# How Can Data Mining Help Bio-Data Analysis?

[Extended Abstract]

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
hanj@cs.uiuc.edu

## ABSTRACT

Recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns in large databases. In the mean time, recent progress in biology, medical science, and DNA technology has led to the accumulation of tremendous amounts of bio-medical data that demands for in-depth analysis. The question becomes how to bridge the two fields, *data mining* and *bioinformatics*, for successful mining of bio-medical data. In this abstract, we analyze how data mining may help bio-medical data analysis and outline some research problems that may motivate the further developments of data mining tools for bio-data analysis.

## Keywords

Bio-medical data analysis, data mining, bioinformatics, data mining applications, research challenges

## 1. INTRODUCTION

In the past two decades we have witnessed revolutionary changes in biomedical research and bio-technology and an explosive growth of bio-medical data, ranging from those collected in pharmaceutical studies and cancer therapy investigations to those identified in genomics and proteomics research by discovering sequencing patterns, gene functions, and protein-protein interactions. The rapid progress of biotechnology and bio-data analysis methods has led to the emergence and fast growth of a promising new field: *bioinformatics*.

On the other hand, recent progress in data mining research has led to the developments of numerous efficient and scalable methods for mining interesting patterns and knowledge in large databases, ranging from efficient classification methods to clustering, outlier analysis, frequent, sequential and structured pattern analysis methods, and visualization and spatial/temporal data analysis tools.

The question becomes how to bridge the two fields, *data mining* and *bioinformatics*, for successful data mining in biomedical data. Especially, we should analyze how data mining may help efficient and effective bio-medical data analysis and outline some research problems that may motivate the further developments of powerful data mining tools for biodata analysis. This is the motivation of this talk.

## 2. HOW DATA MINING MAY HELP BIO-DATA ANALYSIS?

Here we list a few interesting themes on data mining that may help bio-data analysis.

**1. Data cleaning, data preprocessing, and semantic integration of heterogeneous, distributed bio-medical databases.**

Due to the highly distributed, uncontrolled generation and use of a wide variety of bio-medical data, data cleaning, data preprocessing, and the semantic integration of such heterogeneous and widely distributed biomedical databases, such as genome databases and proteome databases, have become an important task for systematic and coordinated analysis of bio-medical databases. This has promoted the research and development of integrated data warehouses and distributed federated databases to store and manage the primary and derived bio-medical data, such as genetic data. Data cleaning and data integration methods developed in data mining, such as [9; 3], will help the integration of bio-medical data and the construction of data warehouses for bio-medical data analysis.

**2. Exploration of existing data mining tools for bio-data analysis.**

With years of research and developments, there have been many data mining, machine learning, and statistics analysis systems and tools available for use in bio-data exploration and bio-data analysis. Comprehensive surveys and introduction of data mining methods have been compiled into many textbooks such as [11; 6; 7]. There are also many textbooks on bioinformatics, such as [2; 8; 5; 4]. General data mining and data analysis systems have been constructed for such analysis, such as SAS Enterprise Miner, SPSS, SPlus, IBM Intelligent Miner, Microsoft SQLServer 2000, SGI MineSet, and Inxight VizServer. There are also some bio-specific data analysis software systems, such as GeneSpring, Spot Fire, VectorNTI, COMPASS, and SMA (Statistics for Microarray Analysis) in R. These tools are evolving as well. For bio-data analysis, it is important to train researchers to master and explore the power of these well-tested and popularly used data mining tools and packages. A lot of routine data analysis work can be done using such tools.

With sophisticated bio-data analysis tasks, there is much room for research and development of advanced, effective, and scalable data mining methods in bio-data analysis. Some interesting topics in this direction are illustrated as follows.

**3. Similarity search and comparison in bio-data.**

One of the most important search problems in bio-data analysis is similarity search and comparison among bio-sequences and structures. For example, gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the two classes of genes. This can be done by first retrieving the gene sequences from the two tissue classes, and then finding and comparing the frequently occurring patterns of each class. Usually, sequences occurring more frequently in the diseased samples than in the healthy samples might indicate the genetic factors of the disease; on the other hand, those occurring only more frequently in the healthy samples might indicate mechanisms that protect the body from the disease. Similar analysis can be performed on microarray data and protein data to identify similar and dissimilar patterns. Moreover, since bio-data usually contains noise or non-perfect matches, it is important to develop effective sequential or structural pattern mining algorithms in the noisy environment, such as that recently reported in [12].

**4. Association analysis: identification of co-occurring bio-sequences or other correlated patterns.**

Currently, many studies have focused on the comparison of one gene to another. However, most diseases are not triggered by a single gene but by a combination of genes acting together. Association and correlation analysis methods can be used to help determine the kinds of genes or proteins that are likely to co-occur in target samples. Such analysis would facilitate the discovery of groups of genes or proteins and the study of interactions and relationships among them.

**5. Frequent pattern-based cluster analysis.**

Most cluster analysis algorithms are based on either Euclidean distances or density [6]. However, bio-data often consists of a lot of features which form a high dimension space, and it is crucial to study differentials with scaling and shifting factors in multi-dimensional space and discover pair-wise frequent patterns and cluster bio-data based on such frequent patterns. One interesting study taken microarray data as examples is in [10].

**6. Path analysis: linking genes or proteins to different stages of disease development.**

While a group of genes/proteins may contribute to a disease process, different genes/proteins may become active at different stages of the disease. If the sequence of genetic activities across the different stages of disease development can be identified, it may be possible to develop pharmaceutical interventions that target the different stages separately, therefore achieving more effective treatment of the disease. Such path analysis is expected to play an important role in genetic studies.

**7. Data visualization and visual data mining.**

Complex structures and sequencing patterns of genes and proteins are most effectively presented in graphs, trees, cubes, and chains by various kinds of visualization tools. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration. Visualization and visual data mining therefore play an important role in biomedical data mining.

**8. Privacy preserving mining of bio-medical data.**

Although information exchange is important, hospitals and research institutes may still be reluctant to give out precious bio-medical data due to confidentiality, liability, and other concerns. Thus it is important to develop privacy preserving data mining methods, such as [1], to maximally protect privacy while achieving effective data mining.

## 3. CONCLUSIONS

Both data mining and bioinformatics are fast expanding research frontiers. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for scalable and effective bio-data analysis. We believe that the active interactions and collaborations between these two fields have just started and a lot of exciting results will appear in the near future.

## 4. REFERENCES


[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD'00*, pp. 439–450, Dallas, TX, May 2000.

[2] A. Baxevanis and B. F. F. Ouellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (2nd ed.)*. John Wiley & Sons, 2001.

[3] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining database structure; or how to build a data quality browser. In *SIGMOD'02*, pp. 240–251, Madison, WI, June 2002.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probability Models of Proteins and Nucleric Acids*. Cambridge University Press, 1998.

[5] W. J. Ewens and G. R. Grant. *Statistical Methods in Bioinformatics: An Introduction*. Springer-Verlag, New York, 2001.

[6] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.

[8] A. M. Lesk. *Introduction to Bioinformatics*. Oxford University Press, 2002.

[9] V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *VLDB'01*, pp. 381–390, Rome, Italy, Sept. 2001.

[10] H. Wang, J. Yang, W. Wang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *SIGMOD'02*, pp. 418–427, Madison, WI, June 2002.

[11] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2001.

[12] J. Yang, P. S. Yu, W. Wang, and J. Han. Mining long sequential patterns in a noisy environment. In *SIGMOD'02*, pp. 406–417, Madison, WI, June 2002.