Artificial Intelligence in Medicine (2004) **31**, 117—136

Artificial Intelligence in Medicine

http://www.intl.elsevierhealth.com/journals/aiim

Bayesian network multi-classifiers for protein secondary structure prediction

Víctor Robles[a,*], Pedro Larrañaga[b], José M. Peña[a], Ernestina Menasalvas[a], María S. Pérez[a], Vanessa Herves[a], Anita Wasilewska[c]

[a]*Department of Computer Architecture and Technology, Technical University of Madrid, Madrid, Spain*
[b]*Department of Computer Science and Artificial Intelligence, University of the Basque Country, San Sebastián, Spain*
[c]*Department of Computer Science, University of Stony Brook, Stony Brook, NY, USA*

Received 1 March 2003; received in revised form 2 May 2003; accepted 16 January 2004

KEYWORDS
Multi-classifier;
Supervised classification;
Machine learning;
Stacked generalization;
Bayesian networks;
Protein secondary structure prediction;
Pazzani-EDA

Summary Successful secondary structure predictions provide a starting point for direct tertiary structure modelling, and also can significantly improve sequence analysis and sequence-structure threading for aiding in structure and function determination. Hence the improvement of predictive accuracy of the secondary structure prediction becomes essential for future development of the whole field of protein research.

In this work we present several multi-classifiers that combine the predictions of the best current classifiers available on Internet. Our results prove that combining the predictions of a set of classifiers by creating composite classifiers is a fruitful one. We have created multi-classifiers that are more accurate than any of the component classifiers. The multi-classifiers are based on Bayesian networks. They are validated with 9 different datasets. Their predictive accuracy results outperform the best secondary structure predictors by 1.21% on average.

Our main contributions are: (i) we improved the best know predictive accuracy by 1.21%, (ii) our best results have been obtained with a new semi naïve Bayes approach named Pazzani-EDA and (iii) our multi-classifiers combine results of previously build classifiers predictions obtained through Internet, thanks to our development of a Java application.

1. Introduction

Prediction of a secondary structure of a protein from its amino acid sequence remains an important and difficult task. Not only can successful predictions provide a starting point for direct tertiary structure modelling, but they can also significantly improve sequence analysis and sequence-structure threading for aiding in structure and function determination [1].

Since early attempts to predict secondary structure, most effort have focused on development of mappings from a local window of residues in the sequence to the structural state of the central residue in the window (see Fig. 1). A large number of methods for estimating such mappings have been developed.

[*] Corresponding author. Tel.: +34-91-336-73-80; fax: +34-91-336-73-73.
E-mail addresses: vrobles@fi.upm.es (V. Robles), ccplamup@si.ehu.es (P. Larrañaga), jmpena@fi.upm.es (J.M. Peña), emenasalvas@fi.upm.es (E. Menasalvas), mperez@fi.upm.es (M.S. Pérez), vherves@fi.upm.es (V. Herves), anita@cs.sunysb.edu (A. Wasilewska).

0933-3657/$ — see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.01.009

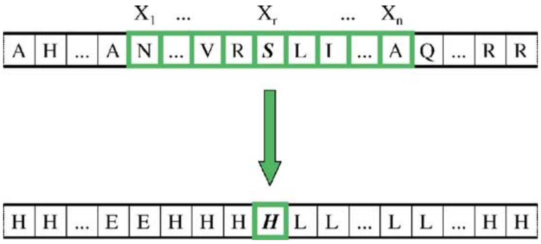

X_1 ... X_r ... X_n
A H ... A N ... V R S L I ... A Q ... R R
H H ... E E H H H H L L ... L L ... H H

Figure 1 Mappings from a local window of residues in the sequence to the structural state of the central residue in the window.

Methods predicting protein secondary structure have improved substantially in the 1990s through the use of machine learning methods and evolutionary information from the divergence of proteins in the same structural family. At the alignment level, the increase of the size of databases and the ability to produce profiles that include remote homologs using PSI-BLAST [2] have also contributed to performance improvement [3,4].

In this paper we present a protein secondary structure prediction multi-classifier system based on the stacked generalization paradigm [5] in which a number of classifier layers are designed to be part of a global multi-classifier, where the upper layer classifiers receive the class predicted by its immediately previous layer as input. The multi-classifier system has been programmed as a JSP Web application using several Java classes.

During the past several years, in a variety of application domains, researches in machine learning have reignited the effort to learn how to create and combine an ensemble classifiers. This research has the potential to apply accurate composite classifiers to real world problems by intelligently combining known learning algorithms.

Classifier combination falls within the supervised learning paradigm. This task orientation assumes that we have been given a set of training examples, which are customarily represented by feature vectors (training records). Each training example is labelled with a class target, which is a member of a finite, and usually small set of class labels. The goal of supervised learning is to predict the class labels of examples that have not been seen.

Combining the predictions of a set of component classifiers has shown to yield accuracy higher than the most accurate component on long variety of supervised classification problems [6].

We have used nine main datasets to train and test our approach: a training set, the PDB_SELECT list [7] of March 2002 (HS1771), and eight test sets (RS126 [8], CB513[9] and 6 different datasets from the EVA project [10]).

We have developed a two layer classification system in which we use a set of protein secondary structure prediction servers of Internet as layer-0 single classifiers, and we induce, over predictions made, different Bayesian network structures that acts as a consensed voting system at layer-1.

The rest of the paper is organized as follows. Section 2 explains the multi-classifier schema. Section 3 describes the datasets for the level-0 classifiers. In Section 4 we present in detail the statistics used to compare the secondary structure servers and the multi-classifiers. Section 5 describes the six level-0 classifiers used in construction of our multi-classifiers. Section 6 describes how to obtain the datasets for the level-1 classifiers (i.e. multi-classifiers). Section 7 contains a description of the level-1 multi-classifiers. In Section 8 we shortly discuss the experimental results. Finally, Section 9 contains the conclusion and the future research plans.

2. Multi-classifier schema

We present a multi-classifier for protein secondary structure prediction based on a straightforward approach that has been termed *stacked generalization* by Wolpert [5]. In its most basic form, its layered architecture consists of a set of component classifiers that form the first layer. Wolpert [5] calls the component classifiers the level-0 classifiers and the combining classifier, the level-1 classifier. In this work we introduce 7 different level-1 classifiers. See Fig. 2 for the schema of our stacked generalization classifiers.

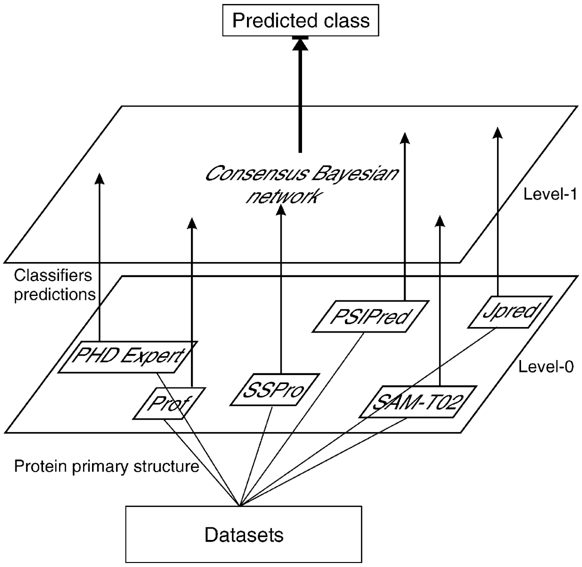


Figure 2 Multi-classifier schema.

Stacked generalization is a framework for classifier combination in which each layer of classifiers is used to combine the predictions of the classifiers of its preceding layer. A single classifier at the topmost level outputs the ultimate prediction. In our approach, we use a two-level system that has a Bayesian network as this single, combining classifier and this Bayesian network is used to perform the last classification step.

3. Datasets for protein secondary structure prediction

We have used nine different datasets to develop and test our multi-classifiers. They are:

- **HS1771**: The dataset HS1771, with 1771 sequences, has been used for training and testing our multi-classifiers. This dataset is the PDB_SELECT list [7] published on March, 2002. The PDB_SELECT lists are intended to save time and effort by offering a representative selection of the PDB database, that is currently a factor of eight smaller than the entire database.
- **CB513**: This dataset of 513 sequences was developed by Cuff and Barton [9] with the aim of evaluating and improving protein secondary structure prediction methods. It is, perhaps, one of the most used independent dataset in this field.
- **RS126**: This original set of 513 sequences by Rost and Sander [8], currently correspond to a total of 23,363amino acids positions (this number has varied slightly over the years due to changes and corrections in the PDB [11]).
- **EVA1, . . ., EVA6**: 6 novel test sets are provided by the datasets available from the real-time evaluation experiment called EVA [10], which compares a number of prediction servers on a regular basis using the sequences deposited in the PDB every week. In particular we have used all the datasets labelled ''common1'' to ''common6'' published on 19/10/2002.

4. Statistics in the PSSP problem

To validate the results obtained in the methods of secondary structure prediction, a set of statistic factors has been defined.

In order to compute these statistics, a 3 × 3-sized confusion matrix has been used, where the rows show the states of the actual secondary structure (obtained through the DSSP program [12]) and the columns describe the states of the secondary structure predicted by the classifier. Table 1 shows the elements of the confusion matrix.

Table 1 Confusion matrix for the PSSP problem

Observed	Predicted			
	H	*E*	*L*	
H				obs_H
E				obs_E
L				obs_L

The following values are calculated from the confusion matrix:

- obs_H: represents the number of residues observed in state helix (*H*), that is, the states *H* that appear in the real structure,
- obs_E: represents the number of residues observed in state β strand (*E*),
- obs_L: represents the number of residues observed in state *coil* (*L*),
- prd_H: represents the number of residues predicted in state helix (*H*),
- prd_E: represents the number of residues predicted in state β strand (*E*),
- prd_L: represents the number of residues predicted in state *coil* (*L*),
- N_{res}: represents the total number of residues of the chain, that is, the length of the sequence.

In a mathematical representation, we observe that:

M_{ij} denotes the number of residues observed in state *i* and predicted in state *j*, with $i, j \in \{H, E, L\}$

The total number of residues observed in state *i* is:

$$obs_i = \sum_{j \in \{H,E,L\}} M_{ij} \tag{1}$$

The total number of residues predicted in state *j* is:

$$prd_j = \sum_{i \in \{H,E,L\}} M_{ij} \tag{2}$$

and the total number of states in the sequence is:

$$N_{res} = \sum_i obs_i = \sum_j prd_j = \sum_{i,j} M_{ij} \tag{3}$$

4.1. Three-state prediction accuracy: Q_3

This is the measure used traditionally for evaluating the accuracy of secondary structure prediction. This parameter represents the total number of residues correctly predicted. In order to calculate it, the states helix, β strand and *coil*correctly predicted are added (sum of all M_{ii}), dividing this sum by the total number of residues of the observed sequence (N_{res}) and expressing the result as percentage. Q_3 is obtained as:

$$Q_3 = 100\frac{1}{N_{res}}\sum_{i=1}^{3} M_{ii} \tag{4}$$

4.2. Per-state percentages

To define the accuracy for a particular state (helix, strand, *coil*), there are two possible variants:

- Percentage of all residues observed in a particular state (*%obs*).

$$Q_i^{\%obs} = 100\frac{M_{ii}}{obs_i} \tag{5}$$

In this way, for example, it is possible to calculate the percentage of residues observed in the state helix H.

- Percentage of all residues correctly predicted in a particular state (*%prd*).

$$Q_i^{\%prd} = 100\frac{M_{ii}}{prd_i} \tag{6}$$

For example, for a particular state $i = H$, it is possible to calculate the percentage of residues correctly predicted in the state helix H.

4.3. Information index

The information index is given by:

$$info = \ln\left(\frac{P_{prd}}{P_{obs}}\right) \tag{7}$$

where P_{obs} describes the probability for finding one particular string of N_{res} residues with obs_i residues being in structure i out of all combinatorial possible ones, and P_{prd} is the probability for a particular realization of the confusion matrix M. The resulting information index is

$$info = \frac{info^{\%obs} + info^{\%prd}}{2} \tag{8}$$

with

$$info^{\%obs} = 1 - \frac{\sum_{i=1}^{3} prd_i \ln prd_i - \sum_{i,j=1}^{3} M_{ij}\ln M_{ij}}{N_{res}\ln N_{res} - \sum_{i=1}^{3} obs_i \ln obs_i} \tag{9}$$

$$info^{\%prd} = 1 - \frac{\sum_{i=1}^{3} obs_i \ln obs_i - \sum_{i,j=1}^{3} M_{ij}\ln M_{ij}}{N_{res}\ln N_{res} - \sum_{i=1}^{3} prd_i \ln prd_i} \tag{10}$$

4.4. Matthew's correlation coefficients

Matthew's correlation coefficients [13] are not influenced by the percentage of true positives (number of elements of the structure i correctly predicted divided by the number of the elements of the structure i) in a sample, being the best way of evaluating different methods. The result is a number between -1 and 1, where value 1 represents a perfect coincidence, value -1 a total inequality and value 0 indicates that the prediction has not correlation with the results.

Although the correlation coefficient is an useful measure of the accuracy of the prediction, this coefficient does not evaluate the similarity between the prediction and the protein. In order to know the accuracy of the prediction, the segment overlap measure is taken into account. Matthew's correlation coefficients are defined by the following formula:

$$C_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}} \tag{11}$$

with

$$p_i = M_{ii}, n_i = \sum_{j\neq i}\sum_{k\neq i} M_{jk}, o_i = \sum_{j\neq i} M_{ji}, u_i = \sum_{j\neq i} M_{ij} \tag{12}$$

$i, j \in \{H, E, L\}$, where:

- n_i: contains the number of different states observed in i and predicted as state j, being j different from i. For example, for the state $i = H$, n_i represents the number of states H observed in the sequence and predicted as L or E.
- u_i: contains the number of residues observed in the state i and predicted in a state different from i. For example, for the state $i = H$, u_i represents the number of states H observed in the sequence and predicted as E or L.
- p_i: represents the number of residues observed in the state i and correctly predicted.
- o_i: represents the number of residues observed in a state different from i and predicted as i.

4.5. SOV: Segment OVerlap measure

Statistics applied previously are general statistics, and thus, can be applicable to every classification problem. However, the segment overlap (SOV) is a measure, developed by Rost [14] and modified by Zemla [15], which specifies the specific objectives of the secondary structure prediction.

Unlike the measure Q_3, which considers the residues in an individual fashion, *SOV* measures the accuracy taking the different segments of a sequence into account. *SOV* provides the measure of the segment overlap for an only state (*H*, *E* or *L*) or for all three states.

If for example we consider the state *coil* (*L*), the measure *SOV* calculates the accuracy of the segments prediction in such state. A segment is considered a part of the sequence where the state *i* appears consecutively (in this case, *L*). Therefore, 100% is obtained when the segments of the observed sequence are equal to the predicted sequence. When the *SOV* is calculated for all the three states, the segments of the three states (helix, β strand and coil) are taking into account.

4.5.1. Per-stage segment overlap

This value is given by:

$$SOV(i) = 100\frac{1}{N(i)}\sum_{S(i)}\frac{minov(s_1,s_2)+\delta(s_1,s_2)}{maxov(s_1,s_2)}len(s_1) \tag{13}$$

where:

- s_1 and s_2: are the observed and predicted secondary structure segments (in state *i*, which can be either *H*, *E* or *L*),
- $len(s_1)$: is the number of residues in the segment s_1,
- $minov(s_1,s_2)$: is the length of actual overlap of s_1 and s_2, i.e. the extent for which both segments have residues in state *i*,
- $maxov(s_1,s_2)$: is the length of the total extent for which either of the segments s_1 or s_2 has a residue in state *i*,
- $\delta(s_1,s_2)$: is the integer value defined as being equal to the following:

$$\delta(s_1,s_2) = min\left\{\begin{array}{l} maxov(s_1,s_2)-minov(s_1,s_2) \\ minov(s_1,s_2) \\ int(0.5*len(s_1)) \\ int(0.5*len(s_2)) \end{array}\right\} \tag{14}$$

- $\sum$: is taken over all the pairs of segments (s_1,s_2), where s_1 and s_2 have at least one residue in state *i* in common,
- $N(i)$: is the number of residues in state *i* defined as follows:

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

The two sums are taken over *S* and *S'*. *S*(*i*) is the number of all pairs of segments (s_1,s_2), where s_1 and s_2 have at least one residue in state *i* in common. *S'*(*i*) is the number of segments S_1 that do not produce any segment pair.

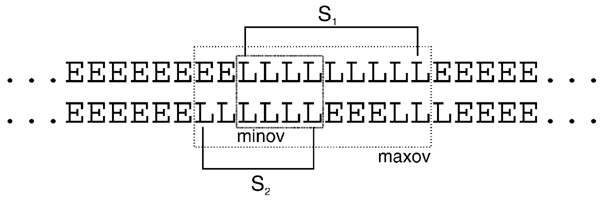


Figure 3 Fragment of an observed sequence and a predicted sequence with the elements of the *SOV* formula.

Fig. 3 shows a fragment of an observed sequence and a predicted sequence, where the elements of the *SOV* formula are depicted.

4.5.2. Segment OVerlap quantity measure for all three states

This value is obtained by applying this formula:

$$SOV = 100\frac{1}{\sum_i N(i)}\sum_i\sum_{S(i)}\frac{minov(s_1,s_2)+\delta(s_1,s_2)}{maxov(s_1,s_2)}len(s_1) \tag{16}$$

where $\sum_i N(i)$ is a sum of $N(i)$ over all three conformational states ($i=$ helix, strand, *coil*).

5. Level-0 composite classifiers

After an exhaustive search over Internet, we have found 9 secondary structure prediction servers. We have selected, with our own experimental results, the best 6 servers as the *level*-0 classifiers. The Table 2 shows all the contacted servers, with its location and prediction method. Also, Fig. 4 shows the geographical location of the servers.

5.1. JPred

JPred [16] is an interactive protein secondary structure prediction Internet server. The server allows a single sequence or multiple alignment to be submitted, and returns predictions from six secondary structure prediction algorithms that exploit evolutionary information from multiple sequences. A consensus prediction is also returned.

All the secondary structure prediction methods used, require either, multiple sequences or an alignment of multiple sequences. Thus, if a single sequence is submitted, an automatic process creates a multiple sequence alignment, prior to prediction [16].

Six different prediction methods: DSC [17], PHD [8], NNSSP [18], PREDATOR [19], ZPRED [20] and MULPRED [21] are then run, and results from each method are combined into a simple file format.

Table 2 Secondary structure prediction servers on Internet

Server	*Secondary structure prediction servers*	
	Location	*Prediction method*
JPred	University of Dundee, Scotland	Consensus
PHD	Columbia University, USA	Neural networks
Prof	University of Wales, UK	Neural networks
PSIPRED	University College London, UK	Neural networks
SAM-T02	University of California, Santa Cruz, USA	Homology
SSPro	University of California, Irvine, USA	Neural networks

A consensus prediction based on a simple majority method of NNSSP, DSC, PREDATOR and PHD is provided by the JPred server.

5.2. SSPro

SSPro is a fully automated system for the prediction of protein secondary structure. The system is based on an ensemble of bidirectional recurrent neural networks (BRNNs) [22,23]. BRNNs are graphical models that learn from data the transition between an input and an output sequence of variable length. The model is based on two hidden Markov chains, a forward and a backward chain, that transmit information in both directions along the sequence, between the input and the output sequences. Three neural networks model respectively the forward state update, the backward state update and the input and hidden states to output transition. BRNNs are trained in a supervised fashion using the gradient descent algorithm. The error signal is propagated through the model using the BPTS (backpropagation through structure) algorithm, an extension of BPTT (backpropagation through time), used in unidirectional recurrent neural networks.

A set of 11 bidirectional recurrent neural networks is trained on the dataset. The networks contain roughly 70,000 adjustable weights, have normalized exponentials on the outputs and are trained using the relative entropy between the target and output distributions. The final predictions are obtained averaging the network outputs for each residue.

5.3. PHD

PHD [24] was the first method to incorporate evolutionary information (in the form of multiple sequence alignment data) in the prediction of protein secondary structure. The first step in a PHD prediction is generating a multiple sequence alignment. The second step involves feeding the alignment into a neural network system. Correctness of the multiple sequence alignment is as crucial for

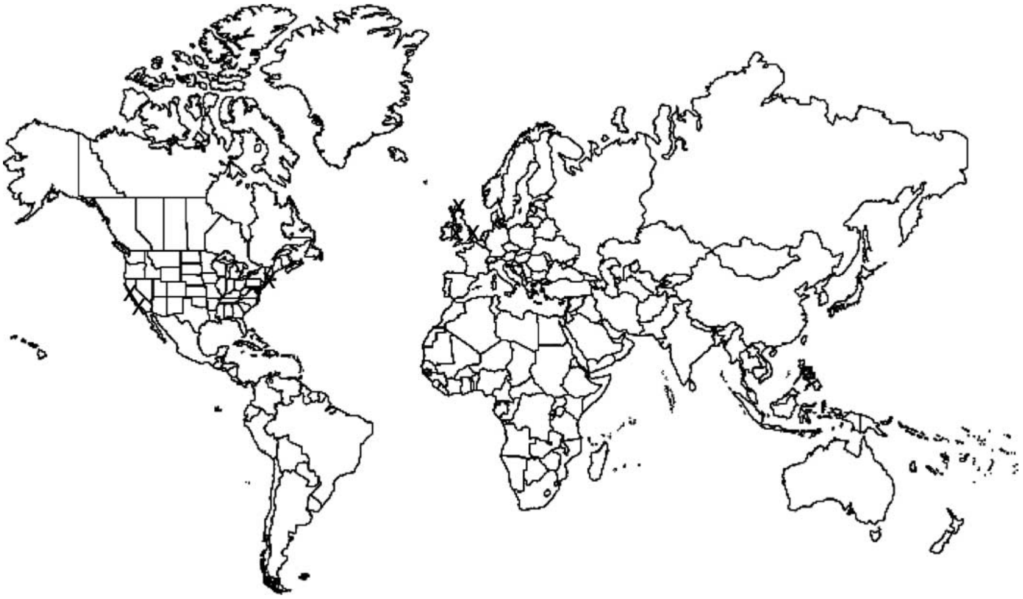


Figure 4 Geographical disposition of the secondary structure prediction servers combined in this work.

prediction accuracy as that the alignment contains a broad spectrum of homologous sequences.

The PHD methods process the input information on multiple levels. The first level is a feed-forward neural network with three layers of units (input, hidden, and output). Input to this first level sequence-to-structure network consists of two contributions: one from the local sequence, i.e., taken from a window of 13 adjacent residues, and another from the global sequence. Output of the first level network is the 1D structural state of the residue at the centre of the input window. The second level is a structure-to-structure network. The next level consists of an arithmetic average over independently trained networks (jury decision). The final level is a simple filter.

5.4. PSIPRED

PSIPRED [3] is a simple and reliable secondary structure prediction method. It use a two-stage neural network to predict protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST. The prediction method is split into three stages: generation of a sequence profile, prediction of initial secondary structure, and finally the filtering of the predicted structure.

5.5. PROF

The Prof server [25] is a classifier for protein secondary structure prediction which is formed by cascading (in multiple stages) different types of classifiers using neural networks and linear discrimination. To generate different classifiers it has been used GOR formalism-based methods extended by linear and quadratic discriminations [26,27] and neural network-based methods [28,24]. The theoretical foundation for Prof comes from basic probability theory which states that all of the evidence relevant to a prediction should be used in making that prediction.

5.6. SAM-T02

The SAM-T02 [29] method is used for iterative SAM HMM construction, remote homology detection and protein structure prediction. It updates SAM-T99 by using predicted secondary structure information in its scoring functions.

The SAM-T02 server is an automatic method that uses two-track hidden Markov models (HMMs) to find and align template proteins from PDB to the target protein. The two-track HMMs use an amino-acid alphabet and one of several different local-structure alphabets.

The SAM-T02 prediction process consists of several parts:

- Finding similar sequences with iterative search using SAM-T2K.
- Predicting local structure properties with neural nets.
- Finding possible fold-recognition templates using 2-track HMMS (the SAM-T02 method).
- Making alignments to the templates.
- Building a specific fragment library for the target (with fragfinder).
- Packing fragments and fold-recognition alignments to make a 3D structure (with undertaker).

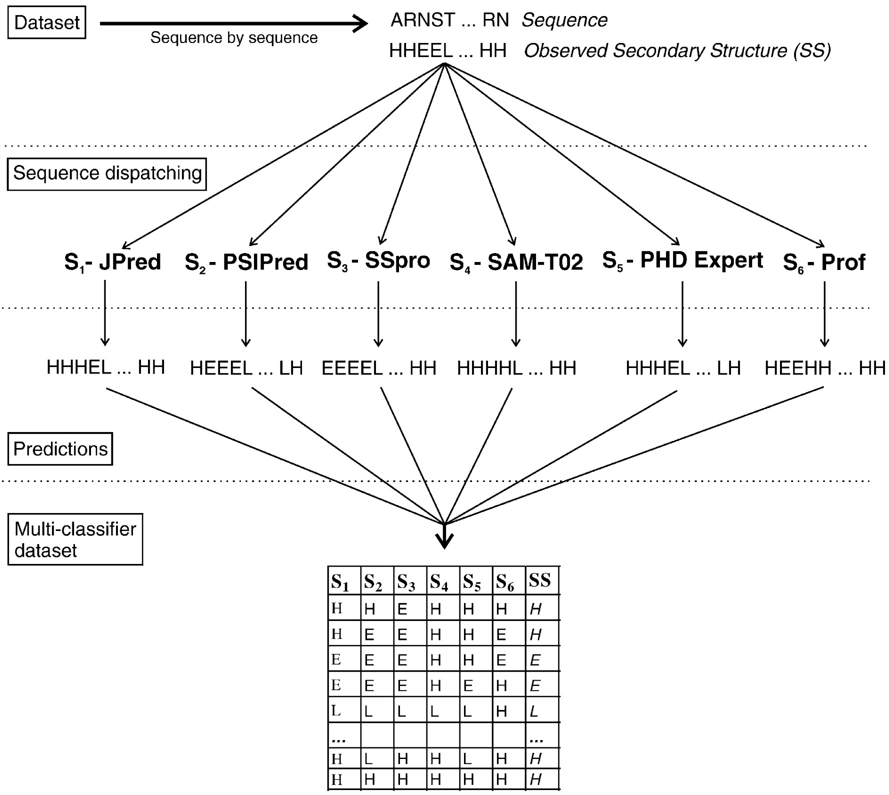
6. Obtaining the datasets for multi-classifiers training

The process of creating an appropriate dataset for both training and evaluation of the multi-classifiers, as shown in Fig. 5, has been achieved in following steps:

(1) High-quality datasets of proteins with known secondary structure are selected. The most representative dataset designed by different groups are HS1771, CB513, RS126 as well as the six data sets gathered by EVA project [10]. From all of them only HS1771 has been used for the training phase of a multi-classifier. This dataset has been selected because it is the most complete of the nine datasets. The remaining ones will be used in the testing phase of the algorithm.
(2) These sequences of proteins are submitted to the six web servers and the process waits for their replies. These replies came as either web pages or e-mail messages.
(3) The replies, once they have been received, are processed. The prediction for the secondary structure of the protein is extracted from the body of the message or from the contents of the web page.
(4) The results are stored in a new dataset to be processed by a multi-classifier. For each of the aminoacids of the protein an instance of the dataset is inserted with all the predictions from the servers and the actual value of its secondary structure.

7. Level-1 classifiers based on Bayesian networks

As exposed in [30] we have used Bayesian networks as the consensed voting system. Thus, for



S_1	S_2	S_3	S_4	S_5	S_6	SS
H	H	E	H	H	H	*H*
H	E	E	H	H	E	*H*
E	E	E	H	H	E	*E*
E	E	E	H	E	H	*E*
L	L	L	L	L	H	*L*
...						...
H	L	H	H	L	H	*H*
H	H	H	H	H	H	*H*

Figure 5 Obtaining the multi-classifier dataset.

building the level-1classifiers, we have used three different Bayesian network structures: naïve Bayes, *Interval Estimation Naïve Bayes* (IENB) and the idea of Pazzani of joining attributes in naïve Bayes.

7.1. Naïve Bayes

The naïve Bayes classifier [31,32] is a probabilistic method for classification. It performs an approximate calculation of the probability that an example belongs to a class given the values of predictor variables. The simple naïve Bayes classifier is one of the most successful algorithms on many classification domains. In spite of its simplicity, it is shown to be competitive with respect to other more complex approaches in several specific domains.

This classifier learns from training data the conditional probability of each variable X_k given the class label c. Classification is then done by applying Bayes rule to compute the probability of C given the particular instance of $X_1, \ldots, X_n$,

$$P(C = c | X_1 = x_1, \ldots, X_n = x_n) \tag{17}$$

Naïve Bayes is founded on the assumption that variables are conditionally independent given the class. Therefore, posterior probability of the class variable is formulated as follows,

$$P(C = c | X_1 = x_1, \ldots, X_n = x_n) \propto P(C = c) \prod_{k=1}^{n} P(X_k = x_k | C = c) \tag{18}$$

This equation is highly appropriate for learning from data, since probabilities $p_i = P(C = c_i)$ and $p^i_{k,r} = P(X_k = x^r_k | C = c_i)$ may be estimated from training data. The result of the classification is the class with highest posterior probability.

7.2. Interval estimation naïve Bayes

Interval estimation naïve Bayes (IENB) [33]belongs to the semi naïve Bayes approaches that correct the probabilities produced by the standard naïve Bayes. In this approach, instead of calculating the point estimation of the conditional probabilities from data, as simple naïve Bayes does, confidence intervals are calculated. After that, the search for the best combination of values into these intervals is performed. The goal of this search is try to relieve the assumption of independence among variables the simple naïve Bayes does. This search is carried out by a heuristic search algorithm and is guided by the accuracy of the classifiers.

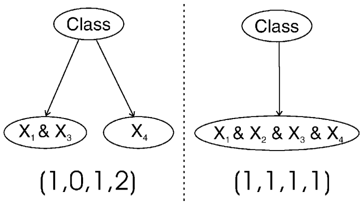


Figure 6 Two Pazzani structures and their corresponding individuals.

To deal with the heuristic search EDAs—estimation of distribution algorithms—have been selected. EDAs [34] are non-deterministic, stochastic and heuristic search strategies that belong to the evolutionary computation approaches. In EDAs, a number of solutions or individuals is created every generation, evolving once and again until a satisfactory solution is achieved. In brief, the characteristic that most differentiates EDAs from other evolutionary search strategies, such as GAs, is that the evolution from a generation to the next one is done by estimating the probability distribution of the fittest individuals, and afterwards, by sampling the induced model. This avoids the use of crossover or mutation operators, and, therefore, the number of parameters that EDAs requires is reduced considerably.

7.3. Pazzani

Pazzani in [35] proposes to improve the naïve Bayes classifier by searching for dependencies among attributes. He develops two algorithms for detecting dependencies among attributes: *Forward Sequential Selection and Joining* (FSSJ) and *Backward Sequential Elimination and Joining* (BSEJ).

In this paper, we propose to make a heuristic search of the Pazzani structure with the target of maximize the percentage of successful predictions. We perform this heuristic search with EDAs. The resulting algorithm is called Pazzani-EDA algorithm and a multi-classified build upon it, the MC-Pazzani-EDA multi-classifier.

Fig. 6 contains an example of two Pazzani structures and their corresponding individuals.

Thus, for a dataset with *n* attributes, individuals will have *n* genes, each one with an integer value between 0 an *n*. The value 0 represents that the corresponding attribute is not part of the Pazzani structure. A value between 1 and *n* means that the corresponding attribute belongs to that group in the Pazzani structure.

7.4. Level-1 classifiers

After all the predictions outputs from the six web servers are collected a preliminary study of the predictive accuracy is performed. The statistics presented on Section 4are calculated. A detailed discussion of these results is included in the next section.

Taking these results into account, we have built and trained the following seven multi-classifiers based on Bayesian networks:

- Naïve-Bayes with the:
  - best 4 severs (MC-NB-4),
  - best 5 severs (MC-NB-5),
  - best 6 severs (MC-NB-6).
- Interval estimation naïve-Bayes with the:
  - best 4 severs (MC-IENB-4),
  - best 5 severs (MC-IENB-5),
  - best 6 severs (MC-IENB-6).
- MC-Pazzani-EDA with the best 6 servers.

The best six servers (sorted by their accuracy) are: PSIPRED, SSPro, SAM-T02, PHD Expert, Prof and JPred. Further details are shown on Tables 3—11.

8. Results

Statistics of the predictions performed by the six selected servers (described in Section 5) and the seven multi-classifiers are presented in a form of tables: Tablest3,t4,t5,t6,t7,t8,t9,t10,t11. We present one table for each of the datasets (HS1771, CB513, RS126, EVA1, EVA2, EVA3, EVA4, EVA5and EVA6).

These results show that the best classifier for these datasets is PSIPRED. Although its predictions of the secondary structure are of the highest accuracy, it has been further improved by our multi-classifier architectures.

Improvements in terms of overall accuracy are presented on Table 12. These results are compared to PSIPRED predictions on each of the datasets. Our results show that the best multi-classifier algorithm is MC-Pazzani-EDA with and absolute improvement of 1.21% compared to PSIPRED. If we consider a relative improvement, taking the theoretical maximum accuracy of 88% into account, MC-Pazzani-EDA outperforms the best classifier by 13.30%. This algorithm gets better results in all but one (EVA6) of the datasets. In Fig. 7, we show the final structure obtained by MC-Pazzani-EDA. The four better servers have been included in this structure.

8.1. Detailed analysis of HS1771 results

Here are some details of a deeper analysis of the results obtained for HS1771, the most complete dataset for protein second structure prediction.

Table 3 Statistics for HS1771 dataset

Servers	HS1771 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	78.90	83.40	69.18	80.29	83.34	78.19	75.72	0.41	0.74	0.67	0.61	77.86	71.53	71.65	74.62
PHD Expert	77.37	80.01	72.87	77.55	84.73	71.58	74.81	0.38	0.73	0.64	0.58	72.81	72.92	70.22	73.44
Prof	74.57	71.56	71.01	78.95	86.50	67.27	70.93	0.34	0.69	0.60	0.55	65.32	69.95	69.43	69.74
GOR	54.31	56.67	45.28	57.13	54.43	42.34	61.49	0.08	0.31	0.27	0.31	52.15	54.41	49.60	50.21
SOPM	65.66	70.34	54.85	67.48	67.40	57.52	68.28	0.19	0.51	0.44	0.45	64.38	64.28	60.81	62.52
SSPro	78.04	82.65	66.26	80.44	83.84	76.83	74.17	0.40	0.74	0.64	0.59	74.70	70.09	69.81	71.78
JPred	71.68	63.56	54.57	87.52	87.04	73.13	64.40	0.30	0.64	0.54	0.52	62.78	63.23	67.82	65.91
SAM-T02	77.54	84.39	74.97	73.20	81.41	72.16	77.14	0.39	0.73	0.66	0.58	75.54	73.56	68.15	73.05
Predator	66.29	65.42	45.83	77.83	71.00	64.42	63.90	0.20	0.52	0.44	0.45	63.64	54.73	61.95	61.07
MC-NB-4	79.91	84.15	75.06	78.89	85.97	76.39	76.74	0.43	0.77	0.69	0.62	79.05	73.68	69.13	73.18
MC-NB-5	79.92	87.09	77.47	75.81	82.50	72.66	81.62	0.43	0.77	0.68	0.62	83.47	76.07	66.38	72.19
MC-NB-6	79.72	86.40	76.77	76.09	82.87	73.01	80.64	0.43	0.77	0.68	0.62	82.28	75.29	66.51	72.11
MC-IENB-4	79.96	83.94	75.50	78.89	86.30	76.06	76.74	0.44	0.77	0.69	0.62	79.09	74.66	69.13	73.29
MC-IENB-5	80.00	86.96	77.19	76.16	82.65	73.49	81.24	0.44	0.77	0.68	0.62	82.67	75.97	66.70	72.37
MC-IENB-6	79.74	86.47	76.79	76.09	82.87	73.11	80.65	0.43	0.77	0.68	0.62	82.39	75.37	66.51	72.15
MC-Pazzani-EDA	80.25	85.41	79.05	76.76	84.80	71.80	80.94	0.44	0.77	0.69	0.63	80.16	77.30	67.00	72.28

Table 4 Statistics for CB513 dataset

Servers	CB513 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	79.95	83.52	70.31	82.16	84.67	80.14	76.37	0.43	0.76	0.68	0.63	78.98	69.81	73.42	76.48
PHD Expert	77.61	78.92	73.32	78.82	85.58	72.00	74.84	0.39	0.73	0.65	0.59	73.36	69.83	71.27	74.98
Prof	77.13	74.19	74.23	81.03	88.65	71.14	73.10	0.39	0.73	0.64	0.58	68.58	69.83	72.24	73.74
GOR	55.36	56.20	47.53	58.82	55.91	43.02	62.55	0.08	0.33	0.28	0.33	51.45	51.87	50.89	51.60
SOPM	66.84	69.78	57.24	69.56	69.29	58.81	68.98	0.20	0.53	0.46	0.46	65.31	62.70	62.39	64.35
SSPro	79.07	82.72	66.90	82.56	85.90	78.43	74.53	0.42	0.76	0.65	0.61	76.22	68.11	72.27	74.39
JPred	73.37	65.21	56.18	89.07	89.41	77.01	65.40	0.34	0.67	0.58	0.54	64.65	61.00	69.16	68.03
SAM-T02	78.17	83.99	75.37	74.96	82.82	72.90	77.23	0.40	0.75	0.66	0.59	75.47	71.11	69.35	74.01
Predator	80.04	78.18	71.88	85.87	84.37	83.40	75.83	0.42	0.72	0.71	0.65	73.49	68.47	72.55	74.88
MC-NB-4	80.57	85.35	76.18	79.05	85.45	76.80	78.62	0.45	0.78	0.70	0.63	81.82	74.88	68.78	74.26
MC-NB-5	80.61	88.13	78.93	76.20	82.06	73.62	83.15	0.45	0.78	0.70	0.63	85.42	77.35	66.18	73.25
MC-NB-6	80.53	87.58	78.49	76.49	82.48	74.12	82.34	0.45	0.77	0.70	0.63	84.64	77.15	66.61	73.39
MC-IENB-4	80.63	85.15	76.67	79.05	85.80	76.56	78.62	0.45	0.78	0.70	0.63	81.79	75.66	68.78	74.4
MC-IENB-5	80.68	87.97	78.67	76.53	82.19	74.42	82.79	0.45	0.78	0.70	0.63	84.07	77.20	66.45	73.41
MC-IENB-6	80.52	87.62	78.44	76.49	82.46	74.15	82.34	0.45	0.77	0.70	0.63	84.65	77.14	66.61	73.39
MC-Pazzani-EDA	80.99	86.24	80.29	77.37	84.97	72.48	82.27	0.45	0.78	0.70	0.64	82.64	78.57	66.74	73.37

Table 5 Statistics for RS126 dataset

Servers	RS126 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	81.01	84.35	72.62	83.01	84.58	80.44	78.88	0.45	0.77	0.7	0.65	81.47	69.42	74.34	76.24
PHD Expert	76.92	78.33	73.65	77.62	84.77	69.46	75.93	0.38	0.73	0.63	0.57	71.42	68.86	70.32	72.57
Prof	76.95	74.18	75.20	79.82	88.07	69.30	74.79	0.38	0.73	0.63	0.58	69.44	68.66	70.23	71.70
GOR	55.39	56.21	48.46	58.41	53.75	44.17	63.68	0.08	0.33	0.29	0.32	51.02	51.47	51.39	50.89
SOPM	66.03	67.98	57.30	69.17	66.37	57.80	70.06	0.19	0.52	0.45	0.45	64.14	60.96	62.65	62.43
SSPro	77.01	80.84	64.38	80.85	82.81	74.62	74.32	0.38	0.74	0.61	0.58	75.63	64.87	68.83	70.24
JPred	73.82	65.53	56.46	88.78	89.32	77.56	66.68	0.34	0.68	0.58	0.54	65.66	59.88	67.33	66.55
SAM-T02	78.81	84.93	77.11	75.36	82.94	72.56	79.28	0.42	0.76	0.67	0.60	77.83	72.62	68.58	73.30
Predator	80.06	79.71	69.38	85.85	83.67	82.62	76.89	0.42	0.73	0.69	0.64	71.48	65.29	69.86	71.42
MC-NB-4	80.21	84.30	74.44	80.43	85.12	78.08	77.84	0.44	0.78	0.69	0.63	81.30	74.73	67.21	71.90
MC-NB-5	80.39	87.05	77.09	77.77	81.83	74.73	82.31	0.44	0.77	0.69	0.63	84.84	77.30	64.41	70.53
MC-NB-6	80.21	86.46	76.54	77.98	82.23	75.12	81.43	0.44	0.77	0.69	0.62	84.23	77.37	64.66	70.50
MC-IENB-4	80.29	84.10	74.96	80.43	85.58	77.83	77.84	0.44	0.78	0.69	0.63	81.77	76.39	67.21	72.08
MC-IENB-5	80.55	86.88	76.89	78.31	81.91	76.05	81.91	0.45	0.77	0.69	0.63	84.61	77.29	64.98	70.96
MC-IENB-6	80.22	86.55	76.47	77.98	82.23	75.16	81.43	0.44	0.77	0.69	0.62	84.22	77.38	64.43	70.37
MC-Pazzani-EDA	81.65	85.36	82.43	78.85	85.67	71.96	83.85	0.47	0.79	0.71	0.65	80.95	83.36	64.64	70.67

Table 6 Statistics for EVA1 dataset

Servers	EVA1 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	75.48	80.83	66.10	75.46	79.72	72.85	73.03	0.35	0.69	0.62	0.55	73.52	69.94	66.88	70.17
PHD Expert	75.26	79.37	73.36	72.64	80.58	68.53	74.23	0.35	0.69	0.63	0.54	67.78	74.05	66.84	69.81
Prof	72.11	71.88	68.67	74.00	81.42	63.82	69.56	0.30	0.64	0.57	0.50	61.84	69.98	65.84	66.59
GOR	54.44	62.23	43.03	53.30	55.07	41.01	61.69	0.08	0.32	0.26	0.30	50.63	56.06	48.92	49.49
SOPM	63.11	70.41	52.29	62.10	63.58	54.55	66.94	0.16	0.46	0.42	0.40	57.84	64.63	57.91	57.89
SSPro	74.48	81.14	63.51	74.10	77.49	72.54	72.64	0.33	0.67	0.60	0.53	67.90	71.07	65.80	67.03
JPred	68.82	59.23	54.07	84.37	83.32	67.67	62.57	0.25	0.58	0.51	0.47	54.99	64.78	65.11	61.79
SAM-T02	74.65	82.03	73.33	68.95	78.00	68.65	74.77	0.34	0.68	0.63	0.52	70.12	72.98	63.55	67.70
Predator	61.72	62.72	37.08	72.91	64.24	56.28	61.41	0.13	0.43	0.35	0.39	54.12	52.55	58.24	55.09
MC-NB-4	76.90	79.46	71.17	77.49	85.12	74.35	71.02	0.38	0.71	0.65	0.57	72.91	69.63	65.61	69.42
MC-NB-5	76.99	82.33	73.60	74.05	81.62	69.66	76.52	0.38	0.72	0.64	0.56	75.28	71.93	62.58	67.84
MC-NB-6	76.96	81.65	73.11	74.69	82.17	70.48	75.58	0.38	0.71	0.65	0.57	74.34	73.06	63.56	68.22
MC-IENB-4	76.86	79.07	71.59	77.49	85.37	73.70	71.02	0.38	0.71	0.65	0.57	72.67	69.78	65.61	69.37
MC-IENB-5	76.95	82.15	73.31	74.23	81.71	70.21	76.10	0.38	0.71	0.65	0.56	73.99	71.30	62.32	67.61
MC-IENB-6	76.99	81.68	73.18	74.70	82.16	70.65	75.58	0.38	0.71	0.65	0.57	74.32	73.17	63.57	68.26
MC-Pazzani-EDA	77.51	80.94	75.67	75.27	83.85	70.37	75.48	0.39	0.72	0.66	0.57	74.69	77.57	64.80	69.50

Table 7 Statistics for EVA2 dataset

Servers	EVA2 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	75.49	80.91	65.80	75.55	79.61	72.88	73.12	0.35	0.69	0.62	0.55	74.16	68.96	67.26	70.40
PHD Expert	75.27	79.42	73.29	72.65	80.62	68.47	74.25	0.35	0.69	0.63	0.54	68.31	73.15	66.75	69.91
Prof	72.11	72.00	68.45	73.99	81.47	63.72	69.56	0.30	0.64	0.57	0.50	62.41	69.08	65.66	66.65
GOR	54.40	61.97	43.15	53.36	55.14	40.68	61.77	0.08	0.32	0.26	0.30	50.60	55.48	48.97	49.53
SOPM	63.03	70.28	52.12	62.09	63.58	54.15	66.94	0.16	0.46	0.41	0.40	57.47	63.77	57.74	57.55
SSPro	74.45	81.21	63.33	74.05	77.50	72.38	72.62	0.33	0.67	0.60	0.53	68.25	70.18	65.52	66.98
JPred	68.86	59.35	54.02	84.34	83.40	67.54	62.63	0.25	0.58	0.51	0.47	55.25	64.08	65.41	61.86
SAM-T02	74.77	82.34	73.29	68.98	77.99	68.76	74.98	0.34	0.68	0.63	0.53	70.34	72.12	63.38	67.61
Predator	61.71	62.56	37.04	73.01	64.29	55.98	61.46	0.13	0.43	0.35	0.39	54.25	51.90	58.57	55.29
MC-NB-4	76.96	79.42	71.35	77.54	85.28	74.32	71.03	0.38	0.71	0.66	0.57	72.97	69.92	65.80	69.32
MC-NB-5	77.04	82.35	73.61	74.13	81.81	69.51	76.55	0.38	0.72	0.64	0.57	76.04	71.34	62.19	67.57
MC-NB-6	76.99	81.68	73.23	74.67	82.32	70.24	75.64	0.38	0.71	0.65	0.57	75.15	73.10	63.03	67.84
MC-IENB-4	76.91	79.00	71.78	77.54	85.54	73.62	71.03	0.38	0.71	0.66	0.57	72.70	70.07	65.80	69.26
MC-IENB-5	77.01	82.19	73.35	74.30	81.88	70.08	76.15	0.38	0.72	0.65	0.56	74.79	70.84	62.08	67.48
MC-IENB-6	77.02	81.74	73.26	74.68	82.30	70.41	75.64	0.38	0.71	0.65	0.57	75.12	73.18	63.03	67.88
MC-Pazzani-EDA	77.60	81.62	75.38	75.09	83.13	70.92	76.03	0.39	0.72	0.66	0.57	75.07	78.28	64.45	69.61

Table 8 Statistics for EVA3 dataset

Servers	EVA3 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	77.65	83.58	67.72	76.70	81.45	75.13	75.09	0.38	0.71	0.65	0.58	79.27	69.03	69.19	72.65
PHD Expert	76.50	80.82	71.86	74.60	82.23	70.99	73.83	0.36	0.70	0.64	0.56	74.88	73.35	67.79	72.28
Prof	73.78	72.87	69.72	76.54	83.61	65.37	70.23	0.32	0.66	0.59	0.53	64.88	70.23	65.90	66.74
GOR	54.76	57.06	47.45	56.02	58.95	38.96	60.34	0.08	0.32	0.28	0.30	50.63	57.38	48.90	50.31
SOPM	66.61	71.56	58.42	65.79	70.85	55.55	68.08	0.20	0.53	0.46	0.44	65.00	67.17	59.30	62.64
SSPro	76.38	82.06	64.84	76.41	80.92	73.47	73.36	0.35	0.69	0.62	0.56	74.57	69.71	66.63	69.74
JPred	71.16	65.33	53.89	84.78	83.93	71.79	63.89	0.28	0.61	0.55	0.50	62.81	64.18	66.35	65.15
SAM-T02	77.21	85.07	74.89	70.88	80.22	71.56	76.93	0.38	0.71	0.67	0.57	76.96	73.83	65.27	71.69
Predator	63.88	63.27	42.73	74.34	71.67	56.40	60.75	0.16	0.48	0.39	0.40	60.54	52.75	57.89	57.19
APSSP	92.49	94.78	93.26	90.43	88.66	96.14	93.58	0.72	0.88	0.93	0.86	79.79	84.15	79.25	74.74
MC-NB-4	78.60	81.91	73.06	77.98	85.79	75.69	73.20	0.40	0.73	0.68	0.59	77.30	70.46	68.27	72.13
MC-NB-5	78.86	84.72	75.81	75.00	82.81	71.41	78.63	0.41	0.74	0.67	0.60	81.36	72.37	65.30	71.77
MC-NB-6	78.51	83.65	74.97	75.35	83.33	71.39	77.31	0.40	0.73	0.67	0.50	80.21	73.27	65.47	71.40
MC-IENB-4	78.56	81.63	73.31	77.98	86.03	74.96	73.20	0.40	0.73	0.68	0.59	77.25	71.56	68.27	72.12
MC-IENB-5	78.88	84.59	75.34	75.34	82.92	72.31	78.15	0.41	0.73	0.68	0.60	80.06	72.15	65.60	71.90
MC-IENB-6	78.56	83.74	75.03	75.36	83.37	71.54	77.31	0.40	0.73	0.67	0.59	80.23	73.34	65.47	71.44
MC-Pazzani-EDA	79.55	85.06	76.38	76.01	83.39	73.54	78.74	0.42	0.74	0.69	0.61	80.33	73.19	66.28	72.27

Table 9 Statistics for EVA4 dataset

Servers	EVA4 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	77.92	83.74	68.71	76.98	81.82	76.24	75.00	0.39	0.72	0.66	0.59	79.29	70.42	70.02	73.56
PHD Expert	76.53	80.85	72.09	74.64	82.54	71.86	73.34	0.37	0.70	0.65	0.56	74.65	73.79	68.19	72.57
Prof	73.93	72.98	70.23	76.65	84.01	66.21	70.13	0.32	0.66	0.60	0.53	65.39	70.91	66.79	67.70
GOR	54.70	57.86	46.82	55.61	58.19	40.20	60.14	0.08	0.32	0.28	0.30	51.22	57.26	48.56	50.36
SOPM	66.50	71.82	57.95	65.70	70.11	57.14	67.72	0.20	0.52	0.47	0.44	64.81	67.44	59.38	62.61
SSPro	76.67	82.26	65.59	76.87	81.54	74.49	73.18	0.36	0.70	0.63	0.57	74.75	70.47	67.65	70.48
JPred	71.29	65.61	54.60	84.88	84.14	72.97	63.71	0.29	0.61	0.55	0.51	62.94	64.65	67.05	65.63
SAM-T02	77.40	85.03	75.73	71.04	80.75	72.15	76.73	0.39	0.72	0.67	0.57	76.71	74.77	65.94	72.24
Predator	63.61	63.59	42.36	74.06	70.87	57.49	60.42	0.16	0.48	0.39	0.40	60.40	52.97	57.72	57.31
MC-NB-4	78.78	82.42	73.88	77.64	85.78	76.14	73.49	0.41	0.74	0.69	0.59	78.03	71.16	68.22	72.38
MC-NB-5	79.04	85.19	76.48	74.83	82.88	72.29	78.75	0.41	0.74	0.68	0.60	82.19	72.91	65.49	72.08
MC-NB-6	78.67	84.06	75.75	75.12	83.32	72.36	77.39	0.41	0.73	0.68	0.59	81.07	73.72	65.74	71.99
MC-IENB-4	78.74	82.16	74.12	77.64	86.00	75.54	73.49	0.41	0.74	0.68	0.59	77.97	72.14	68.22	72.38
MC-IENB-5	79.08	85.06	76.09	75.17	82.99	73.15	78.32	0.41	0.74	0.68	0.60	81.04	72.73	65.84	72.26
MC-IENB-6	78.71	84.14	75.78	75.12	83.36	72.48	77.39	0.41	0.74	0.68	0.59	81.10	73.80	65.74	72.02
MC-Pazzani-EDA	79.71	85.12	78.53	75.44	84.07	71.86	79.46	0.43	0.75	0.69	0.61	81.39	73.69	65.54	71.88

Table 10 Statistics for EVA5 dataset

Servers	EVA5 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	77.95	83.03	67.68	77.86	82.95	74.80	74.57	0.38	0.72	0.65	0.59	79.33	70.64	70.37	73.32
PHD Expert	76.60	80.42	70.22	75.89	83.68	70.36	73.05	0.36	0.71	0.63	0.56	74.91	73.63	68.67	72.91
Prof	73.61	72.07	68.94	77.31	85.12	64.24	69.40	0.32	0.66	0.58	0.53	66.00	70.98	67.32	68.21
GOR	54.40	56.87	47.04	55.49	58.72	38.10	60.36	0.08	0.31	0.27	0.31	51.08	57.11	49.19	49.86
SOPM	65.64	70.77	56.18	65.12	70.15	53.89	67.06	0.18	0.51	0.44	0.43	64.60	66.16	58.14	61.32
SSPro	76.74	81.64	64.43	77.81	82.72	73.47	72.64	0.36	0.71	0.62	0.57	74.95	70.10	68.76	70.46
JPred	70.79	63.69	53.53	85.98	85.71	69.67	63.12	0.28	0.61	0.53	0.51	62.32	65.15	66.97	65.17
SAM-T02	77.40	84.92	74.59	71.40	81.49	71.05	76.35	0.38	0.72	0.66	0.57	75.80	74.96	66.54	72.00
Predator	63.00	62.76	41.39	73.50	70.77	54.25	60.10	0.15	0.47	0.37	0.39	60.82	53.03	57.70	56.89
MC-NB-4	79.09	83.42	73.07	77.63	85.76	74.93	74.57	0.41	0.74	0.68	0.60	79.15	70.58	68.61	72.43
MC-NB-5	79.09	86.20	75.54	74.42	82.14	71.16	79.88	0.41	0.74	0.67	0.60	82.29	72.08	65.16	71.42
MC-NB-6	78.77	85.32	74.60	74.69	82.59	71.05	78.71	0.40	0.74	0.66	0.60	81.22	71.87	65.11	71.08
MC-IENB-4	79.13	83.27	73.47	77.63	86.04	74.53	74.57	0.41	0.74	0.68	0.60	79.01	71.29	68.61	72.40
MC-IENB-5	79.14	86.08	75.22	74.73	82.30	71.92	79.48	0.41	0.74	0.67	0.60	81.59	71.83	65.46	71.56
MC-IENB-6	78.79	85.38	74.61	74.69	82.60	71.16	78.71	0.40	0.74	0.67	0.60	81.28	71.93	65.13	71.11
MC-Pazzani-EDA	79.67	85.65	76.43	75.60	83.87	72.12	79.16	0.42	0.75	0.68	0.61	81.67	73.30	65.88	71.70

Table 11 Statistics for EVA6 dataset

Servers	EVA6 dataset														
	Q_3	$Q_H^{\%obs}$	$Q_E^{\%obs}$	$Q_L^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%prd}$	$Q_L^{\%prd}$	*info*	C_H	C_E	C_L	SOV_H	SOV_E	SOV_L	SOV
PSIPRED	77.56	83.50	66.79	77.20	81.88	75.67	74.31	0.38	0.71	0.64	0.58	78.75	71.15	69.42	72.93
PHD Expert	76.28	80.58	70.16	75.21	82.81	70.38	73.22	0.36	0.70	0.63	0.56	74.07	73.32	67.71	72.03
Prof	73.40	72.51	68.15	76.82	84.43	64.64	69.49	0.31	0.66	0.58	0.53	65.18	69.82	66.88	67.58
GOR	54.70	59.37	45.43	54.8	57.75	39.35	61.07	0.08	0.31	0.26	0.31	52.12	56.46	48.14	50.04
SOPM	65.05	71.23	54.17	64.53	68.29	54.73	66.93	0.18	0.50	0.43	0.43	63.60	65.39	57.85	60.69
SSPro	76.30	82.16	63.85	76.83	81.31	73.69	72.79	0.35	0.7	0.61	0.57	74.36	70.39	67.63	69.96
JPred	70.51	63.45	53.55	85.62	85.30	69.96	62.95	0.28	0.61	0.53	0.50	61.8	64.96	66.24	64.53
SAM-T02	76.76	84.62	73.65	70.81	80.57	70.69	76.00	0.37	0.71	0.65	0.56	75.88	74.35	65.87	71.40
Predator	62.73	63.30	39.72	73.49	69.23	55.26	60.25	0.15	0.46	0.36	0.39	60.82	52.40	58.40	57.24
MC-NB-4	77.05	82.47	73.38	73.66	82.16	71.47	74.86	0.38	0.71	0.66	0.56	77.92	72.08	67.12	71.23
MC-NB-5	77.47	85.23	75.74	71.85	79.78	68.15	79.79	0.39	0.72	0.65	0.57	81.13	74.10	64.31	70.66
MC-NB-6	77.65	84.60	75.17	72.75	80.99	68.91	78.69	0.39	0.72	0.65	0.58	79.97	73.77	64.36	70.71
MC-IENB-4	77.07	82.24	73.83	73.66	82.41	71.04	74.86	0.38	0.71	0.66	0.56	77.62	73.06	67.12	71.25
MC-IENB-5	77.50	85.11	75.45	72.09	79.88	68.90	79.41	0.39	0.72	0.66	0.57	80.37	73.63	64.54	70.70
MC-IENB-6	77.66	84.63	75.21	72.75	80.99	68.98	78.69	0.39	0.72	0.65	0.58	79.98	73.82	64.36	70.74
MC-Pazzani-EDA	75.86	84.36	76.47	69.22	76.48	66.12	80.00	0.36	0.69	0.65	0.55	79.81	76.38	63.62	69.53

Table 12 Comparative results of the 7 multi-classifiers proposed (improvements related to PSIPRED server)

Dataset	MC-NB			MC-IENB			MC-Pazzani-EDA
	4 srv.	5 srv.	6 srv.	4 srv.	5 srv.	6 srv.	all srv.
HS1771	1.01	1.02	0.82	1.06	1.10	0.84	1.35
CB513	0.62	0.66	0.58	0.68	0.73	0.57	1.04
RS126	−0.80	−0.62	−0.80	−0.72	−0.46	−0.79	0.64
EVA1	1.42	1.51	1.48	1.38	1.47	1.51	2.03
EVA2	1.47	1.55	1.50	1.42	1.52	1.53	2.11
EVA3	0.95	1.21	0.86	0.91	1.23	0.91	1.90
EVA4	0.86	1.12	0.75	0.82	1.16	0.79	1.79
EVA5	1.14	1.14	0.82	1.18	1.19	0.84	1.72
EVA6	−0.51	−0.09	0.09	−0.49	−0.06	0.10	−1.70
Average	0.68	0.83	0.68	0.69	0.88	0.70	1.21

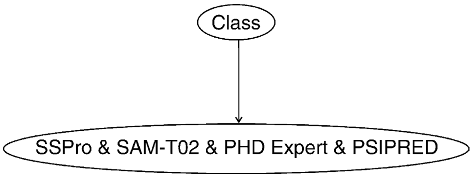


Figure 7 Final structure obtained by MC-Pazzani-EDA.

The most interesting results have been achieved for β strand prediction. PSIPRED predicts accurately 69.18% of the cases while MC-Pazzani-EDA gets 79.05% giving an improvement of 9.87%. As a drawback, *coil* structures are classified correctly in a 76.76% of the cases, instead of the previous 80.29% (although predictions have a better quality 80.94% versus PSIPRED 75.72%).

Another important remark is that the information index is better for all the multi-classifiers compared to PSIPRED. Multi-classifiers have information indexes between 0.43 and 0.44 while PSIPRED has a index of 0.41.

Matthews' correlation coefficients are also better for all multi-classifier approaches.

The best improvements achieved by multi-classifiers are focus on α helix and β strands. *Coil* prediction is less accurate than some of the best servers.

9. Conclusions and future research

On this paper several multi-classifiers based on Bayesian networks have been proposed for the problem of protein secondary structure prediction. Although significant improvements are achieved using simple classifiers (like naïve Bayes), the best results are obtained with innovative methods. These methods have been designed as wrapper approaches for existing Bayesian network classifiers. Interval Estimation Naïve Bayes (IENB) performs an estimation of the best classification probabilities inside of the boundaries of confidence intervals. Another new approach is the design of a variant of Pazzani classification method (greedy search), using heuristic search for selecting the most appropriate features for the classification procedure. Both new approaches use EDAs (estimation of distribution algorithms) to deal with heuristic search.

The multi-classifier system has been programmed as a JSP Web application using several Java classes. This system provides the following features:

- It compares the existing prediction servers worldwide. Statistics of their accuracy and other quality measures are extracted.
- An appropriate selection of datasets for protein secondary structure prediction has been selected.

These datasets have been used to train/test meta-classifiers commented above. The results obtained by these methods have outperformed existing state-of-the-art classifiers by 1.21% getting the best results ever obtained for this problem (80.99% from CB513—the most commonly used—and 80.25% for HS1771—the most complete—).

There are open issues still ahead:

- To evaluate new classification methods as second level strategies.
- To publish the meta-classifiers as an open-access web service.
- To create a portal to access existing web servers for PSSP prediction. This service would provide users with a single point to access multiple servers.

References

[1] Schmidler S, Liu J, Brutlag D. Bayesian segmentation of protein secondary structure. J Comput Biol 2000;2(1—2): 233—48.

[2] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;25(17):3389—402.

[3] Jones D. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292: 195—202.

[4] Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary strucure in three and eight classes using recurrent neural networks and profiles, Proteins (2001) 228—235.

[5] Wolpert DH. Stacked generalization. Neural Networks 1992;5:241—59.

[6] Ho T, Srihati S. Decision combination in multiple classifier systems. IEEE Trans Pattern Anal Machine Learning 1994; 11:63—90.

[7] Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the brookhaven protein data bank. Protein Sci 1992;1:409—17.

[8] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584—99.

[9] Cuff J, Barton G. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508—19.

[10] Rost B, Eyrich V. EVA: large-scale analysis of secondary structure prediction. Proteins 2001;5:192—9.

[11] Berman H, Westbrook J, Feng Z, Gilliand G, Bhat T, Weissig H et al. The protein data bank. Nucl Acid Res 2000;28: 235—42.

[12] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577—637.

[13] Matthews F. The structure, function and evolution of cytochromes. Prog Biophys Mol Biol 1985;45:1—56.

[14] Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994; 235:13—26.

[15] Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999; 34:220—3.

[16] Cuff J, Clamp M, Siddiqui A, Finlay M, Barton G. JPRED: A consensus secondary structure prediction server. Bioinformatics 1998;14:892—3.

[17] King R, Sternberg M. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 1996;5:2298—310.

[18] Salamov A, Solovyev V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. J Mol Biol 1995;247:11—5.

[19] Frishman D, Argos P. 75% accuracy in protein secondary structure prediction. Proteins 1997;27:329—35.

[20] Zvelebil M, Barton G, Taylor W, Sternberg M. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195: 957—61.

[21] Barton G, Taylor W. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1988;195:957—61.

[22] Baldi P, Brunak S, Frasconi P, Pollastri G, Soda G. Bidirectional dynamics for protein secondary structure prediction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99). Stockholm, Sweden, 1999.

[23] Baldi P, Brunak S, Frasconi P, Pollastri G, Soda G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;15:937—46.

[24] Rost B, Sander C, Schneider R. PHD—an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 1994;10:53—60.

[25] Ouali M, King R. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 2000;9:1162—76.

[26] Garnier J, Osguthorpe D, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978; 120:97—120.

[27] Gibrat J, Garnier J, Robson B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J Mol Biol 1987;198:425—43.

[28] Qian N, Sejnowski T. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 1988;202:865—84.

[29] Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, et al. Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction, ProteinsMayo.

[30] Sierra B, Serrano N, Larrañaga P, Plasencia E, Inza I, Jimenez J et al. Using Bayesian networks in the construction of a bi-level multi-classifier. A case study using intensive care unit patient data. Artif Intell Med 2001;22:233—48.

[31] Duda R, Hart P. Pattern classification and scene analysis. Wiley; 1973.

[32] Hand D, Yu K. Idiot's Bayes—not so stupid after all? Inter Stat Rev 2001;69(3):385—98.

[33] Robles V, Larrañaga P, Peña J, Menasalvas E, Pérez M. Interval estimation naïve Bayes. In: Lecture notes in computer science, advances in intelligent data analysis. Berlin, Germany; 2003, in press.

[34] Larrañaga P, Lozano J. Estimation of distribution algorithms. A new tool for evolutionary computation. Kluwer Academic Publisher; 2002.

[35] Pazzani M. Searching for dependencies in Bayesian classifiers. In: Learning from data: artificial intelligence and statistics V; 1997, pp. 239—48.