BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 15 2005, pages 3227–3233
doi:10.1093/bioinformatics/bti524

Sequence analysis

HYPROSP II—A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence

Hsin-Nan Lin, Jia-Ming Chang, Kuen-Pin Wu, Ting-Yi Sung and Wen-Lian Hsu*

Institute of Information Science, Academia Sinica, Taipei 115, Taiwan

Received on December 29, 2004; revised on March 10, 2005; accepted on May 29, 2005

Advance Access publication June 2, 2005

ABSTRACT

Motivation: In our previous approach, we proposed a hybrid method for protein secondary structure prediction called HYPROSP, which combined our proposed knowledge-based prediction algorithm PROSP and PSIPRED. The knowledge base constructed for PROSP contains small peptides together with their secondary structural information. The hybrid strategy of HYPROSP uses a global quantitative measure, match rate, to determine whether PROSP or PSIPRED is to be used for the prediction of a target protein. HYPROSP made slight improvement of Q_3 over PSIPRED because PROSP predicted well for proteins with match rate $>80\%$. As the portion of proteins with match rate $>80\%$ is quite small and as the performance of PSIPRED also improves, the advantage of HYPROSP is diluted. To overcome this limitation and further improve the hybrid prediction method, we present in this paper a new hybrid strategy HYPROSP II that is based on a new quantitative measure called local match rate.

Results: Local match rate indicates the amount of structural information that each amino acid can extract from the knowledge base. With the local match rate, we are able to define a confidence level of the PROSP prediction results for each amino acid. Our new hybrid approach, HYPROSP II, is proposed as follows: for each amino acid in a target protein, we combine the prediction results of PROSP and PSIPRED using a hybrid function defined on their respective confidence levels. Two datasets in nrDSSP and EVA are used to perform a 10-fold cross validation. The average Q_3 of HYPROSP II is 81.8% and 80.7% on nrDSSP and EVA datasets, respectively, which is 2.0% and 1.1% better than that of PSIPRED. For local structures with match rate $>80\%$, the average Q_3 improvement is 4.4% on the nrDSSP dataset. The use of local match rate improves the accuracy better than global match rate. There has been a long history of attempts to improve secondary structure prediction. We believe that HYPROSP II has greatly utilized the power of peptide knowledge base and raised the prediction accuracy to a new high. The method we developed in this paper could have a profound effect on the general use of knowledge base techniques for various prediction algorithms.

Availability: The Linux executable file of HYPROSP II, as well as both nrDSSP and EVA datasets can be downloaded from http://bioinformatics.iis.sinica.edu.tw/HYPROSPII/

Contact: hsu@iis.sinica.edu.tw

*To whom correspondence should be addressed.

1 INTRODUCTION

Protein secondary structure prediction is to predict protein secondary structure based only on its sequence, where each amino acid is assigned a structure state, helix (H), strand (E) or coil (C). Protein secondary structure prediction plays an important role in tertiary structure prediction as it can be used to generate templates for tertiary structure predictions. Fischer and Eisenberg (1996) improved the tertiary structure prediction accuracy from 59.0 to 71.0% by using PHD to predict secondary structures. In Yang and Wang's paper (2003), the tertiary structure prediction accuracy was reduced from 79.0 to 71.9% after switching off the secondary structure prediction in the prediction procedure. McGuffin and Jones (2003) reported that the predicted secondary structure information definitely contributes to a better performance for tertiary structure prediction.

For a better prediction of secondary structure, Rost and Sander proposed a novel prediction method PHD, which uses evolutionary information and has gained significant improvements (Rost and Sander, 1993, 1994; Rost, 2001). Jones (1999) improved the prediction by using PSI-BLAST searches over large databases to obtain better evolutionary information. These two prevailing methods are based on the neural network approach and can achieve an accuracy of $\sim$80%. The advantage of the neural network approach is that evolutionary information, amino acid and structure propensities as well as global sequence compositions can all be taken into account. A drawback of this approach is that, it is unclear how the additional evolutionary information affects the prediction accuracy. The inside of neural network algorithms is hard to understand and to translate into useful knowledge. Machine learning approaches other than the neural network are also used for secondary structure prediction (Hua and Sun, 2001; Kim and Park, 2003), and they have different limitations.

As local structural libraries are frequently encoded in short segments of protein sequences (Alm *et al*., 2002; Yang and Wang, 2003), another line of prediction approach is to use local structure-based sequence databases. This motivated us to design a knowledge-based prediction algorithm PROSP (Wu *et al*., 2004), which uses a peptide sequence-structure knowledge base and a voting scheme for prediction. In order to combine the strength of machine learning approaches, we proposed a hybrid prediction method called HYPROSP (Wu *et al*., 2004), which combines PROSP and PSIPRED. We used a quantitative measure called match rate to determine whether PROSP or PSIPRED should be used to predict the

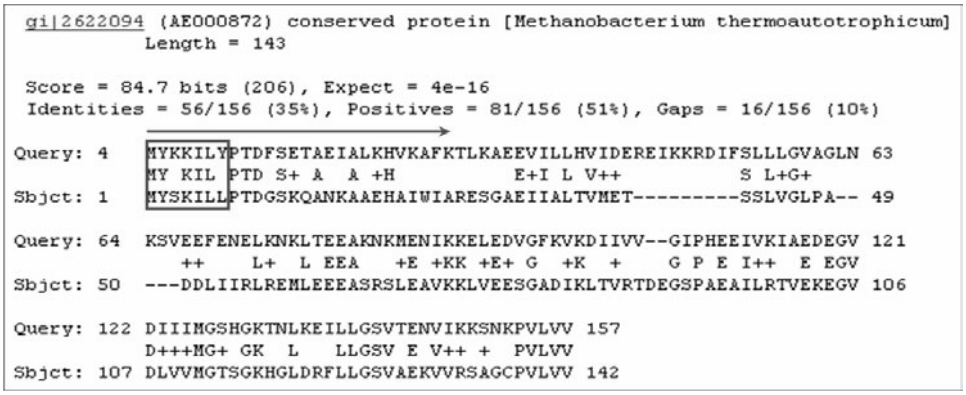

Fig. 1. An example of HSP found by PSI-BLAST. The first peptide pairs as marked by the box are similar, and we assign the secondary structure element of each amino acid in MYKKILY to its counterpart in MYSKILL.

structure of a target protein (i.e. a protein whose structure is unknown and targeted for prediction). The match rate defined in HYPROSP (referred to as the global match rate in this paper) is a global measure for the amount of structural information that a target protein can extract from the knowledge base. Our experiments show that the prediction accuracy of PROSP has a significant positive correlation with the global match rate. The hybrid strategy of HYPROSP is as follows: if the global match rate of a target protein is at least 80%, we use PROSP to predict the protein; otherwise, we use PSIPRED.

HYPROSP made a slight improvement of Q_3 [i.e. the average of $Q_3(p)$] over PSIPRED in several datasets. However, there are two limitations. First, the proportion of proteins with global match rate $>80\%$ is often not large enough, so the improvement could be diluted. Second, as the prediction accuracy, Q_3, of PSIPRED has also been improved from 76% as reported in EVA web site to 79% using version 2.45 on the nrDSSP dataset, the marginal advantage of HYPROSP becomes small. To reduce the effect of these two limitations, we introduce two new concepts: (1) we consider a new quantitative measure called local match rate as opposed to the global match rate defined in HYPROSP; (2) we propose a new hybrid strategy called HYPROSP II, which combines the results of PROSP and PSIPRED based on their confidence levels. This new method achieves much better Q_3 than both HYPROSP and PSIPRED. The Q_3 of HYPROSP II on the nrDSSP and EVA datasets are 2.0% and 1.1% better than that of PSIPRED, respectively.

2 METHODS

2.1 Constructing a peptide sequence-structure knowledge base (SSKB)

Our knowledge base is constructed from a structure database, e.g. DSSP, that contains peptide sequences and their structural information. The success of knowledge-based prediction approaches depends heavily on the size of the knowledge base. In order to amplify the knowledge base, we use PSI-BLAST (Altschul *et al*., 1997) to find in a chosen sequence database (e.g. NCBI nr) proteins remotely homologous to those in a structure database so that peptides of these remotely homologous proteins would inherit the structures of their counterparts in the structure database.

Taking a protein sequence as input, PSI-BLAST can generate a large number of significant local pairwise alignments called high-scoring segment pairs (*HSPs*) between the input protein and homologous proteins, as well as a profile called position-specific scoring matrix (*PSSM*). Most structure prediction methods use the PSSM profile as the source of evolutionary information. Since we assume that the counterpart sequence (denoted by 'Sbjct' in the PSI-BLAST output) in an HSP has a similar structure to the input sequence (denoted by 'Query' in the PSI-BLAST output) we use the HSPs instead, which provide explicit information of sequence variations. Peptides in HSPs will be chosen according to a similarity criterion (explained later) to be included in the knowledge base.

To construct the peptide sequence-structure knowledge base (SSKB), we use proteins of a structure database and select those proteins with $<25\%$ sequence identity among each other. PSI-BLAST is used to search homologous proteins from a sequence database of each protein, where the parameter j is 3 (three iterations), e is 10 (E-value < 10) and the sequence database is NCBI nr. If the input protein (Query) has homologous proteins, PSI-BLAST will return a number of HSPs. An example of an HSP is shown in Figure 1, where the homologous protein (Sbjct) is gi|2622094 and the alignment score is 84.7.

Given an HSP, we choose 'similar' peptides to be included in the knowledge base as follows. Use a sliding window of length w (where w is chosen to be 7 according to our previous work on HYPROSP) in HSPs to define peptide segments (in short, peptides). Define the similarity level between two corresponding peptides in an HSP as the number of exact matches and positive signs in the aligned amino acids. Two peptides are considered similar if the similarity level between those two peptides is at least k. k is chosen to be 3 by Wu *et al*. (2004). For example, the two peptides in the box shown in Figure 1 have the similarity level 5 and are considered similar. If two peptides are similar, the peptide in Sbjct would inherit the structure of its counterpart in Query. Note that if two similar peptides contain one or more gaps, then they are discarded. Besides the sequence and structural information, their confidence score will also be stored in the knowledge base. The confidence score is defined as follows: Let p_f and q_f denote a pair of similar peptides, where p_f is in Query and q_f is in Sbjct. We assign q_f the structure of p_f (per amino acid) with a confidence score $S(p_f q_f) = (t \times s)/7$, where t (≥ 2) is the similarity level between p_f and q_f and s is the alignment score. Intuitively, larger t and s generate a larger confidence score $S(p_f q_f)$. Finally, for each such q_f, we store the record [q_f, structure of p_f, $S(p_f q_f)$] in SSKB.

When adding a new peptide to the knowledge base, if an identical peptide is found, we simply add the new confidence score to the corresponding structure of each amino acid in the peptide record regardless of whether their structural information is identical or not. Table 1 illustrates an example of a peptide record, where the peptide MYSKILL is added into the knowledge base twice (note that only MYKKILY is illustrated in Figure 1) since it is similar to both MYKKILY and MYSSIIL and inherits their structures. To determine the representative structure of a peptide record, we choose the structure type with maximum score at each position. The representative structure of this example is 'CCHHHHC'. After all HSPs of known structure proteins are

Table 1. Example of a peptide record

Peptide	Alignment score	Similarity level	Confidence score	Structure
A				
MYKKILY	85	5	60.7	CCHHHHC
MYSSIIL	76	4	43.4	HHHCCCC

Peptide fragment	M	Y	S	K	I	L	L
B							
H	43.4	43.4	104.1	60.7	60.7	60.7	0.0
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	60.7	60.7	0.0	43.4	43.4	43.4	104.1

(A) Two peptides with known structure that are similar to the peptide MYSKILL. (B) A peptide record MYSKILL in the knowledge base constructed from MYKKILY and MYSSIIL. Each record stores the confidence scores of three secondary structure states at each position.

scanned, we can generate tens of millions diverse peptide records with their structural information in the SSKB.

2.2 PROSP: a structure prediction method based on the SSKB

The construction of our secondary structure prediction procedure, PROSP (Wu *et al*., 2004), consists of three parts:

(1) Construct the knowledge base SSKB;

(2) Use PSI-BLAST to find all peptides similar to those of the target protein;

(3) Use similar peptides found in the SSKB to vote for the dominant structure of each amino acid in the target protein.

To predict the secondary structure of a target protein p, we first use PSI-BLAST to find all HSPs. The parameters and the sequence database used in PSI-BLAST are the same as those used in the construction of SSKB. For each HSP, we use a sliding window of length 7 to divide the aligned sequences into peptides. Define the similarity level between two peptides the same way as before. Each amino acid at position x of p is associated with three variables: $H(x)$, $E(x)$ and $C(x)$, which are the confidence levels corresponding to the three secondary structure elements, H, E and C, respectively. The structure at x is predicted to be H, E or C depending on $\mathbf{Max}\{H(x), E(x), C(x)\}$. We use similar peptides obtained from HSPs to calculate $H(x)$, $E(x)$ and $C(x)$. Let p_f and q_f be a pair of similar peptides with a similarity level t and alignment score s in an HSP, where p_f is a peptide in the target protein p and q_f is a peptide of a sequence in the NCBI nr database. If q_f is not in SSKB, then it is ignored. Otherwise, all structural information contained in q_f is added to p_f, and is updated

$$H(p_f[i]) \leftarrow H(p_f[i]) + H(q_f[i]) \times s \times t/7,$$
$$E(p_f[i]) \leftarrow E(p_f[i]) + E(q_f[i]) \times s \times t/7,$$
$$C(p_f[i]) \leftarrow C(p_f[i]) + C(q_f[i]) \times s \times t/7,$$

where $p_f[i]$ and $q_f[i]$ for $1 \leq i \leq 7$ denote the i-th position of p_f and q_f, respectively. Repeat the above calculation for all similar peptides containing position x and assign the structure at x according to $\mathbf{Max}\{H(x), E(x), C(x)\}$.

2.3 Two match rates

In this section, we define the global match rate for a target protein and the local match rate for a residue in a target protein. Given a target protein p, we obtain all of its similar peptides using HSP. A pair of similar peptides is denoted by p_f and q_f, where p_f is a peptide in the target protein p. Let Q_f be the collection of all those q_f's. Note that not all q_f's are in SSKB. The global match rate of the target protein p is defined as follows:

$$\text{Global match rate} = \frac{|\ Q_f \cap \text{SSKB}\ |}{|\ Q_f\ |} \times 100\%$$

The global match rate represents the percentage of peptides of the target protein that can find similar peptides in the knowledge base. Intuitively, when the global match rate is higher, the structural information obtained for prediction is more reliable. HYPROSP uses PROSP to predict proteins whose global match rate is at least 80% and relies on PSIPRED for those <80%. However, it is relatively hard for target proteins to attain a global match rate >80% when the knowledge base is not big enough.

To improve HYPROSP and to further utilize the knowledge in SSKB, we consider a new quantitative measure called local match rate, which is defined on each position x of the target protein p. Let $Q_f(x)$ be the collection of all similar peptides q_f's containing the position x. We define the local match rate as follows:

$$\text{Local match rate}(x) = \frac{|\ Q_f(x) \cap \text{SSKB}\ |}{|\ Q_f(x)\ |} \times 100\%$$

The local match rate can be regarded as the confidence level of using PROSP to predict the structure at position x. Note that we can have high local match rates at many positions even though the global match rate is low.

2.4 HYPROSP II: a hybrid method based on local prediction confidence

In HYPROSP (Wu *et al*., 2004), the hybrid strategy is to use either the result of PROSP or that of PSIPRED for prediction. In contrast, HYPROSP II predicts the structure at each position by combining the results of these two methods. Given a target protein p, we can obtain two prediction results by PROSP and PSIPRED. The source code of PSIPRED was modified to report not only the prediction result but also three confidence values: $psi\_h(x)$, $psi\_e(x)$ and $psi\_c(x)$. For each amino acid at location x of p, the confidence of PROSP is defined as follows:

$$pro\_h(x) = \frac{\text{Local match rate}(x) \times H(x)}{H(x) + E(x) + C(x)}$$
$$pro\_e(x) = \frac{\text{Local match rate}(x) \times E(x)}{H(x) + E(x) + C(x)}$$
$$pro\_c(x) = \frac{\text{Local match rate}(x) \times C(x)}{H(x) + E(x) + C(x)}$$

And the final predicted structure at position x is determined by the following hybrid function called *hyprosp*_II(x):

$$hyprosp\_\text{II}(x) = \begin{cases} H & \text{if } pro\_h(x) + psi\_h(x) \text{ is max;} \\ E & \text{if } pro\_e(x) + psi\_e(x) \text{ is max;} \\ C & \text{if } pro\_c(x) + psi\_c(x) \text{ is max.} \end{cases}$$

In case the sums of these three structure states are equal, the selection priority will be C followed by H and then E, which is based on their occurrence frequencies in the structure database.

3 IMPLEMENTATION

HYPROSP II is developed under Linux Redhat 9.0; it is implemented as a C++ MPI application suit that runs on a PC cluster of 13 nodes; each node contains a Pentium-4 Xeon 2.8 GHz CPU with 2 GB main memory and a 30 GB hard disk.

3.1 Datasets

Two datasets are used to evaluate HYPROSP II. We download 25 288 proteins from the DSSP database (dated September 22, 2004), and separate these proteins into 46 745 protein chains. Each protein chain

is then checked to find out whether it has homologous protein chains or not by PSI-BLAST and pairwise sequence alignment. If homologous protein chains with sequence identity >25% are found, only one of them is retained. Moreover, we filter out protein chains of length <80. At the end, we have a non-redundant DSSP dataset, called nrDSSP, that contains 3925 unique protein chains (with <25% mutual sequence identity).

Another dataset, EVA, containing 2217 protein chains is downloaded from the EVA server http://cubic.bioc.columbia.edu/eva/doc/ftp.html, which has long been regarded as a benchmark to evaluate protein secondary structure algorithms. The original EVA dataset contains 3107 protein chains (latest list: 2004/05/09); however, only 2217 of them can be identified among 46 745 protein chains in the DSSP database. The 'EVA dataset' mentioned in the rest of the paper refers to these 2217 protein chains.

The DSSP dataset has eight structure states, H, I, G, E, B, S, T and '–' (blank). We follow EVA's convention to reduce them into three states: H, G, I to H; E, B to E; and other states to C. The proportions of the three structural states, H, E and C are 36.36, 22.21 and 41.43%, respectively, in the nrDSSP dataset, and 36.28, 22.19, 41.53%, respectively, in the EVA dataset.

3.2 Experiment design

We use nrDSSP and EVA datasets to test the performance of HYPROSP II. All experiments are performed in a 10-fold cross validation. In a 10-fold cross validation, we divide the dataset into 10 subsets in which one of the 10 subsets is used as the testing set, the other 9 subsets are pooled together to form the training set, and the procedure is repeated 10 times for each subset to be chosen as the testing set in turn. For each chosen testing set, the knowledge base SSKB is reconstructed from proteins in the corresponding training set. The knowledge bases generated by using nrDSSP and EVA datasets contain in average 48 298 002 and 31 342 754 peptide records, respectively. We use Q_3 and SOV (Segment OVerlap measure) to evaluate the performance of different prediction methods.

Finally, we download the latest three months of proteins (September 2004–November 2004) from EVA server (Koh *et al*., 2003) as testing dataset and use the nrDSSP dataset as training dataset to evaluate the performance of HYPROSP II and compare with other methods.

3.3 Experimental results on the nrDSSP dataset

The prediction accuracy Q_3 using nrDSSP dataset with respect to the global match rate for HYPROSP II, PROSP and PSIPRED is shown in Figure 2. Note that, the accuracy Q_3 shown at global match rate $\geq k$% represents the Q_3 of proteins with global match rates at least k%. For example, we can find from Figure 2 that the Q_3 of all proteins (i.e. global match rate at least 0%) in the nrDSSP dataset are 81.8, 70.6 and 79.9% using HYPROSP II, PROSP and PSIPRED, respectively, and 85.2, 81.5 and 81.2% for prediction of proteins with global match rates at least 80%. In addition, Figure 2 shows that the Q_3 using HYPROSP II increases for proteins with increasing global match rate. Note that there is a monotone positive correlation between prediction accuracy of PROSP and the global match rate (Wu *et al*., 2004). The average Q_3 and SOV scores of HYPROSP, HYPROSP II and PSIPRED are listed in Table 2.

To further compare HYPROSP II and PSIPRED, we examine the prediction accuracies of 3925 protein chains in the nrDSSP dataset. Among them, HYPROSP II surpasses PSIPRED in 2634 chains

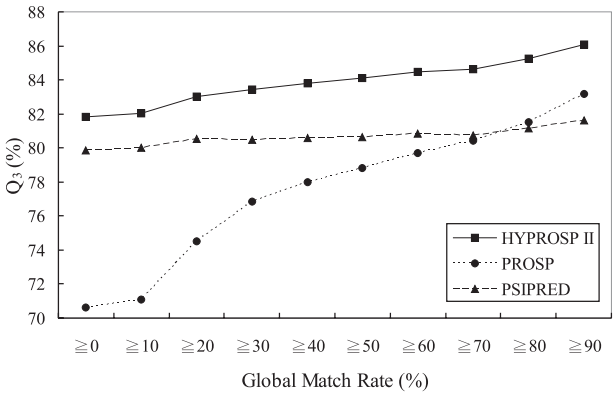

Fig. 2. Q_3 of HYPROSP II, PROSP and PSIPRED with respect to global match rate on the nrDSSP dataset. Note that the prediction accuracy of HYPROSP can be easily derived from the figure: Its Q_3 is the same as PSIPRED if target proteins have global match rate <80%; on the other hand, its Q_3 is the same as PROSP if target proteins have global match rate of at least 80%.

(67%) with an average improvement of 3.6%. HYPROSP II is inferior to PSIPRED in 928 protein chains (24%) with an average reduction of 1.7%. And the remaining 363 proteins chains (9%) have the same accuracies for both methods.

3.4 Experimental results on the EVA dataset

Figure 3 shows the experimental results using the EVA dataset. Note that the Q_3 marked at global match rate $\geq k$% represents the Q_3 for proteins with global match rate at least k%. Because the size of SSKB constructed from the training set of the EVA dataset is much less than that constructed from the nrDSSP dataset (31 342 754 versus 48 298 002), the average Q_3 improvement of HYPROSP II over PSIPRED for all proteins is only 1.1%. Nonetheless, the average improvement Q_3 over PSIPRED is 4.0% for proteins with the global match rate of at least 80%. And at global match rate 80%, PROSP turns better than PSIPRED. It means that even though the size of SSKB constructed from the EVA dataset is not big enough, the prediction accuracy for those proteins with global match rates of at least 80% is still satisfactory. The average Q_3 and SOV scores of HYPROSP, HYPROSP II and PSIPRED are listed in Table 2.

3.5 Experimental results on the new EVA dataset

There are 27 testing proteins obtained from the latest three months on the EVA server. We use the nrDSSP dataset to construct SSKB for predicting the structure of these proteins whose sequence identities are all <25% against the nrDSSP dataset. The average global match rate of these proteins is only 24.7%, where the largest one is <75%. The average Q_3 and SOV scores are 77.5 and 74.3%. The average Q_3 improvement of HYPROSP II over PSIPRED is >2.1%. Table 3 shows the comparisons of Q_3 and SOV scores of HYPROSP II, PSIPRED, PROFsec, PHDpsi (Przybylski and Rost, 2002), SABLE2 (Porollo *et al*., 2003) and PROF_king (Ouali and King, 2000), which are reported on the EVA server except for HYPROSP II and PSIPRED.

4 DISCUSSION

There are two factors affecting the performance of HYPROSP II, which are discussed in this section.

Table 2. The average Q_3 and SOV scores of HYPROSP, HYPROSP II and PSIPRED on nrDSSP and EVA datasets

	Q_3	Q_3H_O	Q_3H_P	Q_3E_O	Q_3E_P	Q_3C_O	Q_3C_P	SOV	SOVH	SOVE	SOVC
A. nrDSSP dataset											
HYPROSP II	81.8	81.2	80.9	71.9	78.9	80.0	79.4	78.1	80.8	76.7	73.4
Errsig	0.1	0.3	0.3	0.4	0.3	0.2	0.2	0.2	0.3	0.4	0.2
HYPROSP	80.0	79.6	79.3	72.2	73.6	76.7	78.3	76.4	79.4	76.2	70.8
Errsig	0.1	0.4	0.3	0.4	0.4	0.2	0.2	0.2	0.4	0.4	0.2
PSIPRED	79.9	78.4	80.6	70.4	73.4	77.2	78.0	76.9	78.5	74.8	71.9
Errsig	0.1	0.4	0.3	0.4	0.3	0.2	0.1	0.2	0.4	0.4	0.2
B. EVA dataset											
HYPROSP II	80.8	79.8	80.0	69.1	76.6	79.5	78.5	77.0	79.5	74.2	72.7
Errsig	0.2	0.5	0.4	0.5	0.5	0.2	0.2	0.2	0.5	0.5	0.3
HYPROSP	79.8	78.3	80.1	70.6	72.7	77.0	78.2	76.6	78.6	74.7	71.6
Errsig	0.2	0.5	0.4	0.5	0.5	0.2	0.2	0.2	0.5	0.5	0.3
PSIPRED	79.8	77.8	80.6	70.2	72.8	77.2	78.1	76.9	78.1	74.5	72.2
Errsig	0.2	0.5	0.4	0.5	0.5	0.2	0.2	0.2	0.5	0.5	0.3

Errsig is the significant difference margin for each score and is defined as the standard deviation (σ) over the square root of the number of proteins ($\sqrt{N}$). Q_3H/E/C and SOVH/E/C values are the specific Q_3 and SOV scores of the predicted helix, strand and coil regions, respectively. Q_3H_O (Q_3E_O and Q_3C_O, respectively) represents correctly predicted helix (strand and coil, respectively) residues (percentage of helix observed), and Q_3H_P (Q_3E_P and Q_3C_P, respectively) represents correctly predicted helix (strand and coil, respectively) residues (percentage of helix predicted).

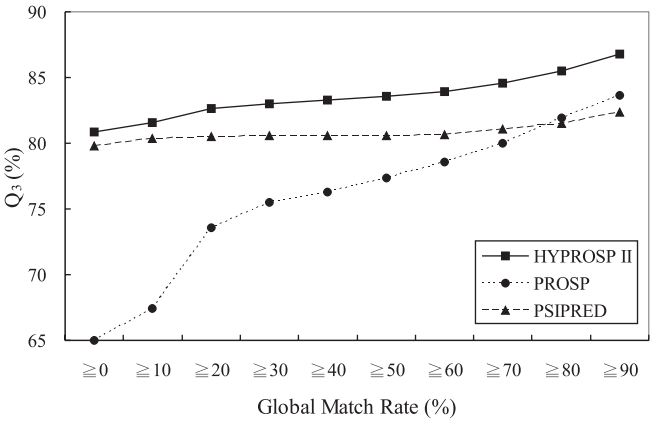

Fig. 3. The Q_3 of HYPROSP II, PROSP and PSIPRED with respect to global match rate using EVA dataset. Note that the prediction accuracy of HYPROSP can also be easily derived from the figure: Its Q_3 is the same as PSIPRED if target proteins have global match rate <80%; on the other hand, its Q_3 is the same as PROSP if target proteins have global match rate of at least 80%.

4.1 The effect of dataset size on the performance

To analyze the effect of dataset size on the performance, we calculate the numbers of proteins with different global match rates. In Table 4 we show the proportions of proteins in nrDSSP and EVA datasets with respect to different global match rates. For example, the proportion of proteins with global match rates at least 50% in nrDSSP is 51.8%; namely, over half of proteins in the nrDSSP dataset have global match rates at least 50%. At the same threshold of global match rate, however, the proportion of proteins in the EVA dataset is only 24.5%. The size of SSKB affects the proportion of proteins at different global match rates. Intuitively, the more peptides contained in the knowledge base, the more proteins obtain higher global match rates, and we get better prediction accuracy. If the number of proteins with known structures increases in the future, the knowledge base will be increased, and so will the prediction accuracies of PROSP and HYPROSP II.

4.2 Evaluation of the hybrid function

The performance of a hybrid prediction approach depends on the underlying prediction methods and the hybrid function. In this section, we analyze our hybrid strategy and introduce a measure called hybrid precision to analyze its performance; we use experimental results on the nrDSSP dataset for discussion.

For each position x in a target protein p, let *prosp*(x), *psipred*(x) and *hyprosp*_II(x) denote the prediction results of x by PROSP, PSIPRED and HYPROSP II, respectively. (Recall that *hyprosp*_II(x) is defined in Section 2.4.) We first consider the case where *prosp(x)* = *psipred(x)*. Note that *prosp(x)* = *psipred(x)* implies *hyprosp*_II(x) = *prosp(x)*. Then all three predictions are correct or none is correct. Experimental results show that 71.1% of the entire dataset belongs to this case, and 86.7% of this case can generate correct predictions. However, 13.3% of this case cannot generate correct predictions which is 9.46% (=13.3% × 0.711) of the entire dataset; and any hybrid function based on PROSP and PSIPRED can hardly improve the prediction on this part.

Now we consider the remaining case where *prosp(x)* ≠ *psipred(x)*. HYPROSP II makes a hybrid prediction according to *hyprosp*_II(x). The measure hybrid precision to evaluate the performance of the hybrid function *hyprosp*_II(x) of HYPROSP II is defined as follows:

$$\text{hybrid precision} = \frac{\text{number of residues correctly assigned by } \textit{hyprosp}\_\text{II}(x)}{\text{number of cases, where } \textit{prosp}(x) \neq \textit{psipred}(x)} \times 100\%$$

Hybrid precision indicates the proportion of the data in this case that can generate correct prediction in spite of the underlying prediction

Table 3. The average Q_3 and SOV scores of the latest 27 testing proteins from the EVA server by different methods

	Q_3	Q_3H_O	Q_3H_P	Q_3E_O	Q_3E_P	Q_3C_O	Q_3C_P	SOV	SOVH	SOVE	SOVC
HYPROSP II	77.5	73.8	75.3	58.7	64.7	81.6	77.7	74.3	79.3	65.1	74.5
Errsig	2.2	5.7	6.0	7.5	6.3	2.7	2.6	3.2	5.5	6.4	3.0
PSIPRED	75.4	71.9	71.2	61.3	57.3	76.7	78.4	72.4	73.2	66.2	72.0
Errsig	2.3	6.5	6.3	7.4	7.0	3.0	2.4	3.2	6.3	6.7	2.9
PROFsec	74.4	63.5	67.1	48.4	46.4	76.8	74.0	72.0	79.7	69.1	69.9
Errsig	2.1	6.0	6.4	7.0	7.0	2.5	2.2	3.6	4.7	5.8	3.7
PHDpsi	73.9	69.7	66.0	43.1	43.2	74.7	75.0	70.6	83.0	62.0	68.8
Errsig	2.4	6.2	6.1	6.7	7.1	3.1	2.3	3.6	4.6	6.1	3.7
SABLE2	73.4	64.5	70.6	44.7	45.1	76.1	74.2	69.5	78.0	63.4	69.2
Errsig	2.3	6.2	6.3	6.9	7.2	2.8	2.7	3.4	4.3	6.4	3.3
PROF_king	72.0	57.0	68.7	42.2	38.1	78.0	70.8	67.1	70.5	61.3	65.8
Errsig	2.1	6.0	6.6	7.5	7.2	2.4	3.1	3.3	5.7	7.1	3.5

Table 4. The proportions of proteins in nrDSSP and EVA datasets with respect to different global match rates

(%)	nrDSSP (%)	EVA (%)
≥0	100.0	100.0
≥10	97.7	68.6
≥20	78.5	38.7
≥30	64.5	33.7
≥40	57.7	28.8
≥50	51.8	24.5
≥60	45.5	20.2
≥70	38.1	15.1
≥80	27.3	9.7
≥90	12.3	4.9

Table 5. Prediction results of the case where $prosp(x) \neq psipred(x)$ on nrDSSP

	Proportion in the entire dataset (%)	Hybrid precision (%)
Either *prosp*(*x*) or *psipred*(*x*) is correct	26.4	73.3
Neither *prosp*(*x*) nor *psipred*(*x*) is correct	2.5	5.9

methods generating inconsistent predictions. Table 5 shows the prediction results on the nrDSSP dataset.

The hybrid precision reflects the performance of the hybrid function. When one of the underlying prediction methods can generate correct prediction, the accuracy of our hybrid function *hyprosp*_II(x) is 73.3%. It shows the advantage of our hybrid function. In Figure 4 we illustrate the cubic regression analysis of the relationship between Q_3 of HYPROSP II and hybrid precision. It is clear that higher hybrid precision gets better Q_3.

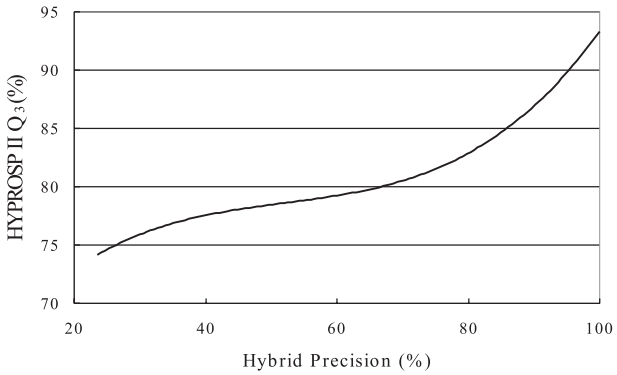

Fig. 4. The cubic regression line shows the relationship between Q_3 of HYPROSP II and hybrid precision.

5 CONCLUSIONS

For a target protein, our previous approach, HYPROSP, selects either PROSP or PSIPRED to predict its secondary structure; the selection is based on a global match rate with a cutoff threshold at 80%. However, such a hybrid approach cannot benefit those target proteins with <80% global match rate. In this paper, we define a new local quantitative measure, local match rate, to further utilize the useful information provided by both PROSP and PSIPRED. According to local match rate, we can define a prediction confidence level for each amino acid by each method, which can be used by HYPROSP II to combine the prediction results of PROSP and PSIPRED effectively. When compared with PSIPRED, the Q_3 of HYPROSP II is 2.0% better than that of PSIPRED, which is statistically significant at $p = 1.2E(-203)$. In contrast to HYPROSP, HYPROSP II performs better than PSIPRED even when the global match rate is zero. This is a great advantage over HYPROSP since the proportion of target proteins with global match rates of at least 80% could be limited. Even if there is no testing protein with global match rate >80% in the new EVA dataset, the average Q_3 improvement of HYPROSP II still achieves 2.1% against PSIPRED.

HYPROSP and HYPROSP II are hybrid prediction methods based on PROSP and PSIPRED. The performances of these hybrid prediction approaches rely largely on the underlying prediction methods and the hybrid function. We introduce a measure hybrid

precision to evaluate the performance of the hybrid function when the underlying prediction methods generate inconsistent predictions. When PROSP and PSIPRED generate inconsistent predictions and one of the predictions is correct, HYPROSP II has a precision of 73.3% using the hybrid function *hyprosp_*II(x). A better hybrid function is desirable to enhance the hybrid precision. However, when neither PROSP nor PSIPRED generates a correct prediction, the hybrid approach can hardly improve the performance, in which case a different strategy is necessary.

ACKNOWLEDGEMENTS

We would like to thank David Jones for providing the PSIPRED program in the public domain ftp://bioinf.cs.ucl.ac.uk/pub/psipred/. Also, we would like to thank S.F. Altschul *et al*. for providing the stand-alone BLAST program ftp://ftp.ncbi.nih.gov/blast/executable/. We would like to thank Dr Frank Hsu for pointing out (Hsu *et al*., 2002) to our attention and useful discussions. This work is partially supported by the thematic program of Academia Sinica under Grant AS9711S7PP and by the National Science Council, Taiwan under Grant NSC93-2213-E-001-024.

Conflict of Interest: none declared.

REFERENCES

Alm,E. *et al*. (2002) Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.*, **322**, 463–476.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Fischer,D. and Eisenberg,D. (1996) Protein fold recognition using sequence-derived predictions. *Protein Sci.*, **5**, 947–955.

Hsu,D.F., Shapiro,J. and Taksa,I. (2002) Methods of data fusion in information retrieval: rank vs. score combination. *DIMACS Technical Report* 58.

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

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

Kim,H. and Park,H. (2003) Protein secondary structure prediction by support vector machines and position-specific scoring matrices. *Protein Eng.*, **16**, 553–560.

Koh,I.Y.Y. *et al*. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.

McGuffin,L.J. and Jones,D.T. (2003) Benchmarking secondary structure prediction for fold recognition. *Proteins*, **52**, 166–175.

Ouali,M. and King,R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.*, **9**, 1162–1176.

Porollo,A., Adamczak,R., Wagner,M. and Meller,J. (2003) Maximum feasibility approach for consensus classifiers: applications to protein structure prediction. In *Proceedings of the Second International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2003)*, Centre for Intelligence Control, Singapore.

Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.

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

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

Rost,B. and Sander,C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Wu,K.P. *et al*. (2004) HYPROSP: a hybrid protein secondary structure prediction algorithm—a knowledge-based approach. *Nucleic Acids Res.*, **32**, 5059–5065.

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