters. The value turns out to be a fraction between 0 and 1 which represents the weight of each structural letter based on this SSKB entry. The calculation is repeated for all similar peptides. The local structure of x in p is given by the letter corresponding to $Max\ \{V^p_1(x), V^p_2(x,), \ldots, V^p_n(x)\}$.

2.1.3. *Confidence measure of the knowledge-based approach*

We use two measures, the local match rate and the global match rate, to represent the amount of information extracted from the knowledge base. At each position, x, of a target protein, we obtain from the HSPs a set of similar peptides, $Q(x)$, that contains the position x. The local match rate at position x, denoted by *LocalMatchRate(x)*, is defined by:

$$LocalMatchRate(x) = \frac{|\mathrm{Q(x)} \cap \mathrm{SSKB}|}{|\mathrm{Q(x)}|} \times 100\%. \tag{3}$$

A higher local match rate implies higher confidence in the result of the knowledge-based method for the position. Note that it is possible for a target protein to have high local match rates in some positions and low local match rates in others.

Unlike the local match rate, the global match rate is a measure that deals with the target protein as a whole. Let Q be the set of peptides obtained from the HSPs of the target protein, i.e., Q is equal to the union of all Q(x) of the protein. The global match rate of a protein, p, is defined as follows:

$$GlobalMatchRate(p) = \frac{|\mathrm{Q} \cap \mathrm{SSKB}|}{|\mathrm{Q}|} \times 100\%. \tag{4}$$

2.2. **Neural network method**

Though the 2-stage neural network has been used extensively for secondary protein structure prediction, we use the 1-stage neural network as our underlying neural-network prediction method as described in this subsection. We compare the performance of using 2-stage neural network prediction method with HYPLOSP in Section 4.4.

2.2.1 Neural network architecture

We use a standard feed-forward back-propagation neural network[21] with a single hidden layer. The layer contains 35 hidden units, which we found to be the most effective number for our training stage.

Each protein sequence in the training set or test set is partitioned into peptides, using a sliding window of length 7. We also perform a PSI-BLAST search to obtain the profile of the sequence, which is called 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 and its secondary structure. Hence, the length of each input vector is 161, i.e., 7×20 for the PSSM segment and 7×3 for its secondary structure. The output reports the results corresponding to the amino acid located at the center of the peptide (the *peptide center*). 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

We use an online back-propagation training procedure to optimize the weights of the network, whereby the weights are randomly initialized and then updated by each input vector. The learning parameters of the hidden layer and the output layer are 0.075 and 0.05, respectively; and the sum of square errors is used during back propagation.

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

2.2.3 Local structure prediction based on the 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 the proteins in the training set, target proteins do not have secondary structure information. Thus, in Step 1 we use HYPROSP II[22] 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 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. **HYPLOSP: a hybrid method for protein local structure prediction**

As our knowledge-based method and neural network method have different strengths, one may outperform the other, depending on the circumstances. To better utilize their respective strengths, we propose a hybrid prediction method, HYPLOSP, which combines the prediction results of both methods at each position along the amino acid sequence based on their confidence scores.

The knowledge-based method generates a set of voting scores, denoted $\{V_1, V_2, \ldots, V_n\}$ for each output letter. We define the confidence score of letter A_i as the normalized voting score multiplied by the prediction confidence, *LocalMatchRate*:

$$Conf\_KB_i = \frac{V_i}{\sum_j V_j} \times LocalMatch\ Rate(x) \tag{5}$$

The neural network also automatically generates a set of confidence scores between 0 and 1. To make both prediction methods have the same scale of confidence scores, we multiple the output score of NN method by 100 to define the confidence scores of the NN method, denoted by $\{Conf\_NN_1, Conf\_NN_2, \ldots, Conf\_NN_n\}^{\dagger}$.

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

$$Conf\_NN_k + Conf\_KB_k = Max_{i \in \{1,2 \ldots n\}}(Conf\_NN_i + Conf\_KB_i)\,. \tag{6}$$

2.4. Structural alphabets and encoding

To evaluate our hybrid prediction method, we use three popular structural alphabets: the *Structural Alphabet of HMMSTR*, *Protein Blocks,* and *STR*, each of which has a relatively small number of letters, We use a non-redundant DSSP database as our dataset (explained in Section 3.1) and encode each amino acid of a protein sequence into structural letters based on its structural information. Such encoding is necessary to construct our knowledge base and the training stage in our neural network method. The definitions of the alphabets and their encodings are presented below.

***Structural Alphabet of HMMSTR* (*SAH*)**[12]. This alphabet, proposed by Bystroff et al. is composed of 11 letters, ten of which represent conformation states in different Φ–Ψ angle regions of a *trans* peptide; the other one corresponds directly to a *cis* peptide. Following Karchin's approach[2], we assign the *cis* residues to one of the other 10 regions according to their Φ–Ψ angles. (In our encoding scheme, SAH is comprised of 10 letters.) Table II shows the Φ–Ψ regions of the SAH alphabet.

To encode each amino acid of a protein sequence into a structural letter, we use the Φ–Ψ angle of an amino acid to compute the Euclidean distance (ED) of the ten letters by the following equation:

$$ED_i = \sqrt{\frac{(\phi_{AA} - \phi_i)^2 + (\psi_{AA} - \psi_i)^2}{2}}\,, \tag{7}$$

where i denotes an alphabet letter; ϕ_{AA} and ϕ_i denote the Φ angles of the amino acid and

---

† For implementation convenience, we further normalize *Conf_KB* and *Conf_NN* in a range of 0 to 94 and assign a corresponding ASCII character to each of them.

letter i, respectively; and ψ_{AA} and ψ_i denote the Ψ angles of the amino acid and letter i, respectively. Note that the angle difference is computed with modulo 360. The letter with the smallest Euclidean distance, i.e., $\min_i \{ED_i\}$, is assigned to the amino acid of the protein.

Table II. Φ−Ψ angle regions of the SAH alphabet

Letters	Φ	Ψ
H	-61.91	-45.20
G	-109.78	20.88
B	-70.58	147.22
E	-132.89	142.43
d	-135.03	77.26
b	-85.03	72.26
e	-165.00	175.00
L	55.88	38.62
l	85.82	-0.03
x	80.00	-170
c	cis residue	

***Protein Blocks* (*PB*)[13].** A.G. de Brevern et al. partition proteins into 5-residue peptides and use an unsupervised learning algorithm to group the peptides into 16 clusters. They also generate a representative for each cluster that models the protein structure with an average *RMSDA* (*Root Mean Square Deviation on Angular values*) of 30 degrees.

PB uses RMSDA as the similarity measure for two 5-residue peptides, each of which is represented by a vector of eight dihedral angles. To assign a structure letter to a peptide p, we calculate the RMSDA with respect to letter i as follows:

$$RMSDA_i = \sqrt{\frac{\sum_{j=1}^{8} (V_j^p - V_j^i)^2}{8}}, \tag{8}$$

where V_j^p is the jth entry of dihedral vector of peptide p, and V_j^i is the jth entry of the dihedral vector of letter i. The angle difference is also computed with modulo 360, and the letter with the smallest RMSDA is assigned to the peptide center. The assignment process is repeated for the protein using a sliding window of length 5.

***STR*[2].** STR is a finer classification of secondary structures. Karchin et al. change the eight secondary structure states to 13 letters by dividing the β-strand E into six classes; the other seven secondary structure states are not changed. A β-strand is assigned the letter "P" if it is surrounded by two parallel strand partners, "A" if it is surrounded by two anti-parallel partners, and "M" if it is surrounded by one parallel partner and one anti-parallel partner. However, a β-strand is assigned the letter "Q" if it is neighbored by only one parallel strand partner, and "Z" if it is neighbored by only one anti-parallel strand partner. If a β-strand does not have any neighboring strand partners, it is assigned "E".

3. Experimental results

3.1. Datasets and experiment design

We carried out two types of experiment. First, we used a large dataset to perform 10-fold cross-validation experiments on each structural alphabet to evaluate our knowledge-based method, neural network method, and the hybrid method, HYPLOSP. To generate the dataset, we downloaded 25,288 proteins from the DSSP database (dated 9/22/2004), which were divided into 46,745 protein chains. We then used PSI-BLAST and pairwise sequence alignment to filter out protein chains with a pairwise sequence identity over 25%. Moreover, protein chains of length less than 80 were removed. Finally, we had a non-redundant DSSP dataset, called *nrDSSP*, containing 3,925 unique protein chains along with their secondary structures. To evaluate our prediction methods, we transformed nrDSSP into structural alphabets of our choice, as described in Section 2.4. In each experiment, the nrDSSP dataset was randomly divided into ten sets. One set was selected as the test set (containing *predicted* secondary structure information) and the other nine were combined as the training set (containing *observed* secondary structure information) for training the neural network and construction of SSKB. This process was repeated for each set in turn to be used as the test set.

Second, we used another dataset, containing new proteins of DSSP reported during the period October 2004 to May 2005, as the test set to compare HYPLOSP with the other two methods. This dataset consisted of fifty-six protein chains after filtering out chains with a sequence identity over 25% to the nrDSSP. Furthermore, all 56 protein chains had a pairwise sequence identity of less than 25%. We compared the HYPLOSP model, which was trained on the nrDSSP dataset, with the following three servers: the HMMSTR[12] server developed by Bystroff et al. for the SAH alphabet, the LocPred[13, 23] server developed by de Brevern et al. for the PB alphabet, and the SAM-T02[4] server developed by Karplus et al. for the STR alphabet. Note that the LocPred server provided three models: Bayesian prediction, sequence families, and a new version of sequence families. We only compared HYPLOSP with the result of the last model, since it was the best of the three.

We used the Q_N and MDA scores as performance measures. Specifically, Q_N was used for the three structural alphabets, while the MDA score was used for SAH and PB only. It was not used for STR, since it lacks torsion angle information.

Our datasets and HYPLOSP's results on the datasets are available at http://bio-cluster.iis.sinica.edu.tw/~caster/hyplosp/.

3.2. Cross-validation results of HYPLOSP

The experimental results of our methods using nrDSSP on the three alphabets are shown in Table III. It can be observed from the table that HYPLOSP improves the knowledge-based (KB) method and the neural network (NN) method. Even when KB and NN methods achieve a similar performance, HYPLOSP still improves their scores, e.g., it improves the MDA scores for SAH and PB by approximately 3-4%.

Table III. Cross-validation results of each method

		Q_N	MDA
	NN	59.53%	58.71%
SAH	KB	56.70%	58.31%
	Hybrid	61.51%	62.69%
	NN	59.54%	55.26%
PB	KB	57.79%	54.57%
	Hybrid	63.24%	58.66%
	NN	58.78%	
STR	KB	58.96%	
	Hybrid	63.07%	

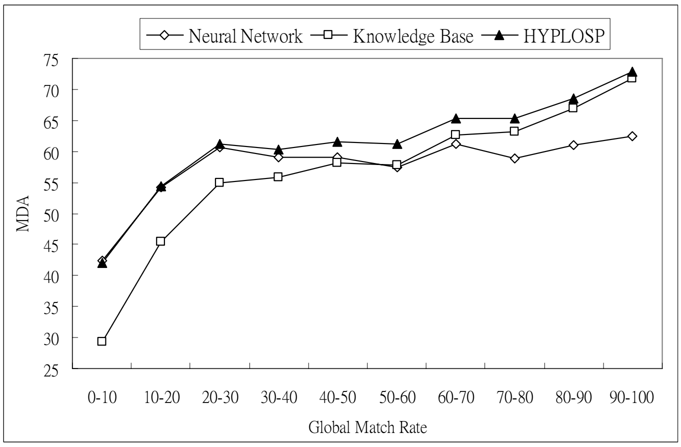

Fig. 2. SAH prediction: MDA of the KB, NN methods, and HYPLOSP with respect to the global match rate.

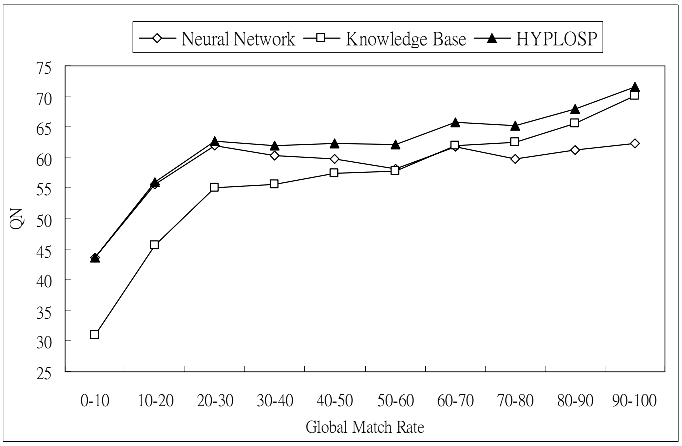

Fig. 3. PB prediction: QN of the KB, NN methods, and HYPLOSP with respect to the global match rate.

Since the global match rate of a protein indicates the amount of information extracted from the knowledge base, we examine the relation between the performance of each method and the global match rate of a protein. For prediction on each structural alphabet, we randomly choose Q_N or MDA to illustrate the performance of the three methods versus the global match rate. We choose MDA for the SAH alphabet, Q_N for the PB alphabet, and Q_N for the STR alphabet (Figures 2, 3 and 4, respectively). We observe that the KB method is more sensitive to the global match rate than the NN method. In the three figures, the KB method shows a positively correlated trend with respect to the global match rate, which contrasts to NN's relatively stable trend. Furthermore, the KB method outperforms the NN method on proteins with a higher global match rate, e.g., higher than 50 in Fig. 2 (which accounts for 52% of the dataset). As more protein structures are determined, the knowledge base will be enlarged. Thus, the number of proteins with higher global match rates will very likely increase, and the KB method and HYPLOSP can then be further improved.

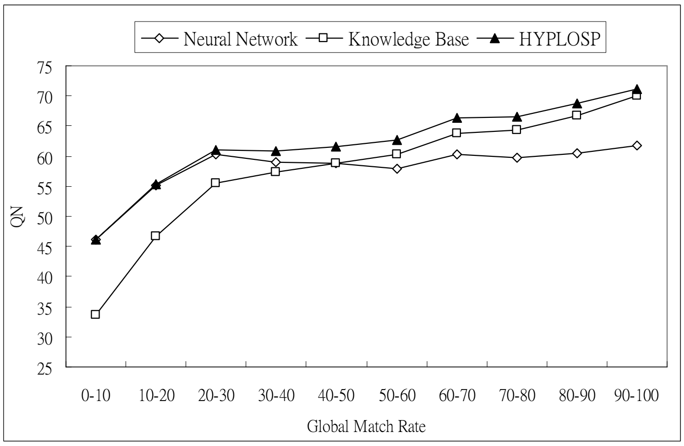

Fig. 4. STR prediction: QN of the KB, NN methods and HYPLOSP with respect to the global match rate.

3.3. Benchmark comparison of HYPLOSP with previous studies

To compare HYPLOSP with existing methods, we followed the methodology adopted in EVA[24] by using new submissions to PDB[25] and avoiding homologous proteins in the training stage since all of these methods may be trained on different datasets. So, the benchmark comparison is based on the second dataset, which contains 56 new protein chains with an average global match rate of only 25.49. The low match rate can be attributed to the following factors: (1) some proteins have only a few HSPs; (2) though there are enough HSPs for some proteins, only a few of their peptides match the SSKB. Fig. 5 shows the number of proteins versus the global match rate.

The experimental results are shown in Table IV. For the SAH alphabet, HYPLOSP outperforms HMMSTR by 4.4% and 5.15% in terms of Q_N and MDA, respectively; for the PB alphabet, it achieves 13.24% and 16.7% improvement over Q_N and MDA, respectively; and for the STR alphabet, it yields a Q_N of 59.03%, which is 0.76% lower than the result of SAM-T02. In addition, for STR, we note that KB has a Q_N of only 47.02% because this dataset has a low average global match rate. In contrast, NN has a Q_N of 58.26%. However, HYPLOSP yields a better Q_N (59.03%), which improves the NN result slightly.

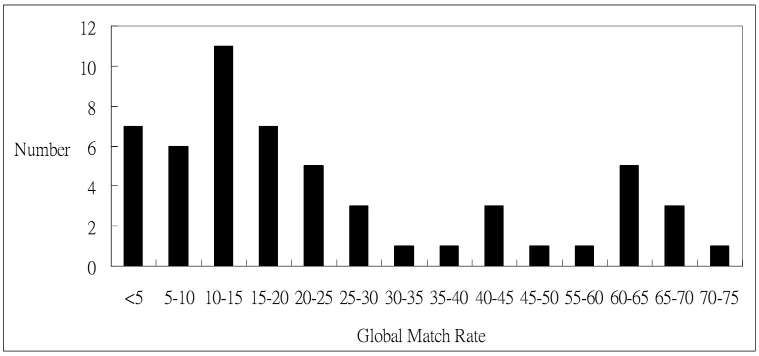

Fig. 5. The number of proteins versus the global match rate

Table IV. Comparison of HYPLOSP with previous studies

		Q_N	MDA
SAH	HMMSTR	53.04%	50.08%
	HYPLOSP	57.44%	55.23%
	Improvement	4.40%	5.15%
PB	LocPred	41.93%	36.11%
	HYPLOSP	55.17%	52.81%
	Improvement	13.24%	16.7%
STR	SAM-T02	59.79%	
	HYPLOSP	59.03%	
	Improvement	-0.76%	

4. Discussion

4.1 Effect of secondary structure information on prediction

Our KB and NN methods use predicted or observed secondary structure element (SSE) information for local structure prediction. We repeat the cross-validation experiments (Section 3.1), with modifications, to investigate the effect of utilizing secondary structure information in local structure prediction.

4.1.1. *Effect of not using secondary structure information*

To examine the effect of not using SSE information, we modify our methods as follows. For the KB method, we do not record SSE while constructing SSKB or while finding similar peptides in SSKB. For the NN method, we do not include the SSE of a peptide in the input of either the training or the testing (prediction) stages; thus, the input becomes a vector of 140 (161－3・7) entries in the network.

Our experimental results are shown in Table V. Compared with the results shown in Table III, the NN and KB methods suffer considerable performance degradation, while HYPLOSP suffers only a moderate decrease. This implies that HYPLOSP is less sensitive to the absence of SSE and can better utilize the results of the NN and KB methods.

Table V. Prediction results of HYPLOSP without secondary structure information

		Q_N	Decrease in Q_N	MDA	Decrease in MDA
SAH	NN	55.72%	3.81%	49.00%	9.71%
	KB	53.14%	3.56%	51.63%	6.68%
	HYPLOSP	60.14%	1.37%	58.29%	4.40%
PB	NN	54.65%	4.89%	46.82%	8.44%
	KB	53.79%	4.00%	48.36%	6.21%
	HYPLOSP	61.91%	1.33%	54.84%	3.82%
STR	NN	52.76%	6.02%		
	KB	52.76%	6.20%		
	HYPLOSP	60.59%	2.48%		

4.1.2. *Effect of using observed secondary structure information*

Using the observed SSE for local structure prediction enables us to estimate the upper bound of the HYPLOSP performance. To do this, we change the "predicted" SSE to the "observed" SSE and repeat the experiments in the testing stage of the KB method, while the neural-network training stage is the same as in Section 2.2.2. We summarize the experimental results in Table VI.

Although our secondary structure prediction method can achieve Q_3 of 81.6%[22], the performance gap between using the predicted SSE and the observed SSE is still larger than the gap between using the predicted SSE and not using SSE. Using the observed SSE (i.e., secondary structure prediction would improve from the current accuracy of 81.6% to an ideal 100%) can enhance the performance significantly. In other words, an improvement in current secondary structure prediction can further improve local structure prediction.

Table VI. Prediction results of HYPLOSP using the observed secondary structure information

	Q_N			MDA		
	Observed SSE	Predicted SSE	Without SSE	Observed SSE	Predicted SSE	Without SSE
SAH	65.81%	61.51%	60.14%	68.85%	62.69%	58.29%
PB	69.14%	63.24%	61.91%	64.41%	58.66%	54.84%
STR	77.15%	63.07%	60.59%			

4.2 Effect of the hybrid mechanism

Our hybrid mechanism defined in HYPLOSP yields better results than the KB and NN methods. To better understand the hybrid effect, we define the *hybrid_benefit* of a protein in terms of Q_N as follows:

$$\text{hybrid\_benefit} = (Q_N \text{ of HYPLOSP}) - (Q_N \text{ of NN}), \tag{9}$$

since the NN prediction accuracy is relatively stable. A negative hybrid_benefit means that the hybrid mechanism's prediction results are worse than those of NN, which happens when the global match rate is low and the KB prediction results are poor. For example, the average global match rate of the second test dataset is approximately 25.49. Despite this low rate, about 89.29%, 80.36%, and 81.82%, respectively, of the proteins in this dataset have a positive hybrid_benefit with an average of 2.05%, 1.56%, and 0.78%, respectively. In Fig. 6, we illustrate the relation between the hybrid_benefit of

STR prediction and the global match rate. (We chose STR because HYPLOSP's performance is slightly inferior on this alphabet.) The figure shows that there is a positive regression curve between the hybrid_benefit and the global match rate.

Fig. 7 further demonstrates the relation between the average hybrid_benefit and the global match rate on nrDSSP. Clearly, there is a positive relation between the hybrid_benefit and the global match rate. This shows, once again, that when more protein structures are determined, the knowledge base will expand and the hybrid mechanism will derive better hybrid_benefit. In this case, the performance of HYPLOSP will improve.

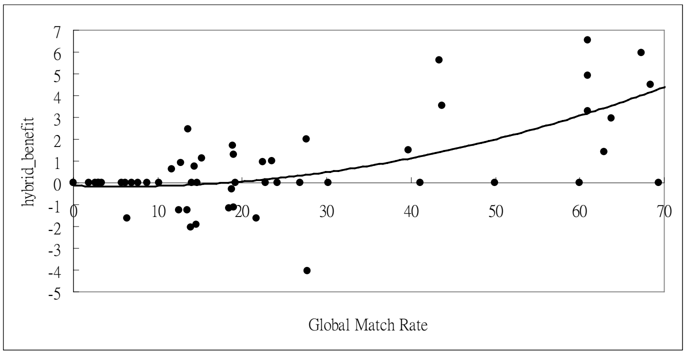

Fig. 6. The positive correlation between the hybrid_benefit of STR and the global match rate

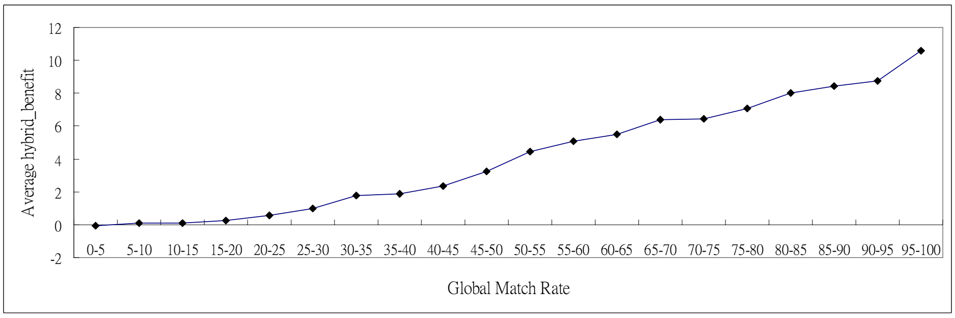

Fig. 7. Analysis of the hybrid_benefit with respect to the global match rate of nrDSSP

Table VII. HYPLOSP's performance in helix (H), strand (E), and coil (C) regions.

	Q_N^H	Q_N^E	Q_N^C	MDA^H	MDA^E	MDA^C
SAH	87.80	55.91	42.35	82.65	68.45	41.51
PB	83.15	68.49	42.64	80.35	65.58	34.4
STR	83.09	52.09	50.91			

4.3. HYPLOSP's accuracy on different secondary structures

Protein structures are generally more regular in helix (H) and strand (E) regions than in coil (C) regions. To investigate whether local structure reflects this fact, we analyze the prediction results of HYPLOSP on the nrDSSP dataset using different alphabets to calculate the HYPLOSP's accuracy, Q_N and MDA, on H, E, and C (Table VII). In the table,

Q_N^H and MDA^H stand for the Q_N and MDA in helix regions, other terms are similarly defined. As expected, helices have the highest accuracy (all greater than 80) and coils have the lowest (all less than 51). Since STR divides strands into 6 letters, the accuracy of strands is slightly lower than that of coils.

Table VIII. QN and MDA of 1-stage neural network (NN-1), 2-stage neural network (NN-2), original version of HYPLOSP, and HYPLOSP-2.

		Q_N	MDA
SAH	NN-1	59.53%	58.71%
	NN-2	59.86%	58.37%
	HYPLOSP	61.51%	62.69%
	HYPLOSP-2	61.02%	60.34%
PB	NN-1	59.54%	55.26%
	NN-2	60.97%	56.85%
	HYPLOSP	63.24%	58.66%
	HYPLOSP-2	64.48%	60.29%
STR	NN-1	58.78%	
	NN-2	59.19%	
	HYPLOSP	63.07%	
	HYPLOSP-2	63.42%	

4.4. Comparison with the standard two-stage neural network approach

Since the two-stage neural network approach is frequently applied to secondary structure prediction, we examine the effect of replacing the 1-stage NN in HYPLOSP with a 2-stage NN. The resulting hybrid system is called HYPLOSP-2. The two-stage neural network architecture and HYPLOSP-2 are described below.

A two-stage neural network for protein secondary structure prediction usually contains the first neural network that maps amino acid sequences (or profiles) to structures, and the second structure-to-structure neural network refines the results[26,27]. In our construction, the first neural network remains the same as described in Section 2.2.1, and the second neural network takes the output of the first network as input. The sliding window length which yields the best performance is 11, and the learning parameters are identical to section 2.2.2. The training and testing procedure of the second neural network are similar to Jones et al.[27], but uses structural letters instead of SSE.

The hybrid strategy of HYPLOSP-2 is basically the same as HYPLOSP. But the confidence score of neural network prediction is obtained from the output of the second network.

We use the nrDSSP dataset to compare the performance of one-stage and two-stage neural network prediction methods, HYPLOSP and HYPLOSP-2. The results are shown in Table VIII. We can observe HYPLOSP outperforms the two-stage NN method in every case, though the two-stage NN method produces slight improvements (0.33% to 1.59%) to the one-stage NN method (shown in Table III). Obviously, incorporating the knowledge-based method is more effective than adding a structure-to-structure neural network.

We further examine the performance of HYPLOSP-2 and two-stage NN method.

Incorporating the knowledge-based method to the two-stage NN method can still improve the performance (even if we do not fine-tune our hybrid mechanism). For example, HYPLOSP-2 improves the Q_N of the two-stage NN results by 0.57%, 1.25%, and 1.51% for SAH, PB, and STR.

5. Conclusion

Since existing local structure prediction methods are limited in performance, we use two different prediction methods: a knowledge-based method and a neural network-based method. To better utilize the advantages of these two methods, we propose a hybrid method, called HYPLOSP, which is alphabet-independent. We use three popular structural alphabets, SAH, PB, and STR, to evaluate the three methods and perform a 10-fold cross-validation test on nrDSSP containing nearly 4,000 protein chains. In addition, we also conduct a benchmark test that compares HYPLOSP with the prediction methods proposed by the authors of SAH, PB, and STR on a dataset of 56 protein chains. HYPLOSP shows promising results in terms of Q_N and MDA accuracy and also demonstrates its alphabet-independent capability. As more protein structures are determined, the knowledge-based method and HYPLOSP can be further improved, as evidenced by the increase in the number of proteins with higher global match rates and the analysis of the hybrid_benefit. We also analyze the relation between prediction accuracy and secondary structure information. The analysis shows that improving current secondary structure prediction accuracy can also enhance local structure prediction.

6. Acknowledgement

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. The authors are grateful to Dr. Ming-Jing Huang in Institute of Biomedical Sciences, Academia Sinica for fruitful discussions.

References

1. R Karchin, M Cline, K Karplus. Evaluation of local structure alphabets based on residue burial. *Proteins* **55**, 508-518 (2004).
2. R Karchin, M Cline, Y Mandel-Gutfreund, K Karplus. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **51**, 504-514 (2003).
3. C Bystroff, Y Shao. Fully automated ab initio protein structure prediction using I-Sites, HMMSTR and Rosetta. *Bioinformatics* **18**, 54-61 (2002).
4. K Karplus, R Karchin, J Draper, J Casper, Y Mandel-Gutfreund, M Diekhans, R Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53**, 491-496 (2003).
5. KT Simons, C Kooperberg, E Huang, D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225 (1997).

6. R Unger, D Harel, S Wherland, JL Sussman. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **5**, 355-373 (1989).
7. C Micheletti, F Seno, A Maritan. Recurrent Oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* **40**, 662-674 (2000).
8. J Schuchhardt, G Schneider, J Reichelt, D Schomburg, P Wrede. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng.* **9**, 833-842 (1996).
9. R Kolodny, P Koehl, L Guibas, M Levitt. Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.* **323**, 297-307 (2002).
10. C Bystroff, D Baker. Prediction of local structure in proteins using a library of sequence-structure motifs*. J. Mol. Biol.* **281**, 565-577 (1998).
11. SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, DJ Lipman. Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17), 3389-3402 (1997).
12. C Bystroff, V Thorsson, D Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* **301**, 173-190 (2000).
13. AG de Brevern, C Etchebest, S Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **41**, 271-287 (2000).
14. R Hughey, A Krogh. SAM: Sequence alignment and modeling software system, version3. Technical report UCSC-CRL-95-7. Santa Cruz, CA: University of California, Santa Cruz, Computer Engineering (1995).
15. R Hughey, K Karplus, A Krogh. SAM: Sequence alignment and modeling software system, version 3. Technical report UCSC-CRL-99-11. Santa Cruz, CA: University of California, Santa Cruz, Computer Engineering (1999). Available from http://www.soe.ucsc.edu/rsearch/compbio/ sam.html
16. AS Yang, LY Wang. Local structure prediction with local structure-based sequence profiles. *Bioinformatics* **19**(10), 1267-1274 (*2003)*.
17. R Kuang, CS Leslie, AS Yang. Protein backbone angle prediction with machine learning approaches. *Bioinformatics* **20**, 1612-1621 (2004).
18. T Tang, J Xu, M Li. Discovering sequence-structure motifs from protein segments and two applications. *Pacific Symposium on Biocomputing* (2005).
19. W Kabsch. and C Sander. Definition of secondary structure of proteins given a set of 3D coordinates. *Biopolymers* **22**, 2577-2637 (1983).
20. KD Pruitt, T Tatusova, DR Maglott. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acids Res.* **33**, 501-504 (2005).
21. D.Rumelhart, G. Hinton, and R. Williams. 1988. Learning internal representations by error propagation. *In Neurocomputing*, 675-695. Cambridge, MA: MIT Press
22. HN Lin, JM Chang, KP Wu, TY Sung, WL Hsu. A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* **21**, 3227-3233 (2005).
23. AG de Brevern, C Benros, R Gautier, H Valadié, H Hazout, C Etchebest. Local back-

bone structure prediction of proteins. *In silico Biology* **4**(3),381-6 (2004).
24. B Rost, VA Eyrich. EVA: large-scale analysis of secondary structure prediction. *Proteins* **5**, 192-199 (2001).
25. HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, PE Bourne. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).
26. B Rost, C Sander. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599 (1993).
27. DT Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202 (1999).

Ching-Tai Chen recieved his B.S. and M.S. degrees from the Department of Computer and Information Science, National Chiao Tung University, Taiwan, in 2000 and 2002, respectively. He is now a PhD student of Bioinformatics Program, Taiwan International Graduate Program (TIGP). His current research interests include protein structure prediction and machine learning applications.

Hsin-Nan Lin recieved his B.S. and M.S. degrees from the Department of Computer and Information Science, National Chiao Tung University, Taiwan, in 1997 and 1999, respectively. He is now a PhD student of Bioinformatics Program, Taiwan International Graduate Program (TIGP). His current research interests include protein structure prediction, Nuclear Magnetic Resonance (NMR) data analysis, and algorithm design.

Ting-Yi Sung is a research fellow in the Institute of Information Science, Academia Sinica, Taiwan. She received her Ph.D. in Operations Research from New York University in 1989. Since then, she has joined the institute. Her current research interests include biomedical literature mining, protein structure prediction, and analysis of high-throughput mass spectrometry data.

Wen-Lian Hsu is a research fellow in the Institute of Information Science, Academia Sinica. He is also an IEEE Fellow. He received his Ph.D. from Cornell University in 1980. Prior to joining the institute as a research fellow in 1989, he was an assistant professor and a tenured associate professor in Northwestern University. His research interests include biological knowledge base, intelligent knowledge management systems, biomedical literature mining, protein structure prediction, and design and analysis of algorithms for biological computing, and natural language processing.