Protein Secondary Structure Prediction: a Literature Review with Focus on Machine Learning Approaches

Leila Khalatbari [a*], Mohammad Reza Kangavari [b]

[a] *Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin,Iran*
[b] *Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran*

Abstract

DNA sequence, containing all genetic traits is not a functional entity. Instead, it is transferred to protein sequences by transcription and translation processes. This protein sequence takes on a 3D structure later, which is a functional unit and can manage biological interactions using the information encoded in DNA. Every life process one can figure is undertaken by proteins with specific functional responsibilities. Consequently protein function prediction is a momentous task in bioinformatics. Protein function can be elucidated from its structure. Protein secondary structure prediction has attracted great attention since it's the input feature of many bioinformatics problems. The variety of proposed computational methods for protein secondary structure prediction is very extensive. Nevertheless they couldn't achieve much due to the existing obstacles such as abstruse protein data patterns, noise, class imbalance and high dimensionality of encoding schemes of amino acid sequences. With the advent of machine learning and later ensemble approaches, a considerable elevation was made. In order to reach a meaningful conclusion about the strength, bottlenecks and limitations of what have been done in this research area, a review of the literature will be of great benefit. Such review is advantageous not only to wrap what has been accomplished by far but also to cast light for the future decisions about the potential and unseen solutions to this area. Consequently in this paper it's aimed to review different computational approaches for protein secondary structure prediction with the focus on machine learning methods, addressing different parts of the problem's area.

Keywords: Protein secondary structure prediction, Machine learning, Neural Networks, Support Vector Machines, Ensemble methods.

4. Introduction

Proteins are building blocks and functional units of living organisms. They play main role in biological interplay. Their extensive functions include metabolism, DNA replication/ modification, transcription/ translation, intracellular signaling, cell-cell communication, protein folding/ degradation, transport, defense and immunity functions, storage, coordinated motions, mechanical support, regulation, generation and transmission of nerve pulse and any other life process one might figure [1].

To answer any question about running changes and procedures in organisms' body, protein functions must be known and tracked. This knowledge helps with early diagnosis of disease and drug and enzyme design. Protein is composed of a sequence of 20 different amino acid molecules and its function is directly and tightly dependant on the structure these molecules adopt. Each amino acid molecule (also called residue) owns its unique and special properties and features. Hence the combination of these molecules with no limitation on their distribution and length along proteins' sequence can create infinite number of proteins with different capabilities and functionalities [2].

There is a four level hierarchy considered for protein structure. Protein's first structure is the very linear sequence of its amino acids. The secondary structure of protein is formed by local compositions between neighboring amino acids through peptide bounds. There are three main secondary structures namely α-helix, β-sheets and coils. As the result of such composition different parts of this chain are exposed by other parts causing different kinds of forces such as repulsion, attraction, hydrophobic,

* Corresponding author. Email: leila.khalatbari@gmail.com

hydrophilic. These forces along with different types of bonds such as and hydrogen bounds and disulfide bridges make protein take on a quite stabilized 3D structure called protein tertiary structure. At the highest level of hierarchy, lies Quaternary structure which describes how several polypeptide chains come together to form a more complex functional protein. Like tertiary structure, quaternary structure is determined by ionic and hydrophobic interactions between amino acids [2].

Since protein function is strongly connected to its structure, it can be determined from 3D structure. However it will be far too challenging of a task. Here secondary structure comes to simplify this task as an intermediate step. This is not the only benefit of Secondary structure prediction. It's also the input feature for many other bioinformatics tasks. There are two main groups of methods to determine secondary structure namely experimental methods and computational methods. The experimental methods were in use before the advent of computational methods and include X-ray crystallography, electron microscopy and nuclear magnetic resonance [2]. The drawbacks of these methods are that they're very time consuming (from several months to years to determine one structure), costly (Up to thousands of dollars) and not applicable to all proteins. On the other hand protein sequencing acceleration made a huge gap between known sequences with undetermined structures. This gap enforced the development of more rapid and yet accurate methods and this was the birth point of computational methods.

The early methods in this research area go back to 1970's. There are reasons why the performance of literature's methods still needs elevation and revision. Major reasons include obscure patterns existing in protein data, noise [3], class imbalance and high dimensionality [4] imposed by using encoding matrixes to convert polypeptide sequence to numerical meaningful vectors. Consequently the variety of computational proposed methods is very broad. Different approaches have been put forward to address different aspects of high complexity of the problem area. To this, some approaches have worked on extracting assorted features and encoding schemes and other preprocessing procedures. Most of the literature's body however is dedicated to creation of statistical models or providing different layers of learning strategies. All kinds of proximity measures, kernels, and multi stage learning schemes are the results of such approaches. An assortment of post processing and refinement strategies has also been developed to enhance the accuracy of problem's solutions.

Therefore a classification and summarization of methods put forth to this problem will be of great benefit since it provides a comprehensive view to the whole extensive literature history. As the result of such study, it will be more feasible to discover the strengths, limitations and bottlenecks of various applied strategies. It can also provide a comparative framework and identify best performing components, employed.

To pursue this end, in this research it is aimed to provide a review of protein secondary structure prediction strategies with the focus on machine learning approaches. Hence the subsequent sections are arranged as following. Section 2 provides an overview and classification of the existing strategies. Section 3 gives details of proposed approaches belonging to each category of methods. Section 4 introduces and describes the available datasets and servers for protein secondary structure prediction. Section 5 provides final conclusion.

5. Classification and Outline of Protein Secondary Structure Prediction Computational Methods

Since early 1970's, solutions have been put forward to the problem of PSSP. Later on, the advent of machine learning approaches and afterwards ensemble methods made remarkable progress in the field. The most frequently used computational approaches explained earlier, are as follow. (i) Information theory-based methods (ii), Hidden Markov models, (iii) Support Vector Machines, (iv) Neural Networks, (v) Distance-based algorithms, (vi) Association Rule Mining and tree based methods, (vii) Methods exploiting feature generation and compound features and (viii) ensemble methods.

Table 1 provides a summary of studied approaches of the literature segregated on the category they belong to. The table is very advantages to obtain a brisk view of the categories of methods, their strength and bottlenecks, their abridged description, frequency of each category employment in PSSP and the distribution of each category's methods along the research timeline. The details of each category of methods are provided in section 3.

Table 1
Abridged description of literature methods

Method Category		Year [Reference]	Authors and Description
Information Theory and Bayes Theory	**Advantage:** Calculation of parameters is dataset-based and straightforward. Clearly identifies what is taken into account and what's neglected for prediction. **Disadvantage:** Sensitivity to sample volume and class imbalance.	[1974][5] [1978][6] [2015][7]	Chou et al. calculate the propensity of each amino acid to form a secondary structure using statistics derived from empirical studies [5]. Carnier et al. calculate propensity of each amino acid to form a secondary structure considering its neighbors[6]. Rithvik et al. carry out a comparative study between DSSP, GORIV and GOR V methods[7].
Support Vector Machines	**Advantage:** Demonstrates great performance usually outperforming neural networks due to its optimization problem solving intrinsic. **Disadvantage:** Sensitivity to kernel parameters choice.	[2012][13] [2011][14] [2003][15] [2010][16] [2006][24]	Zangooei et al. employ a regression based method, SVR, with a fused kernel function [13]. Zangooei et al. employed an SVM based method with fused kernel [14]. MN Nguyen et al. proposed 3 binary SVM classifier and two multiclass SVM based which solve one single optimization problem [15]. Liyu et al. proposed a Two layer multi-SVM with bagging for resampling [16]. He et al. generate a training set from SVM output to train generate rules from decision tree [24].
Neural Networks	**Advantage:** Adaptable architecture composing of neural synopsis capable of complex modeling **Disadvantage:** Prone to over fitting and poor generalization if parameters and architecture are not selected optimally.	[2008][17] [2010][18] [2012][19] [2012][20] [2014][21] [2014][22]	Wang et al. Proposed extreme learning machine with probability based combination method to combine final results and a helix filtering [17]. Babaei et al. applied a combination of multi-layer bidirectional recurrent neural network and modular reciprocal recurrent network based on pruned multi-layer perceptron [18] [19]. Alirezaee et al. combine the prediction of four feed forward neural network and tree-based classifier and sampling o address class imbalance problem [20]. Johal et al. perform a comparative study amongst feed forward neural network, three binary one-versus-one and three binary one-versus-all SVM classifiers [21]. Dinubhai et al. Trains a three-layer feed forward perception using conjugate gradient minimization algorithm and numerical extracted features to predict PSSP [22].
Hidden Markov Models	**Advantage:** Conditional involvement of dependency of each residue to its neighboring residues' structure. **Disadvantage:** The types of prior distributions that can be placed on hidden states are severely limited.	[2014][8]	Agarwal et al. generate sequence encoding scheme using Markov Model of third order and feed them to SVM for structure prediction [8].

Method Category		Year [Reference]	Authors and Description
Association Rule Mining	**Advantage:** Highly interpretable **Disadvantage:** Missing relations emerged less frequently in data	[2014][12]	Mosses et al. mine an FS-Tree with a modified relative support from data to produce sequence structure mapping by generating rules from the tree [12].
Distance-based algorithms	**Advantage:** Simplicity along with capability of modeling complex decision functions. **Disadvantage:** Poor generalization if the distance measure won't take dissimilarity comprehensively and from different aspects of sample features.	[2008][10] [2012][11]	Gosh et al. provide a comparative study of PSSP using three distance based classifier namely minimum distance, K-nearest neighbor and fuzzy K-nearest neighbor fed with a matrix based sequence representation [10]. Liu et al. proposed NN-DM which is a KNN classifier working with LZ-based distance measure [11].
Feature extraction and compound feature generation	**Advantage:** Focusing on more relevant features of data related to the problem at hand and avoiding irrelevant ones. **Disadvantage:** Information loss, selectivity and sensitivity of the predictor to the extracted features.	[2014][23]	Yaseen et al. provide new encoding scheme based on pseudo-potentials as context-based features and later feed them to a two-layer feed forward neural network [23].
Ensemble methods	**Advantage:** Far higher capability to address high pattern complexity involving ambiguity and uncertainty by taking advantage various components' strength. **Disadvantage:** Does not exhibit better performance compared to single predictors if selected components are not complementary and diverse. Is sensitive to combination rules.	[2008][17] [2010][18] [2012][19] [2012][20] [2011][14] [2006][24] [2003][15] [2010][16] [2014][26] [2015][25]	A group of ensemble approaches combine similarity or distance measures or make use of various methods other than classification such as feature extraction, preprocessing, post processing, parameter optimization and filtering to enhance prediction. [17], [14], [24], [20], [26] Another group of ensemble methods employ same classifiers with different features and tunings. [15], [16], [18], [19], [20] Some other ensemble approaches combine various kernel functions. [14], [24] A major category of ensemble strategies combine different classifier. [24] Patel et al. derive a knowledge base from proteomic sequence-structure database, does the prediction based on the knowledge based and then refines the predictions results using a backpropagation neural network. [26] Bouzianeh et al. focus on combination rules. Puts two single members of multi-class SVM and feed forward RBF neural network into an ensemble framework and investigates the performance of various weighted pooling combination rules. [25]

Next chapter provides more profound details of each method.

6. Description of Each Category of Methods

In this section the approaches of each group of methods belonging to table 1 will be discussed in details.

3.1 Information Theory and Bayes Based Methods

The methods of this group were amongst the earliest computational approaches put forth for the problem of secondary structure prediction. Since these methods work on the basis of conditional probabilities and statistical

parameters driven from sequence data, they are sensitive to dataset volume and class imbalance.

The lowest accuracy and performance of these in comparison with machine learning approaches and ensemble of statistical methods and machine learning based methods is a proof to such claim.

Two of the pioneers and most well-known approaches in this category are GOR and Chou-Fesman. GOR is a developed version of the simpler Chou-Fesman method. Like Chou-Fesman, GOR works using probability parameters derived from the empirical studies of known protein structures obtained by X-ray crystallography. Unlike Chou-Fesman, besides the propensity of a single amino acid to take on a specific secondary structure [5], GOR also takes into account the probability of an amino acid to adopt a secondary structure given that its immediate neighbors have already formed that structure [6][7]. So it works on a Bayesian basis.

3.2 Hidden Markov models

Hidden Markov models have been successfully applied in the aforesaid area. The reason behind its success is the conditional involvement of the dependency of each residue to its neighboring residues' structure in adoption a certain structure. This is usually achieved by employment of high order Markov models. Since the secondary structure adoption of each residue is strongly dependant to the neighbor's residue and their structure, this rationale perfectly suites the nature of the problem and consequently leads to better results and performance.

In pursuance of this mean, In [8] three hidden Markov models are derived for representation and encoding of sequences. Derived Markov models are of third order which means the occurrence of each state in a sequence depends on the occurrences of three previous states. As a base statistic, the probability of an amino acid 's' followed by 't' for the class 'I', is calculated by the division of frequency of 'st' belonging to class 'i' to the sum of 'st' frequencies in sequences for all classes. Transition probabilities are calculated using the frequency of residues at each sequence using a sliding window. Obtaining the calculated transition matrix, input vectors for input and output actual data are prepared. These vectors are then used to train three binary one-against-all SVM with RBF kernel.

3.3 Distance-based algorithms

Distance based approaches encountered in the literature as well. The advantage of this group is the capability of class labels determination of most complicated samples quite easily provided that there isn't huge variety amongst sample features represented by feature vectors. It's worth mentioning that the algorithm is pretty sensitive on the choice of distance measure. If a proper proximity measure adaptive to the local features of a problem is developed, then these group of methods exhibit acceptable performance while still regarded quite simple approaches to employ. The most well-known, studied and vastly used methods among such approaches is K-nearest neighbor algorithm and its other extensive variety of versions. The main reason behind its popularity except for the ones mentioned earlier is its power to model arbitrary and complicated boundaries. Many other classifiers such as rule-based and tree-based classifiers can only model rectilinear class boundaries [9]. The usage of distance-based methods in PSSP literature is described in detail in what follows.

In [10] Three distance-based classifiers namely minimum distance with square Euclidean distance measure, k-nearest neighbor and fuzzy k-nearest neighbor classifiers are applied to the problem of PSSP. The logic behind their employment is not having the prerequisite of presumption a special distribution for the data unlike some other methods. Since in most cases, the presumed distribution does not fit the real one with acceptable precision. To overcome the limitation of minimum distance classifier in separating non-linear data with acceptable accuracy, k-nearest neighbor is used as well. To beat the limitation of k-nearest neighbor to assign different classes the same levels of importance, fuzzy k-nearest neighbor classifier is involved. The encoding scheme to turn symbolic sequences into numeric vectors is a 20*w matrix. The rows of the matrix indicate 20 different amino acids existing in nature. The columns however indicate the amino acids lying in the sliding window frame. All vertical elements are zero but the element which matches the amino acid in window center. In the end the performance of stated classifiers are compared. In [11] NN-CDM is proposed for identifying protein structural classes. This approach passes over the feature extraction stage and uses the amino acid sequences straightly as input data. The reason why they don't make use of any encoding schemes is that performance of methods which extract features from original data is severely connected to the sensitivity of selected features and many related features to secondary or structural class might be lost. Further on, they employ a distance measure based on Lempel-Ziv complexity measure which is known to efficiently identify repeated patterns in a sequence. Ultimately the input sequential data and complexity based distance measure are passed to KNN algorithm for final classification.

3.4 Association Rule Mining

The proposed approaches on the basis of trees and association rules are observed less frequently than other groups of approaches. These methods perform well in capturing dependencies amongst data. In other hand the high comprehensibility of the output rules. Such rules not only fulfill the classification task, but also provide an insight about the data and patterns. These rules can play the role of a knowledge-based and can be employed along with other machine learning approaches to enhance the prediction results. The following rule-based approaches have been put forth for PSSP in the literature:

In [12] a variation of FP-Tree called FS-Tree has been developed. The prefix FS-Tree has been adapted to the intrinsic sequential nature of the problem and can reflect sequential subsequences occurring in a sequential database together with their frequencies instead of frequent item sets from transactional databases. The concept of support in FP-Tree has been altered as well since the probability of observing long sequences in database is less than that of short ones and consequently the support will decrease with the increase in tree level. To remedy that, the new support for FS-Tree structure called relative support for each sequence is calculated considering only the set of all sequences with same length. Here confidence demonstrates the preference of the residue sequence to a specific structural sequence. A rule is a frequent subsequence pair (*P,Q*) for which also a confidence threshold, 'minconf', holds. The confidence of a frequent pair depends on the frequency in which *P* occurs as antecedent. In order to obtain the confidence of a frequent pair the tree is searched in depth first and the frequencies of nodes with antecedent *P* are totalized, extracting rules only if 'minconf' is exceeded. In [9] DT is used in fusion with SVM. In this approach, first SVM is applied to the data of classification task and then a new training set is chosen from the output of SVM as the input data for DT. Consequently the noisy and outlier data are removed. Here SVM plays the role of a filter applied to the original unclean data. Hence the advantage of superior generalization capability of SVM and remarkable comprehensibility of DT are both exploited. Extracted rules from DT are later exploited to secondary structure prediction.

3.5 Support Vector Machines

The major group of machine learning approaches applied for PSSP is SVM and its numerous variations. They have demonstrated a great performance due to their optimization intrinsic. However they are quite sensitive to the choice of their kernel functions and need systematic parameter setting (kernel parameters and error constant) to show their best performance.

In [13] an SVM-based approach specialized to regression, called SVR is applied. In order to optimize the parameters such as mapping points of transforming continuous numerical data into integers and kernel function parameters, non-dominated sorting genetic algorithm (NSGA||) is employed. One-against-one strategy is exploited to build a multi-class classifier. Furthermore, to enhance the prediction performance fusion of three kernels by dynamic weighting technique is utilized. Also in [14], an SVM with a fused kernel function composed of three RBF, polynomial and linear kernels is applied. These kernels are merged using a dynamic weighting strategy assigning a weight to each single kernel in the final kernel equation based on its performance. In [15] five previously studied methods and two newly proposed methods have been discussed. The previously studied methods are based on single-stage SVM and the newly proposed ones are two-stage SVM based methods developed combining the single-stage ones. Single-stage methods contain OAO, OAA, DAG, VW and CS. The decision of secondary structures is made by respectively voting, discriminant function maximum value, DAG and winner-take-all rules. The input of the second stage is the output vector of discriminant function from the first stage. The second layers are the same as the first and the combination rules for the final output are also the same. The tenet of using a two-stage method is to incorporate the information of the neighboring residues' structure on which the structure of the central residue depends. In [16] an ensemble of five SVM classifiers are used. Each individual SVM is of One-Against-All type whose prediction is gained using winner-takes-all method. Bagging is used to resample training data and assign one dataset to each classifier. The final prediction of the ensemble system is defined by majority voting.

3.6 Neural Networks

Neural networks are amongst the first groups of machine learning approaches put forth to address PSSP. Their strength is the adaptable architecture composing of neural synopsis and neurons which can manage complex modeling. However it's prone to over fitting and poor generalization if the parameters and architecture is not attempted to be optimized. PSIPRED and JPRED are two NN-based servers for PSSP; Ever since various types of neural networks have been commonly employed. Neural networks such as single and multilayer feed forward, recurrent, bidirectional and reciprocal are come upon in the literature.

In [17], a four-phase procedure is applied. In the first phase called preprocessing, a novel encoding scheme based on base vectors and probability information of amino acid residues appearing in different structures and different amino acid residues appearing in the same structure is obtained. In the second phase, a fast learning single-hidden layer feed forward neural network, called ELM is fed by getting the previously made vectors as its inputs. The best generalization performance of ELM is obtained through applying 5 fold cross validation. To combine the outcome predictions of binary classifiers, a probability base method is employed. According to this method, the structure of each of the 8 possible state of the prediction outcomes is determined, calculating the four TP_i, TN_i, FP_i, FN_i ratios for each of the three secondary structures i and finding the maximum value amongst them. The last phase called helix post processing, refines the final obtained predictions according to the biological rule that obligates each helix segment to have at least four residues. In [18], advanced kinds of recurrent neural networks are devised. In the final approach a multi-layer bidirectional recurrent neural network (MBR-NN) and a modular reciprocal recurrent neural network (MRR-NN) are combined. The tenet behind this combination is inspired by cognitive process of human perception, cognition and the capability to restore processed information to use it later for discovery in novel

situations; modeling the human cognition, the MBR-NN component is employed in order to capture the neighboring effects of amino acids along the protein chain in structure adoption. In other words, it takes into account the long range interactions of amino acids. The MRR-NN module instead, takes into account the correlations between secondary structure elements. In [19] their previous work of MRR-NN and MBR-NN is extended to involve more previous and next states (m more) into context virtual memory and refine prediction results more efficiently. They have also proposed another alternative network for m>7 in which they imitate the bidirectional architecture of the feed forward sub network in reciprocal module. The recurrent links of the reciprocal hidden layers create the virtual memory to save the adjacent information of the reciprocal inputs. As a result, the global relational data in accordance with the local neighboring effect enhance the prediction performance. Back propagation algorithm is used in training phase and updating weights. In [20] an ensemble of four feed forward neural networks are employed. Each network consists of 260 input neurons proportional to sliding window size, one hidden layer of 25 nodes and 3 output nodes. To solve the class imbalance problem over sampling, under sampling and tree-based tertiary classifier are applied. Additionally, three combination methods termed majority voting, weighted majority voting and genetically weighted majority voting are utilized. In [21] a comparative study has been performed to judge the performance of neural networks against SVM in PSSP. To proceed, sliding windows along with PSSM profiles are used as the encoding schemes. For each machine 6 binary classifiers consist of 3 one-against-one and 3 one-against-all are trained and compared later. The kernel function used for SVM is Gaussian. The neural network employed are 3 layer feed forward networks with resilient back propagation training algorithm. The experiments indicate that the neural networks perform much better and more efficiently in time than SVM. In [22] the objective is to evaluate the performance of neural network in PSSP. To pursue this objective three layer neural network with one hidden layer is employed. It's claimed that amongst all various NN structure, a three layer neural network with minimum number of nodes achieves a good performance. The training algorithm is gradient descent. Three groups of features have been extracted from primary sequences namely Composition, Transition and Distribution.

3.7 Feature extraction and compound feature

Some approaches of PSSP have focused on exploiting various features and statistical data and then feed a machine such as a neural network or SVM with the generated vectors. The significance of these methods lies in the fact that there are no computational method independent from the input feature vectors and the choice of these elements drastically influence the performance of all methods. As the result many studies have been dedicated to extract various types of feature from protein sequence data. A fraction of these methods have a more biological basis and other part of them has a mathematical and computational basis. A computational approach in this regard has been described in below.

In [23] the statistics measuring the favorability of a residue to adopt a certain secondary structure is incorporated as features to address interdependency among secondary structures of neighboring residues. In pursuance of this end, the statistics of singlets (Ri), doublet (RiRi+k), and triplet (RiRi+k1Ri+k2) are derived and used later to calculate pseudo-potentials of a residue adopting a certain secondary structure using a mean force approach. These sequence-structure statistics along with PSSM are exploited to feed and train a two-layer feed forward neural network in three phases. In the first phase called sequence-to-structure, context-base features and PSSM values are given as the input vectors and the secondary structure of the central residue in a sliding window of fixed size is predicted. The second phase conducts a structure-to-structure training to eliminate impossible secondary structures. The last phase carries out a refinement procedure. It modifies context-base features by setting them to absolute favorable if the results obtained from the second phase indicates that the probability of a residue taking on a certain secondary structure is above 90%. Then the modified context-based features are used in a similar manner to the first phase.

3.8 Ensemble Methods

As pointed out earlier, the very high complexity involved in protein data and its prediction problem demands a complex solution. Such solution consists of various components, each of which capable of addressing one issue and obstacle of the problem. The emergence of ensemble approaches was based on such rationale and demand.

Most of the approaches discussed in previous sessions of the literature are recruited in an ensemble framework. Ensemble frameworks are developed in different manners. Figures 1 to 4 illustrate the outline of these schemes. The common main component of all these frameworks is an aggregator. This component is responsible for the aggregation of the divergence outputs of each single component.

Figure 1 illustrates the first ensemble scheme which is a compound of several classifiers of different types such as SVM, NN, KNN, ..etc in a parallel flow. In the end the results of all single classifiers are aggregated into one single decision by aggregator. In [24] the ensemble proposed approach follows such scheme.

The next scheme depicted in figure 2 is a compound of classifiers of different types, this time in a sequential flow. It means that each classifier processes the output of the previous classifier until the expected and acceptable outcome is achieved. The proposed approaches in [24] and [25], explained previously adopt such ensemble framework.

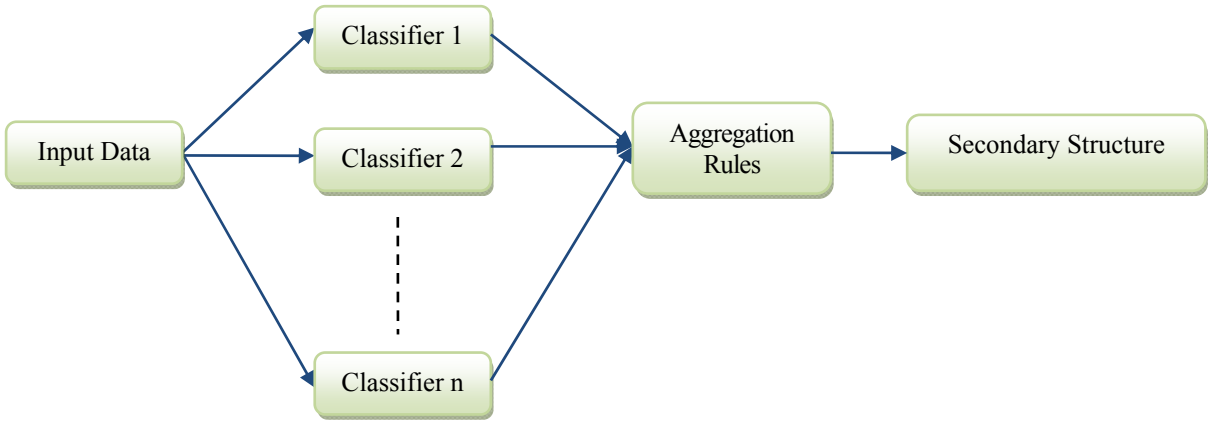

Fig. 1. Ensemble framework consisting of parallel classifiers of different types

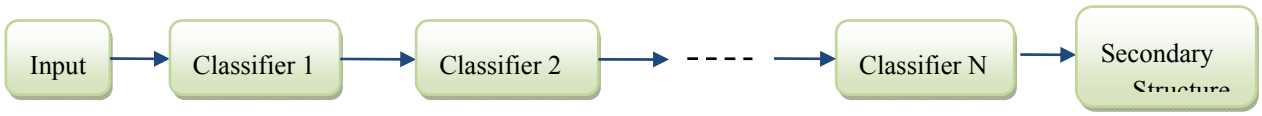

Fig. 2. Ensemble framework consisting of sequential classifiers of different types

Another ensemble scheme is the one in which there exists classifiers of the same type but with different parameter tunings, different proximity measures and various kernel functions. Figure 3 presents such scheme. In [14], [15], [16], [18], [19], [20], [24] this ensemble framework has been exploited.

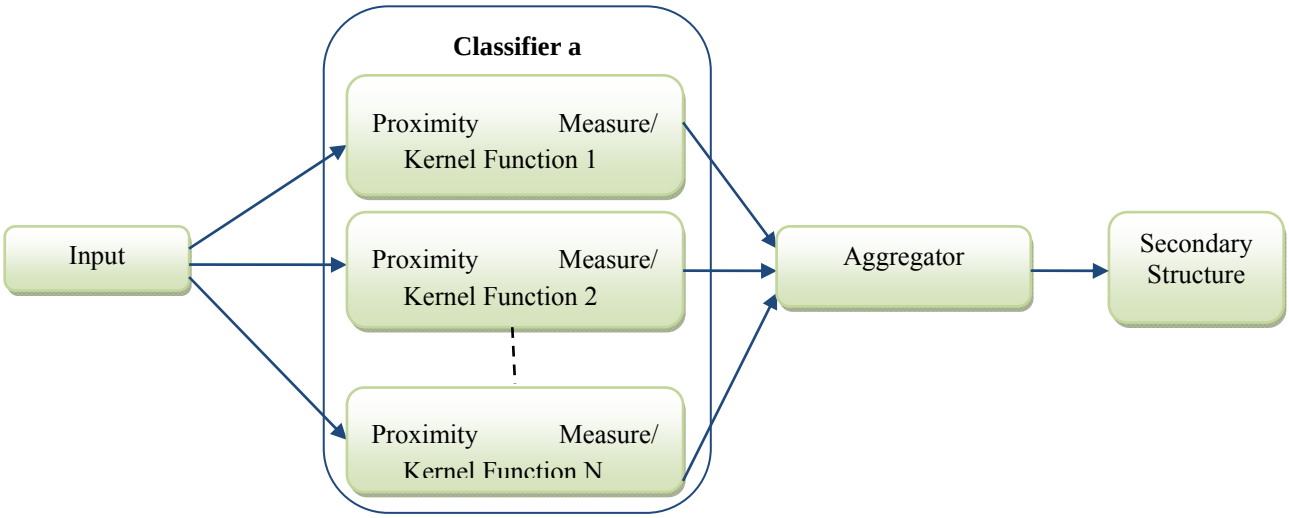

Fig. 3. Ensemble framework consisting of multiple classifiers of the same type but different parameter tunings proximity measures

One other ensemble framework is developed based upon a variety of methods other than classification in order to achieve better performance such as feature extraction, preprocessing, post processing, parameter optimization. Figure 4 displays this scheme. Proposed methods in [14], [17], [20], [24] previously explained, have similar manner as in below figure.

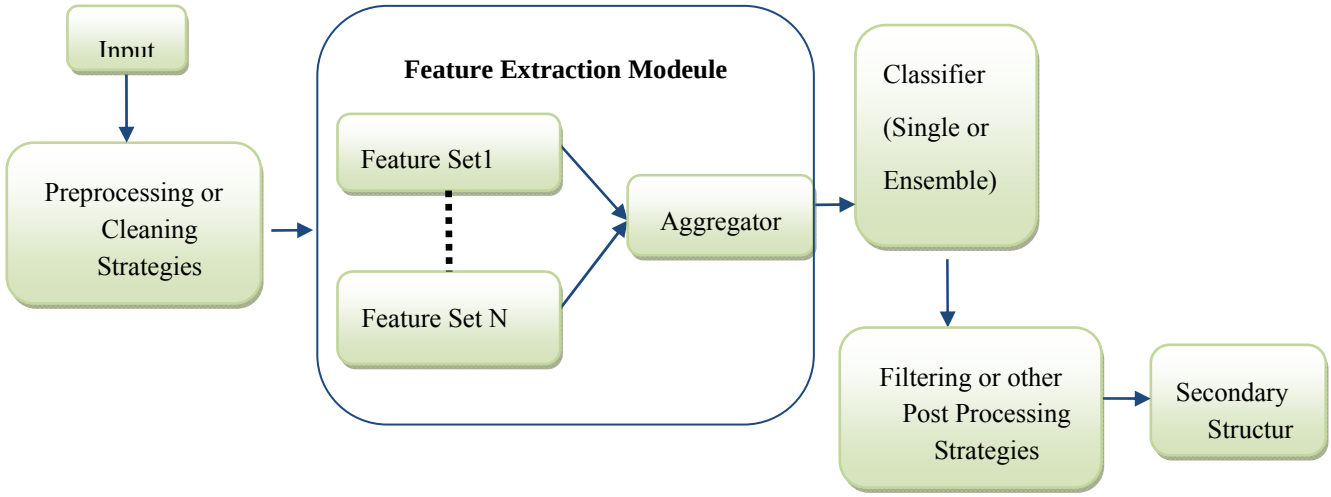

Fig. 4. Ensemble framework consisting of other components other than classifier for further enhancement.

The classification component is not necessarily a machine learning algorithm. It might be a knowledge based driven from sequence statistics as in [26] in which an ensemble of a knowledge-base and neural network has been employed for PSSP. First a knowledge-base is composed using five amino acid long words and their three middle secondary structures along with their count in the training set. Then the prediction is done through two validation processes using the knowledge-base. In the first validation round the three central secondary structure of five AA (amino acid) long word are compared in a 7 AA long window length. If the last two secondary structures of the first word and first two secondary structures of the second word are the same and also the last two secondary structures of the second word and the first two secondary structures of the third word are the same, then a five secondary structure word called W_i is generated. Then these collections of words W are compared for a window of size 9 and the secondary structure of the central amino acid is the more frequent structure occurred at that position in the words of W. Neural network is further used to refine the prediction results of knowledge-base since the less frequent words wouldn't be reflected properly. The exploited network consists of three layers. The input layer takes the binary encoded five AA words and their predicted secondary structures and consists of 40 neurons, 8 for each amino acid. The hidden layer consists of 24 neurons and the output consists of three neurons for each secondary structure. Comparably small hidden layer architecture is used to avoid over fitting and a pruning strategy is applied to remove dormant connections. The refinement process causes a 10-15 % improvement in prediction results.

In [25] an ensemble of neural networks and support vector machines are employed. Different combination schemes are then exploited to aggregate the prediction results of each ensemble member. The neural network ensemble member is one multi-layer perceptron trained using back propagation algorithm with sigmoid activation function. There are four multi-class SVM members each of which solves one optimization problem using quadratic programming. The prediction results of these members later are combined using sum, mean, product, max, min, weighted pooling, decision templates and Dempster-Shafer theory of evidence.

Ensemble methods have proven more efficient than single component methods in PSSP problem.

6. Materials

In this section the most common, popular and vastly used secondary structure prediction datasets will be listed along with their properties and features. Then a list of DSSPs used in different researches is provided. Most well-known PSSP programs and servers are introduced as well. Also the evaluation measures used in the literature will be defined. In the end a comparative view of the literature's methods performance in PSSP is offered.

4.1 Datasets

Table 2 shows the most vastly used datasets in protein secondary prediction problem. The table provides the dataset name along with their description. The reference number of the methods which exploited each data set is included. It's especially beneficial since it shows the popularity and power of dataset in presenting more genuine results of an approach. Also the more popular dataset which are used with more methods can provide a broader comparative framework and consequently it enables the researchers to present the strengths of their method more comprehensively.

Table 2

Common and vastly used datasets for protein secondary structure prediction

Dataset	Description
RS126	Contains 126 non-homologous proteins which mean no two protein sequences share more than 25% sequence similarity. Average protein length in RS126 is 185. There are 23347 amino acids, 32% of which have alpha helix structure, 23% having beta sheet structure and 45% have coil structure. It was first used by Rost and Sander.
CB513	Consists of 513 protein chains, 117 of which are from Rost and Sander's non-redundant proteins and 396 sequences are from the CB396 dataset by Cuff and Barton. No sequences in the dataset share more than 25% sequence identity. It's created by Cuff and Barton.
CB396	Consists of 396 non-homologous proteins with 621184 residues created by Cuff and Barton.
PSIPRED	First used by Jones for PSIPRED model. It contains 3 training and testing set pairs. There are no overlap between protein chains of test sets and train sets. Three test sets are composed of 62, 62 and 63 proteins. The average number of protein chains in three training sets is 1100 and the average number of residues is 215000.
Manesh	Consists of 215 non-homologous proteins with no more than 25% pair0wise sequence identity and 50682 residues used in experiment of Manesh.
CASP9	Contains 203 protein chains with 23298 residues used in CASP9 experiment.
Carugo338	Consists of 338 non-homologous monomeric protein crystal structures extracted from protein data bank in which no pair of sequences, share more than 25% sequence identity.
Barton	Consists of 502 non-homologous protein chains with more than 83000 residues and less than 25% homology. It is generated by Cuff and Barton.
Astral30	Contains 3344 protein chains filtered at 30% sequence identity.
Cull1764	Comprised of 1,764 sequences with a total of 417,978 amino acids with 25% maximum sequence identity. The sequences length is in the range of 30-3000 amino acids.

Figur 5 illustrates the frequency of dataset usage in the studied literature of this review. As seen in figure 5, the most popular and employed dataset is RS126. As the result, taking advantage of this data set gives the opportunity of

having a broad range of methods for comparison. CB512 and SPIRED are the next two frequently used datasets according to figure 5.

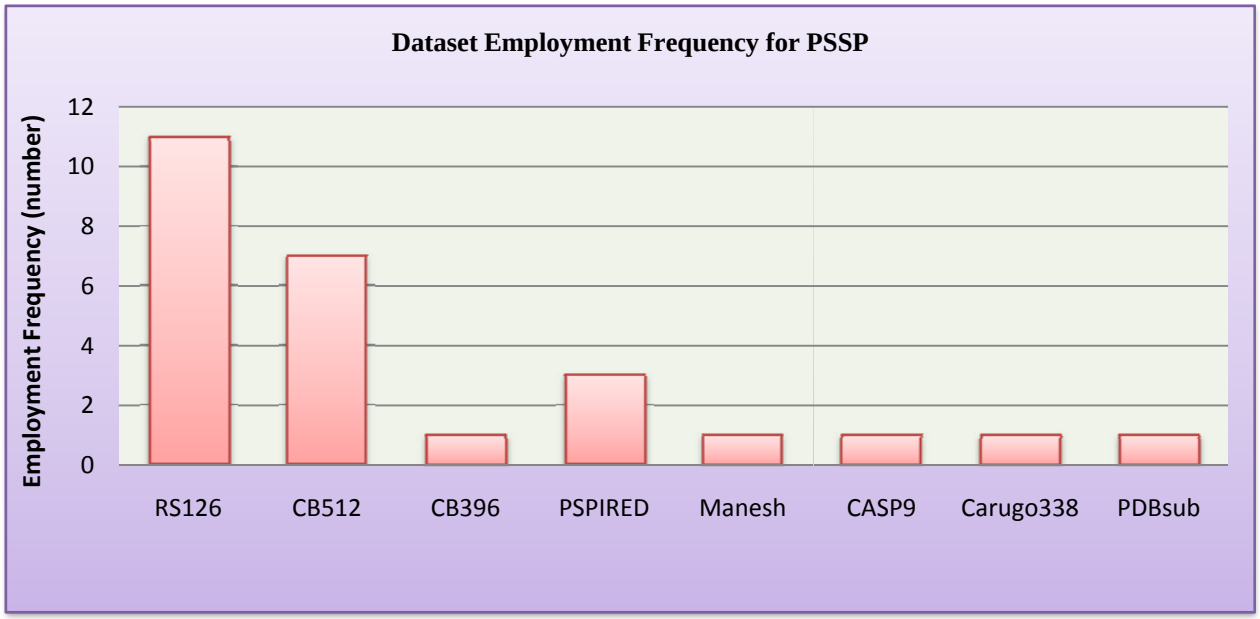

Fig. 5. employment frequency of table 2 datasets in the studied literature of PSSP

4.2 Protein Secondary Structure Prediction Programs and Servers

Table 3 presents the PSSP programs and servers. These programs are the softwares and servers developed from the best performing methods of the literature and are usually benchmarks for comparisons and evaluations. There are different versions of some of these programs which provide more advanced and improved solutions.

Table 3

Well known and benchmark software and servers for protein secondary structure prediction

Protein Secondary Structure Prediction Program and Servers	Description [Reference]
GOR	Information Theory/Baysian based [27]
SPIDER	Iterative Deep Neural Network [28]
JPRED	Neural Network based [29]
PREDATOR	Knowledge based Database Comparison [30]
PSPIRED	Feed Forward Neural Network based [31]
YASSPP	Cascaded SVM-based [32]
SSpro	Homology Analysis based [33]
SABLE	Consensus classifier as an ensemble framework [34]
SAM	Hidden Markov Model based [35]
Porter	Based on ensemble of bidirectional recurrent neural networks [36]

4.3 Conversion Rules for 8 State to 3 State Secondary Structure Translation

In fact there are 8 secondary structures of protein available. These structure include H (alpha-helix), G (3-helix or 310-helix), I (5-helix or p-helix), B (residue in isolated beta-bridges), E (extended sheet), T (hydrogen bond turn), S (bend) and "-" (any other structure). Typically these 8 structures are reduced to three main classes which are of more practice and use. The three reduced structures include helix (H), sheet (B) and coil (C) which are more frequently observed in the literature. There are different rules applied in the literature to perform such mapping. Table 4 shows existing rules to convert 8 structures to 3 states [37] [38].

Table 4

8state to 3 state secondary structure conversion rules

Rules of 8 state secondary structure to 3 state
{H, G} to {H} – { E, B} to {E} – {All other states} to {C}
{H} to {H} – {E} to {E} – {All other states} to {C}
{H, G, I} to {H}- {E} to {E} - {All other states} to {C}
{H, G} to {H} – {E} to {E} – {All other states} to {C}
{H, G, I} to {H} – {E, B} to {E} – {All other states} to {C}
{H,G,I} to {H} – {E,B} to {E} – {All other states} to {L}

4.4 Evaluation Measures of Secondary Structure Prediction

In this section the common evaluation measures applied in the literature of protein secondary structure prediction will be defined and described. These measures are formulated in terms of confusion matrix parameters namely TP, TN, FP and FN described below.

TP: number of correctly predicted residues for each class.

TN: number of correctly predicted residues not belonging to each class.

FP: number of incorrectly predicted residues to belong to each class.

FN: number of incorrectly predicted residues not to belong to each class.

The names, description and mathematical interpretation of each measure can be found in table 5. The most overall and vastly used measure is overall accuracy. Overall accuracy no matter how high, is not solely an indication of the competence of an algorithm. As there are times when the overall accuracy is high and the predictor performs very well except for a particular class in which it exhibits poor results. Hence considering a group of measures gives a real reflection of the weakness points and strengths of a method.

Best predictions according to MCC are close to 1. Predictions at random level are close to 0 and predictions even worse than random are close to -1.

$$MCC = \frac{TP_j TN_j - FP_j FN_j}{\sqrt{(TP_j + FP_j)(TP_j + FN_j)(TN_j + FP_j)(TN_j + FN_j)}} \quad (1)$$

Unlike other discussed measures which are employed in the evaluation of all classification methods, SOV is especially designed for PSSP [39].

$$SOV = \frac{1}{N} \times \Sigma_{j \in \{H,E,C\}} \Sigma_{S(j)} \frac{\min ov(S_1^j, S_2^j) + \delta(S_1^j, S_2^j)}{\max ov(S_1^j, S_2^j)} \times \left|S_1^j\right| \times 100 \quad (2)$$

where N is the total number of residues. $\min ov(S_1^j, S_2^j)$ is the length of overlapping of the two segments s_1^j and s_2^j. $\max ov(S_1^j, S_2^j)$ is the total extent of both segments and is calculated by $\left|s_1^j\right| + \left|s_2^j\right| - \max ov(S_1^j, S_2^j)$. $\delta(S_1^j, S_2^j)$ is defined as in (3) [39]:

$$\delta(S_1^j, S_2^j) = \min(\max ov(S_1^j, S_2^j) - \min ov(S_1^j, S_2^j), \left\lfloor \frac{\left|S_1^j\right|}{2} \right\rfloor, \left\lfloor \frac{\left|S_2^j\right|}{2} \right\rfloor) \quad (3)$$

Table 5

Evaluation measures for protein secondary structure prediction assessment

Measures' names	Measures' Description	Range of Values	Measure's Formulation
Overal Acuuracy **Q_3** **Q_{total}**	The ratio of proteins from test set which have been predicted accurately	0 to 1	$\sum_{J\in(H,E,C)} \frac{TP_j + TN_j}{TP_j + FP_j + TN_j + FN_j}$
Recall **Sensitivity** **Q_{obs}** **Q_i**	The number of proteins belonging to class j which have been predicted accurately.	0 to 1	$\frac{TP_j}{TP_j + FN_j}$
Precision **Q_{pred}**	The number of proteins predicted to belong to class j and are accurate predictions.	0 to 1	$\frac{TP_j}{TP_j + FP_j}$
Specificity	The number of proteins predicted to belong to class j and is accurate predictions.	0 to 1	$\frac{TN_j}{FP_j + TN_j}$
F-measure	The weighted average of precision and recall defined earlier. The values range from 0, worst results to 1, best results.	0 to 1	$2 \times \frac{precision_j \times \mathrm{Re}\, call_j}{precision_j + \mathrm{Re}\, call_j}$
FPR (False Positive Rate)	The ratio of incorrect predictions of residues to belong to class j.	0 to 1	$\frac{TP_j}{FP_j + TN_j}$
MCC **(**Mathew's Correlation Coefficient)	A single measure to evaluate the two measures, Recall and Precision, concurrently.	-1 to 1	Equation (1)
SOV **(**Segment Overlap Score)	Counts the existence of continuous structural elements to be predicted.	0 to 1	Equation (2)

Figure 6, shows the employment frequency of each measure in the studied literature.

As can be seen from figure 6, the three most popular measures are overall accuracy, recall and SOV. Consequently exploiting the measures offers a comprehensive comparative framework. Furthermore, the vast usage of these methods proofs their strength in reflection of the performance measurement.

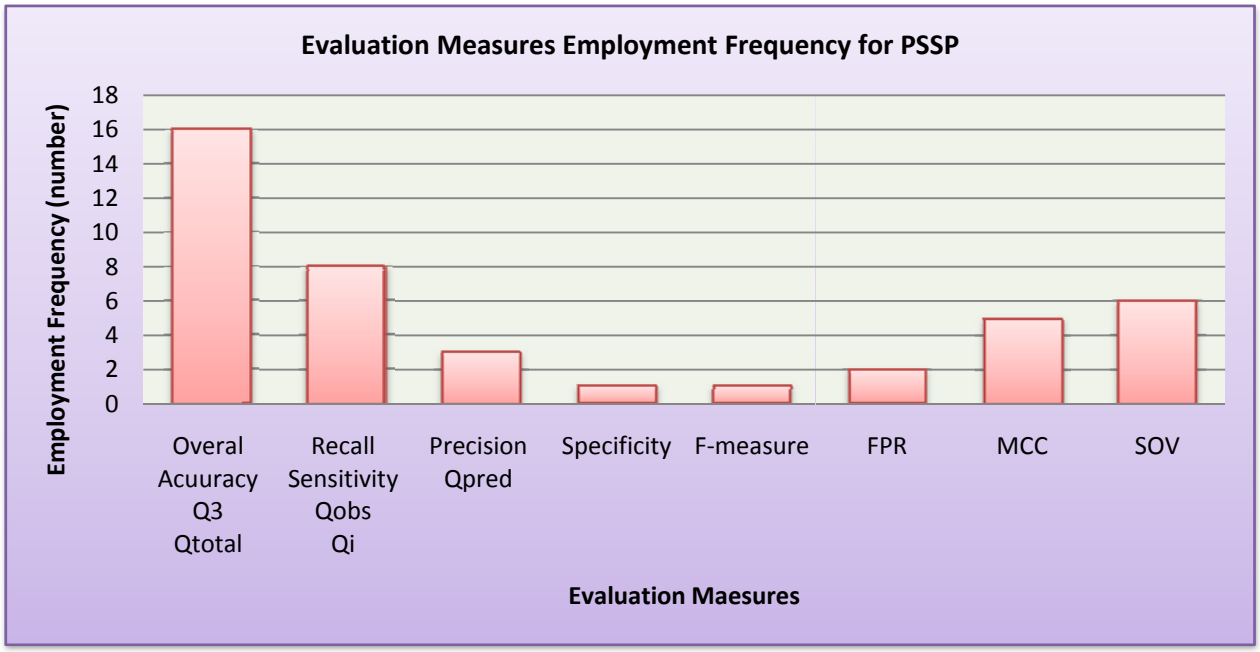

Fig. 1. Evaluation measures employment frequency in the literature of protein secondary structure prediction.

7. Conclusion

Protein secondary structure prediction is momentous task in bioinformatics. The reason behind its significance is that it's either the input feature or intermediate step for other bioinformatics tasks [40] such as structural class prediction, fold recognition, tertiary structure prediction, function prediction and drug and enzyme design. Consequently it is afield which demands great deal of attention from the research body. On the other hand and with universal sequencing project which determined the amino acid sequence of proteins, the gap between known sequences with unknown structure grew increasingly. On the other hand the experimental methods such as X-ray crystallography, electron microscopy and nuclear magnetic resonance for protein secondary structure prediction were far too time consuming, costly and inapplicable to all proteins which made the sequence structure gap, wider.

As the result of the enumerated issues, since 1970's, many endeavors have been made to address the problem of protein secondary structure prediction, using computational methods. Various generations of methods came into existence, each enhancing the performance and obviating the flaws of previous generations. Information theory and statistical methods were amongst the first generation of solutions. The next generation emerged with the advent of machine learning algorithms in this field. Towards the recent years, the ensemble methods made a remarkable improvement in the field of study.

There are a number of complications that make PSSP a laborious problem. These complications include obscure protein data patterns, noise, class imbalance and high dimensionality imposed by encoding schemes of amino acid sequence. These complications along with the multidisciplinary nature of the problem lead to a wide diversity amongst the solutions.

Accordingly a review of the proposed methods which classifies the solutions into major categories and investigates their advantages and drawback is of great benefit. Such study can provide an overview of what has been done by far, reveal trends in methods and materials of the field and cast light to the potential uncovered areas capable of making elevation in the solutions.

To pursue this end, in this study, a variety of methods for PSSP problem were investigated. This investigation lead to a meaningful categorization of methods, identified best and worst performing methods, offered statistics which disclosed the mostly used material of the field such as datasets, benchmarks, algorithms and rules. The extent of provided details and descriptions can be of great help to reach to a beneficial conclusion about what has been done and what needs to be followed.

References

[1] H. del Portillo, A. Gruber, A. Durham, C. Hyung , "Bioinformatics in tropical disease research: A practical approach". editors : F Agüero, G Correa-Oliveira, DS Roos, JC Kissinger - NCBI Electronic Book, 2007.

[2] E. Buxbaum, “Fundamentals of Protein Structure and Function”, Springer, ISBN: 978-0-387-26352-6, 2007.

[3] D. Ofer, and Y. Zhou. "Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training." *Proteins: Structure, Function, and Bioinformatics* 66.4 (2007): 838-845.

[4] A. Rafał. "Dimensionality reduction of pssm matrix and its influence on secondary structure and relative solvent accessibility predictions." *World Academy Of Science, Engineering And Technology* 58 (2009): 657-664.

[5] Y. Chou Peter, and G. D. Fasman. "Prediction of protein conformation."Biochemistry 13.2 (1974): 222-245.

[6] J. Garnier, D. J. Osguthorpe, and B. Robson. "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins." Journal of molecular biology 120.1 (1978): 97-120.

[7] M. Rithvik, and G. N. Rao. "A Comparative Study of Methodologies of Protein Secondary Structure." *Computational Intelligence Techniques for Comparative Genomics*. Springer Singapore, 2015. 37-45.

[8] A. Shivani, P. Agarwal, and D. Mendiratta. "Prediction of Secondary Structure of Protein Using Support Vector Machine." *IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications*. No. 5. Foundation of Computer Science (FCS), 2014.

[9] T. Pang-Ning, and M. Steinbach. "Vipin Kumar, Introduction to Data Mining." (2006).

[10] A. Ghosh, B. Parai, “Protein secondary structure prediction using distance based classifiers”, International Journal of Approximate Reasoning, 2008, 47, 37–44.

[11] T. Liu, X. Zheng , J. Wang, “Prediction of protein structural class using a complexity-based distance measure”,Springer, Amnio Acids, 2012, 38:721–728.

[12] M. Nilson, D. Fernando, M. Carmona, and I. Tischer. "FS-Tree: Sequential Association Rules and First Applications to Protein Secondary Structure Analysis." *Advances in Computational Biology*. Springer International Publishing, 2014. 189-198.

[13] M. HosseinZangooei , S. Jalili, “Protein secondary structure prediction using DWKF based on SVR-NSGAII”, Elsevier, Neurocomputing, 2012, 94 : 87–101.

[14] M. Hossein Zangooi, S. Jalili, “PSSP with dynamic weighted kernel fusion based on SVM-PHGS”, Elsevier, Knowledge-Based Systems, 2011, 27:424-442.

[15] N. Nguyen Minh and J. C. Rajapakse. "Multi-class support vector machines for protein secondary structure prediction." Genome Informatics 14 (2003): 218-227.

[16] L. Liyu, S. Yang, and R. Zuo. "Protein secondary structure prediction based on multi-SVM ensemble." Intelligent Control and Information Processing (ICICIP), 2010 International Conference on. IEEE, 2010.

[17] W. Guoren, Y. Zhao, D. Wang. "A protein secondary structure prediction framework based on the extreme learning machine." Neurocomputing72.1 (2008): 262-268.

[18] S. Babaei, A. Geranmayeh, S. A. Seyyedsalehi. "Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks." Computer methods and programs in biomedicine 100.3 (2010): 237-247.

[19] S. Babaei, A. Geranmayeh, S. A. Seyyedsalehi. "Towards designing modular recurrent neural networks in learning protein secondary structures." Expert Systems with Applications 39.6 (2012): 6263-6274.

[20] M. Alirezaee, A. Dehzangi, E. Mansoori, “Ensemble of neural networks to solve class imbalance problem of protein secondary structure prediction”, International Journal of Artificial Intelligence & Applications (IJAIA), 2012.

[21] J. Anu, R. Kaur, R. Singh. "Protein Secondary Structure Prediction Using Improved Support Vector Machine and Neural Networks." *International Journal of Engineering and Computer Science* 3.1 (2014): 3593-3597.

[22] D. Patel Mayuri , H. B. Shah: "Protein Secondary Structure Prediction Using Neural Network: A Comparative Study" , International Journal of Enhanced Research in Management & Computer Applications, Vol. 3 Issue 4, April-2014, pp: (18-23).

[23] Y. Ashraf, Li. Yaohang "Context-based features enhance protein secondary structure prediction accuracy." *Journal of chemical information and modeling* 54.3 (2014): 992-1002.

[24] J. He, H-J. Hu, R. Harrison, P. C. Tai, Y. Pan, “Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree” , IEEE Transactions, Nanobioscience, 2006, VOL. 5, NO. 1.

[25] B. Hafida, B. Messabih, A. Chouarfia. "Effect of simple ensemble methods on protein secondary structure prediction." *Soft Computing*19.6 (2015): 1663-1678.

[26] S. Patel Maulika, H. S. Mazumdar. "Knowledge base and neural network approach for protein secondary structure prediction." *Journal of theoretical biology* 361 (2014): 182-189.

[27] T.Z. Sen, R.L. Jernigan, J. Garnier, A. Kloczkowski, "GOR V server for protein secondary structure prediction", *Bioinformatics*, 21(11),2787-2788, 2005.

[28] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang and Y. Zhou, “Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning”. Scientific Report, 2015.

[29] C. Cole, JD. Barber, GJ. Barton “The Jpred 3 secondary structure prediction server”. *Nucleic Acids Res* 2008,36(suppl 2):W197-W201.

[30] D. Frishman, P. Argos, “Knowledge-based protein secondary structure assignment”. *Proteins: Structure, Function, and Bioinformatics* 1995,23(4):566–579.

[31] L. McGuffin, K. Bryson, D. Jones, “The PSIPRED protein structure prediction server”. *Bioinformatics* 2000,16(4):404

[32] G. Karypis "YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction". *Proteins: Structure, Function, and Bioinformatics* 2006,64(3):575–586.

[33] A. Randall Cheng, M. Sweredoski, P. Baldi "SCRATCH: a Protein Structure and Structural Feature Prediction Server", Nucleic Acids Research, Web Server Issue vol. 33, 72-76, 2005.

[34] R. Adamczak, A. Porollo, J. Meller: Combining prediction of secondary structure and solvent accessibility in proteins.*Proteins: Structure, Function, and Bioinformatics* 2005,59(3):467–475.

[35] Karplus, Kevin. "SAM-T08, HMM-based protein structure prediction". *Nucleic acids research* 37.suppl 2 (2009): W492-W497.

[36] G. Pollastri, A. Mclysaght: "Porter: a new, accurate server for protein secondary structure prediction". *Bioinformatics*2005,21(8):1719–1720.

[37] B. Rajkumar, O. Duzlevski, and D. Xu. "Profiles and fuzzy K-nearest neighbor algorithm for protein secondary structure prediction." *APBC*. 2005.

[38] L. Christos, et al. "Improving the protein fold recognition accuracy of a reduced state-space hidden Markov model." *Computers in Biology and Medicine* 39.10 (2009): 907-914.

[39] Y. Bingru, et al. "Predicting protein second structure using a novel hybrid method." *Expert Systems with Applications* 38.9 (2011): 11657-11664.

[1] [40] Li, Qiwei, et al. "Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction." (2014): e109832.