# Bioinformatics and Computational Biology

**Gustavo Camps-Valls**
*Universitat de València, Spain*

**Alistair Morgan Chalk**
*Eskitis Institute for Cell and Molecular Therapies, Griffiths University, Australia*

## INTRODUCTION

Bioinformatics is a new, rapidly expanding field that uses computational approaches to answer biological questions (Baxevanis, 2005). These questions are answered by means of *analyzing* and *mining biological data.* The field of *bioinformatics* or *computational biology* is a multidisciplinary research and development environment, in which a variety of techniques from computer science, applied mathematics, linguistics, physics, and, statistics are used. The terms *bioinformatics* and *computational biology* are often used interchangeably (Baldi, 1998; Pevzner, 2000). This new area of research is driven by the wealth of data from high throughput genome projects, such as the *human genome sequencing project* (International Human Genome Sequencing Consortium, 2001; Venter, 2001). As of early 2006, 180 organisms have been sequenced, with the capacity to sequence constantly increasing. Three major DNA databases collaborate and mirror over 100 billion base pairs in Europe (EMBL), Japan (DDBJ) and the USA (Genbank.) The advent of high throughput methods for monitoring gene expression, such as microarrays (Schena, 1995) detecting the expression level of thousands of genes simultaneously. Such data can be utilized to establish gene function (*functional genomics*) (DeRisi, 1997). Recent advances in *mass spectrometry* and proteomics have made these fields high-throughput. Bioinformatics is an essential part of *drug discovery*, *pharmacology*, *biotechnology*, *genetic engineering* and a wide variety of other *biological research* areas.

In the context of these proceedings, we emphasize that *machine learning* approaches, such as neural networks, hidden Markov models, or kernel machines, have emerged as good mathematical methods for analyzing (i.e. classifying, ranking, predicting, estimating and finding regularities on) biological datasets (Baldi, 1998). The field of bioinformatics has presented challenging problems to the machine learning community and the algorithms developed have resulted in new biological hypotheses. In summary, with the huge amount of information a mutually beneficial knowledge feedback has developed between theoretical disciplines and the life sciences. As further reading, we recommend the excellent "*Bioinformatics: A Machine Learning Approac*h" (Baldi, 1998), which gives a thorough insight into topics, methods and common problems in Bioinformatics.

The next section introduces the most important subfields of bioinformatics and computational biology. We go on to discuss current issues in bioinformatics and what we see are future trends.

## BACKGROUND

Bioinformatics is a wide field covering a broad range of research topics that can broadly be defined as the management and analysis of data from generated by biological research. In order to understand bioinformatics it is essential to be familiar with at least a basic understanding of biology. The *central dogma* of molecular biology: DNA (a string of As, Cs, Gs and Ts) encodes genes which are *transcribed* into RNA (comprising As, Cs, Gs and Us) which are then generally *translated* into proteins (a string of *amino acids* – also denoted by single letter codes). The physical structure of these amino acids determines the proteins structure, which determines its function. A range of textbooks containing exhaustive information is available from the NCBI's website (http://www.ncbi.nlm.nih.gov/).

Major topics within the field of *bioinformatics* and *computational biology* can be structured into a number of categories, among which: prediction of gene expression and protein interactions, genome assembly, sequence alignment, gene finding, protein structure prediction, and evolution modeling are the most active

for the data mining community. Each of these problems requires different tools, computational techniques and machine learning methods. In the following section we briefly describe the main objectives in these areas:

1. *Databases and ontologies*. The overwhelming array of data being produced by experimental projects is continually being added to a collection of databases. The primary databases typically hold raw data and submission is often a requirement for publication. Primary databases include: a) sequence databases such as *Genbank, EMBL* and *DDBJ*, which hold nucleic acid sequence data (DNA, RNA), b) microarray databases such as *ArrayExpress* (Parkinson *et. al.* 2005), c) literature databases containing links to published articles such as *PubMed* (http://www.pubmed.com), and d) *PDB* containing protein structure data. Derived databases, created by analyzing the contents of primary databases creating higher order information such as a) protein domains, families and functional sites (*InterPro*, http://www.ebi.ac.uk/interpro/, Mulder *et. al.* 2003), and b) gene catalogs providing data from many different sources (*GeneLynx*, http://www.genelynx.org, Lenhard *et. al.* 2001, *GeneCards*, http://www.genecards.org, Safran *et. al.* 2003). An essential addition is the *Gene Ontology* project (http://www.geneontology.org). The Gene Ontology Consortium (2000), which provides a controlled vocabulary to describe genes and gene product attributes.
2. *Sequence analysis*. The most fundamental aspect of bioinformatics is sequence analysis. This broad term can be thought of as the identification of biologically significant regions in DNA, RNA or protein sequences. Genomic sequence data is analyzed to identify genes that code for RNAs or proteins, as well as regulatory sequences involved in turning on and off of genes. Protein sequence data is analyzed to identify signaling and structural information such as the location of biological active site(s). A comparison of genes within or between different species can reveal *relationships* between the genes (i.e. functional constraints). However, manual analysis of DNA sequences is impossible given the huge amount of data present. Database searching tools such as *BLAST* (http://www.ncbi.nlm.nih.gov/BLAST, Altschul *et. al.* 1990) are used to *search* the databases for similar sequences, using knowledge about *protein evolution*. In the context of genomics, *genome annotation* is the process biological features in a sequence. A popular system is the *ensembl* system which produces and maintains automatic annotation on selected eukaryotic genomes (http://www.ensembl.org).
3. *Expression analysis*. The expression of genes can be determined by measuring mRNA levels with techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, sequence tag reading (e.g., SAGE and CAGE), massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization. Recently the development of protein microarrays and high throughput *mass spectrometry* can provide a snapshot of the proteins present in a biological sample. All of these techniques, while powerful, are noise-prone and/or subject to bias in the biological measurement. Thus, a major research area in computational biology involves developing *statistical tools to separate signal from noise* in high-throughput gene expression (HT) studies. Expression studies are often used as a first step in the process of identifying genes involved in pathologies by comparing the expression levels of genes between different tissue types (e.g. breast cancer cells vs. normal cells.) It is then possible to apply *clustering* algorithms to the data to determine the properties of cancerous vs. normal cells, leading to classifiers to diagnose novel samples. For a review of the microarray approaches in cancer, see Wang (2005).
4. *Genetics and population analysis*. The genetic variation in the population holds the key to identifying disease associated genes. Common polymorphisms such as single nucleotide polymorphisms (SNPs), insertions and deletions (*indels*) have been identified and ~3 million records are in the HGVBase polymorphism database (Fredman et. al. 2004). The international HapMap project is a key resource for finding genes affecting health, disease, and drug response (The International HapMap Consortium, 2005).
5. *Structural bioinformatics*. A proteins amino acid sequence (*primary structure*), is determined from the sequence of the gene that encodes it. This structure uniquely determines its physical structure. Knowledge of structure is vital to