1

Introduction to Bioinformatics

1.1 Introduction

Interesting research fields emerge through the collaboration of researchers from different, sometimes distant, disciplines. Examples include biochemistry, biophysics, quantum information science, systems engineering, mechatronics, business information systems, management information systems, geophysics, biomedical engineering, cybernetics, art history, media technology and others. This marriage between disciplines yields findings which blend the views of different areas over the same subject or set of data.

The stimuli leading to such collaborations are numerous. For example, one discipline may develop tools that generate types of data that require another discipline to analyse. In other cases, one field scratches a layer of unknowns to discover that significant parts of its scope are actually based on the principles of another field, such as the low-level biological studies of the chemical interactions in the cells, which delivered biochemistry as an interdisciplinary field. Other interdisciplinary fields emerged because of their complementary involvement in building different parts of the same target system or in understanding different sides of the same research question; for example, mechatronics engineering aims at building systems which have both mechanical and electronic parts, such as all modern automobiles. Interdisciplinary areas like business information systems and management information systems have emerged due to the high demand for information systems which target business and management aspects; although generic information systems would meet many of those requirements, a customised field focusing on such applications is indeed more efficient given such high demand.

The interdisciplinary field of this book's focus is *bioinformatics.* The motive behind this field's emergence is the increasingly expanding generation of massive raw biological data following the developments in high-throughput techniques in the last couple of decades. The scale of this high-throughput data is orders of magnitude higher than what can be efficiently analysed

Integrative Cluster Analysis in Bioinformatics, First Edition. Basel Abu-Jamous, Rui Fa and Asoke K. Nandi.

in a manual fashion. Consequently, information engineers were recruited in order to contribute to data analysis by employing their computational methods. Cycles of computational analysis, sharing of results, interdisciplinary discussions and abstractions have led, and are still leading, to many key discoveries in biology and medicine. This success has attracted many information engineers towards biology and many biologists towards information engineering to meet in a potentially rich intersection area, which itself has grown in size to establish the field of *bioinformatics.*

1.2 The "Omics" Era

A new suffix has been introduced to the English language in this era of high-throughput data expansion; that is "-omics", and its relatives "-ome" and "-omic". This started in the 1930s when the entire set of genes carried by a chromosome was called the *genome,* blending the words "gene" and "chromosome" (OED, 2014). Consequently, the analysis of the entire genome was called *genomics,* and many known research journals carried the term "genome" or "genomics" in their titles such as Genomics, Genome Research, Genome Biology, BMC Genomics, Genome Medicine, the Journal of Genetics and Genomics (JGG), and others.

The -ome suffix was not kept exclusive for the genome; it has been rather generalised to indicate the complete set of some type of molecule or object. The proteome is the complete set of proteins in a cell, tissue or organism. Similarly are the transcriptome, metabolome, glycome and lipidome for the complete sets of transcripts, metabolites, glycans (carbohydrates) and lipids. In a respective order, large-scale studies of those complete sets are known as proteomics, transcriptomics, metabolomics, glycomics and lipidomics. The -ome suffix was further generalised to include the complete sets of objects other than basic molecules. For example, the microbiome is the complete set of microorganisms (e.g. bacteria, microscopic fungi, etc.) in a given environment such as a building, a sample of soil or the human gut (Kembel *et al.*, 2014). More *omic* fields have also emerged such as *agrigenomics* (the application of genomics in agriculture), pharmacogenomics and pharmacoproteomics (the application of genomics and proteomics to pharmacology), and others.

All of those biological fields of omics involve high-throughput datasets which are subject to information engineering involvements, and therefore reside at the core focus of bioinformatic research. An even higher level of omics analysis involves integrative analysis of many types of omic datasets. *OMICS: a Journal of Integrative Biology* is a journal which targets research studies that consider such collective analysis at different levels from single cells to societies.

More types of high-throughput omic datasets are expected to emerge. The role of bioinformatics as an interdisciplinary field will be more important. This is not only because each of those omic datasets is massive in size when considered individually; it is also because of the size of information hidden in the relations between those generally heterogeneous datasets, which requires more sophisticated computational methods to analyse.

1.3 The Scope of Bioinformatics

The scope of bioinformatics includes the development of methods, techniques and tools which target storage, retrieval, organisation, analysis and presentation of high-throughput biological data.

1.3.1 Areas of Molecular Biology Subject to Bioinformatics Analysis

In a very general statement, each part of molecular biology which produces high-throughput data is subject to bioinformatics analysis. On the other hand, low-throughput data which can be manually analysed do not represent subjects for bioinformatics. The omics fields described in the previous section are indeed included in bioinformatics analysis. This includes aspects of DNA, RNA and protein sequence analysis, gene and protein expression, genetics of diseases including cancers and special phenotypes, analysis of gene regulation, chemical interaction regulation, enzymatic regulation, other types of regulation, analysis of flowing signals in cells, networks of genetic, protein and other molecular interactions, comparable analysis of the diversity of genomes between individuals or organisms in an environment or across different environments, and others.

1.3.2 Data Storage, Retrieval and Organisation

The human genome is a linear thread of more than three billion base-pairs (letters). In 2012, and after more than 4 years after its starting point, the 1000 Genomes Project Consortium announced the completion of sequencing of the complete genomes of 1092 individuals from fourteen different populations (The 1000 Genomes Project Consortium, 2014). Moreover, the genomes of thousands of organisms, other than humans, have been sequenced and stored during the last two to three decades. As for gene expression data, tens of thousands of massive microarray datasets have been generated in the last two decades. Add to that the increasing amounts of data generated for protein expression, DNA binding and other types of high-throughput data. Data generation has not stopped and is expected to increase rapidly due to the massive advances in technologies and cost reduction. Therefore, it is crucial to store such amounts of datasets in an efficient manner which allows for quick and efficient access by large numbers of researchers from different parts of the world simultaneously.

Given the current trend, which is to offer most of the generated high-throughput datasets for public use in centralised databases, it becomes essential to standardise the way in which data are organised, annotated and labelled. This enhances information exchange and mutual understanding between different research groups in the world.

Taken together, the scope of bioinformatics indeed includes designing and implementing appropriate databases for high-throughput biological data storage, building means of data access to those databases such as web services and network applications, organising different levels of data pieces by standard formats and annotations, and, undoubtedly, maintaining and enhancing the availability and the scalability of these data repositories.

1.3.3 Data Analysis

Elaine Mardis, the Professor of Genetics in the Genome Institute at Washington University, and a collaborator in the 1000 Genomes Project, titled her "musing" published in *Genome Medicine* in 2010 as "the $1,000 genome, the $100,000 analysis?" (Mardis, 2010). Mardis discussed the tremendous drop in the cost of sequencing the complete genome of an individual human from hundreds of millions of dollars to a few thousands, and that it is expected to reach the line of $1,000. She mused, based on many facts and observations, that the cost of data analysis, which

does not seem to be dropping, will constitute the major part of the total cost, rather than the cost of data generation.

A small proportion of human genes, out of 20 000–25 000, has been well described and understood, while many gaps in our understanding of the vast majority of them do still exist. Identifying the sequence of a gene from a thousand individuals and measuring its expression profile under many conditions are not sufficient to understand its function. What increases the level of complexity is the fact that genes are highly interrelated in terms of their functions and regulation. Many genes' products work in concert to achieve a common objective; many others perform different related or unrelated tasks in different parts of the cell; many genes' products, if met within the same location, would have conflicts resulting in them negatively affecting each other's function; moreover, many genes' products control, directly or indirectly, the expression of other genes. These are examples of complexities that are not directly seen in raw sequence or expression datasets.

The quest to answer such questions and to unveil more regarding the unknowns is being carried out by large interdisciplinary collaborations, which when blended belong to the field of bioinformatics. Computational methods already existing in the field of machine learning were borrowed to be employed in biological data analysis. However, owing to the high demand, enormous size and various special characteristics, various computational methods have been designed specifically within the area of bioinformatics. Recruiting appropriate existing computational methods and/or designing customised methods for more efficient analysis of the biological data represents a large aspect of bioinformatics.

Furthermore, the gap between the existing methods in bioinformatics and the amount of information hidden in the existing data is large. This calls for more innovative out-of-the-box methods which have the ability to capture the diverse types of hidden information. A key feature of the desired methods is the ability to analyse multiple heterogeneous datasets, which can be of one or multiple types, in order to fetch high-level and low-level comprehensive and collective conclusions (Abu-Jamous *et al.*, 2013). It is expected that such methods would have a level of complexity and sophistication which enables them to delve into the inherently embodied complexities of the biological systems.

1.3.4 Statistical Analysis

In order to rely on the measurements offered by a high-throughput dataset or the results provided by a computational method, quantitative quality measures are required. Owing to the large-scale nature of analysis in bioinformatics, which typically involves large numbers of objects or samples with many stochastic variables, tests and techniques based on statistics represent the most intuitive choice for quality assessment and significance identification.

1.3.5 Presentation

A table with tens of thousands of numbers, a list of thousands of gene names, a network of tens of thousands of gene relations, a string of billions of characters representing the genome of an individual, a list of scores assigned to thousands of genes based on some computational method, a table of gene clusters produced by a partitioning algorithm, and others are examples of ways of data presentation that are not normally comprehensible by human researchers. Thus,

effective techniques for comprehensible data and results presentations are required; designing such techniques is certainly part of the roles of researchers in bioinformatics.

Presentation takes the forms of figures, tables, and text passages. Successful choice of the type of figure or table to be used depends on the ability of that figure to highlight the aspects of interest in a large amount of data in a comprehensible and conclusive way. Colour-coding, symbol-coding, lines, arrows, labels, shapes and others are examples of pieces that need to be brought together cleverly to produce a powerful figure. If such figures can be produced to visualise data or results, they are more desirable than tables, which are in their turn more desirable than text passages. Indeed, supporting text is usually required to describe the figure and explain how it should be read.

Bioinformaticians, as part of their scope of research, have been designing different types of figures and tables that suite the nature of biological high-throughput data. Many of those figures belong to conventional families of figures that are used in various areas of research, and many others are novel or customised versions of the conventional ones.

1.4 What Do Information Engineers and Biologists Need to Know?

Most fruitful bioinformatic studies usually involve collaborations between both biologists and information engineers rather than being carried out solely by one of the two parties. Though, for a successful collaboration, researchers from both sides need to take steps towards each other to reside at the interface between the two fields.

Information engineers are experts in machine learning and related areas, and in many cases, in parts of statistics. They have the ability to design, implement and apply computational methods. An information engineer who works in bioinformatics needs to learn the principles of molecular biology such as the main components and processes of the living cell at its molecular level, and what is known as the *central dogma of molecular biology*. By learning such introductory amounts of molecular biology, the information engineer will be able to comprehend terms like gene, protein, DNA, RNA, messenger RNA, non-coding RNA, chromatin, transcription, translation, gene regulation, gene expression, ribosome, genetic interaction, protein physical interaction, pathway, transport, cellular signalling and others. The information engineer will also be able to understand the structure of the commonly used high-throughput biological datasets such as nucleic acid sequences, microarrays and genetic interaction networks. Moreover, it is important, at least within the specific targeted application, to be aware of the major biological processes taking place and the main questions of research requiring answers. Additionally, familiarity with the statistical properties lying beneath the raw measurements provided by the high-throughput datasets, such as the levels and types of noise, is indeed crucial for reliable analysis. We provide most of what an information engineer needs to know in molecular biology in Parts Two and Three of this book.

In contrast, molecular biologists are familiar with the living cell and its processes. They may be familiar with the structure of high-throughput datasets, but this becomes essential for the type of dataset considered if they are involved in a bioinformatic collaboration. As for the computational methods, they usually need to know them at the *black-box* level of abstraction, that is they need to understand the structure of the raw data provided to the method as an input, the structure of the result generated by the method as an output, and the semantic relation between the input and the output. In most of the cases, they do not need to delve into the mathematical

and logical details that are carried out within the course of the method's application. However, sometimes a more detailed understanding of the method, deeper than the black-box level of abstraction, is very useful for better comprehension of its results.

It can be seen in the preceding paragraphs that both parties need to understand to a good level of detail in the structures of the datasets and the results. This is because the datasets and the results respectively represent the cargoes transferred from biologists to information engineers and from information engineers to biologists across the interface between the two fields. Accompanied with the results, and sometimes with the datasets, statistical measures are provided. Good understanding of the proper interpretation and indications of such measures is certainly required from both parties.

1.5 Discussion and Summary

The massive amounts of high-throughput biological datasets generated in the last two to three decades have motivated biologists to collaborate with information engineers in order to be able to analyse such manually-incomprehensible data. The collaborations have grown considerably, leading to the identification of the interface between biology and informatics as a standalone interdisciplinary field named as *bioinformatics.*

The scope of bioinformatics includes designing, implementing and applying methods for storing, retrieving, organising, analysing and presenting biological high-throughput data. Many of these methods were borrowed from other applications of information engineering, while many others were specially designed for bioinformatics applications. *Omics* datasets, which measure certain types of biological molecules or objects at a large scale, reside at the core focus of bioinformatics. Examples of omics include genomics, proteomics and glycomics, which are large-scale analyses of genes, proteins and glycans (carbohydrates), respectively.

In a successful collaboration, both information engineers and biologists need to have certain levels of knowledge and understanding of each other's fields. Information engineers need to understand the principles of molecular biology and the main biological questions in the context under consideration, while biologists need to learn abstract levels of description of the computational methods involved in the analysis. Both parties need to be familiar with the structure of the raw datasets and the results, as well as the correct indications and consequences of the adopted statistical measures.

References

Abu-Jamous, B., Fa, R., Roberts, D.J. and Nandi, A.K. (2013). Paradigm of tunable clustering using binarization of consensus partition matrices (Bi- CoPaM) for gene discovery. *Plos One*, **8**(2), e56432.

Kembel, S.W., Meadow, J.F., O'Connor, T.K. *et al.* (2014). Architectural design drives the biogeography of indoor bacterial communities. *Plos One*, **9**(1), e87093.

Mardis, E.R. (2010). The $1,000 genome, the $100,000 analysis? *Genome Medicine*, **2**, p. 84.

OED (2014). *Oxford English Dictionary Online*, Oxford University Press, Oxford.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, pp. 56–65.