THE ROLE OF STATISTICAL METHODS IN COMPUTER SCIENCE AND BIOINFORMATICS

Irina Arhipova
Latvia University of Agriculture, Latvia
irina.arhipova@llu.lv

This article discusses the links between computer science, statistics and biology education on the basis of research at the Latvia University of Agriculture. Bioinformatics study is considered from two aspects – as one for biologists learning Information Technologies (IT) to use within their speciality, or for IT specialists learning biology so they can apply their skills to biological problems. The different computer science technologies and statistical methods in bioinformatics are considered. The multidisciplinary approach facilitates the understanding of interrelations between computer science technologies, statistical methods and bioinformatics applications and improves the education at an agriculturally-based university. Therefore the teaching of Information Technology at agriculture-based university should exploit the close relationship with mathematics, statistics and biology.

INTRODUCTION

Now with the strengthening information technology facilities within an agriculture-based university in the Baltic countries, it was inevitable that the subject of Bioinformatics would arise. Bioinformatics can be approached from two aspects – as one for biologists learning Information Technologies (IT) to use within their specialty, or for IT specialists learning biology so they can apply their skills to biological problems. At the Latvia University of Agriculture (LUA) at the Faculty of Information Technologies the master study program "Information technologies in bio systems" is developed. The study course curricula are a combination of mathematics, computer science, computer engineering, statistics, biology and bioinformatics subjects. The master students will be trained to approach multidisciplinary tasks from their own scientific discipline. It is very important to cooperate within study program in different subjects of computer science, statistics and bioinformatics.

In this article there are different computer science technologies and statistical methods in bioinformatics considered. Multidisciplinary approach allows facilitating the understanding of interrelations between computer science technologies, statistical methods and bioinformatics applications and improves the education at an agriculture-based university.

INTERCONNECTION BETWEEN COMPUTER SCIENCE, COMPUTATIONAL STATISTICS ANS BIOINFORMATICS

Bioinformatics is the application of computational tools and techniques to the management and analysis of biological data. The term bioinformatics is relatively new, and it is related to such terms as "computational biology" and others. Bioinformatics would not be possible without advances in computing hardware and software: analysis of algorithms, data structures and software engineering (Lesk, 2002). One reason why computer scientists are attracted to molecular biology is that the way information is encoded in DNA is in some way similar to the way it is coded in computers. While computers on a basic level deal with zeros and ones (bits), DNA carries information as chains of molecules (nucleotides) that come in four different types (Eidhammer, Jonassen, and Taylor, 2004).

Besides data analysis is seen as the largest and possibly the most important area of microarray bioinformatics. Statistical analyses for differentially expressed genes are best carried out via hypothesis tests rather than using a simple fold ratio threshold. More complex data may require analysis via ANOVA or general linear models and may be also include bootstrapping. Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) provide a good way to visualise data without imposing any hierarchy on them. Hierarchical clustering can be used to identify related genes or samples and portray the usage of dendrogram. There are several methods for classifying samples, each with advantages and disadvantages, including: K-nearest neighbour,

centroid classification, linear discriminant analysis, neural network, support vector machines (Stekel, 2003).

It was well understood that computing would play a vital role in the future progress of statistics. Access to elaborate algorithms on computers increased the awareness of more recent methodological developments in statistics. According to the definition proposed by A. Westlake (Lauro, 1996): "Computational statistics is related to the advance of statistical theory and methods through the use of computational methods. This includes both the use of computation to explore the impact of theories and methods, and development of algorithms to make these ideas available to users." Computation in statistics is based on algorithms which originate in numerical mathematics or in computer science. The core topics of numerical mathematics are numerical linear algebra and optimization techniques but practically all areas of modern numerical analysis may be useful. The group of algorithms highly relevant for computational statistics from computer science is machine learning, artificial intelligence (AI), and knowledge discovery in data bases or data mining. These developments have given rise to a new research area on the borderline between statistics and computer science. Besides the difficulties resulting from new problems in various research areas, for example analysis of microarrays in biology, the following three interwoven challenges for computational statistics: handling of problems stemming from new data capture techniques, from the complexity of data structures, and from the size of data (Grossmann, Schimek, and Sint, 2004). The summary of the different subjects of science interrelationship is shown in Figure 1.

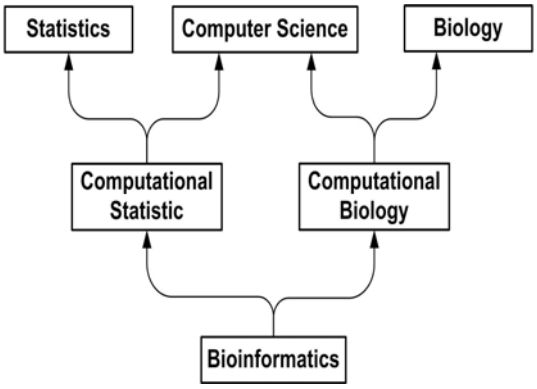

Figure 1: The interrelationship of the different subjects of sciences

Bioinformatics has grown into a large topic, but still one of the most widely used tools in bioinformatics is that for searching a sequence database for all sequences similar to a given query sequence (Waterman, 2000). There are three main bioinformatics problems:

1. Connection with "Dogma": sequence, structure and function.
2. Connection with data: keeping, access and analysis.
3. Biological process simulation: protein structure, (molecular dynamics), biological networks.

BIOINFORMATICS AND COMPUTER SCIENCE CURRICULA

New computer science curriculum standards (Roberts., 2001) included 132 units in the 14 knowledge focus groups (KFGs). Analyzed the core knowledge areas in Computing Curricula 2001, the following core topics are used to bioinformatics subject teaching: Discrete Structures, Programming Fundamentals, Algorithms and Complexity, Net-Centric Computing, Human-Computer Interaction, Intelligent Systems, Information Management, Social and professional Issues, Software Engineering and Computational Science. According to LeBlanc and Dyer, "genomic research intersected with 10 of the 14 knowledge focus groups, involving at least 36 of the 132 units." It means that background in computer science is necessary for the bioinformatics curricula development and specialists' preparation in bioinformatics. For example, computer science topics of algorithms, software engineering and databases are linked with biology topics of cell evaluation and genetics. Linking the courses of biology and computer science topics assumes

the close collaboration of the teaching staff from the different departments of university and is the precondition of the interdisciplinary research.

DATA MINING TECHNIQUES AND STATISTICAL METHODS COMPARISON

A variety of techniques have been developed over the years to explore for and extract information from large data sets. At the end of the 1980s a new discipline, named data mining, emerged. Traditional data analysis techniques often fail to process large amounts of data efficiently. Data mining is the process of discovering valid, previously unknown, and ultimately comprehensible information from large stores of data. Data Mining is the process of extracting knowledge hidden in large volumes of raw data. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by a human analyst. Modern computer data mining systems self-learn from the previous history of the investigated system, formulating and testing hypotheses about the rules which this system obeys. In general, data mining techniques can be divided into two broad categories: predictive data mining and discovery data mining (Mamcenko and Kulvietiene, 2004).

Predictive data mining is applied to a range of techniques that find relationship between a specific variable (called target variable) and the other variables in your data. The following are examples of predictive mining techniques:

- Classification is about assigning data records into pre-defined categories. In this case the target variable is the category and the techniques discover the relationship between the other variables and the category.
- Regression is about predicting the value of a continuous variable from the other variables in a data record. The most familiar value prediction techniques include linear and polynomial regression.

Discovery data mining is applied to range of techniques that find patterns inside your data without any prior knowledge of what patterns exist. The following are examples of discovery mining techniques:

- Clustering is the term for range of techniques, which attempts to group data records on the basis of how similar they are.
- Association and sequence analysis describes a family of techniques that determines relationship between data records.

The particularity of contemporary requirements for the data processing is the following: the data have the unlimited quantity, the data differ (quantitative, qualitative and textual), the results should be particular and comprehensible and the tools for the processing of raw data should be easy to use. The modern technology of Data Mining (discovery-driven data mining) is based on the concept of patterns, reflecting fragments of polydimensional relationships within the data. These patterns are regularities, which are characteristic to the data sub-retrievals that can be reflected compactly in the form, which is easy to comprehend for the human. The search for the patterns is carried out by means of techniques, which are not limited by a priori proposals about the structure of retrieval and type of the value distribution of indicators to be analyzed (Shekhar and Chawla, 2003).

The traditional mathematical statistics, for a long time applying for the role of the basic tool of data analysis, clearly gave up in front of the problems coming into existence. The main reason – the concept of the mean of retrieval that leads to the operations on the fictitious values. The techniques of mathematical statistics proved to be useful mainly for the verification of preliminary defined hypotheses (verification-driven data mining) and for the ‘rough’ research analysis that forms the basis of operative analytical data processing (online analytical processing, OLAP). At the same time the strong correlation exists between data mining and statistical methods, because statistical methods can be divided into the similar categories as data mining techniques: dependence methods and interdependence methods (Sharma, 1996). The objective of the *dependence methods* is to determine whether the set of independent variables affects the set of dependent variables individually and/or jointly. That is, statistical techniques only test for the presence or absence of relationships between the two sets of variables. At the same time there exist such data sets for which it is impossible to designate conceptually the set of variables as

dependent or independent. For these types of data sets the objectives are to identify how and why the variables are related among themselves. Statistical methods for analyzing these types of data sets are called *interdependence methods*. The classification of the data mining and statistical methods is the following:

- Data Mining
  - Predictive techniques: *Classification, Regression.*
  - Discovery techniques: *Association Analysis, Sequence Analysis, Clustering.*
- Statistical methods
  - Dependence methods: *Discriminant analysis, Logistic regression.*
  - Interdependence methods: *Correlation analysis, Correspondence analysis, Cluster analysis.*

DISCUSSION

Today, when the collection of biological data has been increasing at explosive rate, the processing of these data is needed. It would be extremely valuable if the specialists who know what the data mean were also able to imagine ways to collect, process and exploit these data. Biological and agricultural scientists should note that concepts from computer science, discrete mathematics and statistics are being used increasingly to study and describe biological systems. The specialists, who know the subjects of mathematics, statistics and computer programming, are needed for solving the computational problems in biology. Therefore the teaching of computer science at agriculture-based university should exploit the close relationship with mathematics, statistics and biology (Sestoft, 2003).

The idea of joint computer science, statistics and biology subjects is implemented in the master study program "Information technologies in bio systems" at the Latvia University of Agriculture (LUA) at the Faculty of Information Technologies. The aim of the Master's program is to offer education and training that builds upon the growing knowledge of and methodology in information technologies applied to biological systems and processes. Study program aims at computational modeling of biological phenomena and applies techniques from areas such as artificial intelligence, databases, software engineering, theoretical computer science, discrete mathematics, optimization theory, control theory and statistical modeling.

REFERENCES

Eidhammer, I., Jonassen, I., and Taylor, W. R. (2004). *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*. New York: John Wiley and Sons, Ltd.

Grossmann, W., Schimek, M. G. and Sint P. P. (2004). The history of Compstat and key-steps of statistical computing during the last 30 years. *COMPSTAT 2004. Proceedings in Computational Statistics*, (pp. 1-35).

Lauro, C. (1996). Computational statistics or statistical computing, is that the question? *Computational Statistics and Data Analysis*, 23, 191–193.

LeBlanc, M. D. and Dyer, B. D. (2005). Bioinformatics and Computing Curricula 2001. Why Computer Science is well positioned in a post-genomic world, http://genomics.wheatoncollege.edu/LeBlancDyerSIGCSE.html.

Lesk, A. M. (2002). *Introduction to Bioinformatics*. Oxford: Oxford University Press.

Mamcenko, J. and Kulvietiene, R. (2004). IBM intelligent miner for data and its application. *Scientific Proceeding of Riga Technical University, Series-Computer Science. Applied Computer Systems*, 5th Volume, (pp. 81-91).

Roberts, E. (Ed.) (2002). *Computing Curricula 2001: Computer Science Final Report*. New York: IEEE Computer Society.

Sestoft, P. (2003). Teaching information technology in the agricultural sciences. *DinaNews 4*, February, (p. 2). Copenhagen: KVL.

Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley and Sons, Inc.

Shekhar, S. and Chawla, S. (2003) *Spatial Databases: A Tour*. Englewood: Prentice Hall.

Stekel, D. (2003). *Microarray Bioinformatics*. Cambridge: Cambridge University Press.

Waterman, M. S. (2000). Introduction to computational biology. Maps, sequences and genomes. *Interdisciplinary Statistics*. Boca Raton: Chapman and Hall/CRC.