Chapter 8
Statistical Models in Bioinformatics

Stelios Zimeras
University of Aegean, Greece

Anastasia N. Kastania
Athens University of Economics & Business, Greece

ABSTRACT

In recent years, biological research has been witness of a sea change mainly spearheaded by the advent of novel high throughput technologies that can provide unprecedented amounts of valuable data. This has given rise to novel field sharing the popular suffix 'omics'. Genomics/transcriptomics, proteomics, metabolomics, interactomics/regulomics and numerous other terms have been coined to categorize this ever increasing number of new fields. Biomarkers comprise the most critical tools for the early detection, diagnosis, prognosis and prediction of diseases providing key clues for drug development processes. A significant challenge is to define appropriate levels of specificity and sensitivity of new biomarkers in detecting complex diseases. The establishment of new biomarkers is not only an issue of optimizing wet lab experiments but also of designing appropriate and robust data analysis methods. Various approaches, like multivariate analysis methods as well as standard statistical tests have been applied to search for the important features in 'omics' data. Likewise, several methods, e.g. FDA, SVM, CART, nonparametric kernels, kNN, boosted decision stump and genetic algorithms, have been reported. However, it still remains an unsolved challenge to analyze and interpret the enormous volumes of 'omics' data.

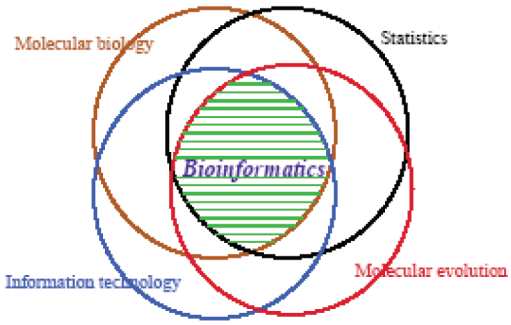
INTRODUCTION

Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature (Figure 1) (Makalowski, 2009). Research in bioinformatics includes method development for storage, retrieval, and analysis of the data.

A statistician – bioinformatician uses a collection of statistical methods for dealing with large biological data sets. In a computer science department - bioinformatics is the marriage of Computer

DOI: 10.4018/978-1-60566-768-3.ch008

Figure 1. Bioinformatics as interaction with wide areas of subjects



Science and Molecular Biology. An artificial intelligence researcher – bioinformatician uses the application of machine learning to biological data. A physicist – bioinformatician uses a collection of methods to solve a protein structure.

Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

The history of computing in biology goes back to the 1920s when scientists were already thinking of establishing biological laws from data analysis by induction (Lotka, 1925). Practical applications of bioinformatics are readily available through the World Wide Web, and are widely used in biological and medical research.

Although bioinformatics is a new term developed in the early 1990s, bioinformatics research started before 1970. Over the past four decades, bioinformatics emerged gradually from a hardly noticeable area to a mainstream discipline in science (Ouzounis and Valencia, 2003).

Analyses in bioinformatics focus on three types of datasets: genome sequences, macromolecular structures, and functional genomics experiments (e.g. expression data, yeast two–hybrid screens). But bioinformatics analysis is also applied to various other data, e.g. taxonomy trees, relationship data from metabolic pathways, the text of scientific papers, and patient statistics. A large range of techniques are used, including primary sequence alignment, protein 3D structure alignment, phylogenetic tree construction, prediction and classification of protein structure, prediction of RNA structure, prediction of protein function, and expression data clustering. Algorithmic development is an important part of bioinformatics, and techniques and algorithms were specifically developed for the analysis of biological data. A number of popular software packages and servers developed in the 1990s are widely used, as indicated by their large numbers of citations (see Figure 2) (Dong Xu, et.al, 2009; Baxevanis and Ouellette, 2001; Higgins and Taylor, 2000).

National health organizations in most developed nations allocate a good share of their public research funding to Bioinformatics, often establishing self-contained Bioinformatics institutions (Figure 3).

Regardless of the definitions, the scope of bioinformatics could address all bio-related issues, the current scope of bioinformatics is mainly at the biomolecular level, particularly on macromolecules (DNA, RNA, and proteins), biological complexes/modules involving a group of genes/proteins, and biomolecular networks/pathways that control various interactions among genes/proteins. A demanding task for bioinformatics is to extract useful biological information and pat-