Biological Data Mining

George Tzanis, Christos Berberidis, and Ioannis Vlahavas
Department of Informatics,
Aristotle University of Thessaloniki, Greece

INTRODUCTION

At the end of the 1980's a new discipline, named data mining, emerged. The introduction of new technologies such as computers, satellites, new mass storage media and many others have lead to an exponential growth of collected data. Traditional data analysis techniques often fail to process large amounts of -often noisy- data efficiently, in an exploratory fashion. The scope of data mining is the knowledge extraction from large data amounts with the help of computers. It is an interdisciplinary area of research, that has its roots in databases, machine learning, and statistics and has contributions from many other areas such as information retrieval, pattern recognition, visualization, parallel and distributed computing. There are many applications of data mining in real world. Customer relationship management, fraud detection, market and industry characterization, stock management, medicine, pharmacology, and biology are some examples (Two Crows Corporation, 1999).

Recently, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies and the introduction of new ones such as the microarrays. These technological advances have assisted the conduct of large scale experiments and research programs. An important example is the Human Genome Project, that was founded in 1990 by the U.S. Department of Energy and the U.S. National Institutes of Health (NIH) and was completed in 2003 (U.S. Department of Energy Office of Science, 2004). A representative example of the rapid biological data accumulation is the exponential growth of GenBank (Figure 1), the U.S. NIH genetic sequence database. (National Center for Biotechnology Information, 2004). The explosive growth in the amount of biological data demands the use of computers for the organization, the maintenance and the analysis of these data.

This led to the evolution of bioinformatics, an interdisciplinary field at the intersection of biology, computer science, and information technology. As Luscombe et al. (2001) mention, the aims of bioinformatics are:

- The organization of data in such a way that allows researchers to access existing information and to submit new entries as they are produced.
- The development of tools that help in the analysis of data.
- The use of these tools to analyze the individual systems in detail, in order to gain new biological insights.

The field of bioinformatics has many applications in the modern day world, including molecular medicine, industry, agriculture, stock farming, and comparative studies (2can Bioinformatics, 2004).

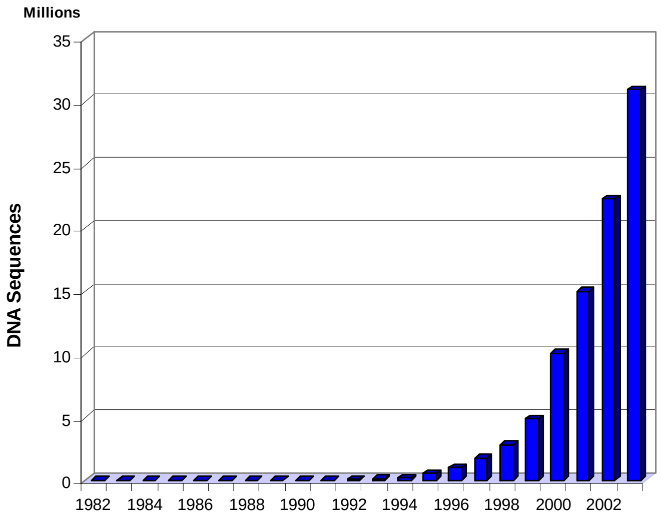

Figure 1: Growth of GenBank (Years 1982-2003).

BACKGROUND

One of the basic characteristics of life is its diversity. Everyone can notice this by just observing the great differences among living creatures. Despite this diversity, the molecular details underlying living organisms are almost universal. Every living organism depends on the activities of a complex family of molecules called proteins. Proteins are the main structural and functional units of an organism's cell. Typical example of proteins are the enzymes that catalyze (accelerate) chemical reactions. There are four levels of protein structural arrangement (conformation) as listed in Table 1 (Brazma et al., 2001). The statement about unity among organisms is strengthened by the observation that similar protein sets, having similar functions, are found in very different organisms (Hunter, 2004). Another common characteristic of all organisms is the presence of a second family of molecules, the nucleic acids. Their role is to carry the information that "codes" life. The force that created both the unity and the diversity of living things is evolution (Hunter, 2004).

- **Primary Structure**. The sequence of amino acids, forming a chain called polypeptide.
- **Secondary Structure**. The structure that forms a polypeptide after folding.
- **Tertiary Structure**. The stable 3D structure that forms a polypeptide.
- **Quaternary Structure**. The final 3D structure of the protein formed by the conjugation of two or more polypeptides.

Table 1: The Four Levels of Protein Conformation.

Proteins and nucleic acids are both called biological macromolecules, due to their large size compared to other molecules. Important efforts towards understanding life are made by studying the structure and function of biological macromolecules. The branch of biology concerned in this study is called molecular biology.

Both proteins and nucleic acids are linear polymers of smaller molecules called monomers. The term sequence is used to refer to the order of monomers that constitute a macromolecule. A sequence can be represented as a string of different symbols, one for each monomer. There are twenty protein monomers called amino acids. There exist two nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), composed by four different monomers called nucleotides. DNA is the genetic material of almost every living organism. RNA has many functions inside a cell and plays an important role in protein synthesis (Table 2). Moreover, RNA is the genetic material for some viruses such as HIV, which causes AIDS.

The genetic material of an organism is organized in long double stranded DNA molecules called chromosomes. An organism may contain one or more chromosomes. Gene is a DNA sequence located in a particular chromosome and encodes the information for the synthesis of a protein or RNA molecule. All the genetic material of a particular organism constitutes its genome.

- **Messenger RNA (mRNA)**. Carries information from DNA to protein synthesis site.
- **Ribosomal RNA (rRNA)**. The main constituent of ribosomes, the cellular components where the protein synthesis takes place.
- **Transfer RNA (tRNA)**. Transfers the amino acids to ribosomes.

Table 2: Some of the Basic Types of RNA.

The central dogma of molecular biology, as coined by Francis Crick (1958), describes the flow of genetic information (Figure 2). DNA is transcribed into RNA and then RNA is translated into proteins. The circular arrow around DNA denotes its replication ability. However, today is known that in retroviruses RNA is reverse transcribed into DNA. Moreover, in some viruses RNA is able to replicate itself. The extended statement of central dogma of molecular biology is depicted in Figure 3.

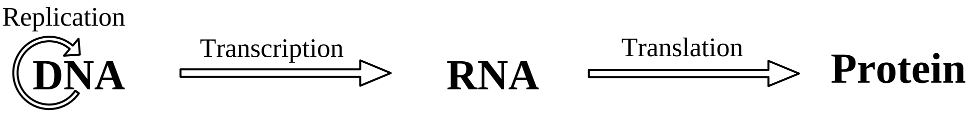

Figure 2: The Central Dogma of Molecular Biology (Initial Statement).

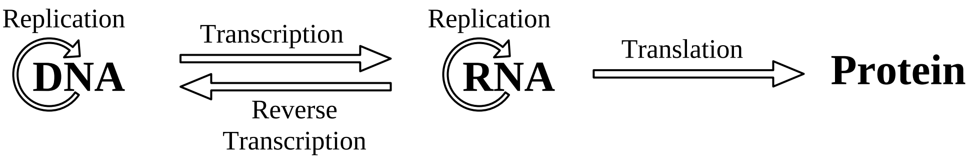

Figure 3: The Central Dogma of Molecular Biology (Extended Statement).

Houle et al. (2000) refer to a classification of three successive levels for the analysis of biological data, that is identified on the basis of the central dogma of molecular biology:

1. Genomics is the study of an organism's genome and deals with the systematic use of genome information to provide new biological knowledge.

2. Gene expression analysis is the use of quantitative mRNA-level measurements of gene expression (the process by which a gene’s coded information is converted into the structural and functional units of a cell) in order to characterize biological processes and elucidate the mechanisms of gene transcription (Houle et al., 2000).
3. Proteomics is the large-scale study of proteins, particularly their structures and functions. (Wikipedia).

These application domains are examined in the following paragraphs.

As many genome projects (the endeavors to sequence and map genomes) like the Human Genome Project have been completed, there is a paradigm shift from static structural genomics to dynamic functional genomics (Houle et al., 2000). The term structural genomics refers to the DNA sequence determination and mapping activities, while functional genomics refers to the assignment of functional information to known sequences. There are particular DNA sequences, that have a specific biological role. The identification of such sequences is a problem that concerns bioinformatics scientists. One such sequence is transcription start site, which is the region of DNA where transcription (the process of mRNA production from DNA) starts. Another biologically meaningful sequence is the translation initiation site, which is the site where translation (protein production from mRNA) initiates.

Although every cell in an organism -with only few exceptions- has the same set of chromosomes, two cells may have very different properties and functions. This is due to the differences in abundance of proteins. The abundance of a protein is partly determined by the levels of mRNA which in turn are determined by the expression or non-expression of the corresponding gene. A tool for analyzing gene expression is microarray. A microarray experiment measures the relative mRNA levels of typically thousands of genes, providing the ability to compare the expression levels of different biological samples. These samples may correlate with different time points taken during a biological process or with different tissue types such as normal cells and cancer cells (Aas, 2001). An example raw microarray image is illustrated in Figure 4.

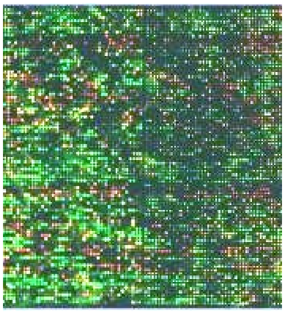

Figure 4: An illuminated microarray (enlarged). A typical dimension of such an array is about 1 inch or less, the spot diameter is of the order of 0.1 mm, for some microarray types can be even smaller. (Brazma et al., 2001).

Serial Analysis of Gene Expression (SAGE) is a method that allows the quantitative profiling of a large number of transcripts (Velculescu et al., 1995). A transcript is a sequence of mRNA produced by transcription. However, this method is very expensive in contrast to microarrays, thus there is a limited amount of publicly available SAGE data.

One of the concerns of Proteomics is the prediction of protein properties such as active sites, modification sites, localization, stability, globularity, shape, protein domains, secondary structure and interactions (Whishart, 2002). Secondary structure prediction is one of the most important problems in proteomics. The interaction of proteins with other biomolecules is another important issue.

MINING BIOLOGICAL DATA

Data mining is the discovery of useful knowledge from databases. It is the main step in the process known as Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996), although the two terms are often used interchangeably. Other steps of the KDD process are the collection, selection, and transformation of the data and the visualization and evaluation of the extracted knowledge. Data mining employs algorithms and techniques from statistics, machine learning, artificial intelligence, databases and data warehousing etc. Some of the most popular tasks are classification, clustering, association and sequence analysis, and regression. Depending on the nature of the data as well as the desired knowledge there is a large number of algorithms for each task. All of these algorithms try to fit a model to the data (Dunham, 2002). Such a model can be either predictive or descriptive. A predictive model makes a prediction about data using known examples, while a descriptive model identifies patterns or relationships in data. Table 3 presents the most common data mining tasks (Dunham, 2002).

Predictive	**Descriptive**
Classification. Maps data into predefined classes. **Regression.** Maps data into a real valued prediction variable.	**Association Analysis.** The production of rules that describe relationships among data. **Sequence Analysis.** Same as association, but sequence of events is also considered. **Clustering.** Groups similar input patterns together.

Table 3: Common Data Mining Tasks.

Many general data mining systems such as SAS Enterprise Miner, SPSS, S-Plus, IBM Intelligent Miner, Microsoft SQL Server 2000, SGI MineSet, and Inxight VizServer can be used for biological data mining. However, some biological data mining tools such as GeneSpring, Spot Fire, VectorNTI, COMPASS, Statistics for Microarray Analysis, and Affymetrix Data Mining Tool have been developed (Han, 2002). Also, a large number of biological data mining tools is provided by National Center for Biotechnology Information and by European Bioinformatics Institute.

Data Mining in Genomics

Many data mining techniques have been proposed to deal with the identification of specific DNA sequences. The most common include neural networks, Bayesian classifiers, decision trees, and Support Vector Machines (SVMs) (Ma & Wang, 1999; Hirsh & Noordewier, 1994; Zien et al., 2000). Sequence recognition algorithms exhibit performance tradeoffs between increasing sensitivity (ability to detect true positives) and decreasing selectivity (ability to exclude false positives) (Houle et al., 2000). However, as Li et al. (2003) state, traditional data mining techniques cannot be directly applied to this type of recognition problems. Thus, there is the need to adapt the existing techniques to this kind of problems. Attempts to overcome this problem have been made using feature generation and feature selection (Zeng & Yap, 2002; Li et al., 2003). Another data mining application in genomic level is the use of clustering algorithms to group structurally related DNA sequences.

Gene Expression Data Mining

The main types of microarray data analysis include (Piatetsky-Shapiro & Tamayo, 2003): gene selection, clustering, and classification.

Piatetsky-Shapiro and Tamayo (2003) present one great challenge that data mining practitioners have to deal with. Microarray datasets -in contrast with other application domains- contain a small number of records (less than a hundred), while the number of fields (genes), is typically in thousands. The same case is in SAGE data. This increases the likelihood of finding “false positives”.

An important issue in data analysis is feature selection. In gene expression analysis the features are the genes. Gene selection is a process of finding the genes most strongly related to a particular class. One benefit provided by this process is the reduction of the foresaid dimensionality of dataset. Moreover, a large number of genes are irrelevant when classification is applied. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied.

Clustering is the far most used method in gene expression analysis. Tibshirani et al. (1999) and Aas (2001) provide a classification of clustering methods in two categories: one-way clustering and two-way clustering. Methods of the first category are used to group either genes with similar behavior or samples with similar gene expressions. Two-way clustering methods are used to simultaneously cluster genes and samples. Hierarchical clustering is currently the most frequently applied method in gene expression analysis. An important issue concerning the application of clustering methods in microarray data is the assessment of cluster quality. Many techniques such as bootstrap, repeated measurements, mixture model-based approaches, sub-sampling and others have been proposed to deal with the cluster reliability assessment (Kerr & Churchill, 2001; Yeung et al., 2003; Ghosh & Chinnaiyan, 2002; Smolkin & Ghosh, 2003).

In microarray analysis classification is applied to discriminate diseases or to predict outcomes based on gene expression patterns and perhaps even identify the best treatment for given genetic signature (Piatetsky-Shapiro & Tamayo, 2003).

Table 4 lists the most commonly used methods in microarray data analysis. Detailed descriptions of these methods can be found in literature (Aas, 2001; Tibshirani et al., 1999; Hastie et al., 2000; Lazzeroni & Owen, 2002; Dudoit et al., 2002; Golub et al., 1999).

One-way Clustering	**Two-way Clustering**	**Classification**
Hierarchical Clustering Self-organizing Maps (SOMs) K-means Singular Value Decomposition (SVD)	Block Clustering Gene Shaving Plaid Models	SVMs K-nearest Neighbors Classification/Decision Trees Voted Classification Weighted Gene Voting Bayesian Classification

Table 4: Popular microarray data mining methods

Most of the methods used to deal with microarray data analysis can be used for SAGE data analysis.

Finally, machine learning and data mining can be applied in order to design microarray experiments except to analyze them (Molla et al., 2004).

Data Mining in Proteomics

Many modification sites can be detected by simply scanning a database that contains known modification sites. However, in some cases, a simple database scan is not effective. The use of neural networks provides better results in these cases. Similar approaches are used for the prediction of active sites. Neural network approaches and nearest neighbor classifiers have been used to deal with protein localization prediction (Whishart, 2002). Neural networks have also been used to predict protein properties such as stability, globularity and shape. Whishart refers to the use of hierarchical clustering algorithms for predicting protein domains.

Data mining has been applied for the protein secondary structure prediction. This problem has been studied for over than 30 years and many techniques have been developed (Whishart, 2002). Initially, statistical approaches were adopted to deal with this problem. Later, more accurate techniques based on information theory, Bayes theory, nearest neighbors, and neural networks were developed. Combined methods such as integrated multiple sequence alignments with neural network or nearest neighbor approaches improve prediction accuracy.

A density based clustering algorithm (GDBSCAN) is presented by Sander et al. (1998), that can be used to deal with protein interactions. This algorithm is able to cluster point and spatial objects according to both, their spatial and non-spatial attributes.

FUTURE TRENDS

Because of the special characteristics of biological data, the variety of new problems and the extremely high importance of bioinformatics research, a large number of critical issues is still open and demands active and collaborative research by the academia as well as the industry. Moreover, new technologies such as the microarrays led to a constantly increasing number of new questions on new data. Examples of hot problems in bioinformatics are the accurate prediction of protein structure and gene behavior analysis in microarrays. Bioinformatics demands and provides the opportunities for novel and improved data mining methods development. As Houle et al. (2000) mention, these improvements will enable the prediction of protein function in the context of higher order processes such as the regulation of gene expression, metabolic pathways (series of chemical reactions within a cell, catalyzed by enzymes) and signaling cascades (series of reactions which occur as a result of a single stimulus). The final objective of such analysis will be the illumination of the way conveying from genotype to phenotype.

CONCLUSION

The recent technological advances, have led to an exponential growth of biological data. New questions on these data have been generated. Scientists often have to use exploratory methods instead of confirming already suspected hypotheses. Data mining is a research area that aims to provide the analysts with novel and efficient computational tools to overcome the obstacles and constraints posed by the traditional statistical methods. Feature selection, normalization, and standardization of the data, visualization of the results and evaluation of the produced knowledge are equally important steps in the knowledge discovery process. The mission of bioinformatics as a new and critical research domain is to provide the tools and use them to extract accurate and reliable information in order to gain new biological insights.

REFERENCES

2can Bioinformatics (2004. March 15). Bioinformatics Introduction. Available: http://www.ebi.ac.uk/2can/bioinformatics/index.html

Aas, K. (2001). Microarray Data Mining: A Survey. NR Note, SAMBA, Norwegian Computing Center.

Brazma, A., Parkinson, H., Schlitt, T. and Shojatalab, M. (2001). A Quick Introduction to Elements of Biology - Cells, Molecules, Genes, Functional Genomics, Microarrays. Retrieved February 2, 2005 from http://www.ebi.ac.uk/microarray/biology_intro.html

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association 97(457), 77-87.

Dunham, M.H. (2002). Data mining: Introductory and advanced topics. Prentice Hall, Upper Saddle River, New Jersey, USA.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. AAAI Press/MIT Press, Menlo Park, California, USA.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286(5439), 531-537.

Ghosh, D. and Chinnaiyan, A.M. (2002). Mixture Modeling of Gene Expression Data from Microarray Experiments. Bioinformatics, 18, 275-286.

Han, J. (2002). How Can Data Mining Help Bio-Data Analysis? In Zaki, M.J., Wang, J.T.L. and Toivonen, H.T.T. (Eds). Proceedings of the 2nd ACM SIGKDD Workshop on Data Mining in Bioinformatics, 1-2.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. and Brown, P. (2000). ‘Gene Shaving’ as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns. Genome Biology, 1(2): research0003.

Hirsh, H. and Noordewier, M. (1994). Using Background Knowledge to Improve Inductive Learning of DNA Sequences. Proceedings of the 10th IEEE Conference on Artificial Intelligence for Applications, 351-357.

Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A. and Lundahl, S. (2004. March 10). Database Mining in the Human Genome Initiative. Whitepaper, Bio-databases.com, Amita Corporation. Available: http://www.biodatabases.com/whitepaper.html

Hunter, L. (2004). Life and Its Molecules: A Brief Introduction. AI Magazine, 25(1), 9-22.

Kerr, M.K. and Churchill, G.A. (2001). Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments. Proceedings of the National Academy of Sciences, 98, 8961-8965.

Lazzeroni, L. and Owen, A. (2002). Plaid Models for Gene Expression Data, Statistica Sinica, 12(2002), 61-86.

Li, J., Ng, K.-S. and Wong, L. (2003). Bioinformatics Adventures in Database Research. Proceedings of the 9th International Conference on Database Theory, 31-46, Siena, Italy.

Luscombe, N.M., Greenbaum, D. and Gerstein, M. (2001). What is Bioinformatics? A Proposed Definition and Overview of the Field. Methods of Information in Medicine, 40(4), 346-358.

Ma, Q. and Wang, J.T.L. (1999). Biological Data Mining Using Bayesian Neural Networks: A Case Study. International Journal on Artificial Intelligence Tools, Special Issue on Biocomputing, 8(4), 433-451.

Molla, M., Waddell, M., Page, D. and Shavlik, J. (2004). Using Machine Learning to Design and Interpret Gene-Expression Microarrays. AI Magazine, 25(1), 23-44.

National Center for Biotechnology Information (2004. June 18). Genbank Statistics. Available: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Piatetsky-Shapiro, G. and Tamayo, P. (2003). Microarray Data Mining: Facing the Challenges. SIGKDD Explorations, 5(2), 1-5.

Sander, J., Ester, M., Kriegel, P.-H. and Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications. Data Mining and Knowledge Discovery, 2(2): 169-194.

Smolkin, M. and Ghosh, D. (2003). Cluster Stability Scores for Microarray Data in Cancer Studies. BMC Bioinformatics, 4, 36.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., & Brown, P. (1999). Clustering methods for the analysis of DNA microarray data (Tech. Rep.). Department of Statistics, Stanford University, Stanford, California, USA.

Two Crows Corporation (1999). Introduction to Data Mining and Knowledge Discovery (3rd ed.).

U.S. Department of Energy Office of Science (2004. June 25). Human Genome Project Information. Available: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995). Serial Analysis of Gene Expression, Science, 270(5235), 484-487.

Whishart, D.S. (2002). Tools for Protein Technologies. In Sensen, C.W. (Ed.), Biotechnology, (Vol 5b) Genomics and Bioinformatics, 325-344, Wiley-VCH.

Wikipedia – Online Encyclopedia (2004. April 12). Available: http://en2.wikipedia.org

Yeung, Y.K., Medvedovic, M. and Bumgarner, R.E. (2003). Clustering Gene-Expression Data with Repeated Measurements. Genome Biology, 4(5), R34.

Zeng, F., Yap C.H.R. and Wong, L. (2002). Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites. Genome Informatics, 13, 192-200.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, R.-K. (2000). Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites. Bioinformatics, 16(9), 799-807.

Terms and Definitions

Data Cleaning (Cleansing): The act of detecting and removing errors and inconsistencies in data to improve its quality.

Genotype: The exact genetic makeup of an organism.

Machine Learning: An area of artificial intelligence the goal of which is to build computer systems that can adapt and learn from their experience.

Mapping: The process of locating genes on a chromosome.

Phenotype: The physical appearance characteristics of an organism.

Sequence Alignment: The process to test for similarities between a sequence of an unknown target protein, and a single (or a family of) known protein(s).

Sequencing: The process of determining the order of nucleotides in a DNA or RNA molecule or the order of amino acids in a protein.

Visualization: Graphical display of data and models facilitating the understanding and interpretation of the information contained in them.