Mining Association Rules Related to Protein-Protein Interactions

Takuya Oyama[1] oyama@isl.intec.co.jp
Kagehiko Kitano[1] kage@isl.intec.co.jp
Kenji Satou[2] ken@jaist.ac.jp
Takashi Ito[3] titolab@kenroku.kanazawa-u.ac.jp

[1] INTEC Web and Genome Informatics Corp., 3-23, Shimoshin-machi, Toyama 930-0804, Japan
[2] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
[3] Cancer Research Institute, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa 920-0934, Japan

Keywords: data mining, protein-protein interaction

1 Introduction

Recently, protein-protein interactions are systematically examined with the yeast two-hybrid method [2, 4]. As a result, a large quantity of information about protein-protein interactions is accumulated. However, general information or knowledge about interactions is not so much identified although a lot of individual interactions are identified using such methods. On the other hand, the technique to extract useful information or knowledge hidden in vast data, which is called "data mining", attracts a great deal of attention, and several researches applying data mining to bioinformatics have already been done [3].

For this reason, we are studying and developing a system to discover knowledge or useful information related to protein-protein interactions from accumulated protein-protein interaction data, using the "association rules discovery" algorithm [1] that is a popular method of data mining.

2 Method

Figure 1 illustrates the outline of the method. The data for mining is created from thousands of two-hybrid interactions between yeasts and features that characterize each protein involved in the interaction. Interactions are obtained from the web sites such as YPD and MIPS, as well as the large-scale two-hybrid experiment by Ito *et al.* Features of each protein are defined from functional, primary structural, and other various viewpoints using data from genome databases on various web sites. Here is a list of six types of protein features used in this system.

(a) In YPD, proteins are classified into dozens of categories based on their cellular role, biochemical function, and so on. Features of the first type are defined by YPD categories that the protein belongs to. That is, if a protein belongs to category C, it is assigned a feature stating "This protein belongs to C".

(b) Enzymes are classified based on their functions into many categories labeled by EC numbers. Features of the second type are defined by the EC numbers which the protein corresponds to.

(c) Many of the proteins correspond to the entries of SWISS-PROT and PIR, and each entry has some keywords representing functional, structural or other features of the protein. Features of the third type are defined by SWISS-PROT/PIR keywords assigned to the protein.

Protein A: Motif "CNMP_BINDING_1" ⇒ Protein B: EC Number "2.7" (conf.=100 %)

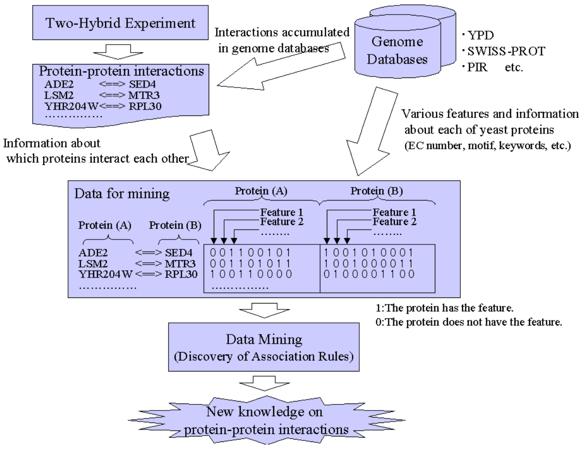
Two-Hybrid Experiment
Interactions accumulated in genome databases
Genome Databases
・YPD
・SWISS-PROT
・PIR etc.
Protein-protein interactions
ADE2 <==> SED4
LSM2 <==> MTR3
YHR204W<==> RPL30
……………
Information about which proteins interact each other
Various features and information about each of yeast proteins (EC number, motif, keywords, etc.)
Data for mining
Protein (A)
Protein (B)
Feature 1
Feature 2
……..
Protein (A)
Protein (B)
ADE2 <==> SED4
LSM2 <==> MTR3
YHR204W <==> RPL30
……………
0 0 1 1 0 0 1 0 1
0 0 1 1 0 1 0 1 1
1 0 0 1 1 0 0 0 0
……………
1 0 0 1 0 1 0 0 0 1
1 0 0 1 0 0 0 0 1 1
0 1 0 0 0 0 1 1 0 0
1:The protein has the feature.
0:The protein does not have the feature.
Data Mining (Discovery of Association Rules)
New knowledge on protein-protein interactions

Figure 1: Mining process.

(d) Features of the fourth type are defined by PROSITE motifs existing on the amino acid residues of the protein.

(e) Features of the fifth type are defined by the "bias" of the amino acid residues. That is to say, if amino acids of particular kinds are rich in a short region on the amino acid residues, the protein is assigned a feature representing it.

(f) We classified homologous segments on the amino acid residues into clusters, based on all-versus-all homology search among all yeast proteins with BLASTP. Features of the sixth type are defined by the clusters that the segments of the protein are classified.

As a result of characterizing all the yeast protein, we got more than two thousand features of above six types. Then we made the data for mining adding features of each protein to protein-protein interactions. In many cases of ordinary mining, physical entities such as genes or proteins are regarded as transactions of data mining. However, in the case of mining from interactions, we think that an interaction itself rather than a protein (or a gene) should be regarded as a transaction. Namely each transaction represents an interaction, and has features of two proteins, which are protein (A) and (B) in Figure 1, related to the corresponding interaction. Mining association rules from that data is expected to discover associations or new knowledge between features of two proteins that are interacting each other.

3 Result

Executing data mining, we got thousands of association rules including many trivial ones. The following is an example of association rules, which shows that all the proteins having the motif of "CNMP_BINDING_1" on their amino acid residues interact with proteins of EC number 2.7.x.x.

References

[1] Agrawal, R., Imielinski, T., and Swami, A., Mining association rules between sets of items in large databases, *Proc. ACM SIGMOD*, 207–216, 1993.

[2] Ito, T., *et al.*, Toward a protein-protein interaction map of the budding yeast : A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins, *Proc. of the National Academy of Science of USA*, 97(3):1143–1147, 2000.

[3] Satou, K., *et al.*, Finding association rules on heterogeneous genome data, *Proc. Pacific Symposium on Biocomputing '97*, 397–408, 1997.

[4] Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403:623–631, 2000.