

Signal Peptide Sequence Analysis and Cleavage Site Prediction

Takehito Yamada¹
ce77732@isc.meiji.ac.jp

Yoichi Hayashi¹
hayashiy@cs.meiji.ac.jp

¹ Department of Computer Science, Meiji University, 1-1-1 Higashimita, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan

Abstract

Neural networks are often used in protein sequence analysis. However, the results are unreliable, mainly because of the “black box” of neural network learning results and the fact that computer scientists treat amino acid sequence data as strings. In this paper, we show that neural networks can identify cleavage sites of signal peptides with high accuracy and reliability, using highly precise annotated datasets and sequence analysis using physicochemical properties.

Keywords: cleavage site, back propagation neural network, signal peptide, signal anchor

1 Introduction

First, we collected a biological dataset with high credibility and analyzed it using statistical techniques. The analysis confirmed the biological meaning of the dataset and did not treat the dataset as strings. Then, we confirmed the existence of specific patterns at the cleavage site of a signal peptide sequence. Finally, we developed neural networks to learn the patterns from the results of analysis and tried to identify the cleavage site of the signal peptide using them.

2 Method and Results

2.1 Data collection

The data used in this study were human signal peptides and signal anchors from SWISS-PROT database ver. 54.0. For cleavage site prediction, signal peptide data with cleavage sites were used as positive control, and signal anchor data without cleavage sites were used as negative control. To improve dataset accuracy, we excluded both uncertainty data (such as “Potential” or “Probable”) and redundancy data with sequence similarity more than 80%.

2.2 Sequence analysis

To analyze protein sequences, we used four physicochemical properties of an amino acid residue: hydrophobicity, hydrophilicity, polarity, and side-chain volume [1]. Each property of the amino acid residue,

the signal peptide, and the signal anchor is expressed numerically and analyzed statistically. The results of statistical analysis revealed different sequence patterns between the signal peptide and the signal anchor. Also, specific patterns were found at the cleavage site. This difference in pattern is one of the vital factors for cleavage site prediction.

2.3 Network architecture

To learn the specific pattern of a cleavage site in a protein sequence, we used a neural network. The neural network is suitable for the recognition of patterns, such as sounds or strings. For the network architecture, we have adopted a multi-layered structure called Back Propagation Neural Network (BPNN). This structure can be built comparatively easily in various neural networks. Figure 1 shows 48 units in the input layer, 12 units in the middle layer, and one unit in the output layer.

2.4 Learning results and accuracy evaluation

We collected 1109 entries (positive 740, negative 369) from SWISS-PROT, randomly divided them into 5 datasets, and tested them with a method called “5-fold cross validation.” After learning, we evaluated the accuracy and the success rate (Table 1).

Table 1: Results of Cleavage Site Identification

dataset	Training		test	
	accuracy	success	accuracy	success
1	96.06	0.97	94.12	0.96
2	96.51	0.97	90.99	0.93
3	96.17	0.97	92.79	0.95
4	95.60	0.97	92.34	0.94
5	95.60	0.97	92.79	0.95

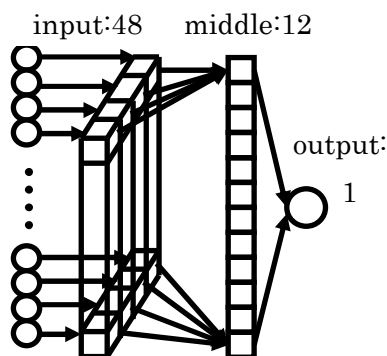


Fig 1. Neural Network Architecture

3 Discussions

We succeeded in the accurate identification of the cleavage site (see Table 1). Our results proved that even if BPNN has a comparatively simple structure, it can predict cleavage site with high accuracy and reliability, given highly precise annotated data collection. This shows the necessity of highly precise data collection and deep knowledge of dataset for protein sequence analysis using neural networks.

Using our algorithms based on neural networks, we will attempt to not only identify signal peptide, but also predict the position of cleavage site. We will also confirm whether highly accurate results can be obtained even if data collection is extended to, for example, other eukaryotes.

References

- [1] G. Schneider, S. Rohlk, and P. Wrede, “Analysis of Cleavage-site Patterns in Protein Precursor Sequences with a Perceptron-type Neural Network”, *Biophys. Biochim. Research Commun.* vol. 194, no. 2, pp. 951-959, 1993.

Acknowledgement

We thank Dr. Yuri Ikeda (Dept. of Electronics & Bioinformatics, Meiji Univ.) for providing cleavage site data.