BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 1 2012, pages 32–39
doi:10.1093/bioinformatics/btr611

Sequence analysis

Advance Access publication November 7, 2011

A novel structural position-specific scoring matrix for the prediction of protein secondary structures

Dapeng Li, Tonghua Li*, Peisheng Cong, Wenwei Xiong and Jiangming Sun

Department of Chemistry, Tongji University, Shanghai 200092, China

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The precise prediction of protein secondary structure is of key importance for the prediction of 3D structure and biological function. Although the development of many excellent methods over the last few decades has allowed the achievement of prediction accuracies of up to 80%, progress seems to have reached a bottleneck, and further improvements in accuracy have proven difficult.

Results: We propose for the first time a structural position-specific scoring matrix (SPSSM), and establish an unprecedented database of 9 million sequences and their SPSSMs. This database, when combined with a purpose-designed BLAST tool, provides a novel prediction tool: SPSSMPred. When the SPSSMPred was validated on a large dataset (10 814 entries), the Q3 accuracy of the protein secondary structure prediction was 93.4%. Our approach was tested on the two latest EVA sets; accuracies of 82.7 and 82.0% were achieved, far higher than can be achieved using other predictors. For further evaluation, we tested our approach on newly determined sequences (141 entries), and obtained an accuracy of 89.6%. For a set of low-homology proteins (40 entries), the SPSSMPred still achieved a Q3 value of 84.6%.

Availability: The SPSSMPred server is available at http://cal.tongji.edu.cn/SPSSMPred/

Contact: lith@tongji.edu.cn

Received on August 8, 2011; revised on October 10, 2011; accepted on October 28, 2011

1 INTRODUCTION

Even prior to the determination of the protein structures of hemoglobin and myoglobin by X-ray diffraction analysis (Kendrew *et al.*, 1958; Muirhead and Perutz, 1963), activity in the field of protein structure study had been increasing steadily; after this point and for some decades, the field attracted intense interest. It is known that the structure of a protein determines its function, and understanding the functions of proteins (such as catalysis, transport, immunity, body defense and so on) is of fundamental importance in the discovery of drugs to treat various diseases. Knowledge of protein structures is therefore highly desirable.

In the protein structure hierarchy, there are four distinct levels—the primary, secondary, tertiary and quaternary. Among these, the protein secondary structure occupies an important position, as it is the basis for the spatial structure of a protein. The secondary structure is formed at an early stage of protein folding, so the study of protein secondary structures is indispensable as the first and the most important step in 3D structure studies. The protein secondary structure has also been found to be instrumental in affecting the performance in predicting the tertiary structure (Clementi *et al.*, 2003; Monge *et al.*, 1994), subcellular localization (Nair and Rost, 2003) and so on.

Unfortunately, experimental methods for the detection of protein secondary structure are time consuming and labor intensive. The great disparity between the known protein sequences stored in the UniProt (Wu *et al.*, 2006) and detected protein structures deposited in the Protein Data Bank (PDB) (Rose *et al.*, 2011) continues to grow larger. To fill this void, the identification of protein secondary structures in terms of the three states of α-helix, β-sheet and random coil has been carried out using their amino acid sequences; this technique has become increasingly prominent. In spite of the many efforts made by researchers over the last few decades, the prediction of protein secondary structures from their amino acid sequences is still difficult.

Looking back on some 40 years of protein secondary structure prediction work, it might be possible to discern two categories that encompass the majority of the research—template-based methods and machine-learning methods. Template-based methods focused on connections between a query sequence and template pool sequences with known structures. The two most successful template-based methods were NNSSP (Salamov and Solovyev, 1997) and PREDATOR (Frishman and Argos, 1997). By comparison, machine-learning methods generated a learning model via the use of a series of proteins with known structures for prediction. In this category, Artificial Neural Networks (Babaei *et al.*, 2010; Chen and Chaudhari, 2007), Support Vector Machines (Chen *et al.*, 2007; Chen *et al.*, 2009; Hu *et al.*, 2004; Nguyen and Rajapakse, 2003; Ward *et al.*, 2003) and Hidden Markov Models (Aydin *et al.*, 2006; Di Francesco *et al.*, 1997; Zheng, 2005) were the most widely used algorithms. Machine-learning methods have been deemed to be the most effective and robust, and have been demonstrated in numerous successful examples that often led to near-perfect predictions.

The development of the now widely adopted machine-learning methods underwent three stages. During the initial stage, simple methods were used for the prediction of structures, and these methods suffered from a lack of data. Predictions were based on sequence compositions and physical and chemical properties; probably, the most famous early methods from that exploratory stage were those proposed by Chou and Fasman (1974a, b), Garnier, Osguthorbe and Robson (GOR) (Garnier *et al.*, 1978) and

*To whom correspondence should be addressed.

Lim (1974). They achieved accuracies of 56–60%, as assessed by Kabsch and Sander (1983a, b). Later, methods were improved in many aspects, and the accuracy performance improved (Deleage and Roux, 1987; Holley and Karplus, 1989; King and Sternberg, 1990; Kneller *et al.*, 1990; Presnell *et al.*, 1992). The second stage arrived with the availability of large families of homologous sequences, which revolutionized the prediction of secondary structures. The combination of sequence alignments generated from a series of protein families with sophisticated computing techniques such as neural networks led to accuracies well in excess of 70%. Many good methods for the prediction of secondary structures from multiply aligned protein sequences emerged in that period, such as PHD (Rost and Sander, 1993), ZPRED (Zvelebil *et al.*, 1987), NNSSP (Salamov and Solovyev, 1997), SSPRED (Mehta *et al.*, 1995), SOPMA (Geourjon and Deleage, 1994), SSP (Solovyev and Salamov, 1994) and DSC (King and Sternberg, 1996). A notable multiple sequence alignment tool was PSI-BLAST (Altschul *et al.*, 1997), which produced a position-specific scoring matrix (PSSM) constructed from a multiple alignment of the top-scoring BLAST responses to a given query sequence. With the continuing development of algorithms and the better usage of sequence alignments, many famous predictors converged on accuracy figures of ∼80%, including PSIPRED (Jones, 1999) (which used PSI-BLAST profiles for prediction), JPRED (Cole *et al.*, 2008; Cuff *et al.*, 1998) (which made consensus predictions), PHD (Rost and Sander, 1993) (which performed an all-in-one prediction) and nnPredict (Kneller *et al.*, 1990) (which used neural networks). The third stage was the use of sequence-structural alignments; several excellent and inspiring articles (Lin *et al.*, 2010; Montgomerie *et al.*, 2006; Pollastri *et al.*, 2007; Zhou *et al.*, 2010) have been published on this work in recent years. These methods focused on sequence similarities and the direct usage of structural information, and achieved accuracies slightly in excess of 80%, which showed that the deep consideration of structural information did impact the performance of the prediction. Zhou *et al.* (2010) dexterously incorporated other secondary structural elements (in the form of a shape string) to improve the predictive performance. Montgomerie *et al*. (2006) directly utilized the secondary structures of the best-matched sequences as the secondary structures of a query. Pollastri *et al*. (2007) generated a set of templates based on a similarity search of the PDB; the templates were further implemented as inputs for an ensemble of recursive neural networks. Lin *et al*. (2010) constructed a dictionary for the storage of short subsequences and their secondary structures, and directly used these structural elements when a query matched the short subsequences.

Here, we present a novel predictor, SPSSMPred, for the prediction of protein secondary structures. SPSSMPred is based on an original structural position-specific scoring matrix (SPSSM) that is generated by sequence alignment, but its elements are secondary structural profiles. The SPSSM can be used to build the relationship between structural profile and protein secondary structure. For the first time, we develop a strategy to construct a database of the secondary structural profiles of 9 million sequences. This database, 9M_database, is one in which every union is an amino acid and its secondary structural profile is derived from the non-redundancy database used in PSI-BLAST. We provide a BLAST tool, 9M-BLAST, to align a query against the 9M-database and results in PSSM and SPSSM simultaneously. A non-redundant dataset is used as the training in the classification algorithm of conditional random fields (CRFs) (Lafferty *et al.*, 2001). The SPSSMPred was tested on newly published protein sequences and benchmark EVA datasets—we achieved results much closer to the expected theoretical limit of secondary structure prediction (Rost, 2003).

2 METHODS

2.1 SPSSPred flowchart

A flowchart for the SPSSPred is shown in Figure 1. To perform a query, 9M-BLAST is first used to search the 9 million sequences and their corresponding structural profiles in the 9M_database, which is established in advance from PDB_99 using PSI-BLAST. As a result, both the PSSM and SPSSM are obtained simultaneously. Finally, from the PSSM and SPSSM, 23 features in total are treated as input access CRFs for modeling prediction, where the training set is the sequences of PDB_30.

2.2 SPSSM

The SPSSM is a distinctive PSSM-like profile composed from three boxes, where the SPSSM scores are used to appraise matched sequences after alignments are stored (Fig. 2, bottom). The score is defined as

$$\text{Score1}(i,s)=\sum_{A(i,j)}\theta(S(i,j),s) \tag{1}$$

where i is the position of an amino acid in the target sequence and s is one of the three state secondary structure elements, H, E or C. $A(i,j)$ denotes a set of all the matched sequence's amino acids at the position i, and j directs matched sequences in $A(i,j)$. $S(i,j)$ represents the corresponding secondary structure element of $A(i,j)$. $\theta(S(i,j),s)$ is defined as

$$\theta(S(i,j),s)=\begin{cases}1 & \text{if } S(i,j)=s\\ 0 & \text{else}\end{cases} \tag{2}$$

The scores for the state of H, E and C are calculated separately, then allotted to the corresponding three boxes. It is clear that the score is the

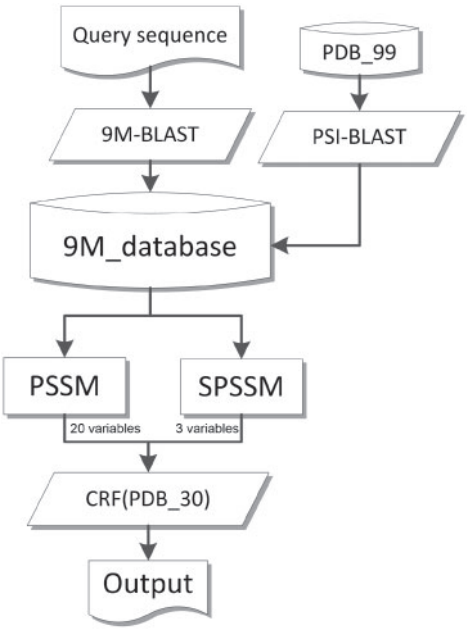

Fig. 1. The flowchart of the SPSSMPred.

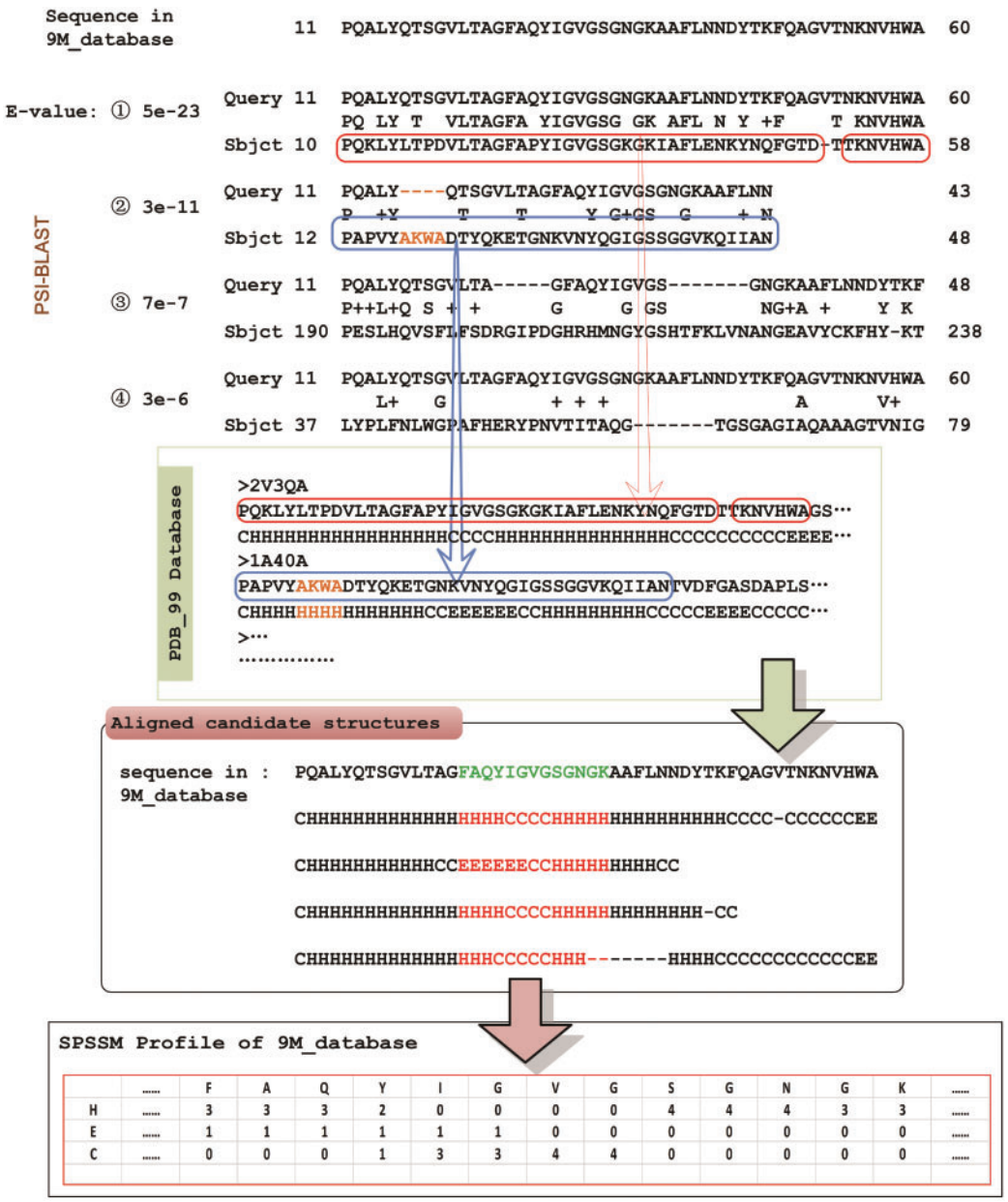

Fig. 2. An example of generation of raw SPSSM in 9M_database. A query in 9M_database is sp|P85173.1, and its listed 49 amino acids (first line). Four aligned 'Sbjcts' examples (in PDB_99) are shown, and two arrow tips point to two obtained sequences and their secondary structures. Then the query and found secondary structural elements are shown in middle. After score, its raw SPSSM is constructed and a part of them (in red) is shown in bottom.

frequency of the secondary structural elements, and that there are three elements for each amino acid in the sequences.

The SPSSM reflects the sequence alignment's shapes, and the properties of the secondary structure. The SPSSM is the expansion of the PSSM with regard to the structural aspects; the SPSSM inherits the concepts of the PSSM, but takes more consideration of the deeper common ground beneath the aligned sequences, where structural information may provide extra clues for regularity.

There are two forms of SPSSM; the raw SPSSM and the normalized SPSSM. The former is mainly stored in the 9M_database, and the latter is one of the results produced by 9M-BLAST.

The raw SPSSM in the 9M_database is obtained by first running PSI-BLAST against the PDB_99 (with three iterations and 500 maximum target sequences) to find homologous sequences relative to the target sequence (Fig. 2). The matched piecewise local sequences are then selected according to *e*-values that are below a given threshold (say 10–5). All the selected sequences are subsequently identified and ranked in accordance by their *e*-values in ascending order, and the top N (default is 10) of these sorted sequences that are considered as containing rich homologous information are reserved (if the number of the selected sequences were less than N, all the selected sequences would be kept). The scores of all amino acids in the 9M_database are then

calculated according to Equation (1), and the final raw SPSSM is constructed.

When 9M-BLAST is executed for a query against the 9M_database, the procedure is similar to that mentioned above. However, there are two differences; one is that 9M-BLAST uses the raw SPSSM instead of the secondary structural elements, and the other is that the output of 9M-BLAST is the normalized SPSSM, which is defined as

$$\text{Score2}(i,s)=\sum_{A(i,j)}P(i,j,s)\Big/\sum_{s\in\{H,E,C\}}\sum_{A(i,j)}P(i,j,s) \quad (3)$$

where $P(i,j,s)$ denotes the value of the corresponding secondary structural profile set of $A(i,j)$ in the 9M_database, where the raw SPSSM has been calculated according to Equation (1).

The normalized SPSSM is utilized as a feature for modeling and prediction. Though there are only three elements in the normalized SPSSM for an amino acid, due to it carrying homological and/or remote homological information on the secondary structure, the profile is often much better able to detect weak relationships than other features that have been used. We have written procedures to construct the raw SPSSM for the 9M_database and the normalized SPSSM as an output of 9M-BLAST.

2.3 Profile encodings

We use widely applied sequence profiles as encodings for our predictor, as well as newly proposed structural profiles. Sequence profiles contain rich sequence evolution information, and have long been proved to be an effective variable for the prediction of secondary structures. On the other hand, structural profiles are very simple, include valuable structural evolution information derived from all the known detected structures and are evidently of significant importance in improving prediction performances. As a result, we utilize 23 variables [PSSM (20 variables) and SPSSM (3 variables). The clarified 23 variables are also shown in Fig. 1.] as our final total encodings, and establish a new relationship between structural profiles and secondary structures in modeling. In the case that the structural profile is not sufficient for encoding, the sequence profile will then take the dominant role in the prediction.

2.4 CRFs

We perform our prediction by applying CRF to the problem of protein secondary structure. CRF is capable of incorporating evidence that contains long-range effects and unknown dependencies without requiring any probabilistic modeling of the observed data, and avoid a fundamental limitation of maximum entropy Markov models (MEMMs) (Liu *et al.*, 2004) and other discriminative Markov models based on directed graphical models, which can be biased toward states with few successor states. CRF is superior to many other machine learning methods in terms of speed. In our approach, CRF is utilized for modeling and prediction.

We used only Unigram template for CRF, the template that we generated considered two upward variables and two downward variables in row, and then, all the variables in column were traversed. We set all the parameters for modeling by default. We applied the CRF++ binary package for MS-Windows. The environment for training and testing was windows 7 64-bit operating system with Intel Core 2 Quad CPU and RAM 6 GB.

2.5 Web servers

There are two servers—SPSSMPred server and 9M-BLAST server—that we have set up for scientific users on our local infrastructure. These are available at http://cal.tongji.edu.cn/SPSSMPred/. The SPSSMPred (version 1) server predicts secondary structure for query sequence(s) in three state forms, and results are provided in the form of a web page and/or an e-mail. The 9M-BLAST server affords four levels of normalized SPSSMs for a query sequence(s), resulting from alignments against the 9M_database.

3 RESULTS AND DISCUSSION

3.1 Dataset construction

Two major datasets were constructed in our approach; these were named PDB_99 and PDB_30. We constructed PDB_99 by using CD-HIT (Li and Godzik, 2006) for single-copy sequences, with sequence identity cut off value at 99% against sequences stored in PDB (as of 2010, containing 70 177 proteins) with a resolution of <2.5 Å and an R-value of <0.3, and using only X-ray structures, the resolution of which is usually higher. This returned 19 876 entries and their secondary structure elements. The three-state secondary structure elements (H: helix, E: sheet and C: random coil) in the PDB_99 were converted using the eight-state define secondary structure of proteins (DSSP) (Kabsch and Sander, 1983a, b) with the following scheme: H, G and I to H; E to E; all others to C (Rost and Sander, 1993). PDB_30 was then generated, in which the sequences were obtained using CD-HIT cut off at 30% against the PDB_99, and any sequence of length <20 amino acids was removed (containing 9062 entries). In both PDB_99 and PDB_30, an amino acid sequence and its secondary structure element were joined as a union.

3.2 The 9M_database

The 9M_database is a BLAST compatible database and is the kernel of the SPSSMPred, in which there are an unprecedented 9 million sequences and corresponding secondary structural profiles. The sequences in the 9M_database were derived from the non-redundant NCBI database (as of 2009, 9 069 431 proteins) applied in PSI-BLAST. The secondary structural profiles of the 9M_database were generated by alignment and score [Equation (1) in Section 2]. Each of the 9 million sequences was aligned against PDB_99 with PSI-BLAST by setting the e-value at four different levels (1e-5, 1e-3, 1e-1 and 10) and other parameters at default. The aligned sequence segments and corresponding unions in PDB_99 were obtained in this way. The secondary structure elements in the matched unions were scored in three boxes that contained the scores of three-state secondary structural elements. These boxes then constituted a secondary structural profile of the original sequence. This procedure was repeated until the profiles of all sequences in the 9M_database were formed.

In constructing the 9M_database, we have created a vast, unparalleled database with integrated sequences and secondary structural profiles. The 9M_database is based on the concept that local similarities in protein sequences typically exhibit conserved structures and also, in addition, that a high degree of robustness of the structure with respect to the sequence variation may represent a remote homology nature in the sequence.

3.3 Structural information in the 9M_database

We calculated the percentage coverage of the structural profiles stored in the 9M_database. The results showed that 94.9% of the 9 million sequences were covered by structural profiles, giving a total of 8 593 661 SPSSMs, and 70.2% of the 3 101 645 645 amino acids (2 176 906 296) overlapped with structural profile information. By comparison, regarding the protein database of known structures, there are only 4 788 328 amino acids of 19 876 proteins stored in PDB_99. This indicates that the 9M_database is extremely expansive not only in sequence diversity but also in structure extension, and acts as an abundant sequence and structural

profile database. Even in the case that some very small parts of the sequences were not aligned by 9M-BLAST, resulting in zero encodings in SPSSMs, PSSMs (see following section) of the SPSSMPred would compensate for the insufficiencies of SPSSMs, and would play an important part in the prediction.

3.4 9M-BLAST: a tool for alignment against the 9M_database

We modified PSI-BLAST slightly to construct 9M-BLAST tool, which can align against the 9M_database for a query. When 9M-BLAST alignment is carried out, both PSSM and SPSSM are obtained simultaneously. It is clear that 9M-BLAST provides more information than BLAST against the non-redundant database, and is expected to be a powerful tool for the prediction of protein structure and function based on sequence.

9M-BLAST is a convenient and practical PSSM and SPSSM creation tool that we constructed to facilitate the process of extracting whole useful sequences and structural profiles. 9M-BLAST enriches the traditional Blast algorithm by providing not only the sequence profile, but also structural profile knowledge of the homology and remote homology (which may be utilized as a modified or enhanced version of Blast in sequence and structure analysis).

3.5 Validations and large-scale predictions

In order to gain an understanding of the performance of the PDB_30 training model of 9062 sequences, we performed a 5-fold cross-validation, as detailed in Table 1. Note that because the 9M_database was derived from a very large, diverse set of databases including PDB, when analyzed using 9M-BLAST, any sequence in the 9062 with an exact match in the 9M_database was discarded for fairness. The 5-fold cross-validation is a relatively strict cross-validation method used to estimate how accurately a predictive model will perform in practice, and is important in guarding against testing hypotheses suggested by the data. The results showed that our program predicted three-state secondary structures with an average Q3 of 93.7% and an average segment overlap measure (SOV) (Rost *et al.*, 1994; Zemla *et al.*, 1999) of 94.6%, which indicated that the training model had strong potential for practical applications. The SOV score treats secondary structure segments as basic units, and can effectively capture structurally important features while reducing the significance of those that are less important. An SOV score of 94.6% was obtained under the 5-fold cross-validation, which indicated that the overall secondary structure segments' regions were distinguished relatively well.

We predicted all the PDB_99 sequences apart from the PDB_30 (we defined it as the rest of PDB_99, containing 10 814 protein chains). When tested with our program using the same strict constraints as in the 5-fold cross-validation (to enable a fair comparison, when using 9M-BLAST, any sequence in the 10 814 with an exact match in the 9M_database—as aligned with the PDB index—was not included), an accuracy of 93.4% was achieved (Table 1). Our method performed satisfactorily for this large quantity of sequences, which indicated that the SPSSMPred is efficient for large-scale predictions. It should be noted that this was an example to confirm the ability of our method on large scale, and more strict evaluation examples such as suggested by Jones (1999) were showed in below.

3.6 Performance on newly measured proteins

We also tested the performance of the SPSSMPred on newly measured proteins (measured using X-ray crystallography). These proteins have not yet been mentioned in the literature relevant to the field of protein structure prediction; we wished to evaluate the ability of our predictor to analyze proteins with structures that have not already been solved. We established two test sets, T_241 and T_141, which were derived from entirely new measured sequences published in January, February and March, 2011 in PDB (1907 proteins). The T_241 dataset included sequences determined using only the X-ray method, while the T_141 dataset was built with stricter conditions; resolution of <2.5 Å, $R<0.3$ and using only X-ray structures. After sequence identity cutting off value at 99% by CD-HIT on the two datasets, 241 proteins remained in T_241, and 141 proteins remained in T_141.

Table 2 shows that impressive scores of 89.6% (Q3) and 89.8% (SOV) for the T_141 test set and 85.3% (Q3) and 85.8% (SOV) for the T_241 test set were obtained. These values approach the theoretical limit of protein secondary structure prediction, which means that not only was an excellent model built, but also the practical performance on not-yet-released proteins was outstanding. In Table 2, we also tested the performance of our program when only PSSM was used for T_141 dataset and T_241 dataset in the same pipeline to make a comparison of the contribution between PSSM and SPSSM. It should be noted that the T_241 results were not as good as those from T_141 and the performance on new measured proteins in Table 2 also showed a decrease of QC and QE. The reason for this could be that T_141 was obtained under stricter conditions, with resolution of <2.5 Å and $R<0.3$; these conditions matched the training model with the same screening parameters. It was also suggested that several β-sheets reported by DSSP were smaller in size and DSSP fragmented the actual β-sheets in many independent ones (Parisien and Major, 2005), which led to a decrease in prediction accuracies of QC and QE. This also indirectly indicated that our model was able to achieve satisfactory prediction results when sufficiently exact measurements were available. Besides,

Table 1. The 5-fold cross-validation on PDB_30 (training set, 9062 proteins) and the prediction of PDB_99 preclude the PDB_30 (remaining 10 814 proteins) (numbers given in percentages, %)

	Q3	QC	QH	QE	SOVall	SOVC	SOVH	SOVE
PDB_30 (5-fold)	93.7	92.4	94.9	94.2	94.6	93.5	95.2	95.6
PDB_99 (preclude PDB_30)	93.4	92.1	94.3	94.3	94.5	93.5	94.9	95.8

Table 2. Performance on new measured proteins of PDB January, February and March, 2011 (%)

	Q3	QC	QH	QE	SOVall	SOVC	SOVH	SOVE
T_141	89.6	87.7	91.5	90.0	89.8	86.9	91.6	91.9
T_241	85.3	82.6	90.1	82.8	85.8	84.2	88.6	84.5
T_141(PSSM only)	73.5	77.4	75.7	62.0	68.4	67.6	70.2	66.5
T_241(PSSM only)	71.7	74.6	73.8	62.3	67.6	68.1	68.6	65.1

Table 3. Performance on low-homology proteins (%)

	Q3	QC	QH	QE	SOVall	SOVC	SOVH	SOVE
40 low homology	84.6	82.3	87.7	83.5	82.7	79.1	85.8	84.3

the performance of a predictor is also dependent on accurate secondary structure assignment from protein atomic coordinates, perhaps β-Spider (Parisien and Major, 2005) could assign more reasonable β-sheet, which may improve unbalance of accuracies on H to E and C.

3.7 Prediction of low-homology proteins

The excellent performance of the SPSSMPred is not limited only to homology sequences; low-homology sequences also benefit from numerous sequences in the 9M_database that contain remote homology information. We cut T_141 against PDB_99 at a 30% sequence identity level; this left 40 sequences to form a low-homologue sequence set. For these 40 sequences (Table 3), the accuracy was Q3 84.6%, with an SOV value of 82.7%. The results showed that even in difficult cases where only a low number of homologues exist, our program is still effective in searching for distant similarities. This is because low homologues behave relatively; one sequence that is quite dissimilar with PDB_99 may find some resemblances in the 9M_database, due to its vast sequence capacity. Such similarities may carry valuable structural profiles with respect to the given low-homology sequence.

3.8 Comparison with results from benchmark EVA datasets

We chose the two latest benchmark EVA datasets (EVAset1 and EVAset2) (Lin *et al.*, 2010) to further evaluate our predictor via a comparison with other excellent existing methods. The EVA set has served for a number of years as a benchmark for protein secondary structure predictors, particularly for CASP competitions (Eyrich *et al.*, 2001). The chosen EVAset1 contained 80 sequence-unique proteins, which is the minimum number of the EVAset2 (containing 212 proteins). These sets are considered as the strictest assessment of the performance of a predictor, since only those sequences without any sequence identities against the previous ones will be kept.

To more strictly assess the performance of the SPSSMPred, we used the two EVA datasets to make a 'blind' test prediction, and the results were compared with other state-of-the-art prediction methods. It is worth mentioning that to make a relatively legitimate and rigorous test, 11 sequences in the training model that coexisted in the EVA sets were removed from the training set. Moreover, during the test encoding process, any sequence alignments found by 9M-BLAST that had an exact match in the 9M_database were not included in the final SPSSM count.

Table 4 shows the results of the test on the two EVA sets, as well as results from other methods. The SPSSMPred achieved Q3 accuracies of 82.0% (SOV 83.0%) and 82.7% (SOV 83.3%) for EVAset1 and EVAset2, respectively. In a comparison with other high-performing secondary structure predictors, the SPSSMPred was superior, giving Q3 scores that were higher by about 3–8%. This resulted from the fact that the application of the 9M_database significantly improved the detection of fragmentary homology, and that even unique sequences in the EVA sets showed some remote similarities.

Table 4. The prediction performance of different methods on the EVA benchmark datasets (%)

	Q3	SOVall	SOVC	SOVH	SOVE
EVAset1 (80 proteins)					
SPSSMPred	82.0±0.8	83.0±0.9	83.3	81.7	85.4
SymPred	78.8±1.4	76.4±1.9	70.4	85.0	76.5
SAM-T99sec	77.2±1.2	74.6±1.5	71.2	80.9	72.5
PSIPRED	76.8±1.4	75.4±2.0	65.2	82.1	72.3
PROFsec	75.5±1.4	74.9±1.9	71.3	78.3	75.9
PHDpsi	73.4±1.4	69.5±1.9	65.2	73.7	73.9
EVAset2 (212 proteins)					
SPSSMPred	82.7±0.8	83.3±1.0	82.4	84.6	81.9
SymPred	79.2±0.9	76.0±1.2	71.3	85.1	77.7
PSIPRED	77.8±0.8	75.4±1.1	70.4	80.6	72.6
PROFsec	76.7±0.8	74.8±1.1	71.8	79.2	76.2
PHDpsi	75.0±0.8	70.9±1.2	67.0	77.0	72.4

It should be noted that Table 4 also shows that the performance of SOVH is not as good as other methods, but SOVC and SOVE are much better compared with other prediction methods. It indicates that our method is able to balance SOVH, SOVE and SOVC quite well especially under non-homology conditions.

3.9 Prediction for an anopheles gambiae odorant-binding protein and a T cell receptor

Protein 3L47 was taken as an example to demonstrate one prediction result by the SPSSMPred for the prediction of almost-no-homology and unique sequences from prior sequences of known structures (Fig. 3). Protein 3L47 is an anopheles gambiae odorant-binding protein that plays a key role in mosquitoes' semiochemical system. Semiochemicals such as pheromones, plant volatiles or animal odors are small hydrophobic molecules that enter the antennae and other sensory organs via pores, and pass across the hydrophilic sensilla lymph surrounding the olfactory neuronal dendrites. The sensilla lymph contains extremely high concentrations of odorant-binding proteins; an understanding of these proteins is highly desirable due to their potential, both for mediating the behavioral expressions of mosquitoes such as host-seeking, mating, blood feeding and oviposition, and for the development of new disease control strategies against mosquitoes (Yang *et al.*, 2011).

Protein 3L47 is an all alpha helix transport protein with a unique chain A. It was released on January 12th 2011 in PDB, and after a search of sequence alignments by PSI-BLAST against the PDB_99, no sequence similarities could be obtained at all under an *e*-value of 1e-5. When a prediction was made using the SPSSMPred, an overall Q3 of 84.4% and an SOV of 85.3% were obtained. 3O4LB is another all β-sheets protein example of genetic and structural basis for selection of a ubiquitous T-cell receptor deployed in Epstein–Barr virus, the overall performance of 3O4LB was Q3 of 90.0%, and Figure 3b showed that we predicted all the core region of β-sheets quite well. This highlights the aptitude of the SPSSMPred in forecasting very dissimilar sequences, which results from its ability to detect remote homology sequences.

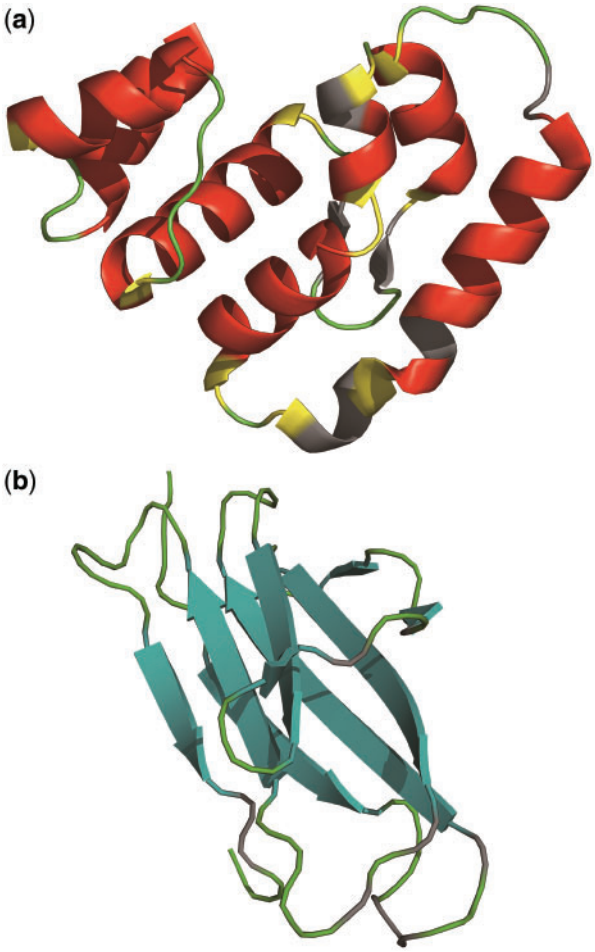

Fig. 3. An example of prediction of low-homology protein, 3L47 (**a**) and 3O4LB (**b**). The red represents helixes which have been predicted correctly. The green represents Random Coils which have been predicted correctly. The blue represents β-sheets which have been predicted correctly. The gray represents structures that have been predicted incorrectly. The yellow represents structures obtained from DSSP that are not matched to the 3D structures but are predicted correctly by the SPSSMPred. The illustration of the 3D structure of 3L47 (a) and 3O4LB (b) were drawn by PyMOL (Schrodinger, 2010).

4 CONCLUSIONS

In this study, we have presented the excellent performance of the SPSSMPred in predicting protein secondary structure. Tests on a proteome-scale set of 10 814 protein chains showed that the overall Q3 accuracy of the SPSSMPred was 93.4%. When tested on the two benchmark sets, the overall Q3 accuracy scores were 82.0 and 82.7%. For the newly published T_141 and T_241—for which no sequences appear in either the training set or PDB_99—the overall Q3 accuracy values still reached 89.6 and 85.3%, respectively. Moreover, after cutting out redundant sequences, the overall Q3 accuracy for the 40 low-homology sequences was 84.6%, which confirmed that the SPSSMPred performed well in the prediction of low-homology sequences.

The SPSSMPred is based on a new methodology for the prediction of protein secondary structure; a methodology that is different from existing state-of-the-art methods. First, there is no doubt that the main contribution comes from the 9M_database, in which a huge number of sequences bring raw SPSSMs that contain homology and remote homology information. In fact, the 9M_database represents almost all the experimental secondary structure information in the PDB, and can be considered as an extension of PDB. When a query is aligned against the 9M_database using 9M-BLAST, a PSSM is returned in the same format as one produced by PSI-BLAST, followed by a normalized SPSSM carrying secondary structural profiles. Second, the SPSSM is at the heart of our methodology. The SPSSM describes the probabilities of the three secondary structural elements of an amino acid in sequence, and includes information on the structural evolution. Third, in this pioneering work a relationship is built between structural profiles and secondary structures; this is the simplest but best model now available. We believe that this methodology could be extended to other structural biological fields, and that it will greatly improve prediction efficiencies.

ACKNOWLEDGEMENT

The authors would like to thank Prof. Wen-Lian Hsu for providing the two EVA sets.

Funding: National Natural Science Foundation of China grants (20675057, 20705024).

Conflict of Interest: none declared.

REFERENCES

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.

Aydin,Z. *et al.* (2006) Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinformatics*, **7**, 178.

Babaei,S. *et al.* (2010) Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. *Comput. Methods Programs Biomed.*, **100**, 237–247.

Chen,C. *et al.* (2007) Prediction of protein secondary structure content using support vector machine. *Talanta*, **71**, 2069–2073.

Chen,C. *et al.* (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.*, **16**, 27–31.

Chen,J. and Chaudhari,N. (2007) Cascaded bidirectional recurrent neural networks for protein secondary structure prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **4**, 572–582.

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

Chou,P.Y. and Fasman,G.D. (1974b) Prediction of protein conformation. *Biochemistry*, **13**, 222–245.

Clementi,C. *et al.* (2003) Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *J. Mol. Biol.*, **326**, 933–954.

Cole,C. *et al.* (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.*, **36**, W197–W201.

Cuff,J.A. *et al.* (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.

Deleage,G. and Roux,B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.*, **1**, 289–294.

Di Francesco,V. *et al.* (1997) Incorporating global information into secondary structure prediction with hidden Markov models of protein folds. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 100–103.

Eyrich,V.A. *et al.* (2001) EVA: continuous automatic evaluation of protein structure prediction servers, *Bioinformatics*, **17**, 1242–1243.

Frishman,D. and Argos,P. (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, **27**, 329–335.

Garnier,J. *et al.* (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.

Geourjon,C. and Deleage,G. (1994) SOPM: a self-optimized method for protein secondary structure prediction. *Protein Eng.*, **7**, 157–164.

Holley,L.H. and Karplus,M. (1989) Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 152–156.

Hu,H.J. *et al.* (2004) Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. *IEEE Trans. Nanobioscience*, **3**, 265–271.

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

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

Kabsch,W. and Sander,C. (1983b) How good are predictions of protein secondary structure? *FEBS Lett.*, **155**, 179–182.

Kendrew,J.C. *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, **181**, 662–666.

King,R.D. and Sternberg,M.J. (1990) Machine learning approach for the prediction of protein secondary structure. *J. Mol. Biol.*, **216**, 441–457.

King,R.D. and Sternberg,M.J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.*, **5**, 2298–2310.

Kneller,D.G. *et al.* (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, **214**, 171–182.

Lafferty,J. *et al.* (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. pp. 282–289.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lim,V.I. (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.*, **88**, 857–872.

Lin,H.N. *et al.* (2010) Improving protein secondary structure prediction based on short subsequences with local structure similarity. *BMC Genomics*, **11** (Suppl. 4), S4.

Liu,Y. *et al.* (2004) Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics*, **20**, 3099–3107.

Mehta,P.K. *et al.* (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.*, **4**, 2517–2525.

Monge,A. *et al.* (1994) An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl Acad. Sci. USA*, **91**, 5027–5029.

Montgomerie,S. *et al.* (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics*, **7**, 301.

Muirhead,H. and Perutz,M.F. (1963) Structure of haemoglobin: a three-dimensional fourier synthesis of reduced human haemoglobin at 5.5 Å resolution. *Nature*, **199**, 633–638.

Nair,R. and Rost,B. (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, **53**, 917–930.

Nguyen,M.N. and Rajapakse,J.C. (2003) Multi-class support vector machines for protein secondary structure prediction. *Genome Inform.*, **14**, 218–227.

Parisien,M. and Major,F. (2005) A new catalog of protein beta-sheets. *Proteins*, **61**, 545–558.

Pollastri,G. *et al.*(2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, **8**, 201.

Presnell,S.R. *et al.* (1992) A segment-based approach to protein secondary structure prediction. *Biochemistry*, **31**, 983–993.

Rose,P.W. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.

Rost,B. (2003) Rising accuracy of protein secondary structure prediction. In Chasman,D.I. (ed.) *Protein Structure Determination, Analysis, and Applications for Drug Discovery*. CRC Press, New York, pp. 207–249.

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

Rost,B. *et al.* (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

Salamov,A.A. and Solovyev,V.V. (1997) Protein secondary structure prediction using local alignments. *J. Mol. Biol.*, **268**, 31–36.

Schrodinger,L.D.D. (2010) The PyMOL Molecular Graphics System, Version~1.3. Available at http://www.pymol.org/citing (last accessed date october 7, 2011).

Solovyev,V.V. and Salamov,A.A. (1994) Predicting alpha-helix and beta-strand segments of globular proteins. *Comput. Appl. Biosci.*, **10**, 661–669.

Ward,J.J. *et al.* (2003) Secondary structure prediction with support vector machines. *Bioinformatics*, **19**, 1650–1655.

Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.

Yang,G. *et al.* (2011) Expression, purification and functional analysis of an odorant binding protein AaegOBP22 from Aedes aegypti. *Protein Expres. Purif.*, **75**, 165–171.

Zemla,A. *et al.* (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.

Zheng,W.M. (2005) Protein secondary structure prediction by combining hidden Markov models and sliding window scores. *Int. J. Bioinform. Res. Appl.*, **1**, 420–428.

Zhou,T. *et al.* (2010) A novel method for accurate one-dimensional protein structure prediction based on fragment matching. *Bioinformatics*, **26**, 470–477.

Zvelebil,M.J. *et al.* (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, **195**, 957–961.