A Probabilistic Inference System for The Prediction of Subcellular Localization Sites of Proteins: Application to *E. coli* Data

Paul Horton [1]
paulh@cs.berkeley.edu

Kenta Nakai [2]
nakai@imcb.osaka-u.ac.jp

[1] Computer Science Division
University of California, Berkeley, CA 94720, USA

[2] Institute of Molecular and Cellular Biology
Osaka University, 1-3 Yamada-oka, Suita 565, Japan

Abstract

We have proposed that the prediction of protein subcellular localization sites can provide a good clue for the characterization of open reading frames of unknown function [1, 2]. Our program, PSORT, has been used by a number of researchers through the Internet [3]. PSORT was originally written in the style of a 'if-then' rule-based system. Although this style has the merit of a great versatility in coding inference pathways, re-optimization of numeric parameters for either given training data or expanded rules needs an expert's manual work. Clearly, this character is not well suited for the rapidly-progressing state of genome analyses. Thus, we have been studying other mathematical models for inference that allow at least semi-automatic optimization of numeric parameters. Last year, we introduced a simple model, the water-flow model, that automatically finds required threshold parameters [4]. This model showed sufficiently high discrimination power for model data. Here, we describe an improved model and report its predictability when applied to more realistic data.

We first collected *E. coli* amino acid sequences of known subcellular localization sites from the PROSITE database (Rel. 31). Excluding hypothetical information, 336 sequences were collected in total. They were classified into the following 8 groups: lipoproteins at the inner membrane (imL), lipoproteins at the outer membrane (omL), inner membrane proteins with a cleavable signal sequence (imS), typical outer membrane proteins (om), periplasmic proteins (pp), inner membrane proteins with a signal- anchor

[2]中井謙太：大阪大学細胞生体工学センター，〒 565 吹田市山田丘 1-3

signal (imU), inner membrane proteins with an internal signal (im), and cytoplasmic proteins including peripheral inner membrane proteins (cp). Since the precise information for topogenic signals is lacking in the database, this classification partly uses the prediction result of PSORT. The number of members varies from 2 to 143 for each group.

Currently, the reasoning tree is the same as the one used in the previous study [1]. However, we employed a probabilistic inference model. Basically, the model is a kind of "water-flow model", as is the model proposed last year [4]. The most important difference is the use of a discrete set of conditional probability values for categorized ranges of the characteristic value at each node. These categories are defined such that each category roughly contains a uniform number of data points. However, since we still leave some room for eyeball inspection in this process, the calculation is semi-automatic.

Although there remains some room for improvement in the characteristic values (for example, the parameters used in MeGeoch's method for signal sequence recognition) at each node, we tested the predictability of the current model by the cross-validation method. The 336 data were randomly divided into a training set of 302 and a testing set of 34. This trial was carried out 10 times and the resultant values were averaged. The overall prediction accuracy marked 79.1% which does not differ much with the accuracy for discriminating the training data. We expect that the method can be applied to a much more complicated eukaryotic problem without terrible difficulty because of its conceptual clearness. Thus, this method seems very promising for upgrading the PSORT system in the near future.

Acknowledgement

This work was supported in part by a Grant-in-Aid, "Genome Informatics" (07249205) for Scientific Research on Priority Areas from The Ministry of Education, Science and Culture in Japan.

References

[1] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in Gram-negative bacteria" *PROTEINS: Structure, Function, and Genetics*, 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.

[3] A. Goffeau, K. Nakai, P. Slonimski, and J.-L. Risler, "The membrane proteins encoded by yeast chromosome III genes" *FEBS letters*, Vol. 325, pp. 112-117, 1993.

[4] K. Nakai, A. Shinohara, and S. Miyano, "Assignment of certainty-factor parameters with a given reasoning tree for the prediction of protein localization sites" *Proc. Genome Informatics Workshop 1994*, pp. 170-171, 1994.