

PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines

Piyali Chatterjee · Subhadip Basu ·
Mahantapas Kundu · Mita Nasipuri ·
Dariusz Plewczynski

Received: 10 September 2010 / Accepted: 19 April 2011 / Published online: 19 May 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract Secondary structure prediction is a crucial task for understanding the variety of protein structures and performed biological functions. Prediction of secondary structures for new proteins using their amino acid sequences is of fundamental importance in bioinformatics. We propose a novel technique to predict protein secondary structures based on position-specific scoring matrices (PSSMs) and physico-chemical properties of amino acids. It is a two stage approach involving multiclass support vector machines (SVMs) as classifiers for three different structural conformations, *viz.*, helix, sheet and coil. In the first stage, PSSMs obtained from PSI-BLAST and five specially selected physicochemical properties of amino acids are fed into SVMs as features for *sequence-to-structure* prediction. Confidence values for forming helix, sheet and coil that are obtained from the first stage SVM

are then used in the second stage SVM for performing *structure-to-structure* prediction. The two-stage cascaded classifiers (PSP_MCSVM) are trained with proteins from RS126 dataset. The classifiers are finally tested on target proteins of critical assessment of protein structure prediction experiment-9 (CASP9). PSP_MCSVM with brainstorming consensus procedure performs better than the prediction servers like Predator, DSC, SIMPA96, for randomly selected proteins from CASP9 targets. The overall performance is found to be comparable with the current *state-of-the art*. PSP_MCSVM source code, train-test datasets and supplementary files are available freely in public domain at: <http://sysbio.icm.edu.pl/secstruct> and <http://code.google.com/p/cmater-bioinfo/>

Keywords Bioinformatics · secondary structure · Physicochemical properties · Protein sequence · Support vector machine

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1102-8) contains supplementary material, which is available to authorized users.

P. Chatterjee
Department of Computer Science and Engineering,
Netaji Subhash Engineering College,
Garia,
Kolkata 700152, India

S. Basu · M. Kundu · M. Nasipuri
Department of Computer Science and Engineering,
Jadavpur University,
Kolkata 700032, India

D. Plewczynski (✉)
Interdisciplinary Centre for Mathematical and Computational
Modelling, University of Warsaw,
5a Street,
02-106 Warsaw, Poland
e-mail: darman@icm.edu.pl

Introduction

Proteins are large biological molecules made of amino acids arranged into a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues which perform all important tasks in organism and participate in every process within cells; such as catalyzing biochemical reactions, structural or mechanical functions, cell signaling, immune responses, cell adhesion etc. Understanding secondary structures from linear sequences help in understanding structure-function relationships of proteins. Many researchers predict the secondary structure of protein to obtain knowledge of three-dimensional structure of protein. Though secondary structure prediction is not suffi-

cient to get three-dimensional structure but it may give insights of overall protein structure. Generating chemically verified ground-truth protein structures using nuclear magnetic resonance (NMR) or X-ray crystallography are expensive both in cost and time sense. With the initiation of the large-scale sequencing projects (such as the Human Genome Project), amino acid sequences for a very large number of proteins have been discovered but their structures are yet to be identified. Here comes the role of computational techniques that may not be as accurate as NMR, but gives an overall idea of protein structure with reasonable accuracy. These are particularly useful for drug design research. However, there is no computational method that predicts protein secondary structures consistently with excess of 80% accuracy. This is tempting for the researchers, especially the machine-learning community, to develop novel methods powered with latest computing facilities and algorithmic advancements.

Protein secondary structure prediction (PSP) has been a well-studied problem for over three decades that produced more and more accurate solutions. Despite such efforts, large gaps in accuracies still exist between the current state-of-the-art and the ground truth annotations for protein sequences. In the current work, we have tried to minimize this gap with the help of two-stage multiclass support vector machine (SVM) classifiers. The general objective of prediction of secondary structure is to classify a pattern of residues (in the form of amino acid sequences) to a corresponding sequence of secondary structure elements, namely α -helix (H), β -sheet (E) and coil (C, the remaining type). Single-stage approaches (sequence to structure) are unable to find the interrelationship between secondary structure elements. This was improved by introducing a second stage prediction strategy which uses the contextual information among the predicted structure elements by the first stage, but the traditional two-stage approaches suffered from the problem of low accuracy that was below 70%. It could not encode long-range interactions between molecules. Next, introducing sequence similarity or homology information, predictive accuracy was increased to around 71%. However, even then the accuracy was not satisfactory enough. To improve the prediction accuracies even further, in our two-stage approach, we have included some physico-chemical properties of amino acid from AAINDEX feature database (<http://www.genome.jp/aaindex/>) in addition to homology information. It is worth mentioning that there are many research experiments in this field involving artificial neural networks (ANN) [1–6] and SVM [7], but in this work, we have designed *multi class SVMs* (MCSVM) for predicting secondary structures. To study the performance, we have investigated the structures of the latest CASP9 target proteins using our developed PSP_MCSVM system.

Algorithmic procedures for prediction of secondary protein structures were started to evolve from late 1970s.

Propensity values of different amino acids for forming various secondary structures play a key role in all these methods. All the algorithmic methods evolved so far for prediction of secondary protein structures can be grouped into three broad categories, viz., first generation, second generation and third generation methods, depending on how residue information from amino acid sequences is used. Each element of an amino acid sequence, representing a specific monomer of the corresponding protein is called a residue. Out of these, the first generation methods are mainly based on single residue statistics, expressed by their propensity values. The second-generation methods are mainly based on a group of adjacent residues, and the third generation methods additionally use information derived from the sets of homologous sequences. Homologs are proteins having similar structures due to their shared ancestry. The homologous sequence similarity is very useful information which is often termed as homology information or evolutionary information.

The work by Chou and Fasman [8] represents an important piece of first generation method. Search for contiguous regions of residues with a high probability of forming a secondary structure feature is central to this work. In another pioneering attempt, Qian and Sejnowski [1] (second generation), considered a window of 13 amino acids, out of which the secondary structure of the central amino acid was predicted by a neural network using the identities of 12 neighboring amino acids, with a 20 bit coding system for each amino acid.

Prediction of the secondary structure of the central amino acid on the basis of information theoretic measure is central to a series of *GOR* methods (second generation) developed throughout the period of 1970s to 1980s. *GOR I* method was developed for prediction of three secondary structures, viz., helix, sheet, reverse turn referred to as structural states of residues in [9]. Under this method, for each of these three possible states, structural information supplied by all the neighboring residues of a central amino acid are summed up to form a gross information measure for the said amino acid. The state, for which highest gross information measure is obtained, is finally assigned to the central amino acid. Here, a window of 17 contiguous residues is considered. With this technique, secondary structures of 60% of the total residues could be correctly predicted as reported in [9]. To update *GOR I method*, Gibrat et al. [10] included a new data table, based on directional information values from 75 proteins of known structures. The work of Gibrat et al. was expanded by Garnier and Robson [11] to include the four structural states. To do so, Garnier and Robson just enhanced the data tables keeping the information theory equations and algorithms of *GOR I* method the same. The technique thus developed by Garnier and Robson is known as *GOR II*

method. The structural information about the structure of a central amino acid, supplied by its neighboring residues, is all unconditional in the information theoretic equation of *GOR I* method. All such information is combined with the structural information preconditioned with the existence of a particular amino acid in the central residue position under *GOR-III* equation [11]. In *GOR IV* method [12], the difference between the information sum for each structural state and that for the set of its complementary states is considered before determining the state of a central amino acid. The state, for which the maximum information difference is produced, is finally selected for the central amino acid. On a database of 267 proteins, the mean accuracy of *GOR IV* method was observed as 64.4% for a three-state prediction.

PSI-BLAST [13] is one of the currently available standard software for producing multiple alignments from a given database. Klockzkowski et al. [12] has used *PSI-BLAST* for inclusion of evolutionary information for prediction of protein secondary structures in *GOR V* method (third generation). For this work *PSI-BLAST* is applied on a non-redundant database (Benson et al. 1999) to generate multiple alignments after five iterations. Using full jackknifing, a mean accuracy of 73.5% is observed by application of *GOR V* method. The segment overlap (Zemla et al. 1999), a measure of normalized secondary structure prediction accuracy is observed to be 70.8% with this method. One of the important neural network based second generation methods is *NN-PREDICT* [14]. A two layer feed forward neural network is used there for secondary structure prediction from amino acid sequences. The technique is reported to have obtained the best-case prediction of 79% for the class of all-alpha proteins. *SIMPA96* [15] (second generation) is also another important window based method, which uses training data of short fragments, each of length equal to the window size, collected from protein sequences of known structures but minimal sequence similarity. Rost & Sander [2, 3] (third generation) made some significant enhancement in the neural network based paradigm for protein secondary structure prediction. In their work, they have used *multiple sequence alignment* with appropriate cut-offs to supply more enriched information about the protein secondary structure, compared to what can be supplied with a single sequence to the neural network. The *balanced training* as introduced by Rost and Sander results into a more balanced prediction, without any bias on the overall three-state accuracy. In the work of Rost and Sander, they have also addressed the problem of inadequacy in the length distribution of the predicted helices and strands by introducing a second-level neural network for *structure-to-structure* prediction. For final prediction of a protein secondary structure, a *voting* method is followed with the

responses of 12 different neural networks working in parallel on the same input. The group of networks is referred to as a *jury of networks* by Rost and Sander. In testing the performance of their method, Rost and Sander prepared a database of 130 representative protein chains of known structures. In this database, no two sequences can have more than 25% identical residues. With seven fold cross validation of results, the overall three-state prediction accuracy of the method is observed as 69.7%. It was three percentage points above the highest value (66.4% [2]) reported previously.

PHD [4] (third generation) is another method by Rost and Sander that uses three level neural networks. This method takes sequence profiles, conservation weights, insertion-deletion, spacer (to exploit the total protein length), length of protein, distances of the window with respect to the protein ends as features. The prediction accuracy for RS126 dataset is achieved as 73.5%.

PRED [16] (third generation) is another secondary structure prediction server exploiting multiple sequence alignments that attains the average Q_3 accuracy of prediction of 72.9%. *NNSSP* [17] (third generation) is a scored nearest neighbor method by considering position of N and C terminal in helices and strands. Its prediction accuracy on RS126 dataset is achieved as 72.7%. *PRED-ATOR* [18] (third generation) is quite a different method from the above mentioned method in which propensity values for seven secondary structures and local sequence alignment is used in lieu of global multiple sequence alignment. The prediction accuracy of this method on *RS126* dataset is achieved as 70.3%. *PSIPRED* [5] (third generation) is a neural network based method, which has three components. It conducts homology searches on a different database and uses a different set of proteins for training and testing. The residues are represented by the *PSI-BLAST* scoring matrices. The sequence-to-structure part of the method is a back-propagation neural network with the input window of 15 residues. This neural network has 75 hidden nodes and three output nodes. The output of the sequence-to-structure network is fed to the structure-to-structure network in a window of 15 residues. It achieves 76.5 to 78.3 prediction accuracy on test data. This method has, however, proven to be more successful than the others in the third critical assessment of techniques for protein structure prediction (*CASP3*). *PROF* [6] (third generation) is a neural network technique which used *PSI-BLAST* with gap or without gap to search the sequence in the NR database, to make the position specific profile. *DSC* [19] (third generation) is another method in which amino acid profile, conservation weights, indels, hydrophobicity have been exploited in predicting the protein structure effectively with 71.1% prediction accuracy on *RS126* dataset. Guo et al. [7] propose a high performance method (third genera-

tion) based on the dual-layer support vector machine (SVM) and position-specific scoring matrices (PSSMs) on the *RS126* and *CB513* dataset. The first SVM classifier classifies each residue of each sequence into the three secondary structure classes. The one–against–all strategy is used for the multiclass classification. The outputs generated from first layer SVM is used as inputs to second layer SVM. Seven fold cross validation is used to test the predictive accuracy of the classifier. On the *CB513* dataset, the overall per residue accuracy, the performance metric Q_3 reaches 75.2% while segment overlap (SOV) accuracy increases to 80.0%.

hybrid protein structure prediction (*HYPROSP*) [20] (third generation) proposes a hybrid method which combines knowledge based approach *PROSP*, with the machine learning approach, *PSIPRED*. In *PROSP* (third generation), a knowledge base is constructed with small peptide fragments along with their structural information. A quantitative measure *match rate* is used for a target protein to extract structural information from *PROSP*. If match rate is at least 80% then that information is being used for prediction, otherwise prediction is done by popular tool *PSIPRED*. However, when most of the proteins have match rate less than 80% then the advantage of using *HYPROSP* is less. To overcome this problem, a new method *HYPROSP-II* [21] is introduced where local match rate is used to define the confidence level of the *PROSP* prediction results. They combine the prediction results of *PROSP* and *PSIPRED* using a hybrid function defined on their respective confidence levels. The average Q_3 of *HYPROSP II* is 81.8% and 80.7% on *nrDSSP* and *EVA* datasets respectively.

Cole et al. proposes a prediction server *JPRED3* [22] (third generation) powered by Jnet algorithm in which position-specific scoring matrix and hidden markov model profiles (*HMMER*) are used. It is developed through 7-fold cross validation experiments on Astral Compendium of *SCOP* domain data. By testing on blind dataset of 149 sequences it achieves a final secondary structure prediction Q_3 score of 81.5%.

PSP_MCSVM: materials and methods

In this work, a novel architecture for protein secondary structure prediction is presented by cascading two MCSVM classifiers. In the first stage, a MCSVM is used to predict the secondary structure elements from the input amino acid sequences. The second stage MCSVM, which is cascaded to the first MCSVM, re-classifies the secondary structure elements using the output from the first MCSVM. The first stage network is referred to as *sequence to structure* and the second stage as *structure to structure*. In the

present MCSVM approach, the input amino acid sequences with fixed size window are fed into the first MCSVM classifier (C_1). Then outputs obtained from C_1 are applied to the second MCSVM classifier (C_2) to get the final predictions. Figure 1 depicts the architecture of this two-stage approach.

Multiclass SVM

Support vector machine (SVM) is primarily a binary classification methodology that has been used for pattern recognition and regression task. An SVM is used to construct an optimal hyper-plane for maximizing the margin of separation between the positive and negative data set of pattern classes.

There are two types of approaches for multiclass SVM. The first method called *indirect approach* where outputs

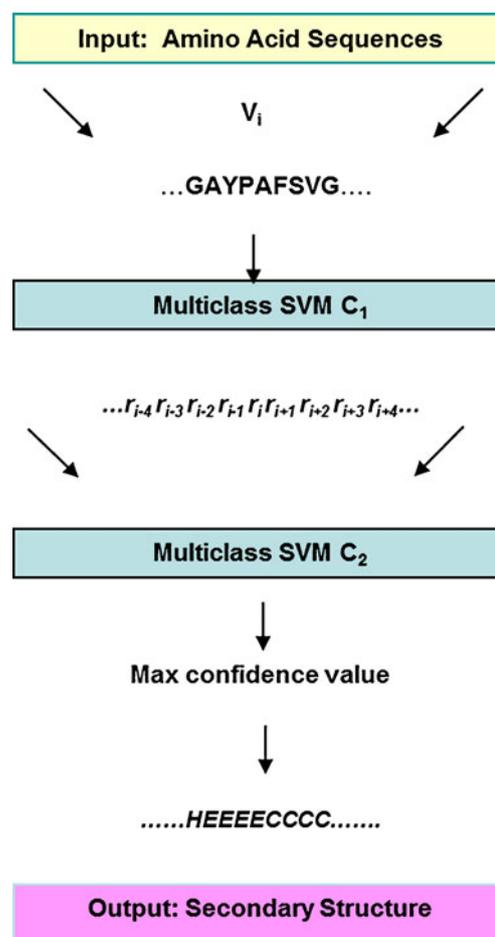


Fig. 1 Architecture of PSP_MCSVM is shown, where v_i is input amino acid sequence and r_j is the amino acid under consideration for prediction by C_1 . The sliding window for amino acids is considered as (r_{j-l}) to (r_{j+l}) ; where $l=4$ in this example. The final output of C_2 is shown as a labeled sequence of secondary structure elements; α -helix (H), β -sheet (E) and coil (C)

from several binary SVMs are combined to generate the final classification. The common techniques for combining the outputs of binary SVMs are one-against-all, one-against-one and directed acyclic graph-support vector machine (DAGSVM). In the second method, called direct approach, separating boundaries for all the classes are found in one step. For each class, a decision rule is defined similar to that of binary SVMs and during testing the test data point is assigned the label of the decision rule that yielded the highest (positive) margin. Suppose a training data set T_D consists of pairs $\{(x_i, y_i), i=1, 2, \dots, n, y_i \in \{-1, 1\}\}$ and $x_i \in \mathbb{R}^n$ where x_i denotes input feature vector for i th sample and y_i denotes the corresponding target value. For a given input pattern x the decision function of an SVM binary classifier is $f(x) = \text{sign}\left(\sum_{i=1}^N y_i \alpha_i K(x, x_i) + b\right)$ where $\text{sign}(u)=1$ for $u>0$ and -1 for $u<0$, b is the bias and $K(x, x_i) = \phi(x)\phi(x_i)$. Here $K(x, x_i)$ is the kernel function, used to map the input feature vector x into higher dimensional feature space to make them linearly separable. There are several kernel functions used in SVM. Some of the popularly used kernel functions, for solving pattern recognition problems, include Gaussian (RBF) kernel: $K(x, x_i) = \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)$ where σ is the standard deviation of the x_i values, polynomial kernel: $k(x, x_i) = (x_i^T x_j + t)^d$ where d is the degree of the polynomial and, linear kernel: $K(x, x_i) = x_i^T x_j$.

To design the MCSVM for the current work, we have used a method based on one-against-all approach proposed by Crammer and Singer where we have three classes for three secondary structures and three two-class rules where k th discriminant function $w_k^T \phi(x) - \gamma_k$ separates training vectors of the class k from the other vectors. The objective function is designed to minimize: $\frac{1}{2} \sum_{k=1}^N w_k^T w_k + c \sum_{i=1}^n \xi_i$ subject to the following constraint:

$$w_k^T \phi(x_i) - w_k^T \phi(x_i) \geq e_k^i - \xi_i, \forall k \neq k_i \tag{1}$$

where k_i is the class or secondary structural type of the residue to which the training data x_i belong to, and $e_k^i = 1 - c_k^i$, where $c_k^i = 1$, if $k_i=k$, and $c_k^i = 0$ if $k_i \neq k$.

Sequence-to-structure prediction

In the first level, features are extracted from the amino acid residues, belonging to the hypothetical sliding window (short sequence of amino acids) placed over the linear chain of amino acids of the protein. Six types of features, viz., position specific scoring matrix (PSSM) of amino acid, its hydrophobicity value, molecular size, polarity, relative frequency of amino acids in α -helix and relative frequency of amino acid in β -sheet are extracted for each residue in the sliding window, for predicting the structure of the central residue in the sliding window under consideration.

For each residue we compute 20 position specific scoring matrix values and five physicochemical properties as features (i.e., 25 features for each residue). The width of the sliding window considered here is 13 (six residues on either side of the central residue under consideration) and the total number of feature values for prediction of the central residue is estimated as $325=25 \times 13$ (i.e., the product of the said number of features and the window size). The MCSVM classifier (C_1) predicts the secondary structure of the central amino acid of the window. By sliding the window over the entire sequence of amino acids and repeating the classification process, secondary structures (sequence of helix, sheet or coil) of entire protein chain are predicted.

Position specific scoring matrix

A position specific scoring matrix (PSSM) is constructed by calculating position-specific scores for each position in the multiple alignments with the highest scoring hits in an initial BLAST search (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). The position specific scores (PSS) is calculated by assigning high scores to highly conserved positions and near zero scores to weakly conserved positions. The profile thus generated is then used to perform a second BLAST search and the results of each iteration are used to refine the profile.

Protein sequences are transformed into FASTA format to get evolutionary information using PSI-BLAST (position specific iterated BLAST). A local BLAST server in Windows environment has been used to get PSSM from PSI-BLAST after three iterations. Each residue in profile has 20 columns. So, the profile information of each amino acid residue is estimated as a 20 element feature-vector.

Physicochemical properties

Physicochemical properties of amino acids have been utilized to form another set of useful features. Those properties are hydrophobicity value, molecular size and polarity. Relative frequencies of amino acids in helix and β -sheet, which are important for predicting the underlying protein structure, are also used as features.

The features for the current experiment are chosen by selecting latest features from significant feature types of AAIndex database release 9.0 (<http://www.genome.jp/aaindex/>). Table 1 gives a brief description of selected features set.

Structure-to-structure prediction

In the second stage of our current work, another MCSVM classifier (C_2) is introduced to get better refinement over the prediction decisions about the secondary structure annota-

Table 1 Description of the selected features from AAIndex database

Sl. No.	Accession number	Brief feature description	Reference
1	JURD980101	Modified Kyte-Doolittle hydrophobicity scale	Juretic et al., 1998
2	GRAR740102	Polarity	(Grantham, 1974)
3	DAWD720101	Molecular size	(Dawson, 1972)
4	PRAM900102	Relative frequency in alpha-helix	(Prabhakaran, 1990)
5	PRAM900103	Relative frequency in beta-sheet	(Prabhakaran, 1990)

tion for the residue sequence. The refinement is based upon the contextual information about the secondary structures, estimated over the neighbors of the central residue within any sliding window. The input supplied to the second level MCSVM consists of the degrees of prediction confidence of C_1 for classifying each residue of the window into any of the three output classes, helix (H), sheet (E) and coil (C). Therefore three prediction confidence features are considered as inputs for the second stage classifier C_2 . The number of outputs for C_2 is again three (i.e., H, E and C).

Degree of confidences for forming three structures, viz., helix, sheet and coil, obtained from the outputs of C_1 have been normalized in the range [0,1]. A sliding window of width 13 is considered again, by taking six residues from either side of the central amino acid under consideration, generating 39 ($=3 \times 13$) features. On the basis of these features, the central residue of each sliding window is reclassified as helix, sheet or coil by the second level MCSVM C_2 .

Results and discussion

To evaluate the classification performance of the developed *PSP_MCSVM* system, protein sequences from two standard data sets, viz., *RS126* and *CASP9*, are considered in our current work.

The *RS126* data set (<http://www.compbio.dundee.ac.uk/~www-jpred/data/>) formed with 126 homologous globular proteins contains structural information about 23,349 residues of these proteins. The percentage of three secondary structures, viz., helix, strand and coil, as occurred in these residues, are 32, 23 and 45 respectively. The other dataset, used for the experimental here, is formed with 96 proteins selectively taken from *CASP9* (<http://predictioncenter.org/casp9/>), originally released in 2010. This is to select only those proteins, of which proper structure related information is available in the critical assessment of techniques for protein structure prediction. The evaluation metrics used here for measuring performances of the *PSP_MCSVM* system on the said data sets are explained below.

Evaluation metrics

Sensitivity measure

Sensitivity (also called recall rate) is the probability of correctly predicting an example of that class. It is defined as follows:

$$\text{Sensitivity} = (\text{TP})/(\text{TP} + \text{FN}), \quad (2)$$

where TP is the number of true positive, TN is total number of true negative, FP is total number false positive and FN is total number of false negative.

Specificity measure

Specificity (also known as Precision) measures the proportion of negatives which are correctly identified. The specificity for a class is the probability that a positive prediction for the class is correct. It is defined as follows:

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}). \quad (3)$$

Accuracy

Accuracy is the overall probability that prediction is correct. It is defined below:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}). \quad (4)$$

False alarm rate

False alarm rate for a class is the probability that an example which does not belong to the class is classified as belonging to the class. It is defined as follows:

$$\text{False Alarm rate} = \text{FP}/(\text{FP} + \text{TN}). \quad (5)$$

Matthews correlation coefficient (MCC)

Matthews correlation coefficient (MCC) is used in machine learning as a measure of quality of binary (two class) classifications. It takes into account true and false positives

and negatives and is generally regarded as a balanced measure which can be used even if the classes are of different sizes. The MCC is a correlation coefficient between the observed and predicted binary classifications.

It returns a value between -1 and $+1$. A coefficient of $+1$ represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction. The corresponding formulation is given below:

$$MCC = (TP \times TN - FP \times FN) / \sqrt{((TP + FN)(TP + FP)(TN + FP)(TN + FN))}. \tag{6}$$

Besides the above metrics, to measure the prediction accuracies of classifiers C_1 , and C_2 , the other metrics, such as, $Q_3, Q_H, Q_E, Q_C, SOV_{all}, SOV_H, SOV_E, SOV_C$ are defined below.

Overall accuracy (Q_3)

It is a measure of prediction accuracy and computes the percentage of residues predicted to be in their correct class (given below). It measures the all true positive data for each class (here, it is 3 for three secondary structural elements) out of total data.

$$Q_3 = \frac{100 \times (TP_H + TP_E + TP_C)}{N}, \tag{7}$$

where TP_H denotes true positive of helix, TP_E denotes true positive of sheet, TP_C denotes true positive of coil, N denotes total data.

Class accuracies (Q_i)

Individual class accuracy measures the percentage of the elements of class i that were predicted correctly out of all elements belonging to that class i .

$$Q_i = 100 \times \frac{TP_i}{N_i}, \tag{8}$$

where TP_i is true positive of class i , and N_i is total data in class i .

Segment overlap measure (SOV)

SOV is a measurement developed by Rost et al. (1994) and modified by Zemla et al. (1999) to reflect the specific goals of secondary structure prediction.

Let S_i be the set of overlapping segments where both segments are in state i

$$S_i = \{(s_1, s_2) : s_1 \cap s_2 \neq \emptyset, \text{ in state } i\} \tag{9}$$

and S'_i be the set of segments in state i for which there is no overlapping segment in state i .

$$S'_i = \{s_1 : s_2, s_1 \cap s_2 = \emptyset, \text{ in state } i\} \tag{10}$$

Let $b(s)$ as the position at which segment s begins, $e(s)$ as the position at which segment s ends, then length of the segment s is defined as:

$$len(s) = e(s) - b(s) + 1. \tag{11}$$

The length of actual overlap of two segments s_1 and s_2 both in state i is defined as:

$$\min ov(s_1, s_2) = \min(e(s_1), e(s_2)) - \max(b(s_1), b(s_2)) + 1. \tag{12}$$

The length of the total extent for which either of the segments s_1 and s_2 has a residue in state i is defined as:

$$\max ov(s_1, s_2) = \max(e(s_1), e(s_2)) - \min(b(s_1), b(s_2)) + 1. \tag{13}$$

$\delta(s_1, s_2)$ is the integer value defined as:

$$\delta(s_1, s_2) = \min \left\{ (\max ov(s_1, s_2) - \min ov(s_1, s_2)), \text{int} \left(\frac{len(s_1)}{2} \right), \text{int} \left(\frac{len(s_2)}{2} \right) \right\} \tag{14}$$

$N(i)$ is the number of residues in state i defined as:

$$N(i) = \sum_{s(i)} len(s_1) + \sum_{s'(i)} len(s_2). \tag{15}$$

The segment overlap for state i is given by

$$SOV_i = 100 \times \frac{1}{N(i)} \times \sum_{s(i)} \frac{\min ov(s_1, s_2) + \delta(s_1, s_2)}{\max ov(s_1, s_2)} \times len(s_1) \tag{16}$$

and the overall segment overlap is given by

$$SOV_{all} = 100 \times \frac{1}{\sum_i N(i)} \sum_i \sum_{s(i)} \frac{\min ov(s_1, s_2) + \delta(s_1, s_2)}{\max ov(s_1, s_2)} \times len(s_1). \quad (17)$$

Before the core portion of the experiment is conducted, the task becomes selection of an appropriate kernel function for the SVM classifiers used here.

Selection of kernel

In this work, $SVM^{multiclass}$ tool has been used to predict three secondary structural elements. $SVM^{multiclass}$ is an implementation of the multiclass SVM described in [23]. Choice of the SVM kernel function is an important decision in the overall classification process. In the current work we experimented with the training data sets for the first stage MCSVM classifier to choose the suitable kernel. The 1st stage MCSVM classifier with linear kernel gives accuracy of 68.6%, the same classifier with polynomial kernel gives ~60% of accuracy and with RBF kernel gives 35% of accuracy, when trained with Subset-II and III dataset (the first fold CV-1 for the 3-fold cross validation experiment discussed in the following sub-section). It is evident from the results (shown in Table S1 and Fig. 2) that the linear kernel with MCSVM outperforms polynomial or RBF kernels, to solve the problem under consideration.

Three-fold cross-validation experiment with RS126 dataset

For three fold cross-validations, the RS126 dataset is partitioned into three equal sized subsets which are referred to as subset-I, II, and III in the subsequent discussion. In each fold two subsets are taken for training and the remaining one set is used for testing. This process is repeated three times such that the train and test sets always remain mutually exclusive in any given experiment. We also use the complete dataset for training over the three experimental setups (to generate three model files).

For testing the performance of MCSVM on RS126 dataset the evaluation metrics Q_3 , Q_H , Q_E , Q_C , SOV_{all} , SOV_H , SOV_E , SOV_C were considered. We generated three model files in three fold cross-validated experiments (CV-1, CV-2 and CV-3) and Tables S2-A, S2-C, S2-E show the *sequence to structure* prediction performance over the test samples. Tables S2-B, S2-D, S2-F give the details for *structure to structure* predictions for three different folds respectively. Here, we achieved 68% to 71% of Q_3 accuracy and 61% to 67% of SOV_{all} measure for all three cross-validation experiments in *sequence to structure* network. The average Q_3 accuracy of 70.51% and the average SOV_{all} measure of 63.82% have been achieved for *sequence to structure* prediction in three fold cross-validation experiments. For some of the proteins, the Q_3 accuracy and SOV_{all} measure has been improved in *structure to structure* prediction. But average Q_3 accuracy and SOV_{all} measure have been decreased for most of the proteins due to low prediction accuracy for the sheet (E) class. Table S3-A and S3-B give the accuracy summary of the two stage classifiers.

Experiments with CASP9 targets

Among 129 proteins of CASP9 targets, we took 96 targets for our experiments to maintain consistency over availability of corresponding structure related information in the website of critical assessment of techniques for protein structure prediction. We studied the performance of our classifiers already trained with different subsets of RS126 dataset on CASP9 target proteins in two ways:

Firstly, we estimate the Q_3 accuracy and SOV_{all} measures on CASP9 targets in *sequence to structure* level and then *structure to structure* level with prediction results given in Table S4-A, S4-B, S4-C, S4-D, S4-E, and S4-F. Additionally, Tables S4-G and S4-H summarize the detailed results by computing average values of the measures. It may however be observed that the average Q_3 accuracy and SOV_{all} measure have been decreased in *structure to*

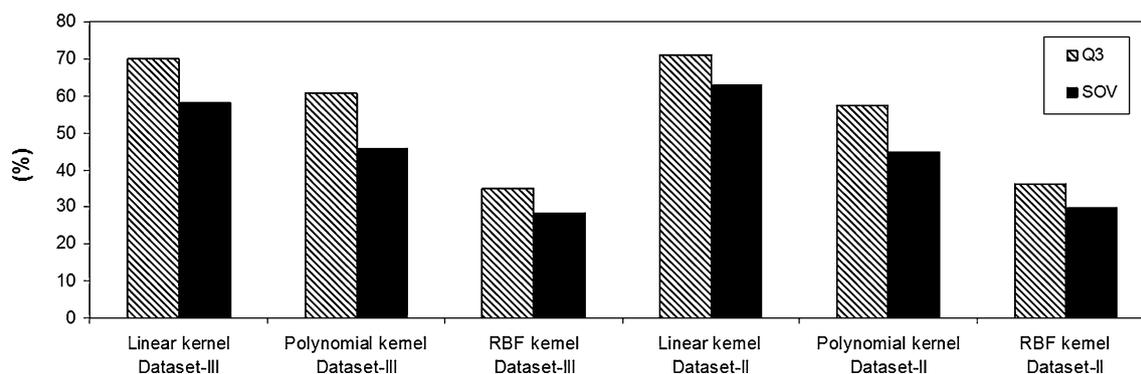
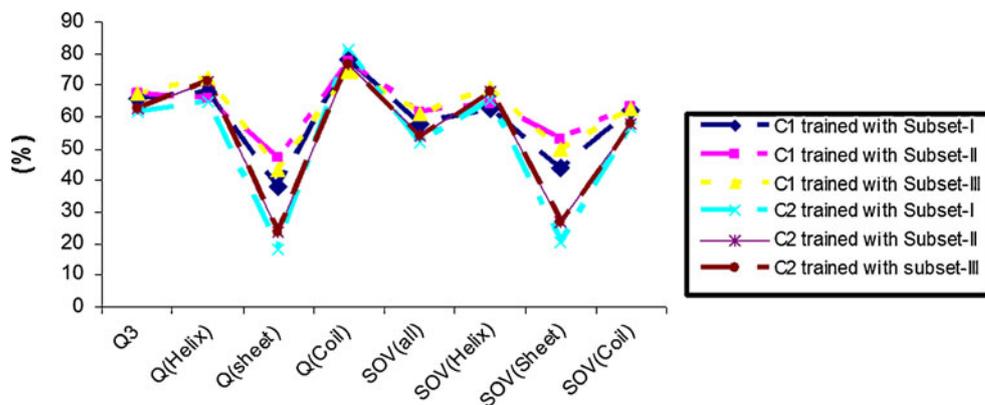


Fig. 2 Comparative study of performance measures of three kernels in 1st stage MCSVM trained with subset-II and III

Fig. 3 Q_3 and SOV measures of CASP9 in two-stage MCSVMs



structure level for most of the proteins due to low prediction accuracy for sheet (E). Figure 3 gives the graphical view of Q_3 and SOV measures on CASP9 targets using the developed two-stage MCSVM.

Secondly, other performance measures like accuracy, sensitivity, specificity, false alarm rate, Matthew correlation coefficient are computed for helix, sheet, and coil in sequence to structure level in three cross-validated experiments. Tables S5-A, S5-B, S5-C, S6-A, S6-B, S6-C, S7-A, S7-B, S7-C describe those performance measures. For the overall classification, it is clearly observed from Tables S8-A, S8-B, S8-C, S9-A, S9-B, S9-C, S10-A, S10-B, S10-C, that the introduction of structure to structure network increases the overall classification performance of classifier C_1 (sequence to structure). Table S11 summarizes the average prediction results of these two classifiers on CASP9 targets.

For helix, classifier C_2 improves the accuracy of classifier C_1 by 9%, 10% and 10% in three-fold cross validated experiments respectively. Classifier C_2 gives better sensitivity (19%, 19%, 16% increase) and specificity (19%, 19%, 22% increase) than classifier C_1 . False alarm rates have become less (5%, 5%, 11% decrease) in classifier C_2 than C_1 . The correlation coefficient for helix prediction ranges from 0.7 to 0.72. These facts are graphically illustrated in Figs. S1-A, S2-A, S3-A, S4-A, S5-A. For sheet, though classifier C_2 improves the accuracy of classifier C_1 by 8%, 8% and 10% in three-fold cross validated experiments respectively but does not give better sensitivity than classifier C_1 . It however gives better specificity (8%, 4% increase) than classifier C_1 . In cross validation experiment CV-3, the specificity is less. False alarm rates and correlation coefficient of sheet are not found to be satisfactory. These have been graphically shown in Figs. S1-B, S2-B, S3-B, S4-B, and S5-B.

From Figs. S1-C, S2-C, S3-C, S4-C, S5-C, it is clear that in the case of coil prediction the classifier C_1 performs well and classifier C_2 increases the prediction results of C_1 like that of helix (see Table S11).

Comparison between proposed method and other popular methods experiments with CASP9 targets

We have also compared the performance of PSP_MCSVM with the contemporary *state-of-the-art* techniques. Figure 4 shows the comparison of the Q_3 , Q_H , Q_E , Q_C , SOV_{all} , SOV_H , SOV_E , SOV_C score from the PSIPRED, PHD, Predator, DSC, SIMPA96 and our method for some randomly selected proteins from CASP9 targets. Due to low prediction accuracy of sheet, the overall accuracy of our classifier does not look so good compared to PSIPRED and PHD. However, in the case of helix and coil prediction, it achieves comparable performance with PHD. Moreover,

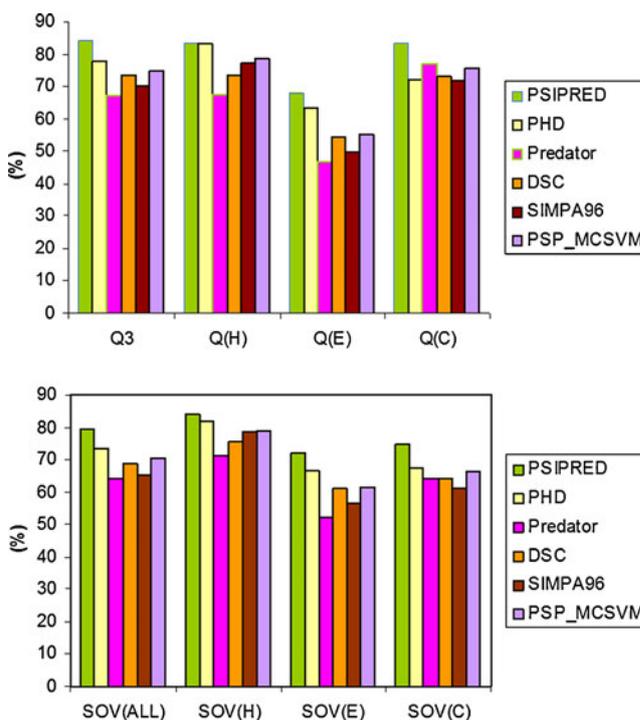


Fig. 4 Bar graph representing a comparison of prediction results by existing methods and our method PSP_MCSVM for randomly selected proteins from CASP9 targets

it achieves better results than Predator, DSC and SIMPA96 in helix and coil prediction.

Conclusions

We finally conclude that the designed feature set alongside *PSP_MCSVM* based methodology effectively predicts the secondary structure in protein chains. The cross-validated experimental setup with standard *RS126* dataset establishes our claims. Target proteins from *CASP9* dataset are also tested to evaluate its prediction accuracy. Different methods examine sliding windows of width 13–17 residues and assume that the central amino acid can be predicted based on the properties of its side groups on either side. The choice of this window-width is often a concern for the researchers. The average length for helices/strands in protein sequences guides the prediction decisions. Ground truth annotations suggest that the lengths of α -helices vary from 5 to 40 residues whereas those for β -strands vary from five to ten residues [18]. Generally, helices are of longer lengths compared to strands. In our previous work [24], different sliding window sizes varying from five to 13 were considered and maximum prediction accuracy for the window width of 13 was observed. Considering that, experiments for the present work has been conducted by fixing the width of sliding window as 13. The *PSP_MCSVM* source code, train/test datasets and supplementary files are available freely in public domain at: <http://code.google.com/p/cmater-bioinfo/>.

The accuracy of the current method may further be improved by incorporating multiple classifiers and a consensus based strategy. We may extend this work to develop a scheme toward combining classification decisions from multiple neural network based classifiers [24] and the current results, obtained from *PSP_MCSVM*. Another important issue is that the choice of dataset for training. In this work, we trained our network using subset of 126 proteins of *RS126* dataset. The higher the number of samples in dataset, the more variations and information in patterns can be accommodated by the machine learning algorithm. Perhaps, for this reason, *PSP_MCSVM* predicted sheet not so accurately whereas our other work [25] did it better. Although *RS126* is popular, commonly used and manually curated dataset but we may extend this work by training our system with a more information contained dataset, consisting of more labeled samples in an attempt to achieve higher prediction accuracy.

Acknowledgments Authors are thankful to the “Center for Micro-processor Application for Training Education and Research”, of Computer Science & Engineering Department, Jadavpur University, for providing infrastructure facilities during progress of the work. P. Chatterjee is thankful to Netaji Subhash Engineering College, Garia

for kindly permitting her to carry on the research work. Research of S. Basu is partially supported by BOYSCAST Fellowship (SR/BY/E-15/09) from DST, Government of India. Presented research was supported by the Polish Ministry of Education and Science (N301 159735 and others) and Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw (grant no. G36-24).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Qian N, Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 202:865–884
2. Rost B, Sander C (1993) Improved prediction of protein secondary structure prediction by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 90:7558–7562
3. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232:584–599
4. Rost B (1996) PHD: predicting 1D protein structure by profile based neural networks. *Methods Enzymol* 266:525–539
5. Jones TD (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195–202
6. Ouali M, King RD (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci* 9:1162–1176
7. Guo J, Chen H, Sun Z, Lin Y (2004) A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins* 54:738–743
8. Chou PY, Fasman GD (1978) Prediction of secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* 47:145–148
9. Garnier J, Osguthrope DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular protein. *J Mol Biol* 120:97–120
10. Gibrat JF, Garnier J, Robson B (1987) Further developments of protein secondary structure prediction using information theory—new parameters and consideration of residue pairs. *J Mol Biol* 198:425–443
11. Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 266:540–553
12. Kloczkowski A, Ting KL, Jerigan RL, Garnier J (2002) Protein secondary structure prediction based on the GOR algorithm incorporating multiple sequence alignment information. *Polymer* 43:441–449
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
14. Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol* 214:171–182
15. Jonathon LM (1997) Exploring the limits of nearest neighbor secondary structure prediction. *Protein Eng* 10:771–776
16. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton JG, Sternberg MJE (1998) JPRED: a consensus secondary structure prediction server. *Bioinformatics* 14:892–893
17. Salamov AA, Solovyev VV (1995) Prediction of protein structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J Mol Biol* 247:11–15

18. Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 27:329–335
19. King RD, Sternberg MJE (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* 5:2298–2310
20. Wu K, Lin H, Chang J, Sung P, Hsu W (2004) HYPROSP: a hybrid protein secondary structure prediction algorithm—a knowledge-based approach. *Nucleic Acids Res* 32:5059–5065
21. Lin H, Chang J, Sung P, Wu K, Hsu W (2005) HYPROSP II: a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* 21:3227–3233
22. Cole C, Barber DJ, Barton JG (2008) The JPRED3 secondary structure prediction server. *Nucleic Acids Res* 36:197–201
23. Crammer K, Singer Y (2001) On the algorithmic implementation of multi-class SVMs. *JMLR*
24. Chatterjee P, Basu S, Kundu M, Nasipuri M, Basu DK (2007) Protein secondary structure prediction through combination of decisions from multiple MLP classifiers. *Proc International Conference MS* 07:206–210
25. Chatterjee P, Basu S, Nasipuri M (2010) Improving prediction of protein secondary structure using physicochemical properties of amino acids. *Proc ISBN* 10. doi:10.1145/1722024.1722036