Enhancing Accuracy for Protein Prediction Secondary Structure by a New Hybrid Method

Youcef Gheraibia
Computing Department
University Md boudiaf
M'sila, 28000, Algeria

Abdelouahab Moussaoui
Computing Department
University Ferhat Abbase
Setif, 19000, Algeria

ABSTRACT
Prediction of protein secondary structure is an important step on the way to spell out its three dimensional structure and its function. This paper describes a new technique for prediction of secondary structure of protein based on contemporary machine learning methodology and data mining approach. More than one method has been developed to predict the protein secondary structure from the amino acids sequence; these methods show that we can achieve accuracy up to 80%. The work in this research is consists of three parts. In the first part, the secondary structure of each amino acid is predict alone with naive bays classifier, this method is based on amino acid preferences for different secondary structure. In the second part, an evolutionary algorithm to ameliorate this prediction is used; this method is based on physicochemical properties of protein regions. In the last part, a fragments bank which contains the protein fragments frequently detected in the Protein Data Bank (PDB) was developed; this method is based on the sequence alignment of protein but with a reduced database. The results of this research shows that the proposed method is improved the best know predictive accuracy by 4.5%, and attaint 85% accuracy with different datasets.

Keywords
Protein secondary structure prediction, Bays, Genetic algorithm, K nearest neighbor, Data mining, Amino acids, Hybrid method, supervised learning.

1. INTRODUCTION
Proteins is a polymer chain composed of 20 amino acids connected by peptide bonds, each protein is characterized by the number, nature, and order of amino acids [1]. The proteins are essential in the formation and function of all living cells. A protein may be between 10 and 10,000 amino acids. Amino acid sequences tend to form secondary structures which are a spatial arrangements and regularities due to hydrogen bonds between amino acids. The protein secondary structure prediction from its amino acids sequence is an NP-hard problem in combinatorial optimization, considered as a starting point to tertiary structure perdition and to improve a sequence analysis methods. Secondary structure is the general 3D regular local form, which are essentially three types: α helixes, β sheets and loops [2].

Prediction of secondary structure from primary sequence is to exploit the characteristics of each amino acid; these characteristics are biological knowledge or preference in experimentally determined protein structures. A large number of approaches have been developed to predict local structure of proteins. The first method takes into consideration individual statistics of each amino acid separately (Chou and Fasman) [3] which give accuracy up to 50%. From this generation, many approaches have been proposed to predict protein secondary structures, such as neural networks [4], hidden Markov models [5], support vector machines [6], and so forth. Despite these successes the Accuracy of prediction of these methods is low. By this generation and with predicting protein secondary structure using some physicochemical properties of amino acids, the evolution of machine learning methods reaches 80% of accuracy based on multiple sequence alignment.

In our research, we have combined several methods of secondary structure prediction based on data mining techniques, in a way that these methods are coherent, synergistic, and each one complementary for the other one. First, we use naive bays classifier to generate the initial population; in this step we use the preferences of each amino acid for each secondary structure separately. Then, we improve this population by a genetic algorithm based on the properties of hydrophobic and hydrophilic regions in proteins. And lastly, a fragments bank which contains the protein fragments frequently detected in the Protein Data Bank (PDB) was developed; this method is based on the sequence alignment of protein but with reduced dataset.

The rest of the paper is organized as follows. First we describe testing datasets, representation of protein secondary structure, evaluation of prediction accuracy and theoretical basis in section 2. The hybrid method schema is presented in Section 3. Section 4 describes the initial population generation. Section 4, presents the genetic algorithm for the initial population optimization, and in Section 6 Knn algorithm for the final solution composition is presented. Section 7 shows the research results and its implications. Finally, a conclusion and the future work are presented in Section 8.

2. MATERIALS AND METHODS

2.1 Data-sets testing:
Three different datasets was used to test our novel method. These data sets are:

- Rost & Sander Bank: is a bank of the proteins secondary structures, built by Rost and Sander in their effort to predict the secondary structure, the bank contains approximately 131 structures of proteins but with less than 20% of homology [7].
- Cuff and Barton Bank: this dataset contain approximately 513 structures, this bank is mostly used in evaluating the secondary structure prediction methods [8].

- Protein Data Bank (PDB): The data bank of proteins of Research Collaborator for Structural Bioinformatics, named Proteins Dated Bank or (PDB), is a database of the three-dimensional structure (Structure 2D includes immediately). These structures are determined by crystallography with x-rays or spectroscopy RMN. These data are deposed in the PDB by biologists and biochemists of the whole world and belong to the domain public. Their consultation is free and can be done directly since the Web site of bank. The PDB contains approximately 167.729 Structures [9].

Table 1. Dataset ingredient

Composition / Dataset	Number of proteins	The proportion of α-helix (%)	The proportion of B-sheet (%)	The proportion of coil (%)
Rost and Sander	131	32.63%	20.68%	46.69%
Cuff and Barton	513	32.18%	24.58%	43.24%
PDB	167.729	34.36%	21.32%	44.32%

2.2 Representation of Protein Structures

There are several systems to present secondary structure assignment. DSSP (Define Secondary Structure of Proteins) [10], is the most broadly one, using secondary structure definition. With DSSP there are 8 different categories of secondary structure, these categories are: H (Alpha-helix), G (3-helix), I (5-helix), E (extended-strand), B (isolated-strand), T (turn), S (bend), and coil (''C''). The 8- structures classes were reduced then into 3 classes.

In this work we have used the frequently used reduction process as shown the table 2 below:

Table 2. DSSP categories

8 categories	3 categories
H, G, I	H
E, B	E
T, S, C	C

For the other protein presentation, FASTA format was used to represent the primary structure, each amino acids are represented with one letter, for the properties of amino acids a code of two letters ''H'' for hydrophilic and ''P'' for hydrophobic was used as well.

3. THEORETICAL BASIS

The exploration of knowledge is the extraction of useful knowledge starting from a great quantity of information; dataminig is used to increase the certainty and to reduce the costs. After the perfection of the production data processing, the ambitions of the companies increase for the use of data processing in decision-making process, dataminig was exploits in various applications in the process of the decision-making such as the reconfiguration of the offers of the products, to increase the sales, and to minimize the losses of errors or of frauds. In this research, three classification techniques were developed and used. The naive bays classification is an algorithm based on the theory of Thomas Bays (conditional probability) with high rate independency on the assumptions (Naive). The Bays classifier belongs to the family of the linear classifiers; it requires only few data to consider the parameters necessary to the classification [11].The genetic algorithms is one of the solutions of the combinative problems inspired from theory of the evolution of Darwin; he starts with a whole of solution at least true and applies the transformations to these solutions in order to improve them. By repeating these transformations, we obtain an approach solution [12]. The transformations are inspired from the biology, the mutation (transformation of an individual giving another individual), the crossing (combination of two individuals) and the selection (the probability of being relative of an individual of the following generation believes according to the performances of the individual for the starting problem).k-nearest neighbor is an algorithm of Supervised classification, we already have a learning base consists of a set of data, predicting the class adapts to a new entry with the method of K nearest neighbors, consists to take the K training nearest points to the new entry [13].

4. DESIGN OF THE SECONDARY STRUCTURE CLASSIFIER

Secondary structures prediction based on different arguments, and each method has its advantages and disadvantages. The method which has been used n this research, is divided into three parts; in each part we exploit a specific characteristic of existing methods.

- Exploit the physicochemical properties of the amino-acids.
- Exploit the conformational preference of the amino-acids.
- Exploit the proteins which the structure is given.

The principle of the research approach is to generate a set of solution with a naive Bays classifier, based on the structural probability of the amino-acids. This set of solution is to be considered as an initial population of a genetic algorithm, which will optimize this population with the physicochemical properties of the amino-acids, the principal sequence of proteins comprises homologous pieces with those of proteins of which the structure is available, the third part of work is to replace the fragment available with their existing solution, as shown in the following figure 1.

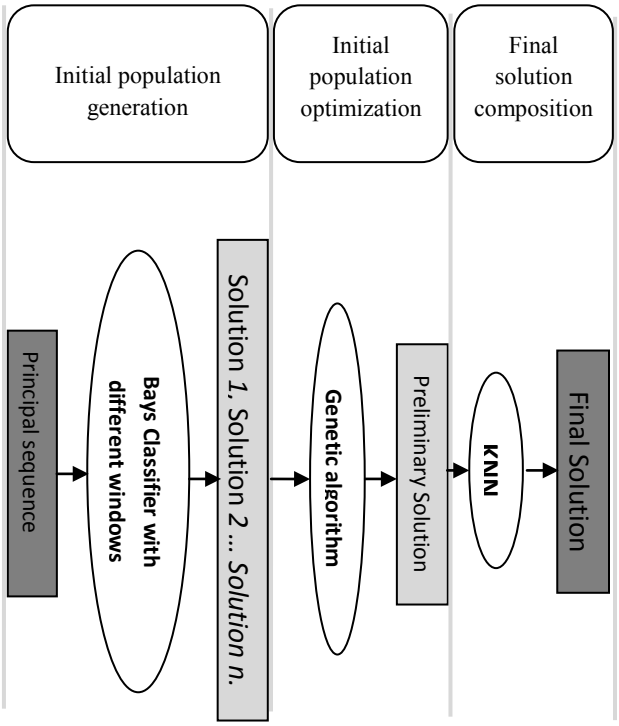


Fig 1: The new hybrid method schema.

5. THE INITIAL POPULATION GENERATION.

Each amino acid has its structural preferences; these preferences are represents by probabilities calculated from the data bank experiments. Bays classification is based on the idea that we can estimate the probability that an instance belongs to a class so this probability of that class corresponds at this instance; the classification is naive because all the probabilities are completely independent. This method is basically depends on making a slip window of the residues (amino-acids), and then the probability of folding up this window is calculating by the three secondary structure, the residue in the middle of the window as shown in figure 2 is represent the target residue, the folding up of this residue is the maximum probability between the three probabilities calculating based on the structural preferences probabilities of the window and the probabilities of the residue targets with the classifier.

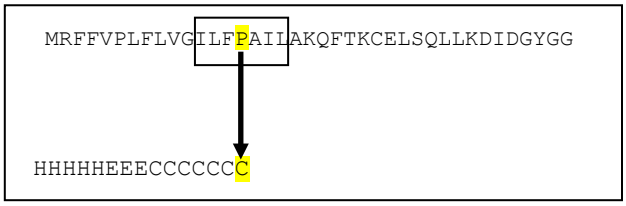


Fig 2: Bays secondary structure prediction.

The probabilities of the fragments represent by the window according to the three secondary structures are:

$$P\,(window/Helix) = \prod P\,(X_i/Helix)$$

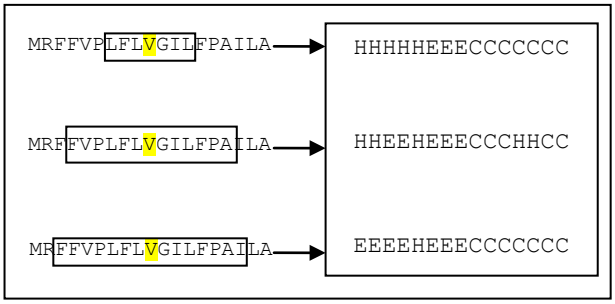
$$P\,(window/Sheet) = \prod P\,(X_i/Sheet)$$

$$P\,(window/Coil) = \prod P\,(X_i/Coil)$$

The folding up of residue represents by the structure corresponds to the maximum probability:

$$P\,(residue/Type) = \frac{P\,(Residue=Helixe)\;P\,(window/Helix)}{\sum P\,(window/Type)}$$

Window size impulses the quality of the prediction, some amino-acids have specific characteristic when they meets, a window of large size poses a problem to introduce amino-acids who do not have any relation with the residue targets, and a very small window is minimizing the interest of the target amino acids. For the benefit of all the sizes of the windows that are proposed in this research, the initial populations are generated with a different size of the window as shown in figure 3.



6. THE INITIAL POPULATION OPTIMIZATION.

The second part of this work acts to optimize the prediction with an evolutionary algorithm (genetic algorithm). Each amino-acid has a set of physic-chemic characteristic for interaction with the other amino-acids or for the interaction with the environment, these characteristic in creep folding up structural of protein. Hydrophobic amino acids are non-polar acids amines, the hydrophobic areas interact in a no covalent way with water, which leaves a strong freedom to make hydrogen bond between the various atoms of the hydrophobic area.

6.1 Presentation of the problem

The speed of a protein finds its conformation as remarkable. For example, a protein of 100 residues this means that the solution has 100 position and each position has 3 possible conformations for each residue (H, E, C), folding up then can adopt $\mathbf{3^{100}}$ different structures which is far from reality.

6.2 Coding

- The representation of each amino-acid of the principal sequence with their Hydrophobic or Hydrophobic characteristic is:

  HHHHHHHHHPPPPPPPPPPPHHHHHHHHHHPP.

- The initial population which is generated with the bays classifiers is already coded with the three state codes of secondary structures as:

  HHEEECHHHHHHCCCCCCCCCCCCCEEEEE.

6.3 Objective Function

The objective function in this research is to choose the individual who in the hydrophobic area are a helix or sheet (maximum probability) and the hydrophilic area is a coil (maximum probability) as shown in figure 4.

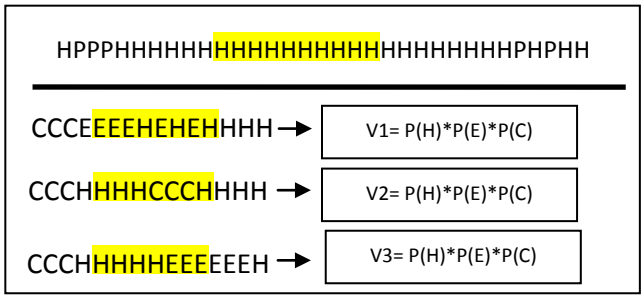


Fig 4: Objective function

$$\text{Fitness (Seqi)} = \prod i \prod j \prod k \, (P(Hk) * P(Ek) * P(Ck))$$

Or:

I: hydrophobic or hydrophilic region.

J: secondary structure type.

K: amino acids.

P (Xi): the probability of the amino acids "i" for the structure X.

The objective function is based on two parameters, the preference probability and the physicochemical property of each amino acid in the principal sequence, the principal advantage of this objective function is that use the proteins history in the same time with the biological properties of amino acids to improve the speed of solution research.

7. THE FINAL SOLUTION COMPOSITION.

Methods of prediction of protein secondary structure by homology seeks the nearest protein to the other protein, these two proteins resemble each other only in specific areas. In this work we develop a library of fragment from the protein data bank, a prediction of protein secondary structure starts with training according to preference probabilities and the physicochemical characteristic, and next we seek the fragments of proteins available in the fragments base, as shown in figure 5.

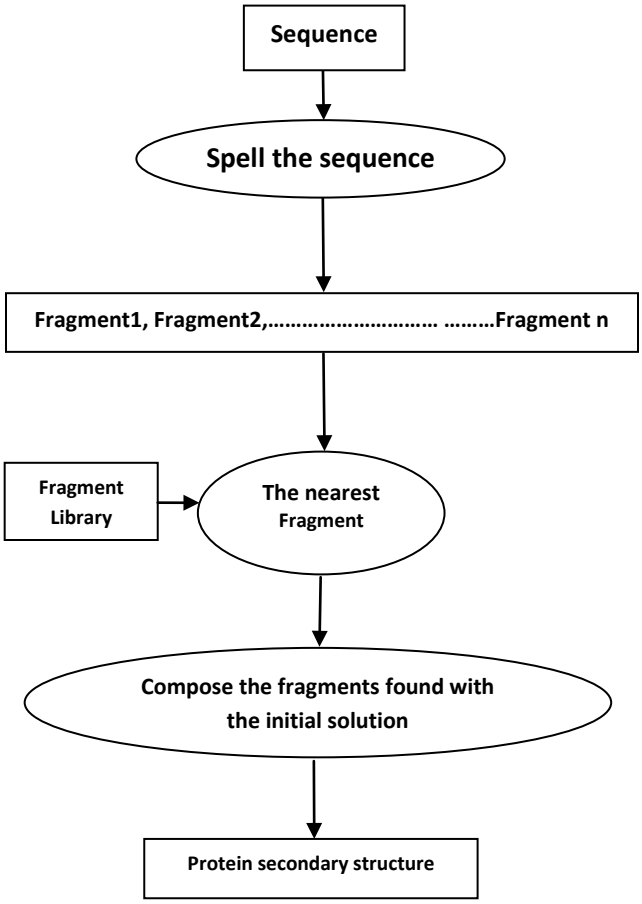


Fig 5: The Knn algorithm for the final solution

A protein fragment is a part of protein in which his amino-acids has a common characteristic, the researcher has uses the notion of the protein fragment in various techniques in the proteomics. The length of fragment plays a crucial role in the role of the whole protein, other researcher have used biological history to have a significant size of the fragments, but the recent studies concentrates on the construction of the variable size fragment, this fragment library was built from the three base protein structure different bases:

- Protein Data Bank.
- Rosta and stone Bank.

  Cuff and Barton Bank.

This library is based on the frequent fragments, but we test fragments randomly, since testing the entire fragment is a combinative problem. The construction of the base is passes by the following stages:

- Break each protein in a set of pieces of fixed size (length=15).

- For each piece calculates a score of occurrence.
- Preserve of the fragment whose score are superior to a NR (NR: number given).

8. RESULTS AND EVALUATIONS.

The objective of this research is to use the historic of the proteins secondary structure given by experimental methods and the evolution of the data mining technique. By using Rost and stander data set, results show that Q3 accuracy reached up to 85.41%. And by Cuff and Barton dataset set, Q3 reached up to 85.58%, and for SOV99 accuracy increased to 85.06%. The best result obtained by reducing dataset (PDB), to more than 160 000 sequences in this bank, the Q3 has reached 85.89%. The results found to be superior to the other methods of protein secondary structure prediction. The method has attaint up to 89.59% of accuracy with Cuff and Barton dataset for the helix secondary structure, the accuracy of all secondary structures with the dataset used is shown in table 3 and table 4.

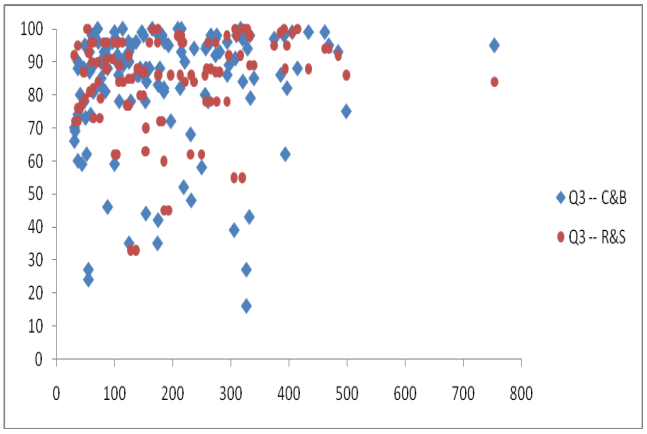


Fig 6: Dot chart illustrating the distribution of accuracy (Q3 %) of the hybrid technique as measured on the (Rost and Sander) and (Cuff and Barton) datasets.

Server	Q3	QH	QE	QC
Psipred	79,95%	83,52%	73,81%	82,54%
PHD Expert	77,61%	79,92%	74,42%	78,53%
SSPRO	79,07%	82,12%	66,9%	82,26%
SAM	78,17%	83,99%	75,58%	74,96%
Predator	80,04%	78,3%	75,98%	85,87%
The hybrid method	85,58%	89,59%	80,19%	86,96%

Table 3. The accuracy of the hybrid method with different protein prediction servers with identical dataset Rost and Sander

Table 4. The accuracy of the hybrid method with different protein prediction servers with identical dataset Cuff and Barton.

Server	Q3	QH	QE	QC
Psipred	79,99%	84.35%	72.62%	83.01%
PHD Expert	76,53%	78.33%	73.65 %	77.62%
SSPRO	75,36%	80.84 %	64.38 %	80.85%
SAM	79,13%	84.93 %	77.11 %	75.36%
Predator	78,31%	79.71%	69.38 %	85.85%
The hybrid method	85.41%	85.55%	84.81%	85.87%

The hybrid prediction method achieves the best classification accuracy when applied to predict the secondary structures of proteins in the Rost and Sander dataset, Cuff and Barton dataset and the PDB dataset. The comparison of our method with the best protein secondary structure servers (table 3, table 4) indicated that the hybrid system have improved the best accuracy with 4.4%.

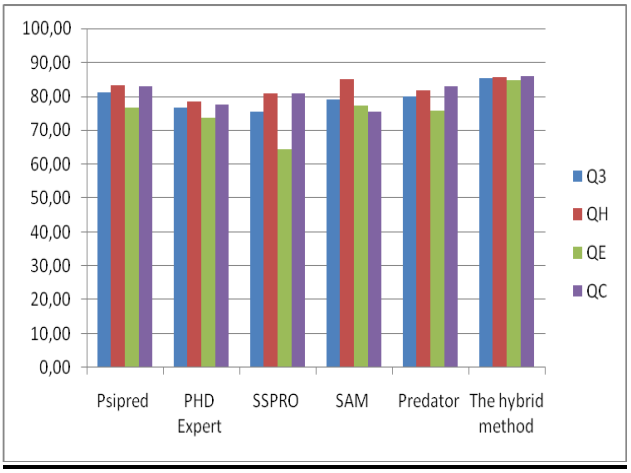


Fig 7: Histogram illustrating the prediction accuracy (Q3, Q Helix, Q sheet, Q Coil) of the best protein secondary structure server result and the hybrid method result with identical data set (Cuff and Barton).

9. CONCLUSION AND FUTURE WORK

In this research we have present several classifiers to propose a novel method for the problem of protein secondary structure prediction. The use of data mining technique and the historic of the experimental technique are showed that the interest of the hybrid systems to solve the bioinformatics problems. In this work, we have used different dataset with low homogeneity such as Rost and sander, Cuff and Barton, and The PDB Bank to

make our experiments, and for method evaluation we have used the Q3 accuracy. According to the result mentioned, we concluded that the hybrid prediction method has improved the prediction accuracy by its complementary layers.

Finally we have plan to continue our research in the following domains:

- Develop an open-access web service for the hybrid method.
- Use the result of hybrid method to improve the tertiary structure prediction.
- Use a heuristic technique to construct the fragments library.
- Parallelization of the two layer of the hybrid method to improve the prediction speed.

10. REFERENCES

[1] Baldi P, S. Brunak, P. Frasconi, G. Pollastri, and G. Soda. Bidirectional Dynamics for Protein Secondary Structure Prediction, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, Sweden (1999)

[2] Bernstein FC, Koetzle TF, Williams GJ, Meyer Jr EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol;112:535-542 (1977)

[3] Chou PY, Fasman GD. "Prediction of protein conformation". Biochemistry 13 (2): 222–245 (1974).

[4] Donald Voet Judith-G Voet Biochimie. 2e édition. *De boeck*

[5] Hall P, Park BU, Samworth RJ. Choice of neighbor order in nearest-neighbor classification". Annals of Statistics 36 (5): 2135–2152. doi:10.1214/07-AOS537 (2008).

[6] Hua, S. J., & Sun, Z. R. A novel method of protein secondary structureprediction with high segment overlap measure: Support vector machine approach. Journal of Molecular Biology, 308(2), 397–407 (2001).

[7] Jean-Jacques Boreux, Eric Parent, Jacques Bernier Pratique du calcul bayésien; Springer ; (2004).

[8] Kabsch W, Sander C.»Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". Biopolymers 22 (12): 2577–637. doi:10.1002/bip.360221211. PMID 6667333. (1983).

[9] Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., et al. Combining local-structure, foldrecognition, and new-fold methods for protein structure prediction. Proteins, 53, 491–496 (2003).

[10] Cuff, J. A., & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins, 40(3), 502–511.

[11] Poli, W. B. Langdon et N. F. McPhee, A Field Guide to Genetic Programming, Lulu.com, ISBN 978-1-4092-0073-4) (2008).

[12] Robert .D et Vian B. Element de biologie cellulaire. Doin, (2008).

[13] Rost B, Sander, C., Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584-599 (1993).