Available online at www.sciencedirect.com
ScienceDirect
Biomolecular Engineering 24 (2007) 237–243
Biomolecular Engineering
www.elsevier.com/locate/geneanabioeng

Application of expert networks for predicting proteins secondary structure

Sarit Sivan [a,*], Orna Filo [a], Hava Siegelmann [b]

[a] Department of Biomedical Engineering, Technion, Israel Institute of Technology, IIT, Haifa 32000, Israel
[b] Department of Computer Science, University of Massachusetts Amherst, Amherst MA 01003, United States

Received 5 November 2006; received in revised form 5 December 2006; accepted 6 December 2006

Abstract

The present study utilizes expert neural networks for the prediction of proteins secondary structure. We use three independent networks, one for each structure (alpha, beta and coil) as the first-level processing unit; decision upon the chosen structure for each residue is carried out by a second-level, post-processing unit, which utilizes the Chou and Fasman frequency values F_α and F_β in order to strengthen and/or deplete the probability of the specific structure under investigation. The highest prediction case was 76%.

Our method requires primitive computational means and a relatively small training set, while still been comparable to previous work. It is not meant to be an alternative to the determination of secondary structure by means of free energy minimization, integration of dynamic equations of motion or crystallography, which are expensive, time-consuming and complicated, but to provide additional constrains, which might be considered and incorporated into larger computing setups in order to reduce the initial search space for the above methods.

Keywords: Proteins; Secondary structure prediction; Expert neural networks; Chou and Fasman frequency parameters

1. Introduction

The knowledge of protein secondary structure is essential for the understanding of both the mechanisms of folding and the biological activity of proteins. X-ray diffraction has been successful in elucidating the three dimensional structure of many crystallized proteins. Although this method can be very accurate, it is expensive and time-consuming. Furthermore many membranes and ribosomal proteins have not yet yielded suitable crystals, so that other approaches must be explored to give the structural information required. Since experimental evidence shows that the native conformation of a protein is coded within its amino acid sequence (Anfinsen et al., 1961), many efforts have been made to predict the protein secondary and tertiary structure from the sequence data.

Following the pioneering work of Pauling and Corey (1951), which suggests that proteins form certain local conformations as helices and strands, many workers used different methods to predict protein secondary structure (Szent-Gyorgyi and Cohen, 1957; Periti et al., 1967; Ptitsyn, 1969; Pain and Robson, 1970; Robson and Pain, 1971). These methods exploit, in different ways, the correlation between amino acid and the local secondary structure, i.e. neighbors effect of no more than 10 amino acids away. The average success of these methods is 50–53% on three types of secondary structures (alpha-helix, beta-sheet, and coil) (Nishikawa, 1983; Kabsch and Sander, 1983a,b). Secondary structure predictions have been performed by various methods. These methods make use of the physicochemical characteristics of the amino acids (Lim, 1974; Ptitsyn and Finkelstein, 1983), sequence homology (Levin et al., 1986; Nishikawa and Ooi, 1986; Zvelebil et al., 1986), pattern matching (Cohen et al., 1983, 1986; Taylor and Thornton, 1983; Rooman et al., 1989; King and Sternberg, 1990; Presnell et al., 1992), statistical analyses of proteins with known structure (Wu and Kabat, 1971, 1973; Chou and Fasman, 1974a,b; Nagano, 1977; Garnier et al., 1978; Maxfield and Scheraga, 1979; Gibrat et al., 1987; Biou et al., 1988; Di Francesco et al., 1997; Fasman, 1989; Garratt et al., 1991; Muggleton et al., 1992), and neural network (Bohr et al., 1988, 1993; Qian and Sejnowski, 1988; Holley and Karplus, 1989;

* Corresponding author at: Julius Silver Institute of Biomedical Sciences, Department of Biomedical Engineering, Technion, Israel Institute of Technology, IIT, Haifa 32000, Israel. Tel.: +972 4 8294150; fax: +972 4 8294599.
E-mail address: sarit@bm.technion.ac.il (S. Sivan).

1389-0344/$ – see front matter
doi:10.1016/j.bioeng.2006.12.001

Kneller et al., 1990; Hirst and Sternberg, 1992; Maclin and Shavlik, 1993; Stolorz et al., 1992; Zhang et al., 1992; Rost and Sander, 1993a,b).

A promising approach in the area of secondary structure prediction is the use of neural network methods (Bohm, 1996). One of the first examples for this method used 48 proteins in the learning dataset, in order to teach the relationship between primary sequence and secondary structure to the neural network (Holley and Karplus, 1989). The overall accuracies achieved in this study and in a similar one (Qian and Sejnowski, 1988) were 63% and 64.3%, respectively, which had no major improvement compared with traditional methods of secondary structure prediction by statistical and knowledge-based methods.

Following the pioneering work of Qian and Sejnowski (1988), many new computational techniques involving neural networks for the prediction of proteins secondary structure were introduced (Holley and Karplus, 1989; Rost and Sander, 1993a,b, 1994; Hua and Sun, 2001; Armano et al., 2005; Lee et al., 2006; Huang et al., 2005; Ceroni et al., 2005; Ruan et al., 2005; Wood and Hirst, 2005; Meiler and Baker, 2003; Hering et al., 2003; Cai et al., 2002, 2003; Kaur and Raghava, 2003; Shepherd et al., 1999, 2003; Pal and Basu, 2001; Petersen et al., 2000; Cuff and Barton, 2000; Chandonia and Karplus, 1995, 1996, 1999; Kawabata and Doi, 1997; Barlow, 1995; Salamov and Solovyev, 1995); the average prediction accuracy achieved varies between 70% and 80%. In order to improve prediction accuracy, several studies applied sophisticated network structures such as hierarchical (Jordan and Jacobs, 1994; Huang et al., 2005; Barlow, 1995), cascade (Wood and Hirst, 2005) and multiple experts networks (Armano et al., 2005). Others combined additional structural information in the network input, for example, amino acid composition (Lee et al., 2006), interaction graphs (Ceroni et al., 2005), tertiary (Meiler and Baker, 2003; Chandonia and Karplus, 1995) and secondary (Rost and Sander, 1993a,b; Shepherd et al., 1999) structure information, information on the probabilities of residues buried in the protein core or on the protein surface (Vieth et al., 1992) and multiple sequence alignment profiles (Rost and Sander, 1993a,b, 1994; Cuff and Barton, 2000). Numerous methods involve preprocessing of protein sequence data using Fourier transform (Shepherd et al., 2003) and binary word encoding (Kawabata and Doi, 1997). Other approaches such as adaptive neuro-fuzzy inference system (Hering et al., 2003) and nearest neighbor algorithm (Salamov and Solovyev, 1995) combine additional classification algorithms with neural networks. Decoding the networks output in order to estimate the probability of finding a secondary structure at a specific position (Chandonia and Karplus, 1999) also provides more accurate prediction.

Our approach is to use three independent expert neural networks, one for each structure (alpha, beta and coil) as the first-level processing unit; decision upon the chosen structure for each residue is carried out by a second-level, post-processing unit, which utilizes the Chou and Fasman statistical frequency values F_α and F_β. This architecture takes into account the 'neighbors' effect and in turn, strengthens and/or depletes the probability of any structure under investigation to be part of a specific secondary structure.

Despite the simplicity of the networks presented in this work, they have the ability to deal with complex classification problems. This advantage was accomplished by separation of the comprehensive problem into three sub-classification items. Implementation of divide-and-conquer algorithms to deal with a complex problem by dividing it into simpler problems whose solutions can be combined to yield an answer to the complex problem was suggested by Jordan and Jacobs (1994).

2. Methods

2.1. Database

The secondary structure assignment used in this study was based on the work of Kabsch and Sander (1983a,b). Their DSSP program was used to classify known structures in the Brookhaven Protein Data Bank (BPDB) as helices and sheets. Residues that are neither helices nor sheets are classified as coil. Following Qian and Sejnowski (1988), we selected a representative sample of proteins that limited the number of almost identical sequence, such as the similar types of hemoglobin.

2.2. Network formulation and training

Three expert nets were applied in this work; each structure (alpha, beta and coil) is represented by a separate network (Fig. 1). All the networks used were feed-forward nets utilizing the back-propagation algorithm and the Sigmoid-Logistic as their activation function. Calculations were carried out using MATLAB. The input vector for each expert net encodes a moving window

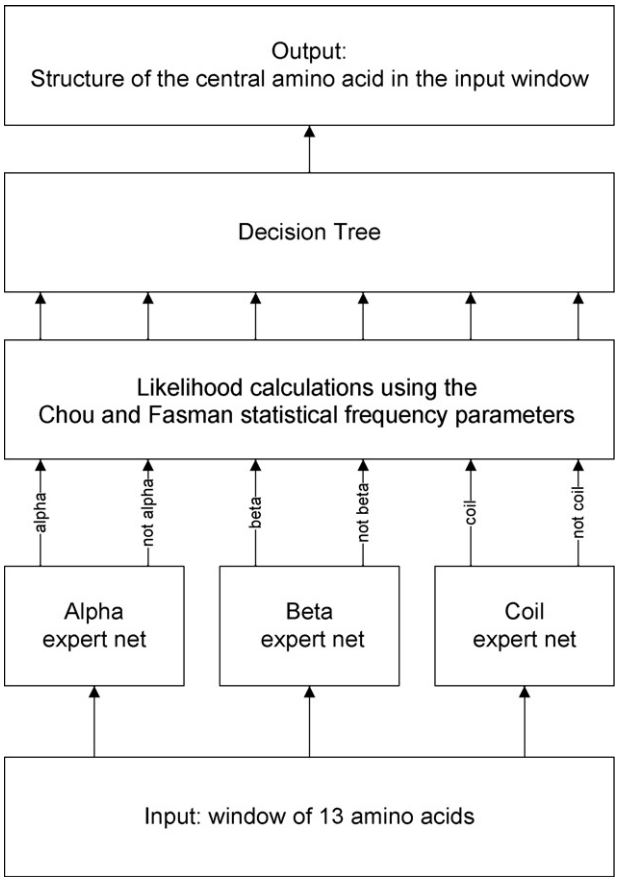

Fig. 1. A schematic description of the expert neural network used for the prediction of proteins secondary structure.