Protein Secondary Structure Analysis Using Neural Network

K. Nakata
nakata@nihs.nihs.go.jp

Division of Chem-Bio Informatics, National Institute of Health Sciences
18-1, Kamiyoga 1-chome, Setagaya-ku, Tokyo 158, Japan

Although various approaches for the protein secondary structure prediction were reported, using the empirical method, the neural network algorithm, the joint prediction method, etc...., the prediction accuracy was not so good. Incorporating specific features on the amino acid sequences into the neural network, we tried to analyze the protein secondary structure.

From PDB database, we picked up three dimensional data of proteins which were used by Qian and Sejnowski [1] and Muskal and Kim [2]. The secondary structures were calculated using Kabsch and Sander method [3]. We made the data sets of 20 amino acids in α helix, β sheet and the others, which consist of nine residue segments including the flanking residues. Using these data sets, we incorporated the following features in the neural network.

a) amino acid
b) two consecutive amino acids
c) two amino acids with an interval of one residue
d) two amino acids with an interval of two residues
e) two amino acids with an interval of three residues
f) three consecutive amino acids
g) hydrophobicity
h) hydrophilicity
i) electric charge and polarity

中田琴子：国立衛生試験所化学物質情報部、〒 158 東京都世田谷区上用賀 1 丁目 18-1

Besides these neighboring residue information, we considered the pairs of residues in β sheet structure. Three consecutive residues in β sheet structure were listed up, with the pairing residues. The feature of residue pairs was incorporated in the neural network, as well. Analyzing all above features, we tried to predict the secondary structure of protein sequences.

Combining all information in the protein sequences, the average accuracy was around 91% in the β sheet structure prediction. We are improving our method for larger protein sequences.

References

[1] N. Qian and T. J. Sejnowski, Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. Vol. 202, 865-884, 1988

[2] S. M. Muskal and S. H. Kim, Predicting Protein Secondary Structure Content: A Tandem Neural Network Approach. J. Mol. Biol. Vol. 225, 713-717, 1992

[3] W. Kabsch and D. Sander, Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers, Vol. 22, 2577-2637, 1983