Genomic Hypothesis Creator: A Multi-Strategic View Creator for Sequences

Osamu Maruyama [1]
maruyama@ims.u-tokyo.ac.jp

Tomoyuki Uchida [2]
uchida@cs.hiroshima-cu.ac.jp

Takayoshi Shoudai [3]
shoudai@i.kyushu-u.ac.jp

Satoru Miyano [1]
miyano@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[2] Faculty of Information Sciences, Hiroshima City University
3-4-1, Ozuka-Higashi, Asa-Minami-Ku, Hiroshima 731-3194, Japan
[3] Department of Informatics, Kyushu University 39
6-1 Kasuga-Kouen, Kasuga 816-8580, Japan

1 Introduction

Genome sequencing projects on almost 70 biological organisms have been constantly processing, and up to now, some of them have already made public their complete genomes, for example, through Internet. These genomes are very carefully analyzed by skillful experts, usually by using some powerful tools for biological sequences like homology search. Even such experts must sometimes take a different view for new discoveries. A view on data, that is, a range or way of vision on data, is always a key point of scientific discovery. Without doubt, such an invention of a new view would be one of the most important part of scientific discovery process.

We have been designing and developing a multi-strategic and discovery-oriented system Genomic Hypothesis Creator based on algorithmic techniques. In the design of the system, we focus on the matter of creating new views automatically and finding a small hypothesis by employing the views which has just been found on the data. Such a view obtained by this system would make it possible for us to understand what the data means, and it would eventually lead us to the goal of discovery.

As indicated in [4], the main bottleneck for scientific discovery applications is not the lack of techniques for data analysis. The problem is to exploit and combine existing algorithms effectively. From this point, our system employs the multi-strategy principle [3] and the plug-in architecture. By linking views created by one of engines in our system, which is called View Designer, Genomic Hypothesis Creator provides a diversity of knowledge discovery tools.

2 Genomic Hypothesis Creator

Genomic Hypothesis Creator consists of four components, which are called Data Collector, View Designer, Hypothesis Generator and Visualizer, respectively (Fig. 1). Our idea of these components comes from the KDD process in [2].

Data Collector: The first component, Data Collector, labors up for collecting data from databases. For this work, we employ and modify the text database management system SIGMA [1] which has been used for service at the Computer Center of Kyushu University. A user can determine collections of data according to user's interests, for which Genomic Hypothesis Creator would find small hypotheses.

View Designer: A view on data provides terms with which we understand and explain the data. Unless we can assume any experts, this process turns to be the most difficult obstacle in discovery. Now we are designing our system on some flexible views like *alphabet indexing*, *regular patterns*, *approximate*

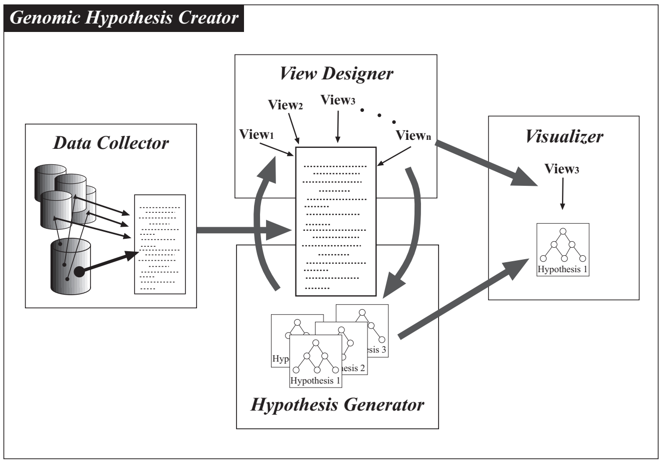

Figure 1: The design concept of Genomic Hypothesis Creator.

string matching, *PROSITE views*, and these combinations. In View Designer, a user can specify a huge view space for automatically creating views on data and select a search strategy over the space.

Hypothesis Generator: The idea of multi-strategy principle has been extensively discussed as an important aspect of knowledge discovery system. In Genomic Hypothesis Creator, it is realized in the following ways: First, the system has a pool of hypothesis generators, from which a user can select one. Second, a user can add a new own hypothesis generator to the pool through a plug-in interface. Currently, hypothesis generators for *decision trees* and *binary decision diagrams* are available.

Visualizer: In some sense, visualization of hypotheses is also an important problem in the discovery process for understanding hypotheses. By using Internet tools, we visualize views and hypotheses which are obtained from View Designer and Hypothesis Generator.

3 Conclusion

We have designed and developed a multi-strategic and discovery-oriented system Genomic Hypothesis Creator by introducing view and view space on data. Our system offers a very wide range of methods for scientific discovery from text databases. The detail concepts of Genomic Hypothesis Creator and experimental results under various conditions will be presented at the conference site.

References

[1] Arikawa, S., Haraguchi, M., Inoue, H., Kawasaki, Y., Miyahara, T., Miyano, S., Oshima, K., Sakai, H., Shinohara, T., Shiraishi, S., Takeda, M., Yamamoto, A., and Yuasa, H., The text database management system SIGMA: An improvement of the main engine, *Proc. Berliner Informatik-Tage*, 72–81, 1989.

[2] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., The KDD process for extracting useful knowledge from volumes of data, *Commun. ACM*, 39:27–34, 1996.

[3] Michalski, R. S., Bratko, I., and Kubat, M., Machine Learning and Data Mining: Methods and Applications, John Wiley & Sons, Ltd., 1998.

[4] Wirth, R., Shearer, C., Grimmer, U., Reinartz, T., Schlosser, J., Breitner, C., Engels, R., and Lindner, G., Towards process-oriented tool support for knowledge discovery in databases, *Proc. PKDD*, Springer-Verlag, 243–53 1997.