

Chapter 1

Introduction

Bioinformatics has been in action for at least three decades. However, there is still a general confusion as to the function of bioinformatics. Some biologists are still treating bioinformatics as tools. Some informatists¹ regard bioinformatics as a career of developing novel algorithms and systems. Because of this, there is a slight difference in definitions. In the literature, one fundamental concept is also missing: that information is a natural, inherent, and dynamic component in all biological systems.

We first examine how bioinformatics is defined in various textbooks. In Attwood and Parry-Smith's book [1] bioinformatics is defined as "the application of computers in biology sciences and especially analysis of biological sequence data". In Baxevanis and Ouellette's book [2] bioinformatics is "a field integrating molecular biology and computational methods". In Higgs and Attwood's book [3] bioinformatics is defined as "the use of computational methods to study biological data". In Baldi and Brunak's book [4] bioinformatics is "the development and application of computer methods for analysis, interpretation, and prediction, as well as the design of experiments". In Mount's book [5] bioinformatics is defined as "the application of computational methods to DNA and protein science". In Augen's book [6] bioinformatics has been extended to include "*in silico* molecular modelling, protein structure prediction, and biological systems

¹ I use informatists to refer to a group of scientists who have the skills to apply the fundamental concepts in computer sciences, applied statistics, applied mathematics, and engineering to generate models.

modelling”. Finally, one of the important concepts in biological research (relationship) has been used in Eidhammer, Jonassen and Taylor’s definition [7], that bioinformatics is “the study of biological information and biological systems – such as the relationship between the sequence, structure and function of genes and proteins”.

We then examine the definitions according to dictionaries and organisations. The Oxford English Dictionary defines bioinformatics as “the science of collecting and analysing complex biological data such as genetic codes”. According to NIH, bioinformatics is defined as “research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data”. The National Center for Biotechnology Information, defines bioinformatics as “the field of science in which biology, computer science, and information technology merge into a single discipline.” NCBI also notes three important sub-disciplines within bioinformatics. The first is the development of new algorithms and statistics for accessing relationships among molecules of large data sets. The second is to analyse and interpret various data types. The outcome of these two is the integration of molecules into systems. This is also the basis of systems biology. The third is to develop and implement tools for efficient access and management of different types of information. This covers various web services and tools for public use. Both NIH and NCBI definitions cover a wide range of activities in bioinformatics.

I have no intention of giving a unique definition of bioinformatics. First, this is unfair for a huge diversity of research interests and points of views in bioinformatics. Second, the field of bioinformatics is still progressing rapidly. Many new methodologies are being developed. This book would like to treat *current* bioinformatics as a multi-discipline, inter-discipline, and cross-discipline science for understanding biological systems, exploring underlying mechanisms of biological complexes, verifying biological hypotheses and providing evidence through *in silico* simulation for further theoretical development. The requirements for bioinformaticists should not be passively taking part in biological research

projects. Instead, they should possess basic multi-disciplinary knowledge to undertake biological research activities independently leading to scientific findings. It is expected that wet laboratory and dry laboratory (*in silico* simulation) will become inseparable in the future for biosciences research.

1.1 Brief history of bioinformatics

Bioinformatics has generally gone through four major stages. In the *first* stage some small-sized databases and fundamental concepts for analysing sequences were established. The theoretical work of some great bioinformaticians laid the foundations. In the *second* stage, sequence analysis algorithms and programs as well as some moderate-sized databases were established. Along with the development of the internet, web services appeared. In the *third* stage, bioinformatics was not solely a market for sequence analysis. The analysis of other molecular data started, such as gene expression data and metabolite data in many medical applications. If we treat the second stage as the stage for natural finding (DNA discovery, protein structure/function annotation and many other hypothesis-based projects), this stage is more application-driven. Many bioinformatics projects have wide support from industry and medical services. The *fourth* stage is for systems-level examination of biological systems. This is a natural development from the third stage where it is difficult to gain a complete picture by analysing individual cases. Integrating molecules from the same data type or different data types has been an urgent task for un-biased understanding of cellular activities.

When looking at the history of bioinformatics, two important pioneering works must be remembered. The first is Pauling and Zuckerkandl's molecular evolution theory developed in the early 1960's [8, 9]. The work illustrated that amino acid sequences of proteins can be used to study evolutionary relationships among organisms. They showed that two proteins with homologous amino acid sequences have similar functions. The work therefore initiated a new field known as "molecular evolutionary". The theory provides theoretical basis for inferring protein

functions based on sequence homology. The technique is called homology alignment [10-15].

The second important work is the computerised protein and DNA sequence databases of Margaret Oakley Dayhoff in the 1970's based on her knowledge of chemistry, mathematics, biology and computer science. From this, she derived evolutionary histories using sequence homology with Pauling and Zuckerkandl's theory. She developed phylogeny for the first time with Richard Eck [16, 17]. The first probability model of protein evolution, referred to as point mutation process, was also her contribution [18]. Her quantitative measure of protein evolution, known as the mutation matrix [15], has been widely used in today's bioinformatics tools.

Based on the successes of Pauling and Dayhoff, rapid progress in bioinformatics started in the 1970's because of the rapid technology development in computers. The progress mainly focused on DNA and protein sequence analysis. Because of the time complexity, the main focus was on improving algorithm speed especially for sequence homology alignment. The comparison of genes within a species or between different species can be used to indicate structural and functional similarity. In 1970, the first sequence homology alignment algorithm was developed and is referred to as the Needleman-Wunsch algorithm [19]. The algorithm aligns two sequences globally using a dynamic programming approach. In this algorithm the comparison between two sequences is based on a binary scoring function. The score is increased by one when the current aligned residues from two sequences match, otherwise zero. In addition, linear gap penalty is used. In the algorithm insertion and deletion is considered. Therefore two sequences with different lengths will be aligned to the same length with inserted gaps. As seen above, all the matching residues have the same score as one and all the mismatching residues have the same score as zero. The first computer program for DNA sequencing was developed in 1977 [20]. The program can be used for effectively assembling sequence data. In 1981, an important concept called sequence motif for sequence analysis was generated [21]. In the same year the Smith-Waterman algorithm was developed [13]. The algorithm also aligns two sequences using a dynamic programming approach to guarantee finding the optimal

local alignment with respect to the substitution matrix and the gap penalty function used. The algorithm is a local alignment algorithm, which is due to the difficulty of obtaining correct alignments in regions of low similarity between distantly related biological sequences. However, the Smith-Waterman algorithm is a slow algorithm requiring a large memory. Because of this, it has been replaced by much more efficient algorithms for instance the FASTP algorithm published in 1983 [14], the FASTP/FASTN algorithm published in 1985 [22] and the BLAST algorithm implemented in 1990 [23].

Contributing to the third generation of bioinformatics are vast activities in analysing gene expression data. A gene is the basic unit of heredity in all living organisms; it is a segment of DNA sequence, a unit coding genetic information which is inheritable [24-26]. In other words, DNA is an organisation of information [27]. Genes are transcribed to RNAs which in turn are translated to proteins. This is controlled by a gene regulation process [24-26]. Gene expression is a process whereby a relevant gene is transcribed and translated to RNAs and proteins respectively according to a regulatory signal. These RNAs and proteins are functional in certain pathways or networks. Gene expression can be measured quantitatively using biotechnology. The measurements can be at the RNA level or protein level depending on techniques used. It is understood in molecular biology that a specific pattern of gene expression in a number of biologically related samples represents the activity of a specific signalling pathway or network. The bioinformatics study of gene expression data was triggered by the generation of DNA microarray data in the 1980's. A DNA microarray is a technology developed particularly for medicine. Each microarray is an array of thousands of DNA oligonucleotides from biologically relevant samples. The samples can be related to a specific disease diagnosis. One group of samples can be disease-free and the other can be disease-related. By analysing the pattern of expression of these DNA oligonucleotides, it is possible to investigate the genetic reason of disease development. Microarray technology evolved from Southern blotting [28] and the first use of DNA microarray expression profiling was in 1987 for identifying genes whose expression is modulated by interferon [29]. The earliest report in analysing microarray expression data of the budding yeast

Saccharomyces cerevisiae using cluster analysis approach was in 1998 by Eisen *et. al.*, [30]. Recent studies in clustering microarray expression data include those looking at renal cell carcinoma [31], inflammatory immune signalling in chronic fatigue syndrome [32], inflammation status in hepatitis C virus-related hepatocellular carcinoma [33], etc. Classification models have also been built for predictive/diagnostic purposes, such as the diagnosis of breast cancer [34], [35], colorectal cancer [36], lung cancer [37], brain cancer [38], ovarian cancer [39], etc.

In the fourth generation of bioinformatics, many researchers turn their eyes to systems biology, which is an inter-discipline and cross-discipline subject in studying biological systems. The major objective of systems biology is to discover new emergent properties of processes at the cellular level and organism level in biological systems in a systematic view. Following this, a number of systems biology institutes have been established and some doctoral training centres have also been created. Although the huge scale of systems biology studies started only a decade ago, the earliest work using the systems biology approach to study biological processes was published in the 1950's [40]. A foundation study of systems biology was completed in the 1960's with the publication of Mesarovic's book [41]. The first systems biology institute was established in 1999 [42] in the Department of Molecular Biotechnology at the University of Washington, aiming to model complex biological systems quantitatively and foster interdisciplinary interactions in the life sciences.

1.2 Database application in bioinformatics

The introduction of database technology into bioinformatics in the early days was brought about by the development of many gene/protein sequencing projects which needed an efficient way for data handling. In the 1930's, electrophoresis was developed for separating proteins in solution using moving boundary or zone electrophoresis [43]. The structure of the alpha-helix and beta-sheet was proposed in the early 1950's [44, 45] and the double helix model for DNA based on x-ray experiment was proposed in 1953 [46]. The first sequenced protein

(bovine insulin) was analysed in 1955 [47]. Herbert Boyer and Stanley Cohen invented DNA cloning or recombinant DNA technology in 1973 [48]. The technology made it possible to manipulate DNA in different species. For instance, some parts in DNA can be removed or replaced and some altered segments can be inserted into DNA. Specific proteins can be produced using gene splicing. In order to analyse the presence of a DNA sequence in a DNA sample the Southern blot was developed in 1975 [28]. The first sequenced DNA was seen in 1977 [49, 50]. In 1980, a multi-dimensional NMR method was developed for protein structure determination [51]. In 1996, the first DNA chip was generated by Affymetrix (NASDAQ: AFFX). The first gene chip product was an HIV genotyping GeneChip. The human genome with 3000 Mbp was produced in 2004 [52]. Based on this simple description of molecular data generation history it can be seen that, on the one hand, technologies are fast developing and, on the other hand, data sizes are dramatically increased, making a huge challenge for data handling, management, mining, i.e. bioinformatics.

In order to fulfil the needs in acquiring data for research, various databases continue to be established thereafter. In 1986, the largest curated protein databank SWISS-PROT was created by the Department of Medical Biochemistry of the University of Geneva and the European Molecular Biology Laboratory (EMBL). In 1988 The National Center for Biotechnology Information (NCBI) was established at the National Cancer Institute. Many successful projects of building data warehouses have well used and well developed database technology in computer sciences for a huge amount of molecular data. Efficiently storing sequence data is one important topic. A number of nucleotide sequence databases and protein sequence databases have therefore been implemented. The well-known nucleotide sequence databases include GenBank [53, 54] referred to as the NIH genetic sequence database, EMBL Nucleotide Sequence Database [55] referred to as the European equivalent to the U.S.'s GenBank, DDBJ (DNA Data Bank of Japan), Human Genome Sequencing Centre at Baylor College of Medicine, IMGT (the International ImmunoGeneTics Database) [56]. The widely used protein sequence databanks are UniProt (United Protein Databases) and Swiss-Prot. The UniProt is a centralised database cooperating with

EBI (European Bioinformatics Institute), PIR (Protein Information Resource), GUMC (Georgetown University Medical Centre), NBRF (National Biomedical Research Foundation), and SIB (Swiss Institute of Bioinformatics). The Swiss-Prot is the major European protein sequence database. In Swiss-Prot, various properties of proteins are stored such as the description of the function of a protein, protein domains structure data, and protein posttranslational modifications data.

1.3 Web tools and services for sequence homology alignment

Since DNA and protein sequencing technologies have been successfully developed, many DNA and protein sequences have been well organised and stored in various databases as mentioned above. One of the urgent tasks is to have tools which can compare two sequences to indicate how similar they are. Based on well-developed homology alignment algorithms, web tools have been developed and are open to the public. For instance, some BLAST tools are implemented in the National Center for Biotechnology Information (NCBI): nucleotide blast, used for searching a nucleotide database for a nucleotide query based on the BLASTn algorithm, protein blast, used for searching the protein database for a protein query based on BLASTp, Position-Specific Iterated – BLAST or psi – BLAST [57, 58], Pattern Hit Iterated – BLAST or phi – BLAST [59], BLASTx, tBLASTn, and tBLASTx. Most of these have been implemented as web tools. All of them deal with predictions indirectly. For instance, a query sequence may have been aligned with a number of database sequences. These database sequences have known structures and functions. If the query sequence has a high returned similarity with these database sequences, the conserved segment corresponding to protein structures or functions in these database sequences can be used for the prediction of the query sequence. A web tool will enable the user to enter a query on the internet while a server of a web tool will conduct all the necessary computing. The computing result will be returned to the user either on the web site or by an email.

The FAST/BLAST series tools are used for aligning a query sequence against many database sequences to find the most similar ones. The

algorithms implemented in all tools consider insertions and deletions. There are also two other classes of web tools implemented in bioinformatics studies, one being prediction using whole protein sequences and another being prediction using sub-sequences or peptides. These two classes of web tools are used for direct protein function prediction. For instance, the tools developed for the prediction of protein localisation [60-62], gene structure prediction [63] and function annotation [64] use whole protein sequences as input to predict protein structures and functions directly.

1.3.1 Web tools and services for protein functional site identification

Protein functional site identification using peptides includes the prediction of protein cleavage sites, protein posttranslational modification sites, binding sites, and turn types. For instance, bioinformatics algorithms and (web) tools have been used to predict proteasomal cleavage sites [65], promiscuous MHC Class-I binding sites [66], RNA binding sites [67, 68], lipoprotein signal sites [69], transcription binding sites [70], active sites [71], ligand binding site [72], miRNA target site [73], protein-protein interaction sites [68], convertase sites [74], SH3 domain interaction sites [75], and signal peptides [76].

In predicting posttranslational modification sites, there are also many web tools being developed, for instance, glycosylation site prediction [77, 78], phosphorylation site prediction [74, 79-83], acetylation site prediction [83], methylation site prediction [84], sumoylation site prediction [85], palmitoylation site prediction [86] and GPI-modification site prediction [87]. Web tools have also been implemented for protein turn prediction [88-90]. Another class of web tools for protein structure prediction uses variable peptide length for prediction. This class of web tools include protein disorder prediction and secondary structure prediction. For predicting secondary structures in proteins, the implemented web tools are PreSSAPro [91], E-SSpred [92] and MUPRED [93], PROTEUS [94], GOR V [95], Porter [96] and logic alignment approach [97]. MeDor [98], DPRoT [99], iPDA [100], PrDOC [101], FoldUnfold [102], Spritz [103], IUPred [104], RONN

[105], DisEmbl [106], TOP-IDP-scale [107], GlobPlot [108] and PONDA [109] are the web tools for disordered protein prediction.

1.3.2 Web tools and services for other biological data

Web tools have also been implemented for other biological data analysis, for instance for RNA data analysis [110], RNA deleterious mutation analysis [111], microarray data interpretation [112], transcriptional regulatory network construction [113] and for gene selection and classification [114], [115]. Web services also cover metabolite data analysis, such as correlating ligand metabolites with pathways [116] and integrating transcripts and metabolites [117]. All these efforts aim to help biologists to enhance their biological experiments and speed up scientific findings.

1.4 Pattern analysis

The third important practice in bioinformatics is pattern analysis. It covers a wide range of topics, methodologies and algorithms. This book will mainly focus on this practice providing a broad introduction and analysis. Compared with the other two subjects mentioned above, pattern analysis deals with many fundamental issues in bioinformatics. If a web tool is more or less computing technique-based, pattern analysis needs some fundamental support from statistics and mathematics. From this, models or web tools can be constructed.

Pattern analysis focuses on the exploration of the underlying mechanism of biological data. It aims to find the rules which govern data distribution. Only by knowing these rules, can proper models be constructed. For instance, in any prediction system, the most important part is a prediction model. Without fully understanding how data are distributed, no accurate or efficient model can be constructed for prediction. In order to build a proper predictor, a rigorous modelling process based on statistical modelling principles must be followed.

Pattern analysis mainly involves two learning mechanisms, i.e. unsupervised learning and supervised learning. The former is for

knowledge discovery, rule extraction and data visualisation, while the latter is for predictive model construction. There are also many different algorithms for each learning mechanism, some being simple leading to coarse but easy-to-interpret models, some being complicated leading to some accurate but difficult-to-interpret models.

In recent years, systems biology and computational systems biology have been paid increasing attention because of their importance in understanding biological systems. Conventionally, biological studies often decompose a system into some very basic and small systems. The study of these decomposed systems may miss important information of complex interplay in cells or organisms. Two trends have emerged in systems biology study. They are top-down compositional analysis, aiming at predicting system dynamics, and bottom-up integrating analysis, aiming at putting molecules into the right classes, pathways, or networks.

1.5 The contribution of information technology

The development, progress, and advances of bioinformatics could never have taken place without the support of IT successes. In 1946, came the announcement of the Turing-complete, a digital computer [118]. It is referred to as Electronic Numerical Integrator And Computer (ENIAC). The main purpose of ENIAC was to calculate artillery firing tables for the U.S. Army's Ballistic Research Laboratory although it can be used to solve various computer programming problems. The advantage of ENIAC is its speed: one thousand times faster than an electro-mechanical machine. Meanwhile, its power in dealing with mathematics for general-purpose programming promoted the spread of using computers in various applications.

In 1958, another revolution occurred in electronics which is closely related to the computer industry. The event was the development of the integrated circuit (IC) which made the manufacture of electronic equipments much faster and cheaper. Later, IC quickly progressed to very large scale IC (VLSI) leading to almost all electronic equipments including computers in use today being packed into a very small space.

Particularly, VSLI has greatly improved the efficiency of the core parts of a computer (CPU – central processing unit) in two ways. First, the size of CPU can be much smaller. Second, the memory is dramatically increased.

Because of the huge progress in electronics and computers (nowadays referred to as hardware in contrast to programming codes as software), using computers to store sequence data has become a convention. However, the following events have also made bioinformatics research feasible.

In 1969, Unix systems appeared in the Bell laboratory, which provided a powerful platform for large scale computing. The next important event was the emergence of the internet. The first internet (1st generation) was called Advanced Research Projects Agency Network (ARPAnet) established by the United States Department of Defence. ARPAnet was first established on November 21, 1969 linking the IMP at UCLA and the IMP at SRI. The 2nd generation was connecting desk PCs through telephone lines. The 3rd generation was using wireless connections to laptop computers. The 4th generation (the current one) is using mobile phone internet through cellular networks [119].

Two important network applications are email and file transfer. Email was invented in 1971. File transfer protocol (ftp) was invented in 1973. These two applications have become the most important composition parts of modern bioinformatics services. Almost all the web services and tools mentioned above include these two applications.

The other important developments in computer sciences include personal PC, window systems, Linux, Netscape, Perl programming language, Java and Java Script Programming languages; all have played important roles in promoting fast bioinformatics progress and development.

1.6 Chapters

This book is composed of 18 chapters. Except for chapters 1 and 20, the rest are divided into three parts. Chapters 2, 3, 4, 5, and 6 constitute part 1 and mainly discuss the issue of unsupervised learning. Chapter 2

introduces general concepts of unsupervised learning. Chapters 3, 4, 5, and 6 separately discuss most commonly used approaches, namely probability density estimation, principal component analysis, cluster analysis, multi-dimensional scaling, and self-organising map. Although they have some overlapped functions, each uses a distinct statistical assumption about data. All these four approaches can be used for different aspects of knowledge discovery. Chapters 7, 8, 9, 10, 11, 12, and 13 constitute part 2 and are used to cover supervised learning algorithms including linear/quadratic discriminant analysis, K-nearest neighbours, decision trees, neural networks, vector machines, and hidden Markov models. Specifically, chapter 13 focuses on an important issue in handling biological data, i.e. feature or variable selection. Chapters 14 and 15 constitute additional components for part 2 and will focus on peptide classification or functional site prediction problems. Chapters 16 and 17 constitute part 3 and will discuss computational systems biology studies including causal networks and S-systems. Chapter 18 discusses the future research directions.

Chapter 2 will focus on the general concepts of knowledge discovery approaches in bioinformatics. The chapter will discuss the principle of unsupervised learning approach and briefly introduce various unsupervised learning algorithms. The chapter will also introduce some applications of using unsupervised learning approaches to explore knowledge from large-scale biological data. Chapter 3 will introduce a useful approach in statistical learning, i.e. probability density estimation for most data analysis projects. This approach is commonly used as primary data analysis aiding proper selection of modelling algorithms. Various algorithms and procedures will also be discussed. Chapter 4 will introduce principal component analysis (PCA) and the Sammon mapping algorithm for biological data dimension reduction. PCA can lead to two outcomes, data reduction and data visualisation. In bioinformatics, PCA is commonly used to visualise data using the first and second principal components. Chapter 5 will discuss how to partition biological data through the use of various clustering algorithms. Data partitioning is commonly used in bioinformatics to visualise how data are clustered. From this, typical biological functions can be extracted. Chapter 6 will introduce the self-organising map as a neural learning algorithm which is

capable of visualising, clustering data and reducing dimensionality of data.

Chapter 7 will briefly discuss the use of supervised learning approaches in bioinformatics. Some linear algorithms will be discussed first followed by nonlinear algorithms. Chapter 8 discusses linear/quadratic discriminant analysis and K-nearest neighbour algorithm as simple learning algorithms. Chapter 9 will discuss decision trees and the random forest algorithm as well as their applications to bioinformatics for exploring human-like decision-making systems. Chapter 10 will discuss neural networks which are one of the powerful nonlinear algorithms. Because neural networks have been widely used in bioinformatics applications, various cases will be discussed. Chapter 11 will discuss recent development in nonlinear classification approaches including basis function neural networks, support vector machine and relevance vector machine. Because they have the advantage of better generalisation capability and interpretation using support/relevance vectors, their applications to bioinformatics projects have gained an increasing interest. Chapter 12 will discuss hidden Markov models which have been intensively used in sequence analysis. Chapter 13 will introduce various approaches of feature selection which are critical in analysing biological data such as gene expression and metabolite data for extracting the most informative biomarkers.

Chapter 14 will discuss the coding problem which is important to the analysis of sequence data, where residues are commonly non-numerical attributes. Several coding mechanisms will be discussed and compared. Chapter 15 will focus on one specific subject in bioinformatics, i.e. peptide classification where the main topics including data selection, organisation, target definition, and modelling procedures.

Chapter 16 will discuss how to use causal network principle and Bayesian network for constructing gene networks. Chapter 17 will discuss the developments in computational systems biology. The focus will be mainly on metabolite data analysis. Chapter 18 will outline the future research directions in bioinformatics.