BIOINFORMATICS
Principles and Applications

ZHUMUR GHOSH
Scientist
Centre of Excellence in Bioinformatics
Bose Institute
Kolkata

BIBEKANAND MALLICK
Assistant Professor
Department of Life Science
National Institute of Technology
Rourkela, Odisha

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

YMCA Library Building, Jai Singh Road, New Delhi 110001

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published in India
by Oxford University Press

First published 2008
Second impression 2009

ISBN-13: 978-0-19-569230-3
ISBN-10: 0-19-569230-6

Typeset in Times
by Datapage (India) Private Limited, Chennai
Printed in India by Radha Press, Delhi 110031
and published by Oxford University Press
YMCA Library Building, Jai Singh Road, New Delhi 110001

Contents

PART III: Bio-algorithms and Tools

PART IV: Protein Modelling and Drug Design

1
A Word on Bioinformatics

'Bioinformatics is like an amoeba; it comes in various shapes and sizes.'

NANCY LORENZI

INTRODUCTION

What is bioinformatics? Is it biology or informatics or an optimized blend of both? Ask ten scientists and you will get ten different responses. There would be common elements — computers and biological database — but the definition depends on who defines it. This is the reason bioinformatics is compared to an amoeba.

The word 'bioinformatics' is a shortened form of 'biological informatics'. An unprecedented wealth of biological data has been generated by the Human Genome Project (HGP) and sequencing projects in other organisms. The huge demand for the analysis and interpretation of these data is being managed by the evolving science of bioinformatics. Bioinformatics is defined as the application of computational and analytical tools to capture and interpret the biological data.

This emerging field is turning out to be a well-opted career choice of the twenty-first century. It is fuelled by the major gene-sequencing projects, now underway, that are creating a demand for experts who understand both biology and computing and can interpret the vast amount of data generated by this type of research. The HGP, for example, has yielded data on more than three billion DNA sequences.

Bioinformatics is often focused on obtaining biologically oriented data — such as nucleic acid (DNA/RNA) and protein sequences, structures, functions, pathways, and interactions — organizing these data into databases, developing methods to get useful information from these databases, and devising methods to integrate the related data from disparate sources. These computer databases and algorithms are developed to speed up and enhance biological research. Functional genomics, biomolecular structure, proteome analysis, cell metabolism, biodiversity, downstream processing in chemical engineering, drug design, and vaccine design are some of the areas in which bioinformatics is an integral component.

Bioinformatics can give *in silico* answers to questions such as:

- Does the protein sequence dictate the functionality of the protein? If so, how?
- Are the genes turned on in a cancer cell different from those in a healthy cell?

And many more.

Bioinformatics is related to life and the story of life begins with DNA. So, before going into details about various aspects of bioinformatics, it is essential to bridge it with DNA and its relatives — genes, RNA, and protein, i.e., the building blocks of a living cell. As you might have noticed, you are a mixture of your biological parents and their preceding generations. May be you have your mother's nose and your father's eyes. How does this happen? Something must carry that information between the generations. This carrier of information of inheritance is nothing but DNA.

DNA, the carrier of information of inheritance, is like a reference book written in a strange language 'genish', which consists of only four alphabets A, T, G, and C. This book of life, the DNA, approximately 800 bibles long, contains everything about building and maintaining a living organism and directs all the events performed by a cell. DNA carries the genetic information of a cell and consists of thousands of genes (Figure 1.1). Precisely, the human genome contains several thousand genes, distributed between the 23 pairs of chromosomes in a cell. The genes are the recipes for proteins, the building blocks and workers in the body. Different genes are active in different types of cells, e.g., a liver cell does not express the same genes as a brain cell. This in turn means that different cell types, depending on their functions, produce different sets of proteins. Some proteins are vital for the survival of a cell and their corresponding genes are therefore active in all cell types and are known as 'housekeeping genes'.

The gene consists of three major structures:

- The gene regulatory segment, which contains structures involved in the initiation and regulation of transcription.
- Exons, the protein coding part of the gene.
- Introns, the non-coding part of the gene.

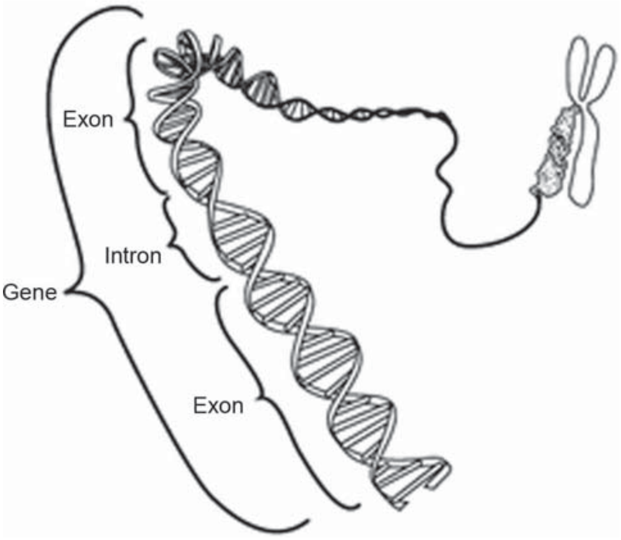

Figure 1.1 Structure of eukaryotic gene

The flow of information from the genes determines the protein composition and thereby the functions of the cell. DNA is situated in the nucleus of the cell, organized into chromosomes. Every cell must contain genetic information, so the DNA is duplicated before a cell divides; this process is known as *replication*.

Earlier, we considered DNA as a reference book. When you need to keep some information from a book, you make a copy of the pages (genes) you are interested in and return the book to the library. This way you do not risk losing or destroying the book. In all eukaryotic cells, DNA never leaves the nucleus; instead, the genetic recipe (the genes) is copied into RNA, which in turn is decoded (translated) into proteins in the cytoplasm. The DNA itself is not translated into proteins directly for several reasons. One is security. The cytoplasm is a dangerous environment for DNA; the daily transcription of genes to proteins would be harmful to the DNA, which has to stay intact to maintain life. Therefore, the RNA works as a sort of throw-away version of the DNA (like the copies from the reference book), good for limited work.

Another reason is to regulate the rate of protein synthesis. How does something as seemingly simple as the DNA's long sequence, composed of only four different letters (A, T, G, C) get converted into so many different kinds of protein molecules that perform the daily work in our body? It is important to understand this in detail (see Figure 1.2).

The path from the DNA sequence to the protein sequence is a complex process called the *central dogma* of biology. It is composed of two major steps, as shown in Figure 1.2. The first step is transcription, in which the DNA is converted into a mature messenger RNA (mRNA). The second is translation, in which the base sequence of the mRNA is 'read' and converted into an amino acid sequence. The information contained in the nucleotide sequence of the mRNA is read as three-letter words (triplets) called *codons*.

In 1953, when the structure of the DNA molecule was published by Watson and Crick, two questions were yet to be resolved. The first was the mechanism by which DNA replicated itself and the second, how a sequence of four things (the DNA bases — A, T, G, C) could encode a sequence of 20 things (the amino acids of protein). After a decade's search for the genetic code, finally a universal genetic code was announced in 1966 in Volume 31 of the Cold Spring Harbor Symposia on Quantitative Biology, contributed by Har Gobind Khorana, Severo Ochoa, Matthew Meselson, Marshall Nirenberg, and Heinrich Matthaei. They determined that the genetic code was composed of three-letter 'words' called codons and each codon codes for a specific amino acid. However, four DNA bases, 'read' three at a time, gave 64 possible codons. Since there were only 20 amino acids, this meant that more than one codon could code for the same amino acid. This phenomenon is called *degeneracy*. Amino acids coded by codons are linked together during translation to form a polypeptide chain that is later folded into a protein. In brief, we can say *DNA makes RNA makes protein*.

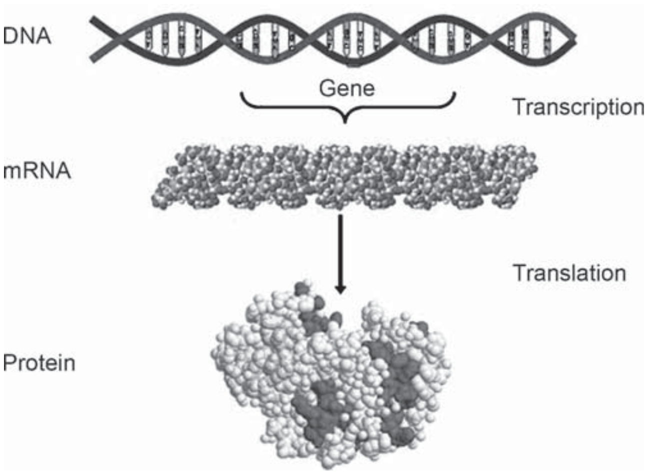

Figure 1.2 Schematic representation of information flow from DNA to protein through transcription and translation

In fact, everything we do in our everyday life emerges from the coordinated activities of a lively, inter-communicating society of protein molecules. Proteins are the building blocks of all cells and execute nearly all the cell functions.

The multiplicity of functions performed by proteins arises from the huge number of different three-dimensional (3D) shapes they adopt. Structurally, proteins are polymers of amino acids, joined together by peptide bonds in a long chain called a *polypeptide chain*. Some proteins consist of more than one polypeptide chain and they frequently associate with each other to form larger protein complexes. Human proteins are composed of only 20 different kinds of amino acids, 10 of which our body can produce by itself (non-essential amino acids) and ten that must be provided through food (essential amino acids). Each type of protein has a unique sequence of amino acids; this sequence, known as its primary structure, determines the shape (secondary and tertiary structure) (see Figure 1.3) and function of the protein.

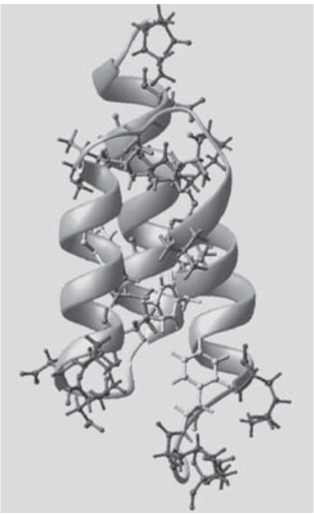

Figure 1.3 Structure of protein

BRANCHES OF BIOINFORMATICS

A living cell is a system where cellular components such as genome, the gene transcript, and the proteins interact with each other, and these interactions determine the fate of the cell, e.g., whether a stem cell is going to become a liver cell or a cancer cell. The characterization of these three types of components and the associated development of analytical methods lead to the establishment of the three closely related branches of bioinformatics: *genomics*, *transcriptomics*, and *proteomics* (see Figure 1.4).

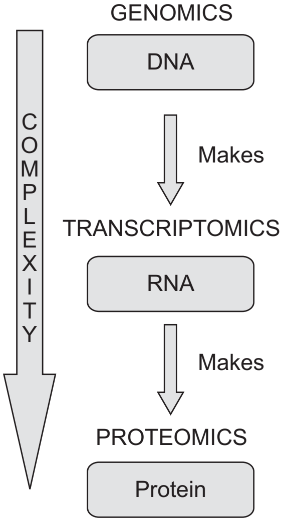

Figure 1.4 The three major branches of bioinformatics

Genomics

Genomics play a significant role in modern biological research in which the nucleotide sequences of all the chromosomes of an organism are mapped and the location of different genes and their sequences are thereby determined. This involves extensive analysis of the nucleic acids through molecular biology techniques before the data are ready for processing by computers. It is a science that attempts to describe a living organism in terms of the sequence of its genome (its constituent genetic material).

Earlier, it was not reliable to estimate the number of genes in an organism based on the number of nucleotide base pairs because of the presence of high numbers of redundant copies of many genes. Genomics has helped to rectify this problem. For example, it is now known that a human being has about 30,000 genes and not 1,00,000, as estimated earlier. Genomics uses the techniques of molecular biology and bioinformatics to identify cellular components such as proteins, rRNA, tRNA, etc., and analyse the sequences attributed to the structural genes, regulatory sequences, and even non-coding sequences. Genomics is closely related to, and sometimes considered a branch of genetics, the study of genes and heredity.

The first automatic DNA sequencer was developed in 1986 by Leroy Hood. This paved the way for the official beginning of the HGP in 1990, which gave a boost to genomics. A large number of bacterial genomes have already been fully sequenced and put in the public domain. *Haemophilus influenzae* was the first bacterium to be sequenced in 1995. The sequencing of bacterial genomes was followed by the first sequenced eukaryotic organism, the unicellular genetic model system — *Saccharomyces cerevisiae* (commonly known as baker's yeast). In December 1998, the first multicellular organism was added to the list, the nematode *Caenorhabditis elegans*, which is now considered as a model organism to provide us with information about unique functions in organisms of greater complexity. The sum of all these information is enormous and its potential in our understanding of life processes can be explored with the help of genomics, almost synonymous with bioinformatics.

Even if one can identify all the genes on a genome, the genes only indicate that, at some point in time, it might be transcribed to produce active cellular components. It contains no time-specific information on when and under what condition the gene will be expressed. For example, a human genome contains about 30,000–60,000 protein-coding genes, but only a subset of them is expressed in a particular cell type at a particular time. However, the state of the cell at time *t* depends much more on those genes expressed at time *t* than the silent ones. Genomics leads to certain developments that facilitate the generation of time-specific gene expression data.

Transcriptomics

Transcriptomics is the study of the transcriptome, which includes the whole set of mRNA molecules (or *transcripts*) in one or a population of biological cells for a given set of environmental circumstances. This study helps us to depict the expression level of genes, often using techniques such as DNA microarrays, that is capable of sampling tens of thousands of different mRNAs at a time. This kind of new technique has helped biologists to routinely monitor the gene expression of cells over time or to compare gene expression between the control cells and treatment cells.

Transcriptomics has a few limitations. The relative abundance of transcripts as characterized by the sequential analysis of gene expression (SAGE) or microarray experiments is not always a good predictor of the relative abundance of proteins. This is because of the following:

1. Differential adaptation to the translational machinery.
2. Differential usage of amino acids of different abundances.
3. The lack of information on post-translational modification of amino acid residues although post-transcriptional modifications such as acetylation, hydroxylation, glycosylation, phosphorylation, and cleavage are fundamental in understanding the interactions of cellular components.

Proteomics

Proteomics represents the earliest attempt to identify a major sub-class of cellular components — the proteins — and their interactions. Proteomics involves the sequencing of amino acids in a protein, determining its 3D structure and relating it to the function of the protein. Before computer processing comes into the picture, extensive data, particularly through crystallography and nuclear magnetic resonance (NMR), is required for this kind of a study. With such data on known proteins, the structure and its relationship to the function of newly discovered proteins can soon be understood. In such areas, bioinformatics has enormous analytical and predictive potential. Metabolic proteins such as haemoglobin and insulin have been subjected to intensive proteomic investigation. It focuses on identifying when and where the proteins are expressed in a cell so as to establish their physiological roles in an organism.

The term 'proteomics' was coined to make an analogy with genomics, and while it is often viewed as the 'next step', proteomics is much more complicated than genomics. Most importantly, while the genome is rather a constant entity, the proteome differs from cell to cell and is constantly changing through its biochemical interactions with the genome and the environment. A single organism has radically different protein expressions in different parts of its body, in different stages of its life cycle and in different environmental conditions. The complete set of proteins existing in an organism throughout its life cycle or, on a smaller scale, the set of proteins found in a particular cell type under a particular type of stimulation, is referred to as the *proteome* of the organism or cell type, respectively.

Scientists feel that the bioinformatics of proteins is crucial since characterizing thousands of proteins and their interactions is a difficult task.

To understand the cellular components and their interactions completely, one needs integrated analyses of proteomic, genomic, and transcriptomic data — and a one-word solution for all this is bioinformatics.

Apart from these three main branches, there are a few other accessory branches of bioinformatics, which we shall now discuss.

Systems Biology

As there has been an explosion of biological data because of the HGP and followed by other sequencing projects, the significant job of extracting relevant information out of this plethora of data is taken up by bioinformatics. This is important for building up meaningful models that fit well with the biological systems. Systems biology is that predictive stem of bioinformatics which, with the aid of mathematical modelling, simulation, and data analysis, generates predictive models of this experimentally generated biological data.

In simpler words, systems biology deals with the system-level understanding of biological systems. Unlike molecular biology, which focuses on molecules such as sequences of nucleic acids and proteins, systems biology focuses on interactions between the various components of a biological system, and how these interactions give rise to the function and behaviour of that system (e.g., the enzymes and metabolites in a metabolic pathway).

Systems biology is a groundbreaking scientific approach that seeks to understand how all the individual components of a biological system interact in time and space to determine the functioning of the system. It allows insight into the large amounts of data from molecular biology and genomic research, integrated with an understanding of physiology, to model the complex function of cells, organs, and whole organisms, bringing with it the potential to improve our knowledge of health and disease. It can also be defined as a system-level approach to biological complexity aimed at providing a complete understanding of the phenomena that occur in living systems beyond the molecular scale**.**

For a system-level understanding of biological systems, a few key points have to be kept in mind. These include:

1. Understanding the structure of the system, such as gene regulatory and biochemical networks, as well as physical structures.
2. Understanding the dynamics of the system. This incorporates both quantitative and qualitative analysis, as well as construction of theory/model of the system with powerful prediction capability.
3. Understanding the control methods of the system.
4. Understanding the design methods of the system.

A central tenant to systems biology is the concept of computer modelling or computer simulation (see Table 1.1 for a list of software). Indeed, this emphasis on computing has given rise to a new discipline called *computational systems biology*. Bioinformatics has now moved beyond the study of individual biological components (genes, proteins, etc.), albeit in a genome-wide context, and moved closer to systems biology, which attempts to study how individual parts (genes, proteins, etc.) interact with each other to create a stable, dynamic, and functional biological system. Computational systems biology aims to develop and use efficient algorithms, data structures, and communication tools to organize the integration of large quantities of biological data, with the goal

Table 1.1 Computational systems biology simulation software packages

Software	Web interfaces
Cell Designer	http://www.celldesigner.org/index.html
CellWare	http://www.cellware.org
Dynetica	http://www.duke.edu/∼you/Dynetica_page.htm
E-Cell	http://www.e-cell.org
Gepasi	http://www.gepasi.org/
SmartCell	http://www.smartcell.embl.de/
Vcell	http://www.vcell.org
CPN Tools	http://www.wiki.daimi.au.dk/cpntools/cpntools.wiki
MesoRD	http://www.mesord.sourceforge.net/index.phtml
SpiM	http://www.pages.cs.wisc.edu/∼larus/spim.html
BioSPI	http://www.wisdom.weizmann.ac.il/∼biospi/index_main.html
CancerSim	http://www.cs.unm.edu/∼forrest/software/cancersim/
Mcell	http://www.mcell.cnl.salk.edu/
SimCell	http://www.wishart.biology.ualberta.ca/SimCell/
AgentCell	http://www.flash.uchicago.edu/∼emonet/biology/agentcell/
Cell++	http://www.compsysbio.org/CellSim/

of modelling the dynamic characteristics of a biological system. Modelled quantities may include steady-state metabolic flux or the time-dependent response of signalling networks. The algorithmic methods used include related topics such as optimization, network analysis, graph theory, linear programming, grid computing, flux balance analysis, sensitivity analysis, and dynamic modelling.

Systems biology has become a viable approach as a result of recent developments in the biological sciences, systems engineering, imaging, mathematics, and computing. It utilizes an iterative cycle of computational modelling and laboratory experiments to understand how the components work together in a system. This opens a new gateway towards the wealth of opportunities across medicine, engineering, and other fields. One of its immediate impacts will be in the pharmaceutical sector where, by means of a more effective drug development process, systems biology will bring innovative drugs to the patients more quickly and cheaply. It will be a vital tool in elucidating many interacting factors that contribute to the causes of common medical conditions, yielding important information on cardiovascular disease and liver function. In the longer term, it will increase our understanding of cancer and other deadly diseases. Systems biology will provide a platform for the development of *synthetic biology*, the design and re-design of biological parts, devices, and systems with applications ranging from materials with enhanced properties to biofuels.

As a whole, systems biology has the potential to affect many areas of biomedical sciences, healthcare, and engineering. It is an example of multi-disciplinary research. Some of its key future outcomes fulfilling certain basic needs are as follows:

- More effective therapeutics that can tackle the underlying causes of disease rather than treating the symptoms.
- Providing bio-industry with the ability to model and manipulate biological processes better so as to provide novel compounds for the chemical, pharmaceutical, and food sectors, thereby improving the competitive edge of these industries.
- A better understanding of healthy ageing and how to maintain a population that remains healthy and productive for longer period.
- The development of predictive (*in silico*) toxicology models of cells and organs, leading to improved drug screens and reduced need for animal testing.

Given the health and economic opportunities that it presents and the substantial international and industrial interest, systems biology will be an indispensable stem of bioinformatics.

Functional Genomics

Since the completion of the first draft of the human genome, the emphasis has been on changing from genes themselves to gene products. This functional relevance to genomic information is obtained by an approach called *functional genomics*. It is the study of genes, their resulting proteins, and the role played by the proteins. Functional genomics aims at determining the functions of different genes. The insertion of the crystal protein genes from *Bacillus thuringiensis* in the genomes of several crop plants for pest resistance was the outcome of functional genomics. It helps us to understand the function of human genes, the genes of the rice plant, and other organisms; identify genes responsible for the production of specific antibodies; and produce vaccines for mass inoculation. It is now possible to identify the genes responsible for pathogenesis in the genomes of parasites and to produce DNA vaccines based on this information. The area concerned with genes responsible for the production of pharmaceutically important compounds is sometimes distinguished as *pharmacogenomics*.

Metabolomics

Metabolomics is the systematic study of the unique chemical fingerprints that specific cellular processes leave behind, specifically, it is the study of their small-molecule metabolite profiles. A primary metabolite is directly involved in normal growth, development, and reproduction. A secondary metabolite is not directly involved in those processes, but usually has an important ecological function. Examples of secondary metabolite include antibiotics and pigments. In a biological organism, the *metabolome* represents the collection of all metabolites, which are the end products of its gene expression. Thus, while mRNA gene expression data and proteomic analysis do not tell the whole story of what might be happening in a cell, metabolic profiling can give an instantaneous 'snapshot' of the physiology of that cell.

The word *metabonomics* is also used, particularly, in the context of drug toxicity assessment. There is some disagreement over the exact differences between 'metabolomics' and 'metabonomics'. In general, the term 'metabolomics' is more commonly used.

Structural Genomics

Structural genomics is aimed at determining the 3D structures of gene products in an efficient and high-throughput mode. Structural biology deals with the thorough understanding of the structure and function of one, or may be a few proteins, whereas structural genomics focuses on determining the structures of large numbers of proteins or other macromolecules without prior regard to function. Structural genomics efforts are producing a wealth of experimental data from NMR studies that are linked to high-quality 3D structures of proteins. When the focus is on proteins, this effort may be called *structural proteomics*.

Nutritional Genomics

A rapidly emerging area, nutritional genomics is the study and manipulation of genes responsible for the synthesis of nutritionally important enzymes or other molecules, often involving entire biosynthetic pathways. This will pave the way for inserting these genes into crop plants to enrich them in special ways. The first example of such a biofortified plant is Golden Rice, in which the biosynthetic machinery for β-carotene (pro-vitamin A) is introduced into the rice genome (*Oryza sativa*) to express a new feature in the rice grain. The genomes of the gene donors for golden rice, daffodil (*Narcissus pseudonarcissus*), and the bacterium *Erwinia uredovora*, have not been worked out. Nor was the genome of the rice plant available till the first successful product was generated.

Cheminformatics

Drug design through bioinformatics is one of the most actively pursued areas of research. Since the majority of drugs are low molecular weight (LMW) compounds and many of them are primarily derived from biological sources, there has always been a great interest in the study of LMW compounds of biological origin. Cheminformatics (or chemoinformatics) deals with products of secondary metabolism, which are often referred to as natural products. The physico-chemical properties and chemical structures for over 100,000 natural products are available in different databases. For most of them, the biological role in the organisms in which they are synthesized is not known, but they have some kind of bioactivity against others. This bioactivity can be turned into an advantage for therapeutic purposes, with the expertise of a pharmacologist. Cheminformatics involves organizing chemical data in a logical form to facilitate the process of understanding chemical properties and their relationship to structures, and making inferences. It also helps us to assess the properties of new compounds by comparing them with known compounds.

Glycomics

It deals with the application of bioinformatic procedures to carbohydrate research. Glycomics is the future field of bioinformatics.

Molecular Phylogeny

Phylogeny is the study of the origin and evolution of organisms. It has been estimated that four million organisms exist on earth, but not even a quarter of this number is currently known to science. So it is necessary to classify and name them properly. This would be very useful to understand the genetic and evolutionary relationships of organisms so that they may be used in a profitable manner in biotechnology and elsewhere. Biologists have constructed elegant systems of classification for the known organisms, though problems persist. All this commendable work, with over three centuries of history, was done using externally visible structural, chemical, or functional attributes of organisms. This constitutes the field of taxonomy, which is called *systematics* when the theory of organic evolution is applied to it.

With the advancements in molecular biology, biologists have used data from the genetic material to characterize organisms and to verify their classification and relationships, inferred on the basis of other evidence. Since it is impractical to use entire genomes for this purpose, nucleotide sequences of genes in the genomes from the mitochondria and chloroplasts are used. These nucleotide sequences are compared using complex computer software. Extensive work was carried out this way, comparing a large number of organisms. A number of systematists would benefit if bioinformatics professionals provided them with computer-based services to analyse their systematic data.

AIM OF BIOINFORMATICS

We have seen the various important ways in which bioinformatics can be used. The aim of bioinformatics is fourfold and includes data acquisition, tool and database development, data analysis, and data integration.

Data Acquisition

Data acquisition is primarily concerned with accessing and storing data generated directly from the biological experiments. The data generated by various sequencing projects have to be retrieved in the appropriate format, and be capable of being linked to all the information related to the DNA samples, such as the species, tissue type, and quality parameters used in the experiments. The data are organized in different databases so that the researchers can access existing information and submit new entries as and when they are produced. Examples of such database are the Entrez Genome of NCBI (for genome data) and the Protein Data Bank (for 3D macromolecular structures data). The information stored in these databases is useless until it is analysed. Thus, the purpose of bioinformatics extends much further.

Tool and Database Development

Many laboratories generate large volumes of data such as DNA sequences, gene expression information, 3D molecular structure, and high-throughput screening. Consequently, they must develop effective databases for storing and quickly accessing data. The other aim is to develop tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences. Programs such as FASTA and PSI–BLAST must consider what comprises a biologically significant match. The development of such resources requires expertise in computational theory along with a thorough understanding of biology.

For each type of data, a different database organization may have to be used. A database must be designed to allow efficient storage, search, and analysis of the data it contains. Designing a high-quality database is complicated by the fact that there are several formats for many types of data and a wide variety of ways in which the scientists may want to use the data. Many of these databases are best built using relational database architecture, usually based on Oracle or Sybase. A strong background in relational databases is a fundamental requirement for working in database development. Having some background in molecular biology techniques used to generate the data is also important. Most critical for the bioinformatics specialist is to have a strong working relationship with the researchers who will be using the database and the ability to understand and interpret their needs into functional database capabilities.

Data Analysis

The third aim is to use these tools to analyse the data and interpret the results in a biologically meaningful manner. Traditionally, biological studies examined individual systems in detail, and compared those with a few related systems. In bioinformatics, we can now conduct a global analysis of all the available data with the aim of unveiling common principles that apply across many systems and highlight novel features. Efficient analysis requires an efficiently designed database. It must allow researchers to place their query effectively and provide them with all the information they need to begin their data analysis. If queries cannot be performed, or if the performance is too slow, the whole system breaks down since scientists will not be inclined to use the database.

Once data are obtained from the database, the user must be able to easily transform it into the format appropriate for the desired analysis tools. This can be challenging, since researchers often use a combination of publicly available tools, tools developed in-house, and third-party commercial tools. Each tool may have different input and output formats. Starting in the late 1990s, there have been both commercial and in-house efforts at pharmaceutical and biotech companies aimed at reducing the formatting complexities. Such simplification efforts focus on building analytical systems with a number of tools integrated within them such that the transfer of data between tools appears seamless to the end user.

Bioinformatics analysts have a broad range of opportunities. They may write specific algorithms to analyse data or they may be expert users of analysis tools, helping scientists to understand how the tools analyse the data and how to interpret the results. Knowledge of various programming languages such as Java, PERL, C, C++, and Visual Basic is useful, if not mandatory, for those working in this area.

Data Integration

Once information has been analysed, a researcher must often associate or integrate it with the related data from other databases. For example, a scientist may run a series of gene expression analysis experiments and observe that a particular set of 100 genes is more highly expressed in a cancerous lung tissue than in a normal lung tissue. The scientist may wonder which of the genes is most likely to be truly related to the disease. To answer the question, the researcher needs to find out more information about those 100 genes, including any associated gene sequence, protein, enzyme, disease, metabolic pathways, or signal transduction pathway data. Such information helps the researcher to narrow down the list to a smaller set of genes. Doing this research requires connections or links between the different databases and a good way to present and store the information. An understanding of database architectures and the relationship between the various biological concepts in the databases is the key to effective data integration.

SCOPE/RESEARCH AREAS OF BIOINFORMATICS

Bioinformatics can be applied to several fields. We shall now take a look at some of these areas.

Genome and Sequence Analysis

Historically, bioinformatics as a concept was invented to describe the task of handling, presenting, and analysing large amounts of sequence data. Today, due to intense efforts at a number of large research centres throughout the world, data can be easily accessed by anyone on the Internet. As a consequence, it is currently almost an everyday activity in most molecular biology labs to screen these sequence databases to find the sequence homologues of a particular gene. This is not only to find homologues within a species but also to look for similar genes in other organisms, thus called orthologues. The discovery of numerous such orthologous groups of genes provides excellent support for the power of using model organisms. Sequence similarities are also used to cluster organisms according to their evolutionary relatedness and thus to create phylogenetic trees, an important tool in taxonomy. In parallel with the DNA sequencing efforts, the determination of the location of genes on chromosomes is today performed in large-scale projects for a number of organisms. These provide information that must be efficiently handled and presented.

From Sequence to 3D Structural Prediction

The function of most macromolecules is closely linked to their 3D structure, which may be most apparent for proteins and some RNA molecules. The experimental determination of these 3D structures is, however, an expensive and slow process. Novel procedures are thus urgently needed for predicting the molecular fold from the primary sequence data. Since the protein structure ultimately carries the information about the enzymatic active site or surface site for protein–protein interactions, knowledge about the protein tertiary structure will be of fundamental importance for the pharmaceutical industry in the future.

Analysis of Genome-wide Biomedical Data and Functional Genomics

In the last couple of years, the advent of biomedical large-scale analysis tools has forever changed the ways of research in biology and medicine. These technologies make it possible to simultaneously study the expressions of thousands of genes, either at the transcript or at the protein level; the thousands of possible protein–protein interactions in a cell; phenotypic analysis of thousands of mutants; etc. All these data, regardless of the type and format, have to be handled, presented, and analysed efficiently. Statisticians are already exploring this challenge for the clustering of similarly regulated genes. This clustering information is currently being evaluated as a potentially useful way of predicting the functions of functionally uncharacterized genes — a research area called *functional genomics*.

Ultimately, the prediction of gene function will involve a more complex procedure — the integrated analysis of many types of large-scale molecular data into one tentative function for the studied gene. The latter task will of course also utilize information gained by applying the described sequence analysis. In addition, genes with similar expression profiles would possibly exhibit consensus sequence elements in their regulatory regions. Identifying these sequences by automated computer methods, which is more difficult than finding clear similarities between the encoded proteins, will be a challenge that can provide extremely useful information.

Mathematical Modelling of Life Processes

The vast amounts of data generated by the genome-wide analytical technologies will not only have to be clustered, but also, more importantly, interpreted in a physiological context. Automated strategies have to be developed to be able to do so in a more sophisticated manner than is currently possible, while handling thousands of information units. This is a formidable task that incorporates modelling all molecular processes in a cell at the molecular level. Initially, this task will be approached by modelling the discrete parts of the cell's physiology like metabolic fluxes or regulatory networks. However, the integration of all this will be the ultimate challenge for bioinformatics and an important part of the final goal of biomedical science, in general, the complete molecular understanding of a living organism.

Database Building and Management

Whatever type of information is being generated, analysed, and finally interpreted, the data have to be presented to the scientific community on the Internet. The presentation of these data are challenging — the problems that arise extend from the formalism of data submission to the intelligent and clear ways of presentation. Database management is thus not only an engineering problem, but also a scientific challenge.

Clinical Applications

The clinical applications of bioinformatics can be viewed in the immediate, short, and long-term spans. The HGPs for sequencing human chromosomes were completed in 2003, producing a database of all the variations in sequences that distinguish us all. The project could have considerable impact on people living in 2020, e.g., a complete list of human gene products providing new drugs and gene therapy for single gene diseases (http://www.ornl.gov/hgmis/medicine/tnty.html).

Basic bioinformatic tools are already accessed in certain clinical situations to aid in diagnosis and treatment plans. For example, PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is accessed freely for biomedical journals cited in Medline, and Online Mendelian Inheritance in Man (OMIM at http://www3.ncbi.nlm.nih.gov/Omim/) — a search tool for human genes and genetic disorders — is used by clinicians to obtain information on genetic disorders in the clinic or hospital setting.

An example of the application of bioinformatics in new therapeutic advances is the development of new, designer targeted drugs such as Imatinib mesylate (Gleevec), which interferes with the abnormal protein made in chronic myeloid leukaemia (Imatinib mesylate was synthesized at Novartis Pharmaceuticals by identifying a lead in a high throughput screen for tyrosine kinase inhibitors and optimizing its activity for the specific kinases). The ability to identify and target specific genetic markers by using bioinformatic tools facilitated the discovery of this drug. In the long term, integrative bioinformatic analysis of genomic, pathological, and clinical data in clinical trials will reveal potential adverse drug reactions in individuals by use of simple genetic tests. Ultimately, pharmacogenomics (using genetic information to individualize drug treatment) is much likely to bring about a new age of personalized medicine. Patients will carry gene cards with their own unique genetic profile for certain drugs aimed at individualized therapy and targeted medicine free from side effects.

Drug Discovery Research

The application of bioinformatics to genomics data could be a potential boon for the discovery of new drugs. During the 1990s, many pharmaceutical and biotech companies became convinced that they could speed up their drug-discovery pipelines by taking advantage of the data from the HGP, funding their own internal genomics programs, and by collaborating with third-party genomics companies. The goal in such practical applications is to use data such as DNA sequence information and gene expression levels to help discover new drug targets. The vast majority of drugs target proteins but there are a handful of drugs, such as some chemotherapeutic agents, that bind to the DNA. In cases where the target is a protein, the drugs themselves are primarily small chemical molecules or, in some cases, small proteins such as hormones that bind to a larger protein in the body. Some drugs are therapeutic proteins delivered to the site of the disease. The extent to which genomics will actually be able to help identify validated drug targets is uncertain. Genomics and bioinformatics are still new areas and the drug development cycle can take up to 10 years. As of 2005, relatively few of the drugs on the market or in the late stages of clinical trials were discovered via genomics or bioinformatics programs.

Bioinformatics Scenario in India

Bioinformatics is a subset of biotechnology and the two of them go hand in hand. Biotechnology hubs are emerging all over the world. America, Europe, Eurasia, South-east Asia, Western Asia, and countries of the Pacific rim have hotspots of biotechnology and bioinformatics. India is particularly suited to become the country of choice for biotechnology and bioinformatics initiatives and endeavours. India holds an advantage over other countries in this respect for those ventures that seek to capitalize on the immense biodiversity available.

Establishment of Centres of Excellence in Bioinformatics

Bioinformatics is growing as an independent discipline and helping immensely to accelerate the growth of biotechnology. Simultaneously, there is enormous growth in the biological data. The Government of India has therefore decided to establish advanced research and training centres in the country by enhancing the existing infrastructure, providing additional manpower and flexibility in their governance. These centres are termed Centres of Excellence (COE) in bioinformatics. The mission of the COEs is to undertake advanced research in bioinformatics, provide doctoral and post-doctoral training, get the required high-end manpower, and develop new solutions so that the bioinformatics industry in India gets support to solve complex biological problems.

Creating Skilled Professionals

We have a responsibility to create properly trained Indian professionals in bioinformatics to keep pace with the international market and avail the emerging opportunities to the fullest.

India was the first to establish a Biotechnology Information System (BTIS) network in 1987, to create an infrastructure that harnesses biotechnology through bioinformatics. The Department of Biotechnology (DBT) has taken up this infrastructure development project and created a distributed network at a low cost. The BTIS network is today recognized as one of the major scientific networks in the world dedicated to providing state-of-the-art infrastructure, education, manpower, and tools in bioinformatics.

The principal aim of the bioinformatics programme is to ensure that India emerges a key international player in the field of bioinformatics. The following are the major thrusts of the programme:

1. To undertake advance research in frontier areas of bioinformatics and computational biology.
2. To develop world-class human resources.
3. To establish an effective academia–industry interface.
4. To pursue and promote international cooperation with leading institutions, organizations, and countries in the world.
5. To create world-class platforms for technology development, transfer, and commercialization.

Training Activities on Bioinformatics

Short-term and long-term training courses in bioinformatics for scientists from different disciplines in biology, statistics, and computer science are important and, over the years, have been found highly useful. These activities, therefore, will be intensified. Experts from other countries will be used as resource persons along with Indian experts. The knowledge base will be upgraded and the knowledge of experts converged from different disciplines to bioinformatics. However, training for proper manpower at research level alone is not sufficient. Considering the importance of the subject, some institutions and university departments have introduced bioinformatics courses at different levels.

Research and Development

Different institutes of repute throughout India have set up the infrastructure to pursue research activities in bioinformatics and biotechnology. Not only are the biological sectors of the institutes undergoing such activities, this inter-disciplinary field (see Figure 1.5) also harnesses multi-disciplinary talents. Hence, chemists, physicists, mathematicians, statisticians, and computer scientists are coming up to work hand in hand with experimental biologists. This is essential for managing data in modern biology and medicine. Table 1.2 lists some of the well-known institutes pursuing research in this field.

The major funding agencies of the Government of India that support bioinformatics initiatives in various states are listed in Table 1.3.

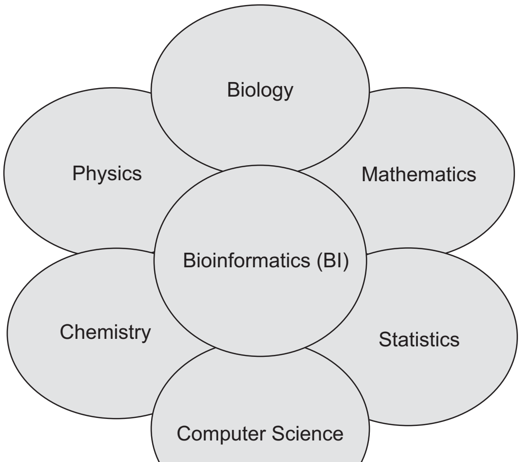

Figure 1.5 The interdisciplinary fields involved in bioinformatics

Table 1.2 Some well-known institutes pursuing research in bioinformatics

Name of the institute	Web interfaces
Indian Institute of Science (IISc), Bangalore	http://www.iisc.ernet.in
Bioinformatics Centre, University of Pune, Pune	http://www.bioinfo.ernet.in
Centre for Cellular and Molecular Biology (CCMB), Hyderabad	http://www.ccmb.res.in
Center for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad	http://www.cdfd.org.in
Central Drug Research Institute (CDRI), Lucknow	http://www.cdriindia.org
Jawaharlal Nehru University (JNU), New Delhi	http://www.jnu.ac.in
Bose Institute, Kolkata	http://www.boseinst.ernet.in
Saha Institute of Nuclear Physics (SINP), Kolkata	http://www.saha.ac.in
Indian Institute of Chemical Biology (IICB), Kolkata	http://www.iicb.res.in
International Centre for Genetic Engineering & Biotechnology (ICGEB), New Delhi	http://www.icgeb.trieste.it/RESEARCH/ND/ndrsprg.htm
Institute of Genomics & Integrative Biology (IGIB), New Delhi	http://www.igib.res.in/
National Centre for Biological Sciences (NCBS), Bangalore	http://www.ncbs.res.in/
Theoretical Physics Department, Indian Association for the Cultivation of Science (IACS), Kolkata	http://www.freewebs.com/zhumur/; http://www.geocities.com/vvekspage/
Indian Statistical Institute, Kolkata	http://www.isical.ac.in

Table 1.3 Major funding agencies of the Government of India that support bioinformatics

Funding agencies	Web interfaces
Department of Science and Technology (DST)	http://www.dst.gov.in
Department of Biotechnology (DBT)	http://www.dbtindia.nic.in
Indian Council of Agriculture Research (ICAR)	http://www.icar.org.in
Indian Council of Medical Research (ICMR)	http://www.icmr.nic.in/
Council of Scientific and Industrial Research (CSIR)	http://www.csirhrdg.res.in
University Grants Commission (UGC)	http://www.ugc.ac.in
Department of Scientific and Industrial Research (DSIR)	http://www.dsir.nic.in/
Defence Research Development Organisation (DRDO)	http://www.drdo.org

A view of the past, present, and future of bioinformatics is presented in Figure 1.6. This is a critical juncture for bioinformatics in India. The Indian bio-companies are presently the heart of the industry and growing fast. However, their performance depends on how well their research and development wings are equipped. The main shortcoming today is the deficit of trained professionals in this field. Once well-trained professionals put in sincere effort, the floodgates of international business will open for India and bioinformatics will really begin to flourish. This milestone is yet to be crossed.

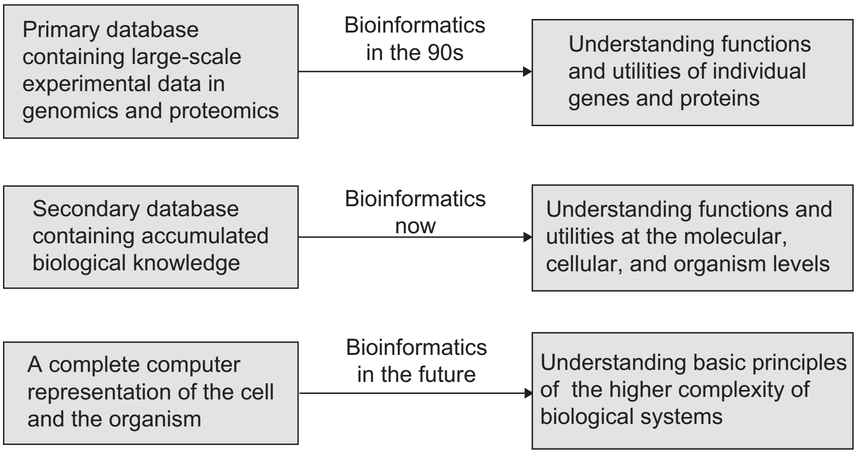
Primary database containing large-scale experimental data in genomics and proteomics
Bioinformatics in the 90s
Understanding functions and utilities of individual genes and proteins
Secondary database containing accumulated biological knowledge
Bioinformatics now
Understanding functions and utilities at the molecular, cellular, and organism levels
A complete computer representation of the cell and the organism
Bioinformatics in the future
Understanding basic principles of the higher complexity of biological systems

Figure 1.6 A view of the past, present, and future of bioinformatics

SUMMARY

Bioinformatics has created lots of fervour in various spheres starting from the academic sectors to the industrial as well as the clinical sectors. Its growing demand has emphasized the need to sow its seeds among the future students. Hence, it has become absolutely essential that they gather a proper idea about 'what is bioinformatics?' This section essentially conveys this answer along with an overview of the subject and its branches. Newcomers in this field may feel the need to get motivated while studying the subject. Such a purpose is also fulfilled in this part along with an idea of the present employment scenario and opportunities in India, which shall be beneficial to the budding professionals.

REVIEW QUESTIONS

1. What is gene? Mention the major parts of a typical eukaryotic gene.
2. What is bioinformatics? Why do people consider it as an interdisciplinary subject?
3. What are the major steps involved in central dogma?
4. What would happen if DNA directly gets translated to protein instead of RNA in the intermediate step?

SUGGESTED READING

Altman R.B. and Dugan J.M., 2003, 'Defining bioinformatics and structural bioinformatics', *Methods Biochem Anal*, **44**: 3–14.

Aoki-Kinoshita K.F. and Kanehisa M., 2006, 'Bioinformatics approaches in glycomics and drug discovery', *Curr Opin Mol Ther*, **8**(6): 514–520.

Bajorath J., 2004, 'Understanding chemoinformatics: A unifying approach', *Drug Discov Today*, **9**(1): 13–14.

Bansal A.K., 2005, 'Bioinformatics in microbial biotechnology: A mini review', *Microb Cell Fact*, **4**: 19.

Benton D., 1996, 'Bioinformatics: Principles and potential of a new multidisciplinary tool', *Trends Biotech*, **14**(8): 261–272.

Blundell T.L. and Mizuguchi K., 2000, 'Structural genomics: An overview', *Prog Biophys Mol Biol*, **73**(5): 289–295.

Brazma A. and Vilo J., 2000, 'Gene expression data analysis', *FEBS Lett*, **480**: 17–24.

Bull A.T., Ward A.C., et al., 2000, 'Search and discovery strategies for biotechnology: The paradigm shift', *Microbiol Mol Biol Rev*, **64**: 573–606.

Chen W.L., 2006, 'Chemoinformatics: Past, present, and future', *J Chem Info Model*, **46**: 2230–2255.

Couzin J., 2003, 'Functional genomics: How to make sense of sequence', *Science*, **299**: 1642.

Dawson K.A., 2006, 'Nutrigenomics: Feeding the genes for improved fertility', *Anim Reprod Sci*, **96**: 312–322.

Dutt M.J. and Lee K.H., 2000, 'Proteomic analysis', *Curr Opin Biotech*, **11**: 176–179.

Fauman E.B., Hopkins A.L., et al., 2003, 'Structural bioinformatics in drug discovery', *Methods Biochem Anal*, **44**: 477–497.

Fenstermacher D., 2005, 'Introduction to bioinformatics', *JASIST*, **56**: 440–446.

Fleischmann R., Adams M., White O., et al., 1995, 'Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd', *Science*, **269**: 496–512.

Ghosh Z. and Mallick B., 2006, 'Golden Grains', *The Statesman, 8th Day*, 15 January, pp. 8–9. (http://www.thestatesman.net/page.arcview.php?clid=30&id=131816&usrsess=1)

Ghosh Z. and Mallick B., 2006, 'Bioinformatics: The career choice of 21st century', *The Statesman, 8th Day*, 27 August 2006 (http://www.thestatesman.net/page.arcview.php?clid=30&id=156172&usrsess=1)

Ghosh Z., Chakrabarti J., and Mallick B., 2007, 'miRNomics — The bioinformatics of microRNA genes', *Biochem Biophys Res Commun*, **363**: 6–11.

Goffeau A., Barrell B.G., Bussey H., et al., 1996, 'Life with 6000 genes', *Science*, **274**: 546–567.

Goldsmith-Fischman S. and Honig B., 2003, 'Structural genomes: Computational methods for structure analysis', *Protein Sci*, **12**: 1813–1821.

Hack C. and Kendall G., 2005, 'Bioinformatics: Current practice and future challenges for life science education', *BAMBED*, **33**: 82–85.

Hann M. and Green R., 1996, 'Chemoinformatics: A new name for an old problem?' *Curr Opin Chem Biol*, **3**: 379–383.

Hood L. and Galas D., 2003, 'The digital code of DNA', *Nature*, **421**: 44–48.

Horner D.S. and Pesole G., 2004, 'Phylogenetic analyses: A brief introduction to methods and their application', *Expert Rev Mol Diag*, **4**: 339–350.

Howard M., 2000, 'The bioinformatics gold rush', *Sci Am*, **283**: 58–63.

Lin J. and Qian J., 2007, 'Systems biology approach to integrative comparative genomics', *Expert Rev Proteomics*, **4**: 107–119.

Lio P., 2003, 'Statistical bioinformatic methods in microbial genome analysis', *Bioessays*, **25**: 266–273.

Mallick B. and Ghosh Z., 2006, 'Bioinformatics: The rising sun', *Indian Science Cruiser*, **20**: 44–50.

Miller W., Makova K.D., Nekrutenko A., and Hardison R.C., 2004, 'Comparative genomics', *Ann Rev Genom Human Gen*, **5**: 15–56.

Mocellin S. and Rossi C.R., 2007, 'Principles of gene microarray data analysis', *Adv Exp Med Biol*, **593:** 19–30.

Mullner S., 2003, 'The impact of proteomics on products and processes', *Adv Biochem Eng Biotech*, **83**: 1–25.

Roos D.S., 2001, 'Bioinformatics: Trying to swim in a sea of data', *Science*, **291:** 1260–1261.

Sardari S. and Dezfulian M., 2007, 'Cheminformatics in anti-infective agents discovery', *Mini Rev Med Chem*, **7**: 181–189.

Singh O.V. and Nagaraj N.S., 2006, 'Transcriptomics, proteomics and interactomics: Unique approaches to track the insights of bioremediation', *Brief Funct Genomics Proteomic*s, **4**: 355–362.

Teufel A., Krupp M., Weinmann A., and Galle P.R., 2006, 'Current bioinformatics tools in genomic biomedical research', *Int J Mol Med*, **17**: 967–973.

Yu U., Lee S.H., Kim Y.J., and Kim S., 2004, 'Bioinformatics in the post-genome era', *J Biochem Mol Biol*, **37**: 75–82.