Dependence of the Accuracy of Protein Secondary Structure Prediction on Long-Range Interactions

Daisuke Kihara$^{1,2}$
dkihara@purdue.edu

$^1$ Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN 47906, USA
$^2$ Department of Computer Science, College of Science, Purdue University, West Lafayette, IN 47906, USA

Keywords: protein secondary structure prediction, long-range interaction, residue contact order

1 Introduction

Secondary structure prediction of a protein from its amino acid sequence has been studied for a long years in bioinformatics. Various techniques including artificial neural network, the support vector machine, hidden Markov models, have been employed to this problem. Although a gradual improvement of the prediction accuracy has been observed, which is due to the sophistication of algorithms and also to the growth of protein structure and sequence databases, the current methods still cannot achieve a near perfect prediction accuracy. It has been discussed that one of the reason for the limitation of the accuracy is the fact that most of the current algorithms neglect long-range interactions of amino acid residues. For example, typical algorithms which use neural network or the support vector machine predict a secondary structure to a local sequence of a certain window size.

Interestingly, contrary to the general consensus and discussion, so far there are not many systematic studies on the influence of long-range interactions on the prediction accuracy of secondary structures. Indeed previous studies [1, 2] rather concluded that the role of long-range interaction are not very significant as it has been believed.

Here we directly address the effect of the long-range interaction on the accuracy of secondary structure prediction by introducing a novel measure of the separation of contacting residues in terms of the relative position in the sequence. Unlike previous studies, we do find that prediction accuracy drops as residues have contacts with more distant residues. These results are published on a recent issue on Protein Science [3, 4].

2 Materials and Methods

The benchmark data set consists of 2777 non-redundant protein sequences selected from PDB database with a sequence identity threshold of 30%. Three secondary structure prediction methods, PSIPRED, Jnet, and PREDATOR are used in this study.

We introduced a novel measure, the residue contact order (RCO), which describe the separation of contacting residues on a primary sequence. The RCO of the residue $i$ is defined as:

$$RCO_i = \frac{1}{n} \sum_{j \neq i} |i - j| \delta_{ij}$$

Here $n$ is the number of contacts between the $i$-th residue and the others ($j$). $\delta_{ij} = 1$ when residue $i$ and $j$ are in contact, and 0 otherwise. Two residues are considered to be in contact if any pair of heavy atoms from each residue locate closer than a threshold value. After the RCO for individual residues are calculated, the average RCO values in a smoothing window is assigned to the center residue of the window.

We run the three prediction algorithms on the benchmark proteins, then residues in the proteins are grouped by their RCO value.
3 Results

The overall residue-based prediction accuracy (Q3) for each algorithm was 79.8% for PSIPRED, 76.3% for Jnet and 69.0% for PREDATOR. Our results show that α helices can be better predicted than β strands, which is consistent with other previous studies. Generally the accuracy is not affected by the size of proteins. We also confirmed that the accuracy is not much affected by the overall contact order and the relative contact order of the proteins.

Now we examine the effect of long-range interactions on the accuracy by using the RCO of residues. Figure 1 shows the Q3 accuracy of the three algorithms relative to the RCO. Here the average RCO value for the secondary structure segments is used. Figure 1 clearly shows that there is a strong correlation between the RCO and the accuracy. The drop of the accuracy is not algorithm-specific. It is also confirmed that this negative correlation is statistically significant by the $\chi^2$ test.

This negative correlation is seen for residues in α helices and β strands, but not in the other (loop) regions. It has been discussed that one of the reason why the prediction accuracy of β strands is lower than α helices is because β strands often make hydrogen bonds with another one which is distant on the primary sequence to form a β sheets. But here not only β strands but also α helices are strongly affected by long-range contacts. As for β strands, the prediction accuracy also drop if it has a low RCO. This corresponds to a situation that the β strand does not have a hydrogen bond partner in its own chain but in a different chain of a protein complex.

3 Discussions

Here we report for the first time that the negative influence on the prediction accuracy of protein secondary structure do actually exist even for modern prediction algorithms. Typically residues which have a high RCO and mispredicted sit on an interface of domains. This type of residues on an interface are not abundant even in a current large database, thus the sequence pattern of them can not be well captured by machine learning techniques used in prediction methods. Contrary to the previous studies, our observation strongly implies that long-range interaction is indeed a dominant factor which determines the local structure of proteins. Therefore it is very important to consider the long-range interactions to make a breakthrough in protein structure prediction.

References


