Extracting Information on Protein-Protein Interactions from Biological Literature Based on Machine Learning Approaches

Kazunari Sugiyama[1]
kazuna-s@is.aist-nara.ac.jp

Kenji Hatano[1]
hatano@is.aist-nara.ac.jp

Masatoshi Yoshikawa[2]
yosikawa@itc.nagoya-u.ac.jp

Shunsuke Uemura[1]
uemura@is.aist-nara.ac.jp

[1] Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
[2] Information Technology Center, Nagoya University, Furo, Chikusa, Nagoya, Aichi 464-8601, Japan

Keywords: information extraction, protein-protein interactions, machine learning, MEDLINE

1 Introduction

DNA arrays of various life forms have been determined, and a large amount of their information is stored in databases such as FASTA [12]. However, by using only information stored in databases, it is impossible to analyze how the molecular network that supports the functions of a cell such as signal transduction is constructed. Consequently, in order to comprehensively understand the entire life system, we need to analyze what kind of interactions each protein conducts with other proteins or molecules in a living body. In the field of life science, knowledge described in literature is actively used with regard to protein-protein interactions. However, in the case of utilizing such knowledge, the main problems are the following: (1) findings described in the form of language can be understood only when researchers read each paper one by one; (2) even if findings in a paper relate to findings in other papers, their correlation can be identified only by researchers who read these papers; (3) if the number of related findings reaches the tens of thousands, the work of indentifying correlative findings sprinkled in papers and organizing them as a knowledge system becomes enormous. Therefore, in order to alleviate these problems, we propose a method for extracting information on protein-protein interactions from biological literature based on machine learning approaches. It is expected that the amount of literature in biology will increase exponentially. Hence, we believe that it would be highly useful to develop such a technique to support biologists' research.

2 Method and Results

Regarding extracting information on protein-protein interactions from biological literature, we can refer to the approaches based on using (1) surface expressions of the sentences within biological literature [1], (2) a full parser [4, 8, 11], and (3) templates [5, 9]. However, in the surface expression approach, it is extremely difficult to define manually comprehensive rules of extracting information on protein-protein interactions. In the full parser approach, it has often been pointed out that the processing speed is slow and that the results have ambiguity. In addition, the template approach can suffer from the same problems that arise in the surface expression and full parser approaches. Therefore, it is necessary to automatically acquire the rules of extracting information on protein-protein interactions without defining the rules manually and without using a full parser. Moreover, biological literature commonly contains (1) many unknown words such as compound names, as well as original proper nouns coined by the paper's authors; (2) non-alphabetical characters such as "()", "-" and so on; and (3) a variety of verbal forms such as indicative, passive, gerundive, and so on. Furthermore, Ding *et al.* [3] found that it is effective to analyze a sentence as a unit for extracting information on the interactions of two objects. On the basis of these points, in order to identify sentences describing information on protein-protein interactions, we employed an approach of automatically constructing the classifier by several machine learning approaches trained by the features that express the characteristics of biological literature. Focusing on the verb in a sentence, we utilize the following sentence features: (1) the verbal form; (2) part of speech information on the previous and following 10 words from the verb; (3) if a noun is found in (2), part of speech information on the previous and following 10 words from that noun. In addition, we also used other information that characterizes a noun: whether the noun contains numerical figures as in "*p53*", non-alphabetical characters like "-", "()" in such usages as "*elF-4a*", "*GLYS(A)*", or upper case letters as in "*Nef*". The dataset we used was 1000 MEDLINE abstracts. In order to construct a classifier using these features extracted from the dataset, we employed the following machine learning approaches: k-nearest neighbor rule [2], decision tree [6], neural network [7], and SVM (Support Vector Machine) [10]. We performed 10-fold cross validation on each learning algorithm with a particular parameter setting and measured precision, recall, and F-measure to verify the effectiveness of each machine learning algorithm. Table 1 shows the experimental results.

Table 1: Accuracy of extracting information on protein-protein interactions.

Machine Learning Algorithm	Precision	Recall	F-measure
k-Nearest Neighbor ($k = 5$)	0.623	0.647	0.635
Decision Tree	0.725	0.774	0.749
Neural Network (with 3 hidden layers of 3 neurons each)	0.762	0.754	0.758
SVM (radial-basis function kernel)	0.881	0.863	0.872

3 Discussion

In this paper, we proposed several machine learning approaches to extracting information on protein-protein interactions from biological literature. Experimental results show that the classifier based on SVM can provide the highest accuracy in this type of task. In future work, we plan to explore the characteristics of each machine learning approach in more detail and refine our approach to develop an information extraction system with even higher accuracy.

Acknowledgments

Kazunari Sugiyama was supported in part by a Grant-in-Aid for the 21st COE Research Program from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Blaschke, C. and Valencia, A., The potential use of SUISEKI as a protein interaction discovery tool, *Genome Informatics*, 12:123–134, 2001.

[2] Cover, T.M. and Hart, P.E., Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, IT-13(1):21–27, 1967.

[3] Ding, J. and Berleant, D., Mining MEDLINE: abstracts, sentences or phrases?, *Proc. of the Pacific Symposium on Biocomputing (PSB 2002)*, 326–337, 2002.

[4] Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A., GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, *Bioinformatics*, 17(Suppl.1):S74–S82, 2001.

[5] Humphreys, K., Demetriou, G., and Gaizauskas, R., Two applications of information extraction to biological science journal articles: enzyme interaction and protein structures, *Proc. of the Pacific Symposium on Biocomputing (PSB 2000)*, 502–513, 2000.

[6] Quinlan, J.R., *C4.5:Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[7] Rumelhart, D.E. and McClelland, J.L., *Parallel Distributed Processing*, MIT Press, 1986.

[8] Sekimizu, T., Park, H., and Tsujii, J., Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts, *Genome Informatics*, 9:62–71, 1998.

[9] Thomas, J., Milward, D., Ouzounis, C., and Pulman, S., Automatic extraction of protein interactions from scientific abstracts, *Proc. of the Pacific Symposium on Biocomputing (PSB 2000)*, 538–549, 2000.

[10] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[11] Yakushiji, A., Tateisi, Y., Miyano, Y., and Tsujii, J., Event Extraction from Biomedical Papers using a Full Parser, *Proc. of the Pacific Symposium on Biocomputing (PSB 2001)*, 408–419, 2001.

[12] `http://fasta.bioch.virginia.edu/fasta_www/`