Discovering Functional Sites of Amino Acid Sequences Using Sorted Variable Generalization

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

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

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

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

Abstract

This research develops a method for discovering functional sites of amino acid sequences using an Inductive Logic Programming (ILP) method with sorted variable generalization. Functional sites provide clues to building a knowledge base for prediction of protein functions from amino acid sequences. The proposed method generates hypotheses of functional sites directly from aligned amino acid sequences using an ILP method extended with sorted variable generalization. The proposed method is shown to be useful for discovering functional sites by an example application to the case of bacteriorhodopsin-like proteins.

1 Introduction

This research develops a method for discovering functional sites of amino acid sequences using an *Inductive Logic Programming* (ILP) method with *sorted variable generalization*. Functional sites provide clues to building a knowledge base for prediction of protein functions from amino acid sequences. Our approach [3] [4] [5] is based on the following assumption: *If there exist any functional sites, then we are able to predict specific functions of a protein from its amino acid sequence*. In order to discover functional sites of amino acid sequences, we use a machine learning technique with a framework of *Inductive Logic Programming* (ILP) [8]. The proposed method generates hypotheses of functional sites directly from aligned amino acid sequences using an ILP method extended with *sorted variable generalization*. *Sorted variable generalization* is a generalization operator of induction to generalize a constant term by replacing it with a *sort* symbol to which the constant belongs [6]. A *sort* is defined as a subset of constants. Therefore we require only to prepare *sorts* for the domain of interest. The proposed method is shown to be useful for discovering functional sites by an example application to the case of bacteriorhodopsin-like proteins. Using the propose method we are able to discover functional sites of amino acid sequences seeming to related to these specific functions.

2 Discovering functional sites

The problem addressed is to discover functional sites of amino acid sequences of proteins with a common specific function. A functional site is a subsequence of an amino acid sequence that only exists in sequences of a certain function. The subsequence has length of one or more and may contain any groups of amino acids. The use of information about secondary structure increases the validity of functional sites with short length. Therefore the problem is reduced to discovering a combination of functional sites existing only in given amino acid sequences with a specific protein function.
The problem is specified as follows:

Given:
- a set of amino acid sequences with a specific protein function
- secondary structures (e.g., alpha helices) of each protein in the set

Find:
a combination of functional sites that *exists* in the given amino acid sequences and *does not exist* in the other amino acid sequences

3 An ILP method with sorted variable generalization

The problem solving process involves two sub processes, *alignment* and *generalization*. The alignment process inputs a set of the given amino acid sequences and outputs aligned sequences which may contain gap symbols representing absence of amino acids. The generalization process inputs the aligned sequences and outputs a generalized sequence pattern covering all the given sequences. The generalized sequence pattern represents functional sites corresponding to the specific protein function. The proposed method uses Horn clauses to represent all information about problem domain. In this representation, an amino acid sequence is represented by a list of symbols each of which represents an amino acid or a group of amino acids [7].
In order to represent generalized sequences, the proposed method uses a subsequence relation *subseq(Subsequence, Sequence)* which represents that *Sequence* contains *Subsequence*, where *Subsequence* may have any length more than one. Any *Subsequence* representing a functional sites is considered to start and end with preserved amino acids (or groups of amino acids). In this representation, a set of functional sites corresponding to a specific function is represented with a conjunction of subsequence relations. Since there exist many sets of functional sites corresponding to a certain function, the proposed method uses the branch-and-bound search algorithm to find a plausible expression of a set of functional sites. This heuristic search utilizes an evaluation function to find a better solution first. Using description length of expressions as the evaluation function.

4 An experiment

We have applied the method to discovering functional sites of bacteriorhodopsin-like proteins. These proteins are divided into two functional groups, *bacterial opsin* and *opsin*. *Bacterial opsin* includes three specific functions, *proton pump*, *chloride pump*, and *sensory rhodopsin*. The goal of this experiment is to discover functional sites of amino acid sequences corresponding to the specific functions of *bacteriorhodopsin*. The data used in the experiment are descriptions in SWISS-PROT database [1]. We use descriptions of TRANSMEM in the database to recognize subsequences corresponding to the trans membrane domains as secondary structures. From these descriptions we prepared a Prolog database as Horn clause representation, and we add clauses to the database to represent the amino acid class hierarchy for *sorted variable generalization*. A discovered sequence corresponding to *proton pump* is *trans_mem_seq(ID, 3, Seq), subseq(['D',_,_,_,_,_,_,_,_,_,'D'], Seq)*. This result coincides with conjectured biological models [2], thus it shows the usefulness of the proposed method.

5 Conclusion

The proposed method generates hypotheses of functional sites directly from aligned amino acid sequences using an ILP method extended with *sorted variable generalization*. The proposed method is shown to be useful for discovering functional sites by an example application to the case of bacteriorhodopsin-like proteins. We have a plan to integrate the proposed method into the developed multistrategy learning environment.

References

[1] Bairoch, A. and Boechmann, B., Nucl. Acids Res., 22:3578-3580 (1994)
[2] Futai, M. ed., Bio-membrane engineering (in Japanese), Maruzen (1991)
[3] Ishikawa, T. and Terano, T., Using analogical reasoning to predict a protein structure., Proc. of Genome Informatics Workshop IV (1993)
[4] Ishikawa, T., Mitaku, S., Terano, T., Hirokawa, T., Suwa, M., and Seah, B-C., Finding functional features of proteins using machine learning techniques, Proc. of Genome Informatics Workshop '94, pp.168-169 (1994)
[5] Ishikawa, T., Mitaku, S., Terano, T., Hirokawa, T., Suwa, M., and Seah, B-C., Building a knowledge-base for protein function prediction using multistrategy learning, Proc. of Genome Informatics Workshop 1995, pp.39-48 (1995)
[6] Ishikawa, T. and Terano, T., Concept-description analogy based on concept simirality, In Proc. of SIG-FAI-9601, JSAI, pp.74-81 (1996)
[7] Lathrop, R., Webster, T., Smith, R., Winston, P., and Smith, T., Integrating AI with Sequence Analysis, in Hunter, L.(ed), Artificial Intelligence and Molecular Biology, AAAI Press, pp.210-258 (1993)
[8] Muggleton, S. and De Raedt, L., Inductive Logic Programming: Theory and Methods, The Journal of Logic Programming, 19:629-679 (1994)