Journal of Theoretical Biology 361 (2014) 182–189

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Knowledge base and neural network approach for protein secondary structure prediction

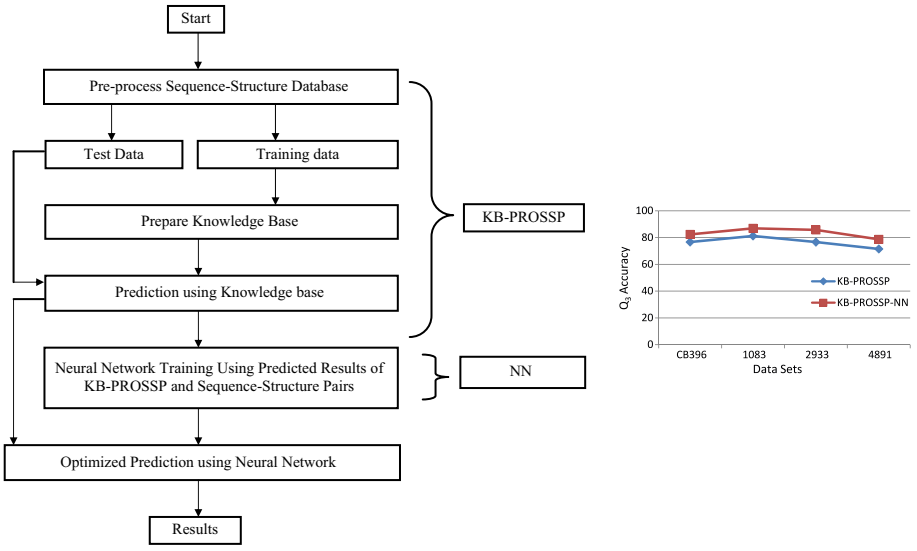
Maulika S. Patel [a,1], Himanshu S. Mazumdar [b]

[a] Department of Computer Engineering, G H Patel College of Engineering & Technology, Vallabh Vidyanagar, Gujarat, India
[b] Research & Development Center, Dharmsinh Desai University, Nadiad, Gujarat, India

HIGHLIGHTS

- A novel neuro-statistical algorithm using knowledge base and neural network for PSSP is proposed.
- Association of 5-residue word with corresponding secondary structure forms the knowledge base.
- Lateral and hierarchical validation is employed for PSSP.
- A Backpropogation neural network is used to model the exceptions in the knowledge base.
- The Q_3 accuracy of 90% and 82% is achieved on the RS126 and CB396 test data sets respectively.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 16 December 2013
Received in revised form
1 August 2014
Accepted 4 August 2014
Available online 14 August 2014

Keywords:
5-Residue words
Knowledge base
Lateral association and validation
Hierarchical validation
Backpropogation neural network

ABSTRACT

Protein structure prediction is of great relevance given the abundant genomic and proteomic data generated by the genome sequencing projects. Protein secondary structure prediction is addressed as a sub task in determining the protein tertiary structure and function. In this paper, a novel algorithm, KB-PROSSP-NN, which is a combination of knowledge base and modeling of the exceptions in the knowledge base using neural networks for protein secondary structure prediction (PSSP), is proposed. The knowledge base is derived from a proteomic sequence–structure database and consists of the statistics of association between the 5-residue words and corresponding secondary structure. The predicted results obtained using knowledge base are refined with a Backpropogation neural network algorithm. Neural net models the exceptions of the knowledge base. The Q_3 accuracy of 90% and 82% is achieved on the RS126 and CB396 test sets respectively which suggest improvement over existing state of art methods.

E-mail addresses: maulika.sandip@gmail.com (M.S. Patel), hsmazumdar@hotmail.com (H.S. Mazumdar).
[1] Tel.: +91 2692 231651; fax: +91 2692 236896.

http://dx.doi.org/10.1016/j.jtbi.2014.08.005
0022-5193/

1. Introduction

Proteins are formed by the transcription and translation of the one dimensional DNA sequences to the three-dimensional (3-D) molecules capable of performing diverse functions. Their secondary structures—helix, strand or a coil are hard to predict. Their folding methods are harder to expound. Experimental techniques such as X-ray Crystallography and Nuclear Magnetic Resonance (NMR) face challenges such as cost, labor, expertise and time. Protein structure identification is necessary to understand the function of protein, required for drug design and also predicting the state of a disease. Often the problem of structure prediction is reduced to the secondary structure prediction problem. Protein secondary structure prediction (PSSP) is significant and relevant as it allows drawing conclusions on fold classification and providing important clues for 3-D structure prediction. Results of secondary structure predictions are used to classify proteins such as all-α or all-β proteins. The research in PSSP dates back to 1970s while computational techniques to address the issue of protein structure prediction gained a lot of attention recently.

The DSSP program (Kabsch and Sander, 1983) classifies each amino acid residue into eight classes, i.e., B, E, G, H, I, S and T. We use '.' instead of the blank entry for clarity. These are typically collapsed into the three standard classes: G, H and I=> Helix (H), B and E=> beta-strand (E) and rest as Coils(C). This grouping suggested by DSSP is assumed to be harder than other possible groupings. The protein sequence comprises of 'R' residues and can have 'm' possible secondary structure states. The problem can be understood as a characterization problem where each of the residues is assigned a secondary state. The per-residue accuracy, Q_m, for m structural states, is defined as the ratio of the number of residues for which the secondary structure is correctly assigned (C) to the total number of residues (R) in a protein sequence (Przybylski and Rost, 2007).

$$Q_m = 100 \times \frac{C}{R} \tag{1}$$

For three class classification, $m=3$ and is referred as Q_3 accuracy. A refined accuracy index, called Q_8, is proposed to evaluate algorithms of secondary structure prediction (Zhang and Zhang, 2001). The per residue accuracy is expected to increase by 3–4% in case of 3 class as compared to 8-class. The interclass distance is comparatively more in case of 3-class and hence the border line misclassifications are less than 8-class where the inter class distance is lesser. However, 8-class classification embeds more specific information. The literature review below discusses the various approaches proposed so far for the PSSP problem highlighting their performance in terms of their Q_3 or Q_8 accuracy.

A detailed review of the prediction methods for globular proteins is given by Rost (2003). The PHD method (Rost and Sander, 1994) makes use of sequence profiles and neural networks to predict with an accuracy of around 70%. The association of knowledge base and neural networks date backs to 1990s (Maclin and Shavlik, 1993). The methods employing neural networks as classifier have largely used an ensemble of neural networks or cascaded networks for increased accuracy. The accuracy of the models using neural networks alone is far less than the estimated upper limit of 88% as suggested by Rost (2003). Recurrent neural networks used by Chen and Chaudhari (2007) and multiple classifiers used by Ouali and King (2000) are able to achieve Q_3 accuracy of 74% and 76%, respectively. Wang et al. (2011) experimented with conditional neural fields, a combination of conditional random fields and neural network and reported a Q_8 accuracy of 64.9%. Support Vector Machines (SVM) gained a lot of consideration by the researchers for PSSP (Reyaz-Ahmed and Zhang, 2007; Kim and Park, 2003). Some methods give a window of amino acid residues as input and the secondary structure of the central amino acid residue is predicted. Rost and Sander (1993) used a window size of 13 while it is 15 in case of Chatterjee et al. (2011). A cascade of support vector machines is used as classifier in Kim and Park (2003). The second stage refines the output of the first stage. Reyaz-Ahmed and Zhang (2007) used a combination of genetic algorithms, neural networks and support vector machines (GNSVM). SSpro2.0 proposed by Pollastri et al. (2002) uses profiles from BLAST and PSI-BLAST along with bidirectional recurrent neural networks. SSpro 2.0 outperformed the above methods with an accuracy of 78.13%. YASPIN, a method proposed by Lin et al. (2005) used a single neural network for PSSP and hidden markov model for optimizing the results. Leong et al. took a rule based data-mining approach that identifies dependencies between amino acids in a protein sequence and generates rules to predict secondary structure. Their method is named as RT-RICO (Relaxed Threshold Rule Induction from Coverings) (Leong et al., 2011). FLOPRED, proposed by Saraswathi et al. (2012), makes use of knowledge-base, a Neural Network based Extreme Learning Machine (ELM) and advanced Particle Swarm Optimization (PSO) techniques to predict protein secondary structures. Wu et al. (2004) used a knowledge base for PSSP and their method is called HYPROSP. The knowledge base is composed of amino acid residue words along with their secondary structure. The method used a measure termed 'match rate' that suggests the amount of information the target protein can extract from the knowledge base. One method which is found to be nearest to our proposed method is the one described by Kountouris et al. (2012). The similarity is the two phase approach, one for prediction and the second for smoothing the predictions. However, the existing method in Kountouris et al. (2012) used 8 BRNNs (Bidirectional Recurrent Neural Network) in the first phase, an average of which is fed to various filtering methods. A trend of using existing prediction servers and building a consensus model is also becoming popular. Such an approach is demonstrated by Yan, Marcus and Kurgan in their method described in Yan et al. (2014). They too used SVM for filtering to achieve the Q_3 accuracy of 85% on a 5-fold cross validation data set.

Prediction accuracy largely depends on three parameters: (1) sequence database size (2) algorithm or method and (3) better database search methods (Rost, 2002). It is anticipated by Rost and Sander (Przybylski and Rost, 2007) that the increase in the protein database sizes will lead to the increase in accuracy of protein structure prediction. We have exploited the availability of the huge sequence structure database, coupled with the development of a novel PSSP algorithm. By finding and using the hidden and embedded association of amino-acid residues together with a Backpropogation neural network, the proposed method has given an accuracy of 90% for a 3-class classification.

2. Materials and methods

The proposed algorithm, KB-PROSSP-NN, uses a hybrid approach. The knowledge based method (KB-PROSSP) and neural network (NN) for optimized results are used in cascade. Fig. 1 presents the flow chart describing the KB-PROSSP-NN method of PSSP while the algorithmic details are explained in Fig. 2. KB-PROSSP-NN is a two phase algorithm. The first phase, KB-PROSSP, is a hierarchical lateral-validation technique for PSSP. The technique uses a knowledge base built from the statistics of association between the 5-residue words and corresponding secondary structure. The second phase uses a pre-trained Backpropogation neural network that corrects the discrepancies of the knowledge base used by KB-PROSSP. The implementation of the proposed method is carried out using the C#.NET environment.