Assignment of Certainty-Factor Parameters with a Given Reasoning Tree for the Prediction of Protein Localization Sites

Kenta Nakai [1]
nakai@nibb.ac.jp

Ayumi Shinohara [2]
ayumi@rifis.kyushu-u.ac.jp

Satoru Miyano [2]
miyano@rifis.kyushu-u.ac.jp

[1] National Institute for Basic Biology
Myodaiji, Okazaki 444 Japan

[2] Research Institute of Fundamental Information Science, Kyushu University
6-10-1 Hakosaki, Higashi-ku, Fukuoka 812 Japan

Abstract

In this age of large-scale sequencing, we have many "potentially expressed" amino acid sequences of unknown function. Characterization of such sequences by computers is undoubtedly useful for further experimental analyses. We have developed a knowledge-based system PSORT for characterizing various sorting signals potentially coded in amino acid sequences and for predicting their final localization sites in cells [1, 2]. The system calculates the probability (certainty factor) of an input protein to be localized at each candidate site. One of the difficulties of our system is that, since it has many adjustable parameters, optimization of them to a given training data is difficult. Therefore, incorporation of recent knowledge into the system has not been easy. We present here a simple scheme for assigning certainty-factor parameters with a given reasoning tree.

Since the size of training data, *i.e.*, sequences of known localization sites, is not large in most cases, we must suppress the number of parameters as possible. In this case, use of our knowledge on the reasoning flow is favorable. Such a flow can be organized into a reasoning tree, in which an input flux is divided into thinner flows on a step-by-step basis according to some characteristic values calculated from the input sequence (Fig. 1). Its final outputs are flows corresponding to candidate localization sites. In this stage, the amount of each flow can be interpreted as the corresponding certainty factor. Thus, the problem is how to find appropriate

[1]中井 謙太：岡崎国立共同研究機構基礎生物学研究所，〒444 岡崎市明大寺町字西郷中 38
[2]篠原歩, 宮野 悟：九州大学理学部基礎情報学研究施設，〒812 福岡市東区箱崎 6-10-1

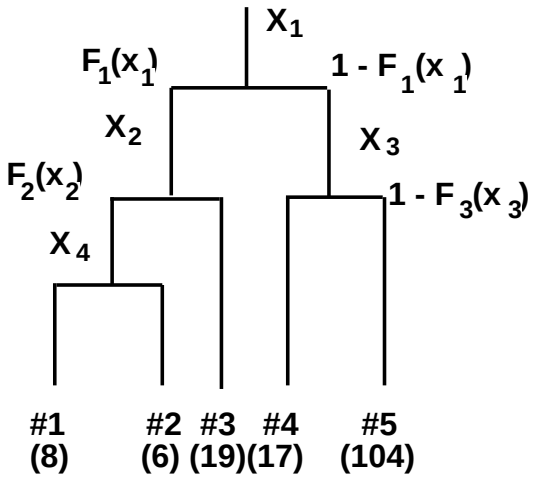

Figure 1: The flow of reasoning goes from the top to the bottom. At each branch point, say, step p, a characteristic value x_p is calculated and the flow is divided in the ratio of $F_p(x_p) : 1-F_p(x_p)$. The bottom nodes correspond to candidate localization sites, where the numbers of data we used are given in parentheses.

functions that transform a characteristic value at each step in an optimized performance for the classification of training data. We used the following formula for that function:

$$F_p(x_p(i)) = \frac{1}{1 + exp(-10 \times (x_p(i) - b_p))}$$

where $x_p(i)$ represents a characteristic value of a sequence i at the step p, *e.g.*, propensity that the input sequence i encodes a membrane protein, and b_p is a threshold value which is obtained by the criterion that can classify the training data at step p with least mistakes. The certainty factor for localizing a candidate site is thus calculated as a probability to choose the corresponding path, *e.g.*, the certainty factor for a protein i to localize at the site #3 is $F_1(i) \times F_2(i) \times (1 - F_4(i))$ in Fig. 1.

To test the validity of our model, we prepared 156 sequences of *Bacillus subtilis* whose localization sites are the prediction results of PSORT. The cross-validation test showed rather good result. Thus, although there is no theoretical proof that our model always gives good results, it will be hopefully used for future improvement of PSORT. Moreover, because of its simplicity, this method may be generally used to interpret unknown sequence data with the latest knowledge of molecular cell biology.

References

[1] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in Gram-negative bacteria," *PROTEINS: Struct., Funct., Genet.*, Vol. 11, pp. 95-110, 1991.

[2] K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," *Genomics*, Vol. 14, pp. 897-911, 1992.