Improved Prediction of Protein Secondary Structures Using Adaptively Weighted Profiles

Gouchol Pok\textsuperscript{1}, Keun Ho Ryu\textsuperscript{2}, and Yong J. Chung\textsuperscript{3}

\textsuperscript{1} Yanbian University of Science and Technology, Department of Computer Science, Yanji, Jilin Province, China
\texttt{pok@ybust.edu.cn}

\textsuperscript{2} Chungbuk National University, Department of Computer Science
Cheongju, Chungbuk, Korea
\texttt{khryu@dblab.chungbuk.ac.kr}

\textsuperscript{3} Chungbuk National University, Department of Biochemistry
Cheongju, Chungbuk, Korea
\texttt{chungyj@chungbuk.ac.kr}

Abstract. Prediction of protein secondary structures from amino acid sequences is a useful intermediate step for further elucidation of native, three-dimensional conformation of proteins. Currently, most predictors are based on machine learning approaches with a short fixed-size input window scanning over the amino acid sequence. The center of the window corresponds to the prediction site where the prediction is performed by utilizing the properties of neighboring amino acid residues. By nature, most machine learning approaches consider feature vectors as position-independent in terms of feature components. As such, for the secondary structure prediction problem, most existing approaches do not take into account the distance of amino acid residues from the center residue. We have studied on how the prediction performance can be affected by imposing different weights on the features according to the distance of residues from the center residue, and in this work, we propose an adaptive weighting scheme to improve prediction accuracy.

1 Introduction

Structural characterization of a protein through prediction is one of the most challenging problems in computational molecular biology. The prediction process itself involves many lesser but still challenging intermediate problems. Among others, one important problem worth of special attention is the prediction of secondary structures from a given amino acid residue sequence. If one knows how a protein is composed of the secondary structural elements, their packing ways give an insight into the possible tertiary structures. The secondary structures refer to the regularly repeated folding patterns sustained by hydrogen bonds and conventionally grouped into three classes: $\alpha$-helices, $\beta$-sheets and coils representing all the other structures without regularities. The arrangement of twenty amino acids in a primary sequence is not randomly composed, but rather shows some kinds of
preferences and correlations between the residues in the sequence. The rule or principle governing these patterns behind the structure formation has been one of the most classical and frequent research issues in computational molecular biology. This is also the area where machine learning approaches have been successfully applied to. The main reason why these machine learning approaches have been widely used is partly because the prediction process is independent of any direct information relating to protein conformation. Following the early works based on the statistical analysis and information theory, which are categorized as the first generation methods, the advent of neural networks provided a useful blackbox model to estimate a unknown function mapping the input sequences to secondary structures. The neural network architecture proposed by Qian and Sejnowski was the fully connected feed-forward network with one hidden layer and a local window of length 13 amino acids. Prediction performance of the approaches based on single sequences and local windows are limited to about 65% accuracy. Qian and Sejnowski also remarked that prediction performance between a single layer network and a network with a hidden layer had no considerable difference. This indicates that the input data do not contain the second order correlations among amino acid residues in the window. The authors experimented further with artificial structures of second order and drew a conclusion that no second order information was present. In addition to this discouraging observation, other main problems with the second generation approaches include that prediction results were not good enough for practical use. In particular, the prediction accuracy for $\beta$-sheets was slightly better than a naive guess. When a protein folds into a natural conformation, amino acids that are far away on the primary chain may come into close contact in the tertiary space. This distal relationship can not be captured by a short window, which seems to be a main reason for poor performance in predicting $\beta$-sheets. This problem can not be addressed by merely considering longer windows of residues because of the over-fitting problem in training neural networks. Therefore, one needs to devise a different way other than merely increasing the window size when incorporating long range interactions into prediction process. The first method that achieved a prediction accuracy of over 70% is the PHD prediction system by Rost and Sander, which later reached up to a level of 74%. While they introduced a number of techniques including early stopping and ensemble of modular networks, the breakthrough in the performance improvement resulted mainly from the use of profiles of the input amino acid sequence derived from the multiple sequence alignment of homologous proteins. As the amino acid sequence profiles contain information pertaining to inter-sequence relationships, one can naturally expect that profile-based approaches outperform the single sequence-based ones. More importantly, as evidenced by the observation that the space of possible amino acid sequences is highly constrained by protein structures, profiles contain a certain kind of long-range information in themselves. However, it should be noted that profiles also discard the information pertaining to intra-sequence correlations among amino acid residues. Once the use of profiles made a breakthrough of 70% level for the prediction accuracy, virtually all new prediction methods since then adopted