MOTIF BASED PROTEIN FUNCTION ANALYSIS USING DATA MINING

Bum Ju Lee, Heon Gyu Lee, Keun Ho Ryu

Database/Bioinformatics Lab., Chungbuk National University
{bjlee, hglee, khryu}@dblab.chungbuk.ac.kr

ABSTRACT Proteins are essential agents for controlling, effecting and modulating cellular functions, and proteins with similar sequences have diverged from a common ancestral gene, and have similar structures and functions. Function prediction of unknown proteins remains one of the most challenging problems in bioinformatics. Recently, various computational approaches have been developed for identification of short sequences that are conserved within a family of closely related protein sequence. Protein function is often correlated with highly conserved motifs. Motif is the smallest unit of protein structure and function, and intends to make core part among protein structural and functional components. Therefore, prediction methods using data mining or machine learning have been developed.

In this paper, we describe an approach for protein function prediction of motif-based models using data mining. Our work consists of three phrases. We make training and test data set and construct classifier using a training set. Also, through experiments, we evaluate our classifier with other classifiers in point of the accuracy of resulting classification.

KEY WORDS: Motif, Protein Function, Classification, Data Mining

1. INTRODUCTION

A motif is a region or portion of a protein sequence that has a specific structure and is functionally significant. Therefore, these motifs are the smallest unit executing the role of protein structure and function. Motifs come out as following features. First, they intend to make core part among protein structural and functional components. Second, they are not capable of independent folding and stability. Third, they are high conserved region in remotely related protein sequences. Finally, they consist of 10~20 residues.

Currently, motif composition is often used to assign putative functions to novel protein sequences based on the known functions of other proteins that share one or more motifs with the novel protein. Many motif databases have been developed. These databases have the use of predicting function and structure of novel apparent protein using relationship between sequence and 3D structure[David, 2001; Attwood 1998; Philipp, 1996]. Existing motif databases such as InterPro[Apweiler, 2001], ProDom[Florence, 2000], BLOCKS[David, 2001], PROSITE[Laurent, 2002], PRINTS[Attwood 1998] and Pfam[Alex, 2002] created the use of each different methods.

Several automated tools for generating a set of motifs that capture conserved sequence regularities among a given set of sequences are available. These tools divided into tow broad classes. The first class of methods relies on local and multiple sequence alignment to extract conserved patterns among set of related sequences. A second class of methods uses a combinational approach to build a dictionary of motifs from a given set of sequences without making any assumptions about the functional family memberships of sequences in question[Xiangyun, 2003, Rigoutsos, 1999]. These pattern discovery methods were introduced to alleviate the problems associated with multiple sequence alignment and algorithms have been steadily appearing in the [Rigoutsos, 1998; Suyama,1995; Wang, 1994; Isidore, 2000].

Essentially, these algorithms seek to determine one or more patterns that represented one or more blocks of related sequences. In some cases, these algorithms are used to compute the cardinality and the boundaries of conserved blocks within groups of related sequences[Henikoff, 1994], build profiles, build HMM [Alex, 2002], or generated regular expression[Laurent, 2002]. The latter are especially useful for extracting sequence regularities among divergent families. Motifs or sequence patterns distil information from groups of related sequences to facilitate detection of weaker sequence similarities. Therefore, pattern based searches are often more sensitive and selective than sequence database search.

In this paper, we describe an approach for protein function prediction of motif-based models using data mining. Our work consists of three phrases. Firstly, we make training and test data set through motif information analysis. A training set of motif information with known functions is used automatically construct classifier. Also, we use attribute-based representation because the choice of attributes plays a critical role in the data mining process. Secondly, we generate classifier for function prediction. Lastly, through experiments, evaluate our classifier with other classifiers in point of the accuracy of resulting classification.

2. RELATED WORKS

Until recently, many of mining techniques based on sequence and structure motifs have been developed. The techniques for pattern discovery based on motifs are as following.

X. Wang et al.[Xiangyun, 2003] suggested a fully automated approach for protein function classification. This method presented a data-driven approach to discovery of rules for assigning protein sequences to functional families on the basis of the presence or absence of specific motifs or combinations of motif. [Giri, 2002] proposed an approach to the problem of automatic motif detection. This approach used methods from Data Mining and Knowledge Discovery to design an algorithm that displays increased sensitivity as compared to existing algorithms, while maintaining good accuracy and also providing additional information about a given protein sequence. Unlike other approaches, this algorithm is not based on statistical methods.

[Bill, 2003] suggested several important time series data mining problem for finding motifs. This approach generalized the definition of time series motifs to allow for don't care subsections, and introduced a novel time and space efficient algorithm to discover motifs. In the [Horng, 2002], the aim is to find the motif correlation in protein sequences. This approach developed a tool to find the correlation about the domain sharing in proteins so as to provide some information for protein or function genomics. The implementation of the tool uses the Apriori mining algorithm to mine the association of functional domain sharing in the protein structures.

The counts of subgraphs in different proteins are then used as input variables for a binary classification task to distinguish between two protein families in the SCOP. The support vector machine approach is used to construct the classifier. This approach suggested that frequent subgraph mining used to identify packing motifs that are highly specific to individual protein families providing opportunities for rapid and automated protein annotation.

Finding recurring residue packing patterns or structural motifs that characterized protein structural families is an important problem in bioinformatics. Many of motif comparison and detection methods use features extracted from the structure motifs. The feature description consists of geometry, topology and properties[Ingvar, 2000]. Geometry features contain coordinates or relative positions of atoms, residues and fragments, and topology feature contains the elements' order along the backbone. Also, properties features contain physico-chemical properties of the elements (e.g., residues). The research of protein structure patterns have developed various comparison methods for prediction. This methods such as MUSTA method[Leibowitz, 2001], SPratt and Spratt2 method[Inge, 2002; Inge, 1999] and Trilogy method[Bradley, 2002]. Specific contents about these methods presented in [Ingvar, 2000].

Though many methods about protein function prediction have been developed. This work progresses toward the development of more exact prediction so far. So, our purpose develops good prediction method of protein function using ADTree classifier.

In the next section, we describe general data mining approach for protein function classification and Alternating Decision Trees. In the section 4, we explain preprocessing and experimental results. Finally, we describe conclusion.

3. OUR CLASSIFICASTION METHOD

In this section we introduce the overview of our experiment design, the algorithm used to classify and the data set.

3.1 Data mining approach

Figure 1 show our data mining approach to predict protein function based on motif feature. Generally, we think that two step are important. There step is to make good data set and to design good algorithm for superior prediction.

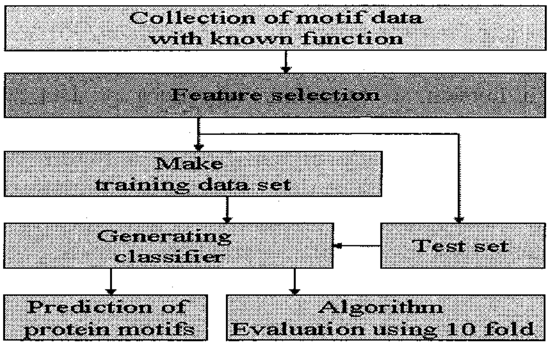
Collection of motif data with known function
Feature selection
Make training data set
Generating classifier
Test set
Prediction of protein motifs
Algorithm Evaluation using 10 fold

Figure 1. Data mining approach to protein function classification.

3.2 Alternating Decision Trees

For protein function prediction, we use ADTrees algorithm for generating classifier. ADTrees are similar to option trees first described by Buntine and further developed by Kohavi et. Al. Option trees were shown to proide significant improvements in classification error compared to single decision trees. A formal definition of alternating trees as weighted votes of simple rules is as following:

(1) A base condition is a boolean predicate over instances. We use $\wedge$ to denote conjunction (AND), $\neg$ to denote negation (NOT) and T to denote the constant predicate that is always true. We use c to denote a set of base conditions.

(2) A precondition is a conjunction of base conditions and negations of base conditions.

(3) A base rule r is a mapping from instances to real numbers which is defined in terms of a precondition

c_1, a base condition c_2, and two real numbers a and b. The base rule maps each instance to a prediction that is defined to be a if $c_1 \wedge c_2$, b if $c_1 \wedge \neg c_2$ and 0 if $\neg c_1$.

(4) An alternating decision tree is a mapping from instances to real numbers which is defined in terms of a set of base rules. The set of base rules must obey the two following conditions:

① The set must include a base rule for which both the condition and the pre-condition are T. The a value of this rule is the prediction associated with the root of the tree.

② A base rule r with precondition d can be in the set only if the set includes a rule r' with precondition c_1 and condition c_2 such that $d = c_1 \wedge c_2$ or $d = c_1 \wedge \neg c_2$. d corresponds to the prediction node that is the direct parent of r.

4. EXPERIMATS AND RESULTS

4.1 Preprocessing

For prediction of protein function and localization, We use data set supported by KDD 2001(http://www.cs.wisc.edu/~dpage/kddcup2001/). Because the data set include some problem such as values out of domain, so we modify and reduce the data set for preprocessing step. In preprocessing step, we need discretization because of transforming attributes into a suitable form and increasing the speed of algorithms. We used an approach of discretization. The specific data contents such as attributes, percentage of missing values and number of distinct labels are shown in Table 1.

Table 1. Features according to each attribute

	Essential	Complex	phenotype	localization	class
Percentage of missing values	0%	31%	17%	0%	0%
Number of distinct labels	3	27	11	10	2

Our experimental configuration is Windows XP Professional, Pentium® 4 CPU 2.80GHz, 1.50GB RAM.

4.2 Experimental Evaluation

We compare ADTree[Yoav, 1999] with C4.5 decision tree, LMT(logistic model tree)[Niels, 2003] and NBTree[Ron, 1996]. We tested all algorithms by 10 fold cross-validation. In the experiments, the parameter setting of the four algorithms is as follows. NBTree and LMT parameters are default values. Confidence factor in C4.5 decision tree parameter fix on 0.1, and other parameter remain default. Boosting iteration parameter in ADTree fix on 18, and other parameter remain default. We used TP, FP, Precision, Recall and F-Measure to evaluate the performance of ADTree algorithm. Table 2 represents detail experimental result of ADTree algorithm. The result of comparison among four algorithms is shown Figure 3. As can be seen from table, the accuracy of ADTree performed better than those of C4.5 decision tree, LMT and NBTree.

Table 2. Detailed accuracy by class using ADTrees algorithm

Class	TP	FP	Precision	Recall	F-measure
Cell growth cell division and dna synthesis	0.758	0.214	0.816	0.758	0.786
Transcription	0.786	0.242	0.722	0.786	0.752

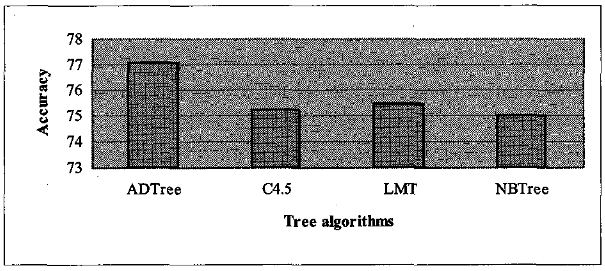

Figure 2. Classification accuracy through comparing algorithms

4.3 Classification accuracy through number of boosting iteration

The number of boosting iterations needs to be manually tuned to suit the dataset. The figure 4 presents the number of boosting iterations and classification accuracy through increasing two by two.

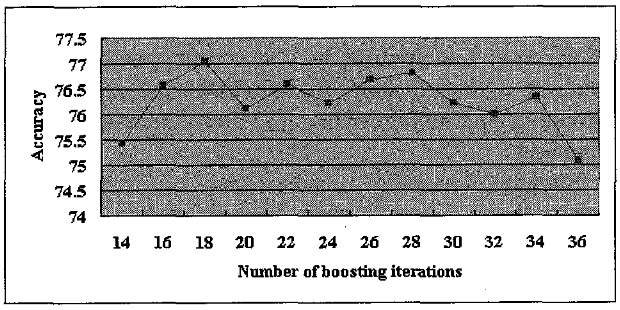

Figure 1. Classification accuracy as a result of number of boosting iterations.

5. CONCLUSION

This paper proposes a framework of classification using ADTrees algorithm for protein function prediction based on motifs. Our framework not only gives a way to construct classifiers, but also helps to solve a problem of protein function prediction.

In our future work, we will focus on building more accurate classifiers by using distinguished techniques, and will perform motif finding by using superior sequential pattern mining techniques

References from Journals:

Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, L. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez,B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. A. Sigrist and E. M. Zdobnov, 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites, Nuleic Acids Research, Vol.29, No.1, pages 37-40.

T. K. Attwood, M. E. Beck, D. R. Flower, P. Scordis, N. Selley, 1998. The PRINTS protein fingerprint database in its fifth year, Nucleic Acids Research, Vol.26, No.1, pages 304-308.

Alex Bateman, Evan Birney, Lorenzo Cerruti, Richard Durbin, Laurence Etwiller, Sean R. Eddy, Sam Griffiths-Jones, Kevin L. Howe, Mhairi Marshall, Erik L. L. Sonnhammer, 2002. The Pfam Protein Families Database, Nucleic Acids Research, Vol.30, No.1, pages 276-280.

Laurent Falquet, Marco Pagni, Philipp Bucher, Nicolas Hulo, Christian J. A. Sigrist, Kay Hofmann, Amos Bairoch, 2002. The PROSITE database, its status in 2002, Nucleic Acids Research, Vol.30, pages 235-238.

Philipp Bucher, Kevin Karplus, Nicolas Moeri, Kay Hofmann, 1996. A Flexible Motif Search Technique Based on Generalized Profiles, Comput. Chem., Vol.20, pages 3-24.

Florence Corpet, Florence Servant, Jerome Gouzy and Daniel Kahn, 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons , Nucleic Acids Research, Vol.28, No.1 pages 267-269.

Xiangyun Wang, Diane Schroeder, Drena Dobbs, Vasant Honavar, 2003. Automated data-driven discovery of motif-based protein function classifiers. Information Sciences 155, 1-18.

I. Rigoutsos, A. Floratos, C. Ouzounix, Y. Gao, L. Parida, 1999. Dictionary building via unsupervised hierarchical motif discovery in sequence space of natural proteins, Proteins 37 (2), 264-277.

Suyama, M., Nishioka, T., Jun'ichi, O., 1995. Searching for common sequence patterns among distantly related protein, Protein Eng, 1075-1080.

Wang, J., Marr, T. G., Shasha, D., Shapiro, B. B., Chirn, G., 1994. Discovering active motifs in sets of related proteien sequences and using them for classification, Nucleic Acids Res., 2769-2775.

Isidore Rigoutsos, Aris Floratos, Laxmi Parida, Yuan Gao, Daniel Platt, 2000. The Emergence of Pattern Discovery Techniques in Computational Biology, Metabolic Engineering 2 159-177.

Henikoff, S., Henikoff, J., 1994. Protein family classification based on searching a database of Blocks, Genomics 19, 97-107.

Ingvar Eidhammer, Inge Jonassen, Wilian R. Taylor, 2000. Protein Structure Comparison and Structure Patterns, Journal of Computational Biology, 7, 685-716.

N. Leibowitz, Z. Fligelman, R. Nussinov, H. J. Wolfson, 2001. Automated mltiple structure alignment and detection of a common sub-structural Motif, Proteins, 43, 235-245.

Inge Jonassen, Ingvar Eidhammer, Darrell Conklin, Willian R. Taylor, 2002. Structure Motif Discovery and Mining the PDB, Bioinformatics 18, 362-367.

Inge Jonassen, Ingvar Eidhammer, Darrell Conklin, 1999. Discovery of local packing motifs in protein structures, Proteins 34, 206-219.

Giri Narasimhan, Changsong Bu, Yuan Gao, Yuning Wang, Ning Xu, Kalai Mathee, 2002. Mining Protein Sequence for Motifs, Journal of Computational Biology, 9, 707-720.

References from Books:

David W. Mount, 2001. Bioinformatics : Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, pages 45-48.

new1. Carl Branden, John Tooze, 1991. Introduction to Protein Structure. published by Garland Publishing, Inc. 13-33

References from Other Literature:

Rigoutsos, I., Floratos, A., 1998. Motif discovery without alignment or enumeration, Proceedings of the 2th annual ACM International Conference on Computational Molecular Biology(RECOMB), New york.

Bill Chiu, Eamonn Keogh, Stefano Lonardi, 2003. Probabilistic Discovery of Time Series Motifs, Conference on Knowledge Discovery in Data, 493-498.

Jorng-Tzong Horng, Hsien-Da Huang, Shih-Hsien Wang, Fan-Mao Lin, Baw-Jhiune Liu, Jenn-Kang Hwang, 2002. Study of Motif Correlation in Proteins by Data Mining, In Proceeding of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '02) 345-350.

P. Bradley, P.S. Kim and B. Berger, 2002. Trilogy: discovery of sequence structure patterns across diverse proteins, Proceedings of the Sixth International Conference on Research in Computational Molecular Biology (RECOMB), 77-88.

Yoav Freund, Llew Mason, 1999. The Alternating Decision Tree Learning Algorithm, Proceedings of the Sixteenth International Conference.

Niels Landwehr, Mark Hall, and Eibe Frank, 2003. Logistic Model Trees, 14th European Conference on Machine Learning, LNCS 2837.

Ron Kohavi 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision tree hybrid. Procedings of the Second Internaltional Conference on Knoledge Discovery and Data Mining.

Acknowledgements (optional)

This works was supported by the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.