Protein Secondary Structure Prediction using Bayesian Inference method on Decision fusion algorithms

Somasheker Akkaladevi[1], Ajay K Katangur[2]

[1] Virginia State University
Department of Computer Information Systems
Petersburg, Virginia 23806, USA
sakkaladevi@vsu.edu

[2]Texas A&M University - Corpus Christi
Department of Computing Sciences
Corpus Christi, Texas 78412, USA
ajay.katangur@tamucc.edu

Abstract

Prediction of protein secondary structure (alpha-helix, beta-sheet, coil) from primary sequence of amino acids is a very challenging task, and the problem has been approached from several angles. Previously research was performed in this field using several techniques such as neural networks, Simulated annealing (SA) and Genetic algorithms (GA) for improving the protein secondary structure prediction accuracy. Decision fusion methods such as the Committee method and Correlation methods were also used in combination with the profile-based neural networks and AI algorithms for achieving better prediction accuracy. In this research we investigate the Bayesian inference method for predicting the protein secondary structure. The Bayesian inference method proposed in this research uses the results from the committee and correlation methods to achieve better prediction accuracy. Simulations are performed using the RS126 data set. The results show that the protein secondary structure prediction accuracy can be improved by more than 2% using the Bayesian inference method.

1. Introduction

Prediction of a secondary structure of a protein from its amino acid sequence remains an important and difficult task. Not only can successful predictions provide a starting point for direct tertiary structure modeling, but they can also significantly improve sequence analysis and sequence-structure threading for aiding in structure and function determination [13]. A protein is a sequence of amino acid residues and can thus be considered as a one dimensional chain of beads where each bead corresponds to one of the 20 different amino acid residues known to occur in proteins. The length of most protein sequence ranges from 50 residues to about 1000 residues but longer proteins are also known, e.g. myosin, the major protein of muscle fibers, consists of 1800 residues [3]. Previously much research was performed on predicting protein secondary structure by many researchers all over the world. Many techniques were used by many researchers to predict the protein secondary structure, but the most commonly used technique for protein secondary structure prediction is the neural network [12].

Around 1988 the first attempts were made to use neural networks to predict protein secondary structure [12]. The accuracy of the predictions made by Qian and Sejnowski seemed better than those obtained by previous methods and was reported to be in the range of 62.7-64.4% [12]. The most successful application of neural networks to secondary structure prediction was obtained by Rost and Sander [13, 14, 15, 16], which resulted in the prediction mail server called PHD [14]. The most significant new feature in the work of Rost and Sander is the use of sequence alignments [13]. For each protein in the data set a set of aligned homologous proteins is found. Instead of just feeding the base sequence to the network they feed the multiple alignment in the form of a sequence profile, i.e., for each position an amino acid frequency vector is fed to the network. This type of configuration is called a profile-based neural network. Using these and a few other methods, the performance of the network is reported to be up to 67.2% [13].

One of the primary goals of the present work is to design a new method combining profile-based neural networks [13], SA [2, 18], GA [2] and the decision fusion algorithms [2]. Researchers previously used the feed forward neural network combined with GA and SA algorithms, and then applied the two decision fusion methods; committee method and the correlation methods and obtained improved results on the prediction accuracy

1-4244-0910-1/07/$20.00 ©2007 IEEE.

[2]. Sequence profiles of amino acids were fed as input to the profile-based neural network. The two decision fusion methods improved the prediction accuracy, but it was noticed that one method worked better in some cases and the other for some other sequence profiles of amino acids as input [2]. Instead of compromising on some of the good solutions from that could have generated from the both the either approaches, it will be better if we can combine these two approaches so that better prediction accuracy can be achieved. This criterion is the basis for our Bayesian inference method [4, 17, 18]. Initially we feed the sequence profiles of amino acids into the profile-based neural network to predict the protein secondary structure. After we predict the protein secondary structure from a profile-based neural network we send the predicted protein secondary structure to SA, GA and then fuse the errors using both the decision fusion methods (committee and correlation methods) running in parallel on two different threads. The Bayesian inference method is then applied on these results for calculating the error values to be back-propagated to the profile-based neural network for weight adjustments. Sections 2 and 3 discuss about protein secondary structure prediction and data set used for prediction. Section 4 discusses on how to evaluate the prediction accuracy. Section 5 gives an introduction to the Bayesian inference method, and section 6 discusses on how to apply the method discussed in section 5 for protein secondary structure prediction. Sections 7 and 8 detail the results and conclusion.

2. Protein Structures - Secondary Structure Prediction

One approach to protein structure prediction is first to predict secondary structure as a stepping stone toward the full structure. The aim is to predict which secondary structural element will be formed by each residue of the protein [15].

The structure of a protein has different levels and it has an energically and structurally optimized form [7]. The *primary structure* is the amino acid of the protein and can be presented by a sequence with 20 letters, where each letter indicates an individual amino acid. The *secondary structure* describes the areas in the primary structure where secondary structure elements occur in the backbone of the protein. The *tertiary structure* is the three-dimensional structure of a single protein chain. In order to predict the tertiary structure [5], the secondary structure must be first predicted. However, secondary structure predictions can be of advantage in other ways. They have recently been shown to be useful in the prediction of regions of the protein likely to undergo structural change [6] and in the classification of proteins for genome analysis [7].

Advances in secondary structure prediction have to some extent been based on developments in machine learning theory; beginning with rule-based approaches, moving onto neural network approaches and AI techniques.

3. Assignment of Secondary Structure

In the problem of the protein secondary structure prediction, the inputs are the amino acid sequence profiles while the output is the predicted structure (also called conformation, which is the combination of alpha helices, beta sheets and loops) [7]. A typical protein sequence and its conformation class are shown below:

```
ADADADADCCQQFFFAAAQQAQQA
 HHHH  EEEE    HHHHHHHH
```

H means Helical, E means Extended, and blanks are the remaining coiled conformations.

A typical protein contains about 32% alpha helices, 21% beta sheets and 47% loops or non-regular structure [13]. Proteins evolved from a common ancestor are called homologous proteins and they usually have similar amino acid sequences and conformations, and hence similar properties and functions. Researchers usually select non-homologous proteins from the protein data bank as working data for structure prediction research. It is possible to predict loop regions with higher accuracy than alpha helices or beta sheets [14].

In order to assess the value of any prediction scheme, it must be possible to quantify accurately how well the scheme performs. Unfortunately, this is not as easy as it might seem; there is the problem of choosing a data set, how to split that data set into testing and training sets, and finally what statistics should be generated.

3.1. Choice of Data Set

Choosing a suitable data-set is a hard problem that requires both knowledge of learning machines and domain specific knowledge. The idea is to choose a representative set of problems with known solutions that can be used to train the network and to test its performance. In this research we used the *seven-fold cross-validation* on the set of 126 non-homologous globular proteins from (Rost & Sander, 1994), which is called the RS126 data set [16]. With seven-fold cross-validation approximately 1/7 of the database is left out while training and the remaining part is used for testing. This is done cyclically seven times, and the resulting prediction is thus a mean over seven different testing sets. No proteins in the RS126 data set have more than 25% pair-wise sequence identity for lengths greater than 80 residues. The RS126 dataset contains 24,395 amino acids with 32% α –helix, 21% β –strand and 47% coil [16].

4. Performance Measures – Calculating the Protein Secondary Structure Prediction accuracy

The protein secondary structure accuracy is calculated by using the three-state per-residue accuracy (Q_3), which gives the percentage of correctly predicted residues in either of the three states (classes), alpha helix, beta strand or loop region [12, 15]:

$$Q_3 = \left[\frac{(P_\alpha + P_\beta + P_{loop})}{T} \right] \times 100\%$$

where P_α, P_β and P_{loop} are number of residues predicted correctly in state alpha helix, beta strand and loop respectively while T is the total number of residues. There are three simple measures for assessing the quality of predicted secondary structure segments (or states): the number of segments in the protein, the average segment length and the distribution of the number of segments with length. Prediction methods need to meet four requirements. Firstly, no significant pair wise sequence identity between proteins used for training and test set (<25%). Secondly, all available unique proteins should be used for testing (since proteins vary in structural complexity, certain features are easier to predict than others). Regardless of which data sets are used for a particular evaluation, a standard set should be used for which results are also reported. Finally, test set should never be used before the method is set up [15].

5. Bayesian Inference method

Results from different methods, algorithms, sources or classifiers can often be combined to give estimates of a better quality solution than could be obtained from any of the individual sources alone. Luo and Kay give a comprehensive survey of the Decision Fusion in [10]; their paper also appears with a collection of other fusion survey papers in [1]. We briefly explain the other decision fusion methods previously applied to this problem as we will be considering the results of these methods for the proposed Bayesian inference method.

Committee Method - A key problem in Decision Fusion is how to enable the different information sources to contribute to a result. Vote based decision fusion methods group individual experts or discriminating functions into a set termed a *committee*. In this approach, the individual experts cast votes for the correct hypothesis. A variety of voting rules have been proposed, in the Majority Vote rule the hypothesis with the most votes is chosen [8, 11]. Initially the amino acid input sequence profile is fed into the profile-based neural network. In the second step the output of the profile-based neural network is fed as input to the GA and SA algorithms. The output of these algorithms is again fed as input to the committee method. The committee method then calculates the new error value and back-propagates it to the profile-based neural network for weight adjustments [2].

Correlation Method - It is similar to the committee method [8], except that the decisions after applying the correlation method are back-propagated to the correlation method as in a profile-based neural network and then the final decision is obtained [2].

Bayesian inference method - Bayesian inference is statistical inference in which evidence or observations are used to update or to newly infer the probability that a hypothesis may be true [4]. Hypotheses with a very high degree of belief should be accepted as true; those with a very low degree of belief should be rejected as false [4].

An example of Bayesian inference is: For billions of years, the sun has risen after it has set. The sun has set tonight. With very high probability (or I strongly believe that or it is true that) the sun will rise tomorrow. With very low probability (or I do not at all believe that or it is false that) the sun will not rise tomorrow [4].

Bayesian inference usually relies on degrees of belief, or subjective probabilities, in the induction process and does not necessarily claim to provide an objective method of induction [4]. Bayes' theorem adjusts probabilities given new evidence in the following way:

$$P(H_0 \mid E) = \frac{P(E \mid H_0)P(H_0)}{P(E)}$$

where H_0 represents a hypothesis, called a null hypothesis, which was inferred before new evidence, E, became available. $P(H_0)$ is called the *prior probability* of H_0. $P(E \mid H_0)$ is called the *conditional probability* of seeing the evidence E given that the hypothesis H_0 is true. It is also called the *likelihood function* when it is expressed as a function of H_0 given E.

$P(E)$ is called the *marginal probability* of E: the probability of witnessing the new evidence E under all mutually exclusive hypotheses. It can be calculated as the sum of the product of all probabilities of mutually exclusive hypotheses and corresponding conditional probabilities [$\sum P(E \mid H_i)P(H_i)$]. $P(H_0 \mid E)$ is called the *posterior probability* of H_0 given E.

The factor $P(E \mid H_0) / P(E)$ represents the impact that the evidence has on the belief in the hypothesis. If it is likely that the evidence will be observed when the hypothesis under consideration is true, then this factor will be large. Multiplying the prior probability of the hypothesis by this factor would result in a large posterior probability of the hypothesis given the evidence. Under Bayesian inference, Bayes' theorem therefore measures how much new evidence should alter a belief in a hypothesis [4]. Multiplying the prior probability $P(H_0)$ by the factor $P(E \mid H_0)$ / $P(E)$ will never yield a probability that is greater than 1. Since $P(E)$ is at least as great as $P(E \cap H_0)$, which equals $P(E \mid H_0).P(H_0)$, replacing $P(E)$ with $P(E \cap H_0)$ in the factor $P(E \mid H_0)$ / $P(E)$ will yield a posterior probability of 1. Therefore, the posterior probability could yield a probability greater than 1 only if $P(E)$ were less than $P(E \cap H_0)$ which is never true [4].

The marginal probability, $P(E)$, can also be represented as the sum of the product of all probabilities of mutually exclusive hypotheses and corresponding conditional probabilities: $P(E \mid H_0)P(H_0)$ + $P(E \mid \text{not } H_0)P(\text{not } H_0)$. As a result, we can rewrite Bayes' theorem as:

$$P(H_0 \mid E) = \frac{P(E \mid H_0)P(H_0)}{P(E \mid H_0)P(H_0) + P(E \mid notH_0)P(notH_0)}$$

The same equation can be represented with simple terminology as follows:

$$P(H_1 \mid D) = \frac{P(H_1) \times P(D \mid H_1)}{P(H_1) \times P(D \mid H_1) + P(H_2) \times P(D \mid H_2)}$$

With two independent pieces of evidence E_1 and E_2, Bayesian inference can be applied iteratively. We could use the first piece of evidence to calculate an initial posterior probability, and then use that posterior probability as a new prior probability to calculate a second posterior probability given the second piece of evidence. Bayes' theorem applied iteratively implies:

$$P(H_0 \mid E_1, E_2) = \frac{P(E_1 \mid H_0) \times P(E_2 \mid H_0)P(H_0)}{P(E_1) \times P(E_2)}$$

This iteration of Bayesian inference could be extended with more independent pieces of evidence. Bayesian inference is used to calculate probabilities for decision making under uncertainty [4].

6. Appling Bayesian Inference for Protein Secondary Structure Prediction

All neural networks used in this research are standard feed- forward networks, and are trained using the back-propagation algorithm [9]. Networks are trained on a set of data for which the desired output is known; this is referred to as the training set. The method used is back-propagation, a well-characterized algorithm for adjusting the weights [9]. After training, the network can be exposed to new data for which the desired output is not known to the network; this is known as the test set. In this research the RS126 dataset is used, which contains 126 sequences with approximately more than 23,300 amino acid positions and 20 amino acids [16].

6.1. Neural Networks for Protein Secondary Structure Prediction using sequence profiles

The profile-based neural network is used in this research. Orthogonal encoding scheme is used to encode all the amino acids [6, 12]. It is well known that homologous proteins have the same three-dimensional fold and approximately equal secondary structures down to a level of 25-30% identical residues [16]. Using profiles at the input level has been shown to yield better results than using profiles at the output level [5, 13].

Figure 1 shows the use of multiple sequence alignments rather than a single sequence as input to a profile-based neural network. At the prediction stage, the database of sequences is scanned for all homologues of the protein to be predicted, and the family profile of amino acid frequencies at each alignment position is fed into the network. A sequence profile of a protein family, rather than just a single sequence, is used as input to the profile-based neural network as shown in Figure 1 for secondary structure prediction [14].

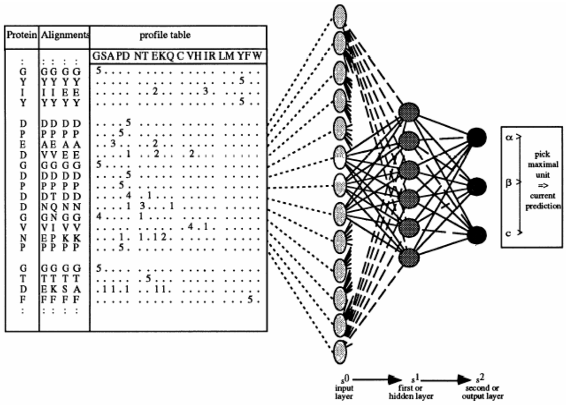
Protein
Alignments
profile table
GSAPD NT EKQ C VHIR LM YFW
α
β
c
pick maximal unit => current prediction
s0 input layer
s1 first or hidden layer
s2 second or output layer

Figure 1. A profile-based neural network for secondary structure prediction

Each sequence position is represented by the amino acid-residue frequencies derived from multiple sequence alignments as taken from the homology-derived structure of proteins. The residue frequencies for the 20-residue types are represented by 3 bits each. To code the N- and C-terminal ends an additional 3 bits are required. The 63 bits originating from one sequence position are mapped onto 63 input units of the profile-based neural network. A window of 13 sequence positions, thus corresponds to 819 (13 × 63) input units. The input signal is propagated through the network with one input, one hidden, and one output layer. The output layer has three units corresponding to the three secondary-structure states (alpha helix, beta strand, and coil), at the central position of the input sequence window. The output values are between 0 and 1. The observed secondary structure states are encoded as 1,0,0 for helix; 0,1,0 for strand; and 0,0,1 for coil [14].

The secondary structure obtained is compared to the structure already known during the training phase. Normally, the prediction for the terminal amino acids is not reliable, because the chain is usually flexible at the beginning and end of proteins, and they are often treated as a special case. The protein secondary structure accuracy is then calculated by using the three-state per-residue accuracy (Q_3) [12, 15].

$$Q_3 = \left[(\frac{(P_\alpha + P_\beta + P_{loop})}{T} \right] \times 100\%$$

where P_α, P_β and P_{loop} are number of residues predicted correctly. Using this approach the secondary structure prediction accuracy (Q_3) is obtained to be 66.8%.

6.2. Combining GA and the profile-based Neural Networks for protein secondary structure prediction

For example, if we have the predicted structure from a profile-based neural network as follows:

PredictedStructure:- -HHHCCEEEECCCHHHHHHHH- -
TrueStructure: - - HHHHHCEEEECCCHHHHHCH- -

The predicted structure is given to GA; the GA does a mutation operation on the predicted structure from the profile-based neural network to generate a new solution (offspring) [2]. The process of mutation is random and can occur at any point in the given structure.

- - HHHCCEEEECCCHHH***E***HHHH - -

After generating the offspring the fitness of this new offspring is calculated by again comparing to the true structure already known by using the Q_3 function. The GA then keeps this solution or throws it away depending on the fitness value, which is in our case is the prediction accuracy Q_3. Similarly we keep on applying crossover and mutation operations to generate new offspring's, evaluate the fitness and then keep the offspring if it has a better fitness value or prediction accuracy. The crossover operation is always done only when we have at least one valid offspring generated from mutation operation, since crossover operation needs a minimum population of 2. For example a crossover operation is shown below between the predicted structure from the profile-based neural network and the mutated offspring which is obtained above from the GA. The crossover happens at the position indicated by the arrow [2].

- - HHHCCEEEECCCHHHHH↓***HHH*** - -
- - HHHCCEEEECCCHHH*E*H**HHH** - -

After crossover the new solutions are:

- - HHHCCEEEECCCHHHHH **HHH** - -
- - HHHCCEEEECCCHHHEH ***HHH*** - -

Even though we applied a crossover operation we did not get any improvement as we generated the same offspring's again. We only input the profiles to the profile-based neural network on the input side. We keep on applying the mutation and crossover operations to generate new solutions until the number of generations are complete, and finally at this point we calculate the error value which is to be back-propagated to adjust the weights of the profile-based neural network. However while testing the network we do not employ the GA as we do not need any adjustments of the weights for the profile-based neural network. The mutation probability for GA in this research is set at 0.25, number of generation's value at 75, population size at 30 and the crossover probability as 100% [2]. Using this approach the secondary structure prediction accuracy (Q_3) is obtained to be 69.2%.

6.3. Combining SA and the profile-based Neural Networks for protein secondary structure prediction

In traditional profile-based feed forward neural network we calculate Q_3 from the secondary structure obtained and then calculate the error to be back-propagated to adjust the weights of the network. But when we apply SA algorithm to the profile-based neural network, the predicted structure from the profile-based neural network is sent to the SA algorithm for further processing by the SA algorithm [2]. The SA algorithm then generates new solutions and compares it with the true secondary structure which is already known to calculate the prediction accuracy Q_3. The error is then calculated from the solution generated by SA algorithm by calculating the value of Q_3. This error value is then back-propagated to adjust the weights of the profile-based neural network. The initial solution is generated by the profile-based neural network shown in Figure 1. However while testing the network we do not employ the SA algorithm as we do not need any adjustments of the weights for the profile-based neural network. The starting temperature for SA in this research is set at 600, the final temperature at 0.20, the temperature cooling rate at 0.84, and the number of iterations per temperature at 20 [2]. Using this approach the secondary structure prediction accuracy (Q_3) is obtained to be 68.3%.

6.4. Prediction of Protein Secondary Structure using the Committee method and the profile-based Neural Network

In the committee based method of applying decision fusion we calculate the secondary structure values using a combined profile-based neural network (PNN) with GA, a combined profile-based neural network with SA, and the independent profile-based neural network. We then feed this output obtained from the profile-based neural network, combined profile-based neural network plus GA and combined profile-based neural network plus SA to the decision fusion algorithm, for fusing the solutions as shown in Figure 2 [2, 11].

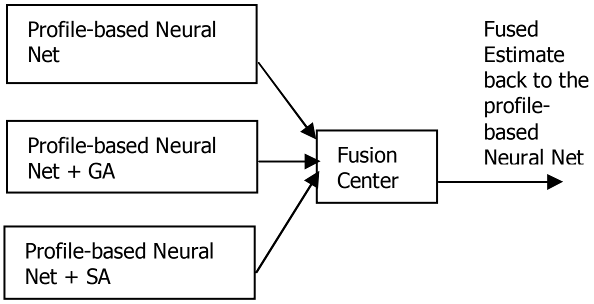
Profile-based Neural Net
Profile-based Neural Net + GA
Profile-based Neural Net + SA
Fusion Center
Fused Estimate back to the profile-based Neural Net

Figure 2. Fusing the various solutions according to the rules in the Fusion Center

The decision fusion algorithm works on the basis of a committee (committee method or voting method), where each individual in the committee decides on the best solution according to pre-determined rules and then cast their vote for the best approach [11]. In the event of a tie, the tie is broken by one more rule, where we have a priority given to each algorithm. The algorithm with the highest priority wins. The Committee fusion algorithm is as outlined below:

1. Given a secondary structure output obtained by profile-based neural network of N_i elements, where $i = 1,2,\ldots\ldots,n$. (Here for 'H' we assume a value of 2, for 'E' a value of 3, and for 'C' a value of 4. These are arbitrarily chosen values)

2. In the same manner, the secondary structure output obtained from GA and SA are represented by G_i and S_i respectively.

3. Calculate the following values:

$$G = \sum_{i=1}^{n} (N_i - G_i)^2 \qquad (1)$$

$$S = \sum_{i=1}^{n} (N_i - S_i)^2 \qquad (2)$$

$$N = 0 \qquad (3)$$

4. Compute N_i - G_i. If N_i - G_i > 0, then (bin+) ← N_i - G_i else if N_i - G_i < 0, then (bin-) ←N_i - G_i, where bin+ and bin- are the so called positive and negative bins. If the result of the operation is zero, then we do not store that in either of the bin.

5. Evaluate bin+ and bin-, the positive and negative bins for G and then, if they are equal or if the positive bin has a higher count compared to the negative bin we assign G as a positive sign ($+G$), else we assign G a negative sign ($-G$). We always consider $N=0$.

6. Repeat steps 4 and 5 to calculate S.

7. Use $max(N, G, S)$ to be the secondary structure for calculating Q_3 used to determine the error for back-propagation, so that the weights of the network can be adjusted.

8. Each algorithm votes for the best solution by comparing its value with the other algorithms values. The algorithm with the majority votes wins the race. In the event of a tie, the tie is broken according to the algorithm's priority and, then the algorithm that wins calculates the prediction accuracy using the function Q_3 to determine the error that is to be back-propagated to the profile-based neural network for weight adjustments.

9. In our case we have assumed the highest priority for profile-based neural network (PNN) secondary structure values, followed by the combination of profile-based neural network and GA (PNN+GA), and then followed by the combination of profile-based neural network and SA (PNN+SA) [2]. However while testing the network we do not employ these algorithms as we do not need any adjustments of the weights on the profile-based neural network. Using this approach the secondary structure prediction accuracy (Q_3) is obtained to be 70.8%.

6.5. Prediction of Protein Secondary Structure using the Correlation method and the profile-based Neural Network

The correlation method of decision fusion is applied next to the problem to further improve on the prediction accuracy. This method is very similar to the committee method outlined in section 6.5, and has some minor changes to it [2, 8].

In this method the algorithm that wins after decision fusion is applied is used to calculate the prediction accuracy using the function Q_3 to determine the error that is to be back-propagated to the profile-based neural network for weight adjustments. After this adjustment of weights on the profile-based neural network again the same previous protein sequence is used for testing purpose, to check whether a better prediction accuracy is achieved or not. Here we keep these new weights if we get an improvement of more than 1.5%, otherwise from the previously calculated prediction accuracies from (PNN), (PNN+GA) and (PNN+SA), we take the method which produces the highest prediction accuracy (which is calculated using the function Q_3) to determine the error that is to be back-propagated to the profile-based neural network for weight adjustments [2]. However while testing the network we do not employ these algorithms as we do not need any adjustments of the weights on the profile-based neural network. Using this approach the secondary structure prediction accuracy (Q_3) is obtained to be 71.4%.

6.6. Prediction of Protein Secondary Structure using the Bayesian Inference method

In this method, we mainly use the committee and correlation methods of decision fusion as discussed in section 6.6 and section 6.7, and then apply the Bayesian inference method on the output generated by these two methods [4, 17]. From the results of both these approaches we have noticed that the committee method using the profile-based neural network gives prediction accuracy (Q_3) of 70.8% compared to the 71.4% produced by the correlation method using the profile-based neural network [2]. In the Bayesian inference approach we use both these methods, by assigning a specific probability value to them, and then generating a new value using the Bayesian equation [4, 18]. This new value obtained is used to decide between the two methods (committee method and correlation method) to be used for calculating the error that is to be back-propagated to the profile-based neural network for weight adjustments. The following Bayesian equation is used to calculate the value for judging between the two methods [4].

$$P(H_1 \mid D) = \frac{P(H_1) \times P(D \mid H_1)}{P(H_1) \times P(D \mid H_1) + P(H_2) \times P(D \mid H_2)}$$

To illustrate, let H_1 corresponds to correlation method, and H_2 corresponds to committee method. Since the correlation method was producing better prediction accuracy compared to the committee method, for our first instance we assume that $P(H_1) = 0.51$, and $P(H_2) = 0.49$ (we assign more probability for choosing correlation method as this method produced better prediction accuracy compared to the committee method).

For example if we obtain a prediction accuracy of 71% using the correlation method and a prediction accuracy of 70.5% using the committee method, then $\mathrm{P}(D|H_1) = 0.71$ and $\mathrm{P}(D|H_2) = 0.705$. Bayesian equation then yields:

$$P = \frac{0.51 \times 0.71}{0.51 \times 0.71 + 0.49 \times 0.705} = 0.5117$$

If the probability obtained is greater than or equal to 0.5, we then use the correlation method for calculating the error that is to be back-propagated to the profile-based neural network for weight adjustments.

If for example, we obtain a prediction accuracy of 69% using the correlation method and a prediction accuracy of 72% using the committee method, then $\mathrm{P}(D|H_1) = 0.69$ and $\mathrm{P}(D|H_2) = 0.72$. Bayesian equation then yields:

$$P = \frac{0.51 \times 0.69}{0.51 \times 0.69 + 0.49 \times 0.72} = 49.93$$

If the probability obtained is less than 0.5, we then use the committee method for calculating the error that is to be back-propagated to the profile-based neural network for weight adjustments.

Overall we choose the correlation method for calculating the error for weight adjustments if the probability obtained is greater than or equal to 0.5, otherwise we use the committee method to calculate the error to be back-propagated to the profile-based neural network for weight adjustments.

Similarly we have tested our method using various values of probability for $P(H_1)$ and $P(H_2)$, and always choosing $P(H_1)$ greater than $P(H_2)$. From the several test cases, we concluded that the values of 0.506 for $P(H_1)$ and 0.494 for $P(H_2)$ produced the greatest prediction accuracy. Using the Bayesian approach we obtained a prediction accuracy of 73.3% (Q_3). This method produced the highest protein secondary structure prediction accuracy compared to all the other methods in our research.

7. Simulation Results

The simulations were performed using code written in JAVA programming language on a 3.6 GHz Intel Pentium IV PC with hyper-threading running Microsoft Windows XP with 2GB of RAM and a 160GB hard disk. We used the multi-threading approach for running the GA and SA algorithms, and the decision fusion methods in parallel.

Table 1 provides the summary of the prediction accuracies achieved using various methods in this research.

Table 1. Comparison of prediction accuracy (Q_3) for various approaches

Approach Used	Prediction Accuracy (Q_3)
Profile-based Neural Network	66.8%
Profile-based Neural Network & GA	69.2%
Profile-based Neural Network & SA	68.3%
Decision fusion (Committee method) using Profile-based Neural Network	70.8%
Decision fusion (Correlation method) using Profile-based Neural Network	71.4%
Bayesian Inference method	73.3%

It is clearly evident from Table 1 that the Bayesian inference method improves the prediction accuracy by 2% compared to that of correlation method and overall a prediction accuracy of 6.5% more than the profile-based neural network, which is a significant achievement.

8. Conclusion

In this research the goal was to improve the protein secondary structure prediction accuracy using the Bayesian inference method. Although there exists a variety of protein structure classification algorithms, we believed that further improvement can be attained by finding the best way to combine several methods to lead to a unified better decision. From the research performed we conclude that applying AI algorithms along with decision fusion techniques improved the prediction accuracy compared to that of prediction by neural networks or AI algorithms individually or combined with profile-based neural networks. The simulations results prove that the Bayesian Inference method improved the prediction accuracy over the other decision fusion methods. The main advantage of using this approach is that, it does not comprise the advantages provided by either committee or correlation methods of decision fusion. The future work comprises of using other decision fusion methods such as the clustering method, the fuzzy set method, and the probabilistic method for further improving on the protein secondary structure prediction accuracy.

References

[1] M. A. Abidi and R. C.Gonzales, eds., Data Fusion in Robotics and Machine Intelligence. Academic Press Inc., 1992.

[2] Somasheker Akkaladevi, Ajay K Katangur, Saeid Belkasim, and Yi Pan, “Protein Secondary Structure Prediction using decision fusion of Genetic Algorithm and Simulated Annealing Algorithm,” International Conference on Neural Networks and Brain, Vol. 1, pp. 467-472, October 13-15, 2005, Beijing, China.

[3] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”, Nucleic Acid Research, 1997, 25:3389-3402.

[4] G.Anandalingam and L. Chen, “Linear combination of forecasts: a general bayesian model,” Journal of Forecasting, vol. 8, pp. 199–214, 1989.

[5] Baldi P, Brunak S, Frasconi P, Pollastri G, Soda G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;15:937–946.

[6] Banavar J.R and Maritan A., “Computational Approach to the Protein-Folding Problem”, Proteins: Structure, Function, and Genetics, 2001, 42: 433-435.

[7] Branden, C. and Tooze, J. (1999). Introduction to Protein Structure. Garland Publishing.

[8] T. K. Ho, J. J. Hull, and S. N. Srihari, “Decision combination in multiple classifier systems,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, pp. 66– 75, January 1994.

[9] Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational properties. Proceedings of the National Academy of Sciences of the USA, 79:2554 -- 2588.

[10] R. C. Luo and M. G. Kay, “Multisensor integration and fusion in intelligent systems,” IEEE Transactions on Systems Man and Cybernetics, vol. 19, pp. 901–931, September/ October 1989.

[11] V.D.Mazurov, A.I.Krivonogov, and V.S.Kazantsev, “Solving of optimization and identification problems by the committee methods,” Pattern Recognition, vol. 4, no. 20, pp. 371–378, 1987.

[12] Qian, N. and Sejnowski, T. (1988). Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology, 202:865-884.

[13] Rost, B. and Sander, C. (1993b). Improved prediction of protein secondary structure by use of sequence structure and neural networks. Proceedings of the National Academy of Sciences of the United States of America, 90:7558-7562.

[14] Rost, B. and Sander, C. (1993c). Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology, 232:584-599.

[15] Rost, B. (1996). predicting 1d protein structure by profile based neural networks. Meth. in Enzym., 266:525-539.

[16] Rost, B., Sander, C., and Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. Jour. Mol. Biol., 235:13-26.

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

[18] Simons K. T., Kooperberg C., Huang E. and Baker D., “Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions”, J. Mol. Biol., 1997, 268: 209-25.