Function Profile: Functional Sites of Proteins Discovered by Generalization of Amino Acid Sequences

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
Yoko Kawata [1] j93410@j.kisarazu.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 paper describes a method for discovering functional sites of amino acid sequences using an Inductive Logic Programming method with Sorted Variable Generalization and a graphical method called Function Profile. The method generates hypotheses for functional sites of given proteins by generalization of their amino acid sequences. The method is validated for discovering functional sites of proteins by applying the method to proteins for which conjectured functional sites are known.

1 Objective

The objective of the research is to discover functional sites of proteins by applying machine learning techniques in artificial intelligence to amino acid sequence databases. A functional site is a subsequence of an amino acid sequence that exists only in sequences of proteins having certain functions. We have developed an *Inductive Logic Programming method with Sorted Variable Generalization* and applied it to discovering functional sites of proteins [3]. The method succeeds to discover candidates of functional sites for proteins having function such as proton pomp, but it does not provide information which candidate is more promising site for protein functions. In this paper, we describe the method *Function Profile* that solves the issue of our previous method. Function profile is a graphical method to find which site of amino acid sequences is closely related to the function of proteins. In order to validate the method for discovering functional sites of proteins, we apply the method to amino acid sequences of proteins for which conjectured functional sites are known. We report some results of preliminary experiments.

2 Method

In order to find candidates of functional sites of amino acid sequences, the method generalize amino acid sequences of proteins with a common specific function in the following procedure. Input data of the method are amino acid sequences of proteins having given functions (positive examples) and amino acid sequences of proteins not having given functions (negative examples).

1. Find a maximum common subsequence of the positive examples by *Sorted Variable Generalization* [3]. A maximum common subsequence is a common subsequence with maximum length for given amino acid sequences. Mismatched sites in a common subsequence are represented by variables (expressed as '_'). An example of a maximum common subsequence is depicted by the following sequences.

```
Example 1        A B C D E
               Z A B F D          positive examples
                 A G H D
               ------------
                 A _ _ D          maximum common sequence
```

2. Find all subsequences of the maximum common subsequence satisfying the following conditions for candidates of functional sites.
   a. Subsequences do not match with any subsequence of all negative examples.
   b. If subsequence X matches with a sub part of subsequence Y, then discard subsequence Y (prefer shorter subsequence). Example 2 shows these subsequences.

```
Example 2       A _ _ D          subsequence X (selected)
                A B _ D E        subsequence Y (discarded)
```

   c. In order to reduce computation time, number of specific sites in subsequences and length of subsequences do not exceed specified limits. (In the preliminary experiments, these limits are specified as 5 and 10.)

3. Count all specific sites in the maximum common subsequence existing in all candidate subsequences and normalize these frequencies by subsequence sizes and number of candidate subsequences. A function profile is made by plotting the normalized frequencies (vertical axis) to the sites of the maximum common subsequence (horizontal axis).

3 Result

In order to validate the method, we apply the method to discovering functional sites of bacteriorhodopsin−like proteins (proton pump) for which conjectured functional sites are known in the literature. The data used in the experiment are found in TMBase database [1]. We use amino acid sequences of TM3 of BAC1_HALS1, BAC2_HALS2, and BACR_HALHA as positive examples for proton pump and use amino acid sequences of other proteins in TMBase as negative examples. The method finds the following candidates of functional sites of proton pump: 'AD__FT', 'D_F_T__L', 'TTPL', 'WL___P__L', 'D____T__LL', ... From these candidates the method draw the following *Function Profile* of TM3 of proton pump.

4 Conclusion

The Function Profile of TM3 of proton pump reveals three active areas (underlined with ^^^^) of TM3 including conjectured functional sites such as two 'D's [2], while the center area is not known in the literature. This result shows the effectiveness of the method to re−discover functional sites of proteins known in the literature and possibility of the method to discover yet unknown functional sites of proteins by application of machine learning techniques.

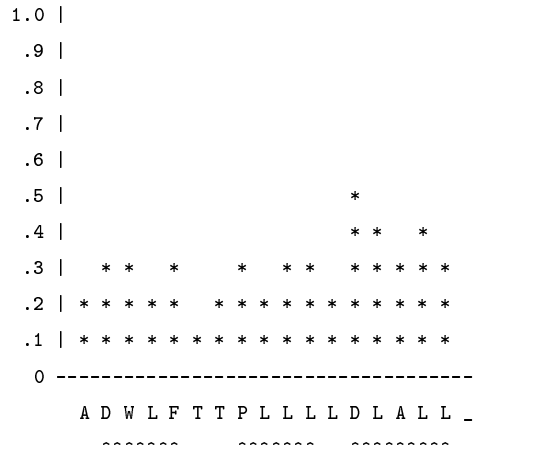

Figure 1. Function Profile of TM3 of proton pump

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., Mitaku, S., Terano, T., Suwa, M., and Hirokawa, T., Discovering Functional Sites of Amino Acid Sequences Using Sorted Variable Generalization. Proc. of Genome Informatics Workshop 1996, pp.178-179 (1996)