Protein Structure Prediction by Fusion, Bayesian Methods

Somasheker Akkaladevi
Virginia State University, USA

Ajay K. Katangur
Texas A&M University – Corpus Christi, USA

Xin Luo
The University of New Mexico, USA

INTRODUCTION

Prediction of protein secondary structure (alpha-helix, beta-sheet, coil) from primary sequence of amino acids is a very challenging and difficult task, and the problem has been approached from several angles. A protein is a sequence of amino acid residues and can thus be considered as a one dimensional chain of 'beads' where each bead correspond 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 (Altschul et al. 1997). Many techniques were used many researchers to predict the protein secondary structure, but the most commonly used technique for protein secondary structure prediction is the neural network (Qian et al. 1988).

This chapter discusses a new method combining profile-based neural networks (Rost et al. 1993b), Simulated Annealing (SA) (Akkaladevi et al. 2005; Simons et al. 1997), Genetic algorithm (GA) (Akkaladevi et al. 2005) and the decision fusion algorithms (Akkaladevi et al. 2005). Researchers used the neural network (Hopfield 1982) 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 (Akkaladevi et al. 2005). Sequence profiles of amino acids are fed as input to the profile-based neural network. The two decision fusion methods improved the prediction accuracy, but noticeably one method worked better in some cases and the other method for some other sequence profiles of amino acids as input (Akkaladevi et al. 2005). Instead of compromising on some of the good solutions that could have generated from either approach, a combination of these two approaches is used for obtaining better prediction accuracy. This criterion is the basis for the Bayesian inference method (Anandalingam et al. 1989; Schmidler et al. 2000; Simons et al. 1997). The results obtained show that the prediction accuracy improves by more than 2% using the combination of the decision fusion approach and the Bayesian inference method.

BACKGROUND

A lot of interesting work has been done on protein secondary structure prediction problem, and over the last 10 to 20 years the methods have gradually improved in accuracy. The most successful application of neural networks (Hopfield 1982) to secondary structure prediction was obtained by Rost and Sander (Rost et al. 1993b; Rost et al. 1993c; Rost 1996; Rost et al. 1994), which resulted in the prediction mail server called PHD (Rost et al. 1993c). Using profile-based neural network and a few other methods, the performance of the network is reported to be up to 67.2% (Rost et al. 1993b).

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) (Banavar et al. 2001; Branden et al. 1999). A typical protein sequence and its conformation class are shown below:

Protein Sequence: ADADADADCCQQFFFAAAQQA-QQA
Conformation Class: HHHH EEEE HHHHHHHH

H stands for Helical, E for 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 (Rost et al. 1993b). It is possible to predict loop regions with higher accuracy than alpha helices or beta sheets (Rost et al. 1993c). The *seven-fold cross-validation* technique is used on the set of 126 non-homologous globular proteins from (Rost & Sander, 1994), which is called the RS126 data set (Rost et al. 1994) for training and testing purpose.

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 (Qian et al. 1988; Rost 1996):

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

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.

PROTEIN SECONDARY STRUCTURE PREDICTION BY VARIOUS APPROACHES

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 (Rost et al. 1994). Orthogonal encoding scheme is used for the input which is sent to the profile-based neural network.

Protein Secondary Structure Prediction using sequence profiles - The profile-based neural network is used for this research. Using profiles at the input level generally has been shown to yield better results than using profiles at the output level (Baldi et al. 1999; Rost et al. 1993b). Using this approach the secondary structure prediction accuracy (Q_3) is 66.8%.

GA and the profile-based Neural Networks for protein secondary structure prediction - The predicted structure from the profile-based neural network is given to GA; the GA does a series of mutation and crossover operations on the predicted structure from the profile-based neural network to generate new solutions (offspring's) (Akkaladevi et al. 2005). After the offspring is generated; 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 accepts or rejects this solution depending on the fitness value, which in this case is the prediction accuracy Q_3. Finally at this point the error value is calculated and back-propagated to adjust the weights of the profile-based neural network. The mutation probability for GA in this research is set at 0.25, number of generation's at 75, population size at 30 and the crossover probability as 100% (Akkaladevi et al. 2005). Using this approach the secondary structure prediction accuracy (Q_3) is 69.2%.

SA and the profile-based Neural Networks for protein secondary structure prediction - The predicted structure from the profile-based neural network is sent to the SA algorithm for further processing by the SA algorithm (Akkaladevi et al. 2005). The SA algorithm 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 than calculated by determining the value of Q_3. This error value is then back-propagated to adjust the weights of 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 (Akkaladevi et al. 2005). Using this approach the secondary structure prediction accuracy (Q_3) is 68.3%.

Prediction of protein secondary structure using the Committee method and the profile-based Neural Network - In the committee based method (Mazurov et al. 1987) of applying decision fusion the secondary structure values are calculated 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. The output obtained from the profile-based neural network, combined profile-based neural network plus GA and combined profile-based neural network plus SA is routed to the decision fusion algorithm, for fusing the solutions as shown in Figure 1 (Akkaladevi et al. 2005).

The decision fusion (Abidi et al. 1992) 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