Finding Functional Features of Proteins using Machine Learning Techniques

Takashi Ishikawa[1]
takashi@j.kisarazu.ac.jp

Shigeki Mitaku[2]
mitaku@cc.tuat.ac.jp

Takao Terano[3]
terano@gssm.otsuka.tsukuba.ac.jp

Takatsugu Hirokawa[2]
hirokawa@cc.tuat.ac.jp

Makiko Suwa[2]
suwa@cc.tuat.ac.jp

Seah Boon Chieng[2]
seah@cc.tuat.ac.jp

[1] Kisarazu National College of Technology
Kiyomidai-higashi 2-11-1, Kisarazu, Chiba 292, Japan

[2] Tokyo University of Agriculture and Technology
Nkacho 2-24-16, Koganei, Tokyo 184, Japan

[3] The University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112, Japan

Abstract

Protein function prediction from amino-acid sequences is one of the major tasks in genome informatics. To predict protein functions of a given amino-acid sequence, we can use similarities among functions and structural features of amino-acid sequences, i.e., motif and homology. Difficulties of the previous function prediction methods are caused by the facts that few already known motif have been found and that proteins of similar sequence may not have similar functions. A main objective of our research is to facilitate to find functional features of proteins using machine learning techniques.

Our hypothesis for the protein function prediction is that a protein function arises from physical structures of the protein. Since the structures of proteins are built with physico-chemical interactions among amino-acids, there might exist some features of amino-acid sequences according to the physico-chemical interactions. We call these features 'functional features'. We know that there exists electric interactions among alpha-helices of bacteriorhodopsin from its tertiary structure of the protein and localization of polar amino-acids in the structure. If the amino-acids localization of bacteriorhodopsin is closely related to the function of the protein, we can use this functional feature to predict protein function.

To create rules to predict protein functions, we use the three machine learning techniques (Fig. 1). The first technique is analogical reasoning to make a assumptions about functional features. For example, if there exists localization of polar amino-acids in some proteins, then the localization might imply relation between the functional features and functions of the protein, using analogical reasoning from the fact about bacteriorhodopsin. The second technique is inductive reasoning to generalize the hypothesis made by analogical reasoning. The goal of the inductive reasoning for protein function prediction is to decide which localization pattern is most useful to classify protein functions. The third technique is deductive reasoning to refine the localization pattern into classification rules. In the deductive reasoning, knowledge about protein functions and structures are used to make logical description of classification rules.

[1]石川 孝：木更津高専情報工学科，〒292 千葉県木更津市清見台東 2-11-1
[2]美宅 成樹，広川 貴次，諏訪 牧子，謝 文清：東京農工大学工学部，〒184 東京都小金井市中町 2-24-16
[3]寺野 隆雄：筑波大学経営システム科学，〒112 東京都文京区大塚 3-29-1

We have carried out some experiments to implement our idea to find functional features of proteins using machine learning techniques. First we have simulated analogical reasoning process to create a hypothesis about functional features of bacteriorhodopsin using ABA framework proposed by authors[1]. In the current stage of our research, this analogical reasoning process is executed by hand simulation, but it will be executed on a computer in the next stage. Next we have analyzed the relation between the functional features and protein protein functions of seven-helices membrane proteins using a cluster analysis method. From this analysis, we have found that amino-acid interval frequencies for polar amino-acids is closely related to some function classes of the classified proteins. The feature of the amino-acid interval frequencies is thought to be a representation of the abstract functional feature: 'localization of amino-acids'. From the result of this cluster analysis, we can use the functional features for the inductive reasoning in the next step.

In the preliminary experiments described above, we have found new functional features to classify protein functions from amino-acid sequences. Specifically, these features can discriminate different functions of proteins that have similar amino-acid sequences in homology analysis. Furthermore, the features can recognize same function proteins that have not similar sequences. From these results we state that our idea is useful to predict protein functions. In the next stage of the research, we have a plan to refine classification rules and to integrate three machine learning techniques.

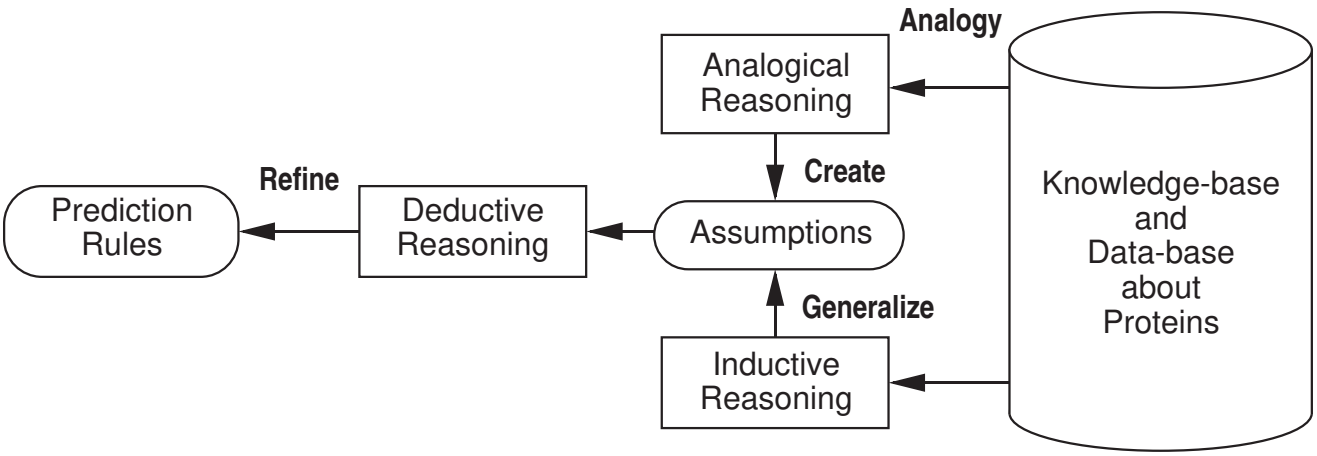

Figure 1. Overview of the method

References

[1] Ishikawa, T. and Terano, T. "Using Analogical Reasoning to Predict a Protein Structure" *Proceedings of Genome Informatics Workshop IV*, 1993.