A Model for Protein Secondary Structure Prediction Meta - Classifiers

Anita Wasilewska
Department of Computer Science, Stony Brook University,
Stony Brook, NY, USA
e-mail: anita@cs.sunysb.edu

***Abstract*—We present here a mathematical model for the Protein Secondary Structure Prediction (PSSP) problems and research. It also represents un effort to build a uniform foundations for PSSP research. The model, and hence the paper, is designed to facilitate and speed up understanding of the long standing PSSP research and its problems also for people who want to get involved in it. We present an abstract definition of a protein and its structures and discuss the Protein Data Banks, and other Proteomic Data Bases as well as three generations of PSSP algorithms and servers (all of them web-accessible). We also discuss the development of most important results, problems and methods of data preparation for PSSP classifiers. Finally, we describe a model for a Meta-Classifier utilizing all, or a subset of PSSP servers and discuss its relationship with the first ever developed, Bayes Network Meta-Classifiers of [13] based on 4 to 6 servers.**

I. Introduction

Techniques for the prediction of protein secondary structure provide information that is useful both in ab initio structure prediction and as an additional constraint for fold-recognition algorithms. Knowledge of secondary structure alone can help the design of site-directed or deletion mutants that will not destroy the native protein structure. However, for all these applications it is essential that the secondary structure prediction be accurate, or at least that, the reliability for each residue can be assessed. Due to the improvement of protein information in databases and use of evolutionary information, todays predictive accuracy is about 80%. In the PPSP research improvement of predictive accuracy by even 1% is considered an important result. The Bayes network based meta (multi)-classifiers presented in [13] improved the best known predictive accuracy by up to 1.21% over the best known classifiers. This result justifies the importance of Meta-Classifiers based approach. The paper is organized as follows. In section II we define a formal mathematical model for four levels of protein structure and use them to formulate a symbolic definition od a protein (section 2.6). In sections III, IV, and V we discuss the most modern PSSP data sets, servers, and first, second and third generation of PSSP algorithms and classifiers, respectively. Finally we describe, in section VII a construction of a Meta Classifier dataset and discuss advantages of building the Meta Classifiers applications based on descriptive data mining algorithms, as opposed to purely statistical methods.

II. Four Levels of Protein Structure; a Protein Model

The name Protein comes from the Greek word PROTEUO which means "to be first (in rank or influence)". Proteins make up about 15% of the mass of the average person. Medicine, Agriculture and Industry benefit the most from protein research. The design of drugs which inhibit specific enzyme targets for therapeutic purposes (engineering of insulin) is an example of an application of protein research in Medicine. In Agriculture, development of new treatments of plant diseases, growth modification and other improvements of crops, in Industry, the synthesis of enzymes to carry out industrial processes on a mass scale, are only few other examples.

A. Protein Primary Structure

Any Protein consists of four levels of Protein Structure: primary, secondary, tertiary, and quaternary. The basic component of all of them, and hence of a protein, are 20 amino-acids:

Alanine (A), Cysteine (C), Aspartic Acid (D), Glutamic Acid (E), Phenylalanine (F), Glicine (G), Histidine (H), Isoleucine (I),Lycine (K), Leucine (L), Methionine (M), AsparagiNe (N), Proline (P), Glutamine (Q), ARginine (R), Serine (S), Threonine (T), Valine (V),Tryptophan (W), and Tyrosine (Y).

All aminoacids have represented by by fixed symbols: , C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S,T, V, W, Y. We denote by **A** the set containing all aminoacids, as represented by their symbols, i.e. $\mathbf{A} = \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. Protein Primary Structure consists of sequences of sequences of aminoacids. The sequences in a sequence of sequences are called sub- units. We define it symbolically as follows.

Definition 2.1: Let **A** set of symbols denoting all aminoacids and $\mathbf{A}^*$ be the set of all finite sequences formed out of elements of A. Elements $\mathbf{A}^*$ of are denoted by x, y, z, with indices if necessary. Any $x \in \mathbf{A}^*$ is called a *protein primary sub- units structure*. Any $x_1, x_2, ...x_n \in (\mathbf{A}^*)^*$ is called a *protein primary structure*.

B. Protein Secondary Structure PSS

Protein secondary structure PSS is a term protein chemist give to the arrangement of the peptide backbone in space. It is produced by hydrogen bindings between aminoacids. The assignment of the PSS categories (hydrogen bindings)

to the experimentally determined three-dimensional (3D) structure of proteins is a non-trivial process and is typically performed by widely used DSSP program ([5]). PSS consists of protein sequence (sub-unit) and its hydrogen bonding patterns, called *secondary structure (SS) categories*. There are 8 different secondary structure categories (*SS categories (states, residues)*) determined by DSSP program: α-helix denoted by H, ε_{10} denoted by G, π - helix (extremely rare) denoted by I, β - strand denoted by E, β - bridge denoted by B, β - turn denoted by T, bend denoted by B, and L denotes the rest of the bindings.

Databases for protein sequences (primary structures) and its sub-units are expanding rapidly due to the genome sequencing projects and the gap between the number of known protein secondary structures and the number of known protein sequences in public domain is widening rapidly. PSSP (Protein Secondary Structure Prediction) research is trying to breach this gap. Known, experimentally determined PSS data is stored in a Protein Data Bank (PDB), widely web-accessible (http://www.rcsb.org/pdb/). Other Proteomic Data Bases are discussed in section III.

PSS prediction methods are normally trained and assessed for only 3 states (residues): H (helix), E (strands) and L (coil) instead of the 8 SS categories. Standard 8-to-3 reduction methods are defined by programs DSSD, STRIDE, and DEFINE [5]), and others.

Improvement of predictive accuracy of different PSSP (Protein Secondary Structure Prediction) programs depends on the choice of the reduction method.

For example, the most difficult to predict 8-to-3 reduction method used in CASP (International Contests for PSSP programs)is the DSSP program method: H(helix) = { G (ε_{10}), H (α -helix)}, E (strands) = { E (β -strand), B (β-bridge)}, L (coil) - all the rest. We write is as a shorthand:

$$E, B \Rightarrow E; \quad G, H \Rightarrow H; \quad Rest => C.$$

Some other methods are as follows.

STRIDE program method: H(helix) as in DSSP, E (strands) = { E (β -strand), B(β -bridge)}, L (coil) - all the rest.

DEFINE program method: H(helix) as in DSSP E (strands) = { E (β -strand)}, L (coil) - all the rest.

Method B: $E \Rightarrow E; H \Rightarrow H; Rest \Rightarrow L; \ EE, HHHH \Rightarrow L$.

Method C: $GGGHHHH \Rightarrow HHHHHHH; \ B, GGG \Rightarrow L; \ H \Rightarrow H; \ E \Rightarrow E$.

Typical PSSP Data is gathered in a form of protein sub-units (sequences) and assigned to them sequences of SS categories as observed empirically in their 3-dimensional structures and assigned DSSP program. Then the 8-to-3 reduction to $\{H, I, L\}$ methods are performed by DSSP, or other programs. The Data is always presented as a pair: protein sequence (sub-unit) and a sequence if its observed SS categories, for example: Protein sub-unit:

$$KELVLALYDYQEKSPREVTHKKGDILTLLN$$

Observed sequence of SS categories $\{H, I, L\}$ after the DSSP reduction:

$$HHHHHLLLLEEEHHHLLLEEEEEELLLHHH$$

Any of such pairs is called Protein Secondary Structure for a given sub-unit.

PSSP goal is to build a tool (a classifier) that on input of a given sub-unit (or protein primary structure) would return a sequence its (predicted) SS categories, and hence determine the Protein Secondary Structure for a given sub-unit.

C. Protein Secondary Structure Formal Definition

Given the set $\mathbf{A}$ of symbols denoting aminoacids and a protein sequence (sub-unit) $x \in \mathbf{A}^*$. Let $\mathbf{S}$ be a set of all secondary structure categories (residues). In particular $\mathbf{S}_3 = \{H, E, L\} \subseteq \mathbf{S}$ is the set of symbols of 3 states (residues): H (helix), E (strands) and L (coil). Let $\mathbf{S}^*, \mathbf{S}_3^*$ be the set of all finite sequences of elements of $\mathbf{S}, \mathbf{S}_3$, respectively. We denote elements of $\mathbf{S}^*$ by e, o, with indices if necessary i.e. we write $e \in \mathbf{S}^*, e1, e2 \in \mathbf{S}^*$ etc

Definition 2.2: Any partial, one to one function $\mathbf{f} : \mathbf{A}^* \xrightarrow{1-1} \mathbf{S}^*, \ i.e. \ \mathbf{f} \subseteq \mathbf{A}^* \times \mathbf{S}^*$ is called a protein *sub-unit secondary structure identification function* and any element $(x, e) \in \mathbf{f}$ is a called the *protein sub- unit secondary structure.*

The element e of $(x, e) \in \mathbf{f}$ is called secondary structure residues (states, categories) sequence and often for short, a protein sub-unit secondary structure. We extend definition 2.2 of the protein sub-unit secondary structure to the definition of protein (sequence) secondary structure as follows.

Definition 2.3: A protein secondary structure identification function is any a partial, one to one function $\mathbf{F} : (\mathbf{A}^*)^* \xrightarrow{1-1} \mathbf{S}^*$ defined as follows. For any $x_1, x_2, ...x_n \in (\mathbf{A}^*)^*$, $x_1, x_2, ...x_n \in Dom\mathbf{F}$ if and only if $\forall(1 \leq i \leq n) \ (x_i \in Dom\mathbf{f})$, where $\mathbf{f}$ is the protein sub-unit secondary structure identification function (definition 2.2), and $\mathbf{F}(x_1, x_2, ...x_n) = \mathbf{f}(x_1), \mathbf{f}(x_2), ...\mathbf{f}(x_n)$.

Any Data Set (DS) of sub- units used in PSS Prediction defines its own identification function $\mathbf{f}_{DS}$ empirically and by DSSP program.

We identify any given Data Set DS with $\mathbf{f}_{DS}$ and write $DS = \{\begin{pmatrix} x \\ e \end{pmatrix} : \ x \in \mathbf{A}^* \cap e \in \mathbf{S}^* \ \cap \ \mathbf{f}_{DS}(x) = e\}$.

For example: if the Data Set DS is such that a protein sub -unit $ARNVSTVVLA$ has the observed SS category sequence $HHHEEECCCH$ we put : $\mathbf{f}_{DS}(ARNVSTVVLA) = HHHEEECCCH$ and write

$$\begin{pmatrix} ARNVSTVVLA \\ HHHEEECCCH \end{pmatrix} \in DS.$$

D. Protein Tertiary and Quaternary Structures

The tertiary structure of a protein is the arrangement in space of all its atoms. The overall 3D shape of a protein molecule is a compromise, where the structure has the best balance of attractive and repulsive forces between different

regions of the molecule. For a given protein we can experimentally determine its tertiary structure by X-rays or NMR. Given the tertiary structure of a protein we extract its secondary structure with the DSSP program. We define symbolically the tertiary structure as follows.

Definition 2.4: Let $s \in (\mathbf{A}^*)^*$ be a protein sequence (primary structure), $\mathbf{T} = (s, e) \in \mathbf{F}$ be the secondary structure of s (definition 2.3) , the element

$$\varphi x = (\mathbf{T}, \mathbf{t}_s)$$

is a *protein tertiary structure* of $s \in (\mathbf{A}^*)^*$. $\mathbf{t}_s$ is the sequences s tertiary *folding function*.

Many globular proteins are made up of several polypeptide chains called sub-units, stuck to each other by a variety of attractive forces but rarely by covalent bonds. Protein chemists describe this as quaternary structure. We define it symbolically as follows.

Definition 2.5: Protein quaternary structure is a pair

$$(\mathbf{Q}, \mathbf{F}_Q)$$

where $\mathbf{Q}$ is a multi-set of tertiary structures, specific for different proteins, (for example $\mathbf{Q} = [\alpha, \alpha, \beta, \beta]$ in a haemoglobin) and $\mathbf{F}_Q$ is the *quaternary folding function.*

E. Protein: Symbolic Definition

A protein $\mathbf{P}$ is build out of protein sequences, protein sub-units, their secondary structures, their tertiary structure, and their quaternary structure.

Definition 2.6: We define a protein $\mathbf{P}$ as follows.
$\mathbf{P} = \{x_1, x_2, ...x_n;\ (x_1, e_1), (x_2, e_2), ..., (x_n, e_n); \alpha_{x_1} = ((x_1, e_1), \mathbf{t}_{x_1}); \alpha_{x_2} = ((x_2, e_2), \mathbf{t}_{x_2}); ...;$
$\alpha_{x_n} = ((x_n, e_n), \mathbf{t}_{x_n}); ([\alpha_{x_1}, \alpha_{x_2}..., \alpha_{x_n}]; \mathbf{F}_{\alpha_{x_1}...\alpha_{x_n}}\}$,
where x_i is protein $\mathbf{P}$ i-th sub-unit, $\mathbf{t}_{x_i}$ is x_i's tertiary folding function, and $\mathbf{F}_{\alpha_{x_1}...\alpha_{x_n}}$ is protein $\mathbf{P}$ quaternary folding function

In PSSP research we deal with protein sub-units x_i, not with the whole protein sequence (primary structures). We write $\mathbf{P}_{x_i}$ when we refer only to the sub-unit x_i of the protein $\mathbf{P}$. We write $\mathbf{P}_(x_i, e_i)$ when we refer to the sub-unit x_i and its secondary structure. We write $\mathbf{P}_{\alpha_{x_i}}$ when we refer to the sub-unit x_i of the protein $\mathbf{P}$and its secondary structure and its tertiary structure.

For example, for a given protein
$\mathbf{P} = \{x_1, x_2, ...x_n;\ (x_1, e_1), (x_2, e_2)...., (x_n, e_n); \alpha_{x_1} = ((x_1, e_1), \mathbf{t}_{x_1}); \alpha_{x_2} = ((x_2, e_2), \mathbf{t}_{x_2}); ...;$
$\alpha_{x_n} = ((x_n, e_n), \mathbf{t}_{x_n}); ([\alpha_{x_1}, \alpha_{x_2}..., \alpha_{x_n}]; \mathbf{F}_{\alpha_{x_1}...\alpha_{x_n}}, \}$ we have that
$\mathbf{P}_{x_i} = x_i$, $\mathbf{P}_(x_i, e_i) = \{x_i, (x_i, e_i)\}$, and
$\mathbf{P}_{\alpha_{x_i}} = \{x_i, (x_i, e_i), \alpha_{x_i} = ((x_i, e_i), \mathbf{t}_{x_i})\}$.

Example: Haemoglobin $\mathbf{H}$ is defined as follows.

$\mathbf{H} = \{x, y; (x, e_x), (y, e_y); \alpha = ((x, e_x), \mathbf{t}_x);$
$\beta = ((y, e_y), \mathbf{t}_y);\ ([\alpha, \alpha, \beta, \beta],)\}$.

III. Proteomic Databases and Data Sets

The most important proteomic databases are: SWISS-PROT + TrEMBL , PIR-PSD, PIR-NREF, and PDB. They are all web-accessible and their short description follows. The SWISS-PROT (http://us.expasy.org/sprot/) is a protein sub-units and protein sequences database with high level of annotations, a minimal level of redundancy and high level of integration with other databases. It contains 124464 entries. The TrEMBL is a computer-annotated supplement of SWISS-PROT that contains all sequence entries not yet integrated in Swiss-Prot. It contains 28210 entries. The PIR-PSD - Protein Information Resource (http://pir.georgetown.edu/) was founded in 1960 by Margaret Dayhoff. It is a comprehensive and annotated protein sequence database in the public domain with 283308 entries. The PIR-NREF, full name PIR-Non-Redundant Reference Protein Database (http://pir.georgetown.edu/) contains all sequences in PIR-PSD, SwissProt, TrEMBL, RefSeq, GenPept, and PDB with a total of 1,186,271 entries. It is mostly used for finding protein profiles with PSI-BLAST program. This Database is mandatory in Protein Secondary Structure Prediction (PSSP) research. The PDB- Protein Data Bank (http://www.rcsb.org/pdb/) contains 3-D biological macromolecular structure data. In 2003 it contained already 20747 Structures. All PSSP data sets start with some PDB sequences (sub-units) with known secondary structures. Then DSSP program provides the secondary structure and its reduction to three categories. This becomes, after transformation (section VI, a learning data for PSSP algorithms. Other very important Datasets are the following. RS126 - a historic, original set of 513 sequences by Rost and Sander [9], currently corresponds to a total of 23,363 entries, CB513 -it contains 513 sub-units with known secondary structure selected by Cuff and Barton in 1999 ([2]). It is one of the most used datasets in PSSP research, HS1771 - it contains 1771 sequences is a family of datasets is formed out of a non redundant PDB (Protein Data Bank) subsets ([4]). The newest one is EVA. It contains 6 novel test sets EVA1, ..., EVA6. They are provided by the datasets available from the real-time evaluation experiment called EVA [11], which compares a number of prediction servers on a regular basis using the sequences deposited in the PDB every week. A lot of authors has their own and "secret" datasets.

IV. PSSP Algorithms

There are three generations in PSSP algorithms. First Generation was based on statistical information of single aminoacids. The most relevant algorithms are Chow-Fasman (1974) and GOR (1978). Both algorithms claimed 74-78% of predictive accuracy, but tested with better constructed datasets were proved to have the predictive accuracy 50% (Nishikawa, 1983). Second Generation algorithms are based on windows (segments) of aminoacids. Typically a window contains 11-21 aminoacids. The main problem they faced was that their predictive accuracy was $< 70\%$ and predicted SS category chains (sequences) were usually too short what lead to the

difficulties with the use of predictions. Third Generation is based on the use of windows on evolutionary information. Use of evolutionary information is the following: 1. Scan a database with known sequences with alignment methods for finding similar sequences 2. Filter the previous list with a threshold to identify the most significant sequences 3. Build aminoacid exchange profiles based on the probable homologs (most significant sequences) 4. The profiles are used in the prediction. The first third generation algorithm is the PHD algorithm developed by Rost and Sander in 1993 ([9]). It is based on multilevel Neural Networks. Many of the second generation algorithms have been updated to third generation. The most important algorithms of today, besides the PHD, are: PREDATOR([8]) based on Nearest-neighbour Classification, DSC [6], NNSSP [12], ZPRED [14] and MULPRED [1]. Due to the improvement of protein information in databases i.e. better evolutionary information, today's predictive accuracy is on and above 80%. It is believed that maximum reachable accuracy is 88%.

V. PSSP Internet Servers

We list here two of 9 secondary structure prediction servers available on the Internet, as they were used and published in detail [13].

JPred ([2]) 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.

Six different prediction algorithms used are: DSC [6], PHD [9], NNSSP [12], PREDATOR [8], ZPRED [14] and MULPRED [1] are then run, and results from each method are combined into a simple file format.

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

SSPro [7] 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). 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 (back propagation through structure) algorithm, an extension of BPTT (back propagation through time), used in unidirectional recurrent neural networks.

A set of 11 bidirectional recurrent neural networks is trained on the data set. 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.

PSIPRED [7] 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.

Prof server [10] 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 differents classifiers it has been used GOR formalism-based methods extended by linear and quadratic discriminations and neural network-based methods [10]. 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.

The SAM-T02 [4] 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.

VI. PSSP Data Preparation

Public Protein Data Sets used in PSSP research contain protein secondary structure sequences. In order to use classification algorithms we must transform secondary structure sequences into classification data tables. In PSSP literature the records in the transformed classification data tables are called *instances.* The mechanism used in this transformation process is called *a window*. A window algorithm has a secondary structure as input and returns a classification table. Its records consist of aminoasids instances (subsequences of the sub-units) and theirs assigned SS categories acting as classification attribute. The window mechanism produces very large datasets . For example, window of size 13 applied to the CB513 dataset of 513 protein sub-units produces about 70,000 records.

VII. PSSP Meta- Classifiers Data Construction

Combining the predictions of a set of classifiers has shown to be an effective way to create composite classifiers that are more accurate than any of the component classifiers. The process of creation of a PSSP Meta- Classifier is as follow.

Select some subset $\mathbf{D_m} = \{D_1, D_2, .., D_m\}$ of the Public Protein Data Set (section III). Choose which datasets from $\mathbf{D}$ you want to use for training and which for testing. For example, in [13] 9 datasets were selected and only one, the HS1771 was used for training and the rest for testing. Select a subset $\mathbf{S_k} = S_1, S_2, ..., S_k$, $1 \leq k \leq 9$, of the the set $\mathbf{S}$ of all 9 servers (section V) available on the Internet. For given $\mathbf{D_m}$ and $\mathbf{S_k}$ construct a meta-classifier data set **MCDS** for training and testing as follows. Submit transformed data (section VI) to all of the $\mathbf{S_k}$ web servers and wait for their replies. These replies came as either web pages or e-mail messages. Process them once they have been received. Extract the prediction for the secondary structure of each aminoacid instance (record) from the body of the message or from the contents of the web page. Store extracted results of the prediction for the secondary structure from each of $\mathbf{S_k}$.

The new **MCDS** data set to be processed by the multi-classifier $\mathbf{MC}_k$, based on the training dataset $TDS \in \mathbf{D_m}$ and of classifiers $\mathbf{S_k}$ is obtained as follows. Attributes are all elements of $\mathbf{S_k}$ i.e. $S_1, S_2, ..., S_k$, classification (decision) attribute is $\mathbf{C}$. The values of an attribute S_i are the stored, extracted results of the prediction; i'e. usually elements of $\{H, E, L\}$. The values of the classification attribute are the actual values of the secondary structure. For each of the aminoacids of the protein sub-units represented by the instance of the training data set $TDS \in \mathbf{D_m}$ we insert all predictions from corresponding servers, followed by the actual value of its secondary structure as the $\mathbf{MC}_k$ record.

The choice of servers is essential to the final results obtained by trained and tested meta classifiers. The experiments conducted in [13] involved 4, 5, and 6 "hand selected" servers. In total of 7 Bayesian network classifiers were built and in the end the one with the best results was chosen.

Observe that we deal with a large amount of data (section VI). The 9 datasets available for training and testing provide a really (about 1,000,000 records) large, already well prepared, standardized, and publicly available set of data to experiment with.

Moreover, the meta classifiers data do not contain missing values, and other then the statistical methods can be used, unlike in the case of PSSP classifiers. Hence it is natural to explore the non-statistical, descriptive methods. Additionally, these experiments could form a basis for a well founded research into comparison of statistical and non-statistical approaches.

We refer interested readers to [13] for detailed methods of evaluating obtained results.

REFERENCES

[1] Barton G, Taylor W. *Prediction of protein secondary structure and active sites using the alignment of homologous sequences*, J Mol Biol, 1988;195, pp. 957- 961.

[2] Cuff J, Barton G. *Evaluation and improvement of multiple seqience methods for protein secondary prediction*, Proteins, 1999;34, pp. 508 -519.

[3] Frishman D, Argos P. *75% accuracy in protein secondary structure prediction*, Proteins, 1997;27, pp. 329-335.

[4] Hobohm U, SCharf M, Schneider R, Sander C. *Selection of a Representaive Set of structures from the Brookhaven Protein Data Bank*, Protein Sci, 1992;1, pp. 409-417.

[5] Kabsch W, Sander C. *Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features*, Biopolymers, 1983;22, pp. 2577-25637.

[6] King R, Sternberg M. *Indentification and application of concepts important for accurate and reliable protein secondary structure prediction*, Protein Sci, 1996;5, pp. 2298-22310.

[7] Cuff J, Clamp M, Siddiqui A, Finlay M, Barton G. *JPRED: A concensus secondary prediction server*, Bioinformatics, 1998;25(14), pp. 892-893.

[8] Frishman D, Argos P. *75% accuracy in protein secondary structure prediction*, Proteins, 1997;27, pp. 329-335.

[9] Rost B, Sander C, Schneider R. *PHD: an automatic mail server for protein seconadary structure prediction*, Comput Appl Biosci, 1994;10, pp. 53-60.

[10] Rost B, Sander C. *Prediction of protein secondaty structure at better then 70% accuracy*, J Mol Biol, 1993; 232, pp. 584-599.

[11] Rost B, Eyrich V. *EVA: large scale analysis of secondary structure prediction*, Proteins, 2001;5, pp. 192-199.

[12] Salamov A, Solovyev V. *Prediction of proteoin secondarystructure by combining nearest- neighbor algorithms and multiple sequence alighment*, J Mol Biol, 1995;247, pp. 5-11.

[13] Victor Robles, Pedro Larranaga, Jose M. Pena, Ernestina Menasalvas, Maria S. Perez, Vanessa Herves, Anita Wasilewska. *Bayesian Network Multi-classifiers for Protein Secondary Structure Prediction*, Artificial Intelligence in Medicine, 2004; 31, pp. 117 - 136.

[14] 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, pp. 957-961.