Bioinformatics

Sean D. Mooney, Jessica D. Tenenbaum, and Russ B. Altman

Contents

E. H. Shortliffe, J. J. Cimino (eds.), *Biomedical Informatics*, https://doi.org/10.1007/978-3-030-58721-5_9

Learning Objectives

After reading this chapter, you should know the answers to these questions:

- Why are sequence, structure, and biological pathway information relevant to medicine?
- Where on the Internet should you look for a DNA sequence, a protein sequence, or a protein structure?
- What are two problems encountered in analyzing biological sequence, structure, and function?
- How has the age of genomics changed the landscape of bioinformatics?
- What are two computational challenges in bioinformatics for the future?

9.1 The Problem of Handling Biological Information

Bioinformatics is the study of how information is represented and analyzed in biological systems, especially information derived at the molecular level. Whereas clinical informatics deals with the management of information related to the delivery of health care, bioinformatics focuses on the management of information related to the underlying basic biological sciences. As such, the two disciplines are closely related—more so than generally appreciated (see ▶ Chap. 1). Bioinformatics and clinical informatics share a concentration on systems that are inherently uncertain, difficult to measure, and the result of complicated interactions among multiple complex components. Both deal with living systems that generally lack straight edges and right angles. Although **reductionist approaches** to studying these systems can provide valuable lessons, it is often necessary to analyze those systems using **integrative models** that are not based solely on first principles. Nonetheless, the two disciplines approach the patient from opposite directions. Whereas applications within clinical informatics usually are concerned with the social systems of medicine, the cognitive processes of medicine, and the technologies required to understand human physiology, bioinformatics is concerned with understanding how basic biological systems conspire to create molecules, organelles, living cells, organs, and entire organisms. Remarkably, however, the two disciplines share significant methodological elements, so an understanding of the issues in bioinformatics can be valuable for the student of clinical informatics and vice versa.

The discipline of bioinformatics continues to be in a period of rapid growth, because the needs for information storage, retrieval, and analysis in biology—particularly in molecular biology and **genomics**—have increased dramatically over the past two decades. History has shown that scientific developments within the basic sciences tend to have a delayed effect on clinical care and there is typically a lag of a decade before the influence of basic research on clinical medicine is realized. It cannot be understated the impact that genomics and bioinformatic approaches are having in the clinic and the point of care. Indeed, chapters focusing on "Translational Bioinformatics" and "Precision Medicine and Informatics" (▶ Chaps. 28 and 30) describe how these foundational advances are leading toward impacts on human health and improved approaches to clinical care.

9.1.1 Many Sources of Biological Data

There are many sources of information that are revolutionizing our understanding of human biology and that are creating significant challenges for computational processing. New technologies are enabling the miniaturization of laboratory experiments, increased automation of experiments and through advanced computer processing, and the interpretation of data quickly. These technologies are producing data at a staggering rate. The data produced can interrogate different views into the **Central Dogma of Biology**, the **metabolome**, the **metagenome** and ancillary molecular processes.

The most dominant new type of information is the **sequence information** produced by genetic studies. This was enabled by the **Human**

Genome Project, an international undertaking intended to determine the complete sequence of human DNA as it is encoded in each of the 23 human chromosomes. The first draft of the sequence was published in 2001 (Lander et al. 2001) and a final version was announced in 2003 coincident with the 50th anniversary of the solving of the Watson and Crick structure of the DNA double helix. The sequence continues to be revised and refined and now the sequence the genomes of many different individuals have been realized. Initially, the 1000 genomes consortium provided >1000 genomes of healthy individuals (1000 Genomes Consortium, 2010), and now datasets exist with >100,000 genomes of individuals with a variety of conditions.[1] Essentially, the entire set of genetically driven events from conception through embryonic development, childhood, adulthood, and aging are encoded by the DNA blueprints within most human cells. Given a complete knowledge of these DNA sequences, we are in a position to understand these processes at a fundamental level and to consider the possible use of DNA sequences for diagnosing and treating disease. This has led to the application of bioinformatics (and other foundational domains) as Translational Bioinformatics and Precision Medicine Informatics (► Chaps. 28 and 30).

Additionally, large-scale experimental methodologies are used to collect data on thousands or millions or more molecules simultaneously. Scientists apply these methodologies longitudinally over time and across a wide variety of organisms or within an organism to observe the development of various physiological phenomena. Technologies give us the ability to follow the production and degradation of molecules, such as the expression (transcription) of large numbers of genes simultaneously, the presence of proteins or metabolites in a biosample, or the populations of microorganisms in a sample.

The first high throughput experiments measured the expression of genes on **gene expression microarrays** (Lashkari et al. 1997). This enabled the study of the expression of large numbers of genes with one another (Bai and Elledge 1997) and to study multiple variations on a genome to explore the implications of changes in genome function on human disease. This work has led to the field of **genomics**, the study of the molecular state of a cell, tissue or organism through the state and activity of its genome. With technology advancements, gene expression can now be measured by directly sequencing messenger RNA molecules in a cell and counting the number of copies of that RNA molecule that is observed.

While some scientists are studying the human genome, other researchers are studying the functions of the genomes of numerous other biological organisms, including important model organisms (such as mouse, rat, fruit fly and yeast) as well as important human pathogens (such as *Mycobacterium tuberculosis* or *Haemophilus influenzae*). The genomes of these organisms have been determined, and efforts are underway to characterize them. These allow two important types of analysis: the analysis of mechanisms of pathogenicity and the analysis of animal models for human disease. In both cases, the functions encoded by genomes can be studied, classified, and categorized, allowing us to decipher how genomes affect human health and disease.

These ambitious scientific projects are not only proceeding at a furious pace, but also are often accompanied by another approach to biology, which produces another source of biomedical information: **proteomics**, the study of the protein gene products of the genome—the proteome. Proteomics enables researchers to discover the state (quantity and configuration) of proteins within an organism. These protein states can be correlated with different physiological conditions, including disease states. Some of these protein states can be used as identifying markers of human disease. Similar approaches are being applied to understanding the diversity, concentration levels and functions of non-DNA, RNA or protein molecules such as metabolites through the study of the small molecules in the **metabolome**.

Using these technologies together, we can now study the **epigenome**, the non-genetic

1 ► https://www.nhlbiwgs.org/ (accessed December 1, 2018).

effects that influence genome function. These include molecules that directly alter the structure of DNA but not its sequence (such as **DNA methylation**) or proteins that bind to DNA and affect how that DNA expresses genes. Epigenomics gives us a more complete picture of how biology functions and what its implications are for human health.

All these technologies, along with the genome-sequencing projects, are conspiring to produce a volume of biological information that at once contains secrets to age-old questions about health and disease and threatens to overwhelm our current capabilities of data analysis. Thus, bioinformatics is becoming critical for medicine in the twenty-first century.

9.1.2 Implications for Clinical Informatics

The effects of this new biological information on clinical medicine and clinical informatics are still evolving. It is already clear, however, that some major changes to medicine will have to be accommodated. These efforts have emerged as important areas of biomedical informatics that have become their own domains, Translational Bioinformatics (▶ Chap. 26) and Precision Medicine and Informatics (▶ Chap. 28) and use of biotechnology data is now common in Clinical Research Informatics (▶ Chap. 27).

1. *Genetic information in the medical record*. With the first set of human genomes now available and prices for gene sequencing rapidly decreasing, it is now cost-effective to consider sequencing every patient genome or at least genotyping key sections of the genomes and integrating that with the medical record.
2. *New diagnostic and prognostic information sources*. One of the main contributions of the genome-sequencing projects (and of the associated biological innovations) is that we are likely to have unprecedented access to new diagnostic and prognostic tools. Diagnostically, the genetic markers from a patient with an autoimmune disease, or of an infectious pathogen within a patient, will be highly specific and sensitive indicators of the subtype of disease and of that subtype's probable responsiveness to different therapeutic agents. Several **genotype**-based databases have been developed to identify markers that are associated with specific **phenotypes** and identify how genotype affects a patient's response to therapeutics. ClinVar[2] and The Human Gene Mutation Database (HGMD)[3] both annotate mutations with disease phenotype. This resource has become invaluable for genetic counselors, basic researchers, and clinicians. Additionally, the Pharmacogenomics Knowledge Base (PharmGKB) collects genetic information that is known to affect a patient's response to a drug (more on PharmGKB is described in Translational Bioinformatics, ▶ Chap. 26).[4]
3. *Ethical considerations*. One of the critical questions facing the genome-sequencing and other related projects is "Can genetic or other molecular information be misused?" The answer is certainly yes. With knowledge of a complete genome for an individual, it may be possible in the future to predict the types of disease for which that individual is at risk years before the disease actually develops. If this information fell into the hands of unscrupulous employers or insurance companies, the individual might be denied employment or coverage due to the likelihood of future disease, however distant. There is even debate about whether such information should be released to a patient even if it could be kept confidential. Should a patient be informed that he or she is likely to get a disease for which there is no treatment? What about that patient's relatives, who share genetic information with the patient? This is a matter of intense debate, and such questions have significant implications for what information is collected and for how and to whom that information

2 ▶ https://www.ncbi.nlm.nih.gov/clinvar/ (accessed November 1, 2018).
3 ▶ http://www.hgmd.org/ (accessed November 1, 2018).
4 ▶ http://www.pharmgkb.org/ (accessed November 1, 2018).

is disclosed (Durfy 1993). Passage of the Genetic Information Nondiscrimination Act in 2008 set initial federal guidelines on use of genetic information.[5] Additionally, the Personal Genome Project (PGP) has been working to define **open consent models** for releasing genetic information.[6] The Clinical Sequencing and Exploratory Research Consortium (CSER) has been tackling the difficult issues in translation of genomic data to the clinic broadly.[7]

9.2 The Rise of Bioinformatics

A brief review of the biological basis of medicine will bring into focus the magnitude of the revolution in molecular biology and the tasks that are created for the discipline of bioinformatics. The genetic material that we inherit from our parents, that we use for the structures and processes of life, and that we pass to our children is contained in a sequence of chemicals known as **deoxyribonucleic acid (DNA)**.[8] The total collection of DNA for a single person or organism is referred to as the **genome**. DNA is a long polymer chemical made of four basic subunits. The sequence in which these subunits occur in the polymer distinguishes one DNA molecule from another and directs a cell's production of proteins and all other basic cellular processes. **Genes** are discreet units encoded in DNA and they are transcribed into **ribonucleic acid (RNA)**, which has a composition very similar to DNA. Genes are transcribed into messenger RNA (mRNA) and a majority of mRNA sequences are translated by complex macromolecular machines, called ribosomes, into protein. Not all RNAs are messengers for the translation of proteins. Ribosomal RNA, for example, is used in the construction of the ribosome, the huge molecular engine that translates mRNA sequences into protein sequences. Additionally, mRNAs can be modified through alternative splicing, degradation, and formation of secondary structures that influence transcriptions. Once expressed, proteins are frequently modified (e.g. phosphorylated), and these modifications can change the function of the protein. This process of DNA being transcribed to RNA and RNA being translated to protein is commonly referred to as the **Central Dogma of Biology**.

Understanding the basic building blocks of life requires understanding the function of genomic sequences, genes, and proteins. When are genes expressed? Once genes are transcribed and translated into proteins, into what cellular compartment are the proteins directed? How do the proteins function once there? Do the proteins need to be modified in order for them to become active? How are the proteins turned off? Experimentation and bioinformatics have divided the research into several areas, and the largest are: (1) DNA and protein sequence analysis, (2) macromolecular structure–function analysis, (3) gene expression analysis, (4) proteomics, (5) metabolomics, (6) metagenomics, and (5) systems biology.

9.2.1 Roots of Modern Bioinformatics

Practitioners of bioinformatics have come from many backgrounds, including medicine, molecular biology, chemistry, physics, statistics, mathematics, engineering, and computer science. It is difficult to define precisely the ways in which this discipline emerged. There are, however, two main developments that have created opportunities for the use of information technologies in biology. The first is the progress in our understanding of how biological molecules are constructed and how they perform their functions. This dates back as far as the 1930s with the invention of electrophoresis, and then in the 1950s with the elucidation of the structure of DNA and the subsequent sequence of discoveries in the relationships

5 ► http://www.genome.gov/10002328 (accessed November 1, 2018).

6 ► http://www.personalgenomes.org/ (accessed November 1, 2018).

7 ► https://cser-consortium.org/ (accessed November 1, 2018).

8 If you are not familiar with the basic terminology of molecular biology and genetics, reference to an introductory textbook in the area would be helpful before you read the rest of this chapter.

among DNA, RNA, and protein structure. The second development has been the parallel increase in the availability of computing power. Starting with mainframe computer applications in the 1950s and moving to modern workstations, and 'the Cloud', there have been hosts of biological problems addressed with computational methods.

9.2.2 The Genomics Explosion

The benefit of the human genome sequence to medicine is both in the short and in the long term. The short-term benefits lie principally in diagnosis; the availability of sequences of normal and variant human genes will allow for the rapid identification of these genes in any patient (e.g., Babior and Matzner 1997). The long-term benefits will include a greater understanding of the proteins produced from the genome: how the proteins interact with drugs; how they malfunction in disease states; and how they participate in the control of development, aging, and responses to disease.

The effects of genomics on biology and medicine cannot be overstated. We now have the ability to measure the activity and function of genes within living cells. Genomics data and experiments have changed the way biologists think about questions fundamental to life. Whereas in the past, reductionist experiments probed the detailed workings of specific genes, we can now assemble those data together to build an accurate understanding of how cells work.

9.3 Biology Is Now Data-Driven

Nearly 30 years ago, the use of computers was proving to be useful to the laboratory researcher. Today, computers are an essential component of modern research. This has led to a change in thinking about the role of computers in biology. Before, they were optional tools that could help provide insight to experienced and dedicated enthusiasts. Today, they are required by most investigators, and experimental approaches rely on them as critical elements. This is because advances in research methods such as **genetic sequencing**, **experimental robotics and microfluidics**, **X-ray crystallography**, **nuclear magnetic resonance spectroscopy**, **cryoelectron microscopy**, **proteomic mass spectrometry** and other high throughput experiments have resulted in experiments that generate massive amounts of data. These data pose new problems for basic researchers on how the data are properly stored, analyzed, and disseminated.

The volume of data being produced by genomics projects is staggering. There are now more than 211 million sequences in **GenBank** comprising more than 285 billion digits. Since 2008, sequencing has bested Moore's law (see ► Chap. 1).[9] But these data do not stop with sequence data: PubMed contains over 28 million literature citations, the **Protein Data Bank** (**PDB**) contains three-dimensional structural data for over 45,538 distinct protein structures, and the **Gene Expression Omnibus** (**GEO**) contains over 2.8 million arrayed samples. These data are of incredible importance to biology, and in the following sections we introduce and summarize the importance of sequences, structures, gene expression experiments, systems biology, and their computational components to medicine.

9.3.1 Sequences in Biology

Sequence information (including DNA sequences, RNA sequences, and protein sequences) is critical in biology: DNA, RNA, and protein can be represented as a set of sequences of basic building blocks (bases for DNA and RNA, amino acids for proteins). Computer systems within bioinformatics thus must be able to handle biological sequence information effectively and efficiently. To that end, the bioinformatics community has developed central databases to store sequence information, data models to represent that information and software analysis tools to process sequence data.

9 ► http://www.genome.gov/sequencingcosts/ (accessed November 1, 2018).

9.3.2 Structures in Biology

The sequence information mentioned in ▶ Sect. 9.3.1 is rapidly becoming inexpensive to obtain and easy to store. On the other hand, the **three-dimensional structure information** about the proteins, DNA, and RNA is much more difficult and expensive to obtain, and presents a separate set of analysis challenges. Currently, only about 45,000 distinct three-dimensional structures of biological macromolecules are known.[10] These models are incredibly valuable resources, however, because an understanding of structure often yields detailed insights about biological function. As an example, the structure of the ribosome has been determined for several species and contains more atoms than any other structure to date. This structure, because of its size, took two decades to solve, and presents a formidable challenge for functional annotation (Cech 2000). Yet, the functional information for a single structure is dwarfed by the potential for comparative genomics analysis between the structures from several organisms and from varied forms of the functional complex. Since the ribosome is ubiquitously required for all forms of life these types of comparisons are possible. Thus, a wealth of information comes from relatively few structures. To address the problem of limited structure information, the publicly funded structural genomics initiative aims to identify all of the common structural scaffolds found in nature and to increase the number of known structures considerably. In the end, it is the physical interactions between molecules that determine what happens within a cell; thus the more complete the picture, the better the functional understanding. In particular, understanding the physical properties of therapeutic agents is the key to understanding how agents interact with their targets within the cell (or within an invading organism). These are the key questions for structural biology within bioinformatics:

1. How can we analyze the structures of molecules to learn their associated function? Approaches range from detailed molecular simulations (Levitt 1983) to statistical analyses of the structural features that may be important for function (Wei and Altman 1998).
2. How can we extend the limited structural data by using information in the sequence databases about closely related proteins from different organisms (or within the same organism, but performing a slightly different function)? There are significant unanswered questions about how to extract maximal value from a relatively small set of examples.
3. How should structures be grouped for the purposes of classification? The choices range from purely functional criteria ("these proteins all digest proteins") to purely structural criteria ("these proteins all have a toroidal shape"), with mixed criteria in between. One interesting resource available today is the Structural Classification of Proteins (SCOP),[11] which classifies proteins based on shape and function.

9.3.3 Genome Sequencing Data in Biology

Advances in sequencing technology are pivotal in enabling the practice of genomic medicine. Whereas the first human genome sequence was carried out over approximately 13 years at a cost of $2.7 billion (Davies 2010), whole human genomes can now be sequenced in a matter of days at a cost that is growing ever-closer to the magic, if somewhat arbitrary, $1000 price tag. This amount is commonly seen as the price at which it becomes feasible to sequence a patient in the course of clinical care, justifiable both clinically and financially. In 2004, and again in 2011, the National Human Genome Research Institute (part of the National Institutes of Health) funded a number of efforts specifically aimed

10 For more information see ▶ http://www.rcsb.org/ (accessed November 1, 2018).

11 ▶ http://scop2.mrc-lmb.cam.ac.uk/ (accessed December 1, 2018).

at increasing speed and decreasing the cost of genome scale sequencing.

Traditional sequencing involves a method referred to as Sanger sequencing. This method typically is applied to sequences ranging from 300 to 1000 nucleotides in a non-high throughput manner.[12] In the early to mid 2000s, several technologies were introduced to sequence large amounts of DNA in parallel. These **high throughput sequencing methods** (of which there are many including sequencing by synthesis, single molecule sequencing, combinatorial probe anchor synthesis, and others) typically involve shorter sequences than Sanger based approaches, but can generate gigabases of sequence in short fragments at low cost (<$0.05 per megabase sequenced). These methods are being used for many applications, including identification of genetic variants in clinical studies, characterizing genome function with specific experiments and sequencing novel species genomes. These studies have already discovered the genetic basis of rare genetic disorders by sequencing entire families (Ng et al. 2010), and we have seen a glimpse of the future of genome sequencing for routine health care in the analysis of a single genome of a healthy man (Ashley et al. 2010). As will be described in detail in the Translational Bioinformatics chapter (► Chap. 26), these sequencing approaches have been put to practice clinically. One emergent area of research is **metagenomics**, the study of microorganism ecosystems using DNA sequencing, including the association of human gut flora populations to disease phenotypes in humans (Qin et al. 2010).

9.3.4 Expression Data in Biology

The development of DNA microarrays led to a wealth of data and unprecedented insight into the fundamental biological machine. The traditional premise is relatively simple; tens of thousands of gene sequences derived from genomic data are fixed onto a glass slide or filter. The sequences for each spot are derived from a single gene sequence and the sequences are attached at only one end, creating a forest of sequences in each spot that are all identical. An experiment is performed where two samples (e.g. groups of cells that are grown in different conditions or for comparisons of normal and cancer tissue), one group is a control group and the other is the experimental group. The control group is grown normally, while the experimental group is grown under experimental conditions. For example, a researcher may be trying to understand how a cell compensates for a lack of sugar. The experimental cells will be grown with limited amounts of sugar. As the sugar depletes, some of the cells are removed at specific intervals of time. When the cells are removed, all of the mRNA from the cells is separated from the cells and converted back to DNA, using reverse transcriptase (a special enzyme that can create a DNA copy from an RNA template). This leaves a pool of cDNA molecules (DNA derived from mRNA is called complementary DNA or cDNA) that represent the genes that were expressed (turned on) in that group of cells. In the development of genomics experimentation, these cDNA molecules would be tagged with florescence and hybridized to slides containing single stranded DNA "probes" that are arrayed in a grid. These microarray "chips" can then be analyzed for color differences between grid points that correspond to specific gene regions. Today, with the advent of high throughput sequencing the RNA/cDNA can be sequenced directly to measure expression levels and using DNA barcoding technology and microfluidics, individual cells can be sequenced alone instead of in pooled samples where all cells' contributions to mRNA is in the same analysis. High throughput single cell sequencing is an exciting advancement which adds orders of complexity to the required computational analysis (Shapiro et al. 2013).

Computers become critical for analyzing these data because it is impossible for a researcher to measure and analyze all of the datasets by hand. Currently scientists are using gene expression experiments to study how cells from different organisms compen-

12 ► http://en.wikipedia.org/wiki/DNA_sequencing (accessed November 1, 2018).

sate for environmental changes, how pathogens fight antibiotics, and how cells grow uncontrollably (as is found in cancer). A challenge for biological computing is to develop methods to analyze these data, tools to store these data, and computer systems to collect the data automatically.

9.3.5 Metabolomics Data in Biology

Genomics and proteomics study the function of the genome and the proteome, while **metabolomics** studies the diversity and function of small molecules in a biosample. These include metabolites such as lipids, carbohydrates, metal ions, hormones, signaling molecules, etc. Interest in the metabolome has increased significantly with the development of separation and mass spectrometry technologies that can identify small molecule molecular mass and identities in a high throughput fashion. Bioinformatics is a key component of both the identification of specific molecules by matching mass spectrometry "fingerprints" with a database of known molecules as well as in the analysis the resulting data. For example, researchers have characterized the metabolome of human colorectal cancers and stool and identified disease enriched metabolites as a possible detectable markers of disease or treatment outcomes (Brown et al. 2016).

9.3.6 Epigenetics Data in Biology

Epigenetics consists of heritable changes that are not encoded in the primary DNA sequence. Several types of epigenetic effects can now be studied in the laboratory, and they have been associated to disease and risks of disease (Goldberg et al. 2007). First, the regional structure of chromosomes affects which regions of the genome can be transcribed, i.e. which regions can be *expressed*. Large proteins, called histones, coordinate the structure of chromosomes and their structure and positions are regulated with protein posttranslational modifications to the histones bound to the DNA. These changes have been associated with spontaneous mutations in cancer, complex genetic diseases, and Mendelian inherited genetic diseases. Second, cytosine bases in the DNA can be methylated and this can affect gene expression. DNA methylation patterns can be passed on when DNA is replicated. Like chromosome structure, these modifications have been associated with human disease (Bird 2002).

9.3.7 Systems Biology

Recent advances in high throughput technologies have enabled a new, dynamic approach to studying biology, that of **systems biology**. In contrast to the historically reductionist approach to biology, studying one molecule at a time, systems biology looks at the entirety of a system including dynamic relationships between the different components. With that said, systems biology is still maturing. As an analogy, consider an airplane. Having a "parts list" for a Boeing 747 does not enable us to understand how those parts work together to make the airplane operate. If the airplane breaks, the parts list alone does not tell us how to remedy the situation. Rather, we need to understand how the parts interact, how one affects another, and how perturbations to one part of the system affect the rest of the system. Similarly, systems biology involves understanding not only the "parts list", i.e. the list of all genes, proteins, metabolites, etc., but also the dynamic networks of interactions among these parts. An integrated simulation of an entire bacterial cell has shown the feasibility of accurate computational simulations of cell physiology (Karr et al. 2012).

Current research in **-omics technologies** have both enabled and catalyzed the advancement of systems biology. However, a systems biology approach goes beyond simply performing these high bandwidth methods for the purpose of biological discovery. Rather, systems biology implies a systematic, hypothesis-driven approach based on omic-scale (very large) hypotheses. Once the interactions in a biological network are understood, one can model that network to make predictions

regarding the system's behavior, particularly in light of specific perturbations. Understanding how the system has evolved to work can also help us understand what goes wrong when the system breaks down, and how to intervene in order to restore the system to normal.

9.4 Key Bioinformatics Algorithms

There are a number of common computations that are performed in many contexts within bioinformatics. In general, these computations can be classified as sequence alignment, structure alignment, pattern analysis of sequence/structure, gene expression analysis, and pattern analysis of biochemical function.

9.4.1 Early Work in Sequence and Structure Analysis

As it became clear that the information from DNA and protein sequences would be voluminous and difficult to analyze manually, algorithms began to appear for automating the analysis of sequence information. The first requirement was to have a reliable way to align sequences so that their detailed similarities and distances could be examined directly. Needleman and Wunsch (1970) published an elegant method for using **dynamic programming** techniques to align sequences in time related to the cube of the number of elements in the sequences. Smith and Waterman (1981) published refinements of these algorithms that allowed for searching both the best global alignment of two sequences (aligning all the elements of the two sequences) and the best local alignment (searching for areas in which there are segments of high similarity surrounded by regions of low similarity). A key input for these algorithms is a matrix that encodes the similarity or substitutability of sequence elements: When there is an inexact match between two elements in an alignment of sequences, it specifies how much "partial credit" we should give to the overall alignment based on the similarity of the elements, even though they may not be identical. Looking at a set of evolutionarily related proteins, Dayhoff (1974) published one of the first matrices derived from a detailed analysis of which amino acids (elements) tend to substitute for others.

Within structural biology, the vast computational requirements of the experimental methods (such as X-ray crystallography and nuclear magnetic resonance) for determining the structure of biological molecules drove the development of powerful structural analysis tools. In addition to software for analyzing experimental data, graphical display algorithms allowed biologists to visualize these molecules in great detail and facilitated the manual analysis of structural principles (Langridge 1974; Richardson 1981). At the same time, methods were developed for simulating the forces within these molecules as they rotate and vibrate (Gibson and Scheraga 1967; Karplus and Weaver 1976; Levitt 1983).

The most important development to support the emergence of bioinformatics, however, has been the creation of databases with biological information. In the 1970s, structural biologists, using the techniques of X-ray crystallography, set up the Protein Data Bank (PDB) specifying the Cartesian coordinates of the structures that they elucidated (as well as associated experimental details) and made PDB publicly available. The first release, in 1977, contained 77 structures. The growth of the database is chronicled on the Web: the PDB now has over 75,000 detailed atomic structures and is the primary source of information about the relationship between protein sequence and protein structure.[13] Similarly, as the ability to obtain the sequence of DNA molecules became widespread, the need for a database of these sequences arose. In the mid-1980s, the GENBANK database was formed as a repository of sequence information. Starting with 606 sequences and 680,000 bases in 1982, the GENBANK has grown by much more than 135 million sequences and 125 billion bases.[14] The GENBANK database of DNA sequence information supports the experimental reconstruction of genomes and acts as a focal point

13 See ► http://www.rcsb.org/ (accessed December 1, 2018).

14 ► http://www.ncbi.nlm.nih.gov/genbank/ (accessed December 1, 2018).

for experimental groups. Numerous other databases store the sequences of protein molecules[15] and information about human genetic diseases.[16]

Included among the databases that have accelerated the development of bioinformatics is the Medline database of the biomedical literature and its paper-based companion Index Medicus (see ▶ Chap. 23).[17] Including articles as far back as 1809 and brought online free on the Web in 1997, Medline provides the glue that relates many high-level biomedical concepts to the low-level molecule, disease, and experimental methods. In fact, this "glue" role was the basis for creating the NCBI suite of databases and software and PubMed systems (see ▶ Sect. 9.5) for integrating access to literature references and the associated databases.

9.4.2 Sequence Alignment and Genome Analysis

Perhaps the most basic activity in computational biology is comparing two biological sequences to determine (1) whether they are similar and (2) how to align them. The problem of alignment is not trivial but is based on a simple idea. Sequences that perform a similar function should, in general, be descendants of a common ancestral sequence, with mutations over time. These mutations can be replacements of one amino acid with another, deletions of amino acids, or insertions of amino acids. The goal of **sequence alignment** is to align two sequences so that the evolutionary relationship between the sequences becomes clear. If two sequences are descended from the same ancestor and have not mutated too much, then it is often possible to find corresponding locations in each sequence that play the same role in the evolved proteins. The problem of solving correct biological alignments is difficult because it requires knowledge about the evolution of the molecules that we typically do not have. There are now, however, well-established algorithms for finding the mathematically optimal alignment of two sequences. These algorithms require the two sequences and a scoring system based on (1) exact matches between amino acids that have not mutated in the two sequences and can be aligned perfectly; (2) partial matches between amino acids that have mutated in ways that have preserved their overall biophysical properties; and (3) gaps in the alignment signifying places where one sequence or the other has undergone a deletion or insertion of amino acids. The algorithms for determining optimal sequence alignments are based on a technique in computer science known as **dynamic programming** and are at the heart of many computational biology applications (Gusfield 1997). ◘ Figure 9.1 shows an example of a Smith-Waterman matrix, the first described local alignment algorithm that utilizes a dynamic programming approach. The algorithm works by calculating a similarity matrix between two sequences, then finding optimal paths through the matrix that maximize a similarity score between the two sequences.

Unfortunately, the dynamic programming algorithms are too computationally expensive to apply to large numbers of sequences, so a number of faster, more heuristic methods have been developed. The most popular algorithm is the **Basic Local Alignment Search Tool (BLAST)** (Altschul et al. 1990). BLAST is based on the observation that sections of proteins are often conserved without gaps (so the gaps can be ignored—a critical simplification for speed) and that there are statistical analyses of the occurrence of small subsequences within larger sequences that can be used to prune the search for matching sequences in a large database. These tools work well for both protein and nucleic acid sequences. Other tools have been developed that are better suited for nucleic acid sequence assembly and mapping of short read high throughput sequencing data including BLAT (Kent 2003), SOAP (Li et al. 2008), and others.

Protein 3D structures can be aligned, visualized and compared in a similar way to linear protein sequences (◘ Fig. 9.2). Tools such

15 ▶ http://www.uniprot.org/ (accessed December 1, 2018).
16 ▶ http://www.ncbi.nlm.nih.gov/omim (accessed December 1, 2018).
17 ▶ http://www.ncbi.nlm.nih.gov/pubmed (accessed December 1, 2018).

a) Pairwise alignment between human chymotrypsin and human trypsin.

```
CTRB_HUMAN    MAFLWLLSCWALLGTTFGCGVPAIHPVLSGLSRIVNGEDAVPGSWPWQVSLQDKTGFHFC
TRY1_HUMAN    MNPLLILTFVA- - - - - - - - - - - - AALAAPFDDDDKIVGGYNCEENSVPYQVSLN- - SGFHFC

CTRB_HUMAN    GGSLISEDWVVTAAHCGVRTSDDVVVAGEFDQGSDEENIQVLKIAKVFKNPKFSILTVNND
TRY1_HUMAN    GGSLINEQWVVSAGHC- YKSRIQVRLGEHNIEVLEGNEQFINAAKIIRHPQYDRKTLNND

CTRB_HUMAN    ITLLKLATPARFSQTVSAVCLPSADDDFPAGTLCATTGWGKTKYNANKTPDKLQQAALPL
TRY1_HUMAN    IMLIKLSSRAVINARVSTISLPTAPP - - ATGTKCLISGWGNTASSGADYPDYPDELQCLDAPV

CTRB_HUMAN    LSNAECKKSWGRRITDVMICAG - - ASGVSSCMGDSGGPLVCQKDGAWTLVGIVSWGSDTC
TRY1_HUMAN    LSQAKCEASYPGKITSNMFCVGFLEGGKDSCQGDSGGPVVCNG - - - - QLQGVVSWGDGCA

CTRB_HUMAN    STSSPGVYARVTKLIPWVQKILLAN -
TRY1_HUMAN    QKNKPGVYTKVYNYVKWIKNTIAANS
```

b) Smith Waterman matrix illustrating the aligned region in A, using the BLOSUM62 mutation matrix (Henikff and Henikoff, 1994).

	G	F	L	E	G	G	K	D	S	C	Q	G	D	S	G	G	P	V	V	C	N	G	Q	L	Q
G	6	-3	-4	-2	6	6	-2	-1	0	-3	-2	6	1	0	6	6	-2	-3	-3	-3	0	6	-2	-4	-2
A	0	-2	-1	-1	0	0	-1	-2	1	0	-1	0	-2	1	0	0	-1	0	0	0	-2	0	-1	-1	-1
S	0	-2	-2	0	0	0	0	0	4	-1	0	0	0	4	0	0	-1	-2	-2	-1	1	0	0	-2	0
G	6	-3	-4	-2	6	6	-2	-1	0	-3	-2	6	1	0	6	6	-2	-3	-3	-3	0	6	-2	-4	-2
V	-3	-1	1	-2	-3	-3	-2	-3	-2	-1	-2	-3	-3	-2	-3	-3	-2	4	4	-1	-3	-3	-2	1	-2
S	0	-2	-2	0	0	0	0	0	4	-1	0	0	0	4	0	0	-1	-2	-2	-1	1	0	0	-2	0
S	0	-2	-2	0	0	0	0	0	4	-1	0	0	0	4	0	0	-1	-2	-2	-1	1	0	0	-2	0
C	-3	-2	-1	-4	-3	-3	-3	-3	-1	9	-3	-3	-3	-1	-3	-3	-3	-1	-1	9	-3	-3	-3	-1	-3
M	-3	0	2	-2	-3	-3	-1	-3	-1	-1	0	-3	-3	-1	-3	-3	-2	1	1	-1	-2	-3	0	2	0
G	6	-3	-4	-2	6	6	-2	-1	0	-3	-2	6	1	0	6	6	-2	-3	-3	-3	0	6	-2	-4	-2
D	-1	-3	-4	2	-1	-1	-1	-6	0	-3	0	-1	6	0	-1	-1	-1	-3	-3	-3	1	-1	0	-4	0
S	0	-2	-2	0	0	0	0	0	4	-1	0	0	0	4	0	0	-1	-2	-2	-1	1	0	0	-2	0
G	6	-3	-4	-2	6	6	-2	-1	0	-3	-2	6	1	0	6	6	-2	-3	-3	-3	0	6	-2	-4	-2
G	6	-3	-4	-2	6	6	-2	-1	0	-3	-2	6	1	0	6	6	-2	-3	-3	-3	0	6	-2	-4	-2
P	-2	-4	-3	-1	-2	-2	-1	-1	-1	-3	-1	-2	-1	-1	-2	-2	7	-2	-2	-3	-2	-2	-1	-3	-1
L	-4	0	4	-3	-4	-4	-2	-4	-2	-1	-2	-4	-4	-2	-4	-4	-3	1	1	-1	-3	-4	-2	4	-2
V	-3	-1	1	-2	-3	-3	-2	-3	-2	-1	-2	-3	-3	-2	-3	-3	-2	4	-4	-1	-3	-3	-2	1	-2
C	-3	-2	-1	-4	-3	-3	-3	-3	-1	9	-3	-3	-3	-1	-3	-3	-3	-1	-1	9	-3	-3	-3	-1	-3
Q	-2	-3	-2	-2	-2	-2	1	0	0	-3	5	-2	0	0	-2	-2	-1	-2	-2	-3	0	-2	5	-2	5
K	-2	-3	-2	1	-2	-2	5	-1	0	-3	1	-2	-1	0	-2	-2	-1	-2	-2	-3	0	-2	-1	-1	1
D	-1	-3	-4	2	-1	-1	-1	6	0	-3	0	-1	6	0	-1	-1	-1	-3	-3	-3	1	-1	1	-4	1
G	6	-3	-4	-2	6	6	-2	-1	0	-3	-2	6	1	0	6	6	-2	-3	-3	-3	0	6	-2	-4	-2
A	0	-2	-1	-1	0	0	-1	-2	1	0	-1	0	-2	1	0	0	-1	0	0	0	-2	0	-1	-1	-1
W	-2	1	-2	-3	-2	-2	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-4	-3	-3	-2	-4	-2	-2	-2	-2
T	-2	-2	-1	-1	-2	-2	-1	-1	1	-1	-1	-2	-1	1	-2	-2	-1	0	0	-1	0	-2	-1	-1	-1
L	-4	0	4	-3	-4	-4	-2	-4	-2	-1	-2	-4	-4	-2	-4	-4	-3	1	1	-1	-3	-4	-2	-4	-2
V	-3	-1	1	-2	-3	-3	-2	-3	-2	-1	-2	-3	-3	-2	-3	-3	-2	4	4	-1	-3	-3	-2	1	-2

Fig. 9.1 Example of sequence alignment using the Smith Waterman algorithm

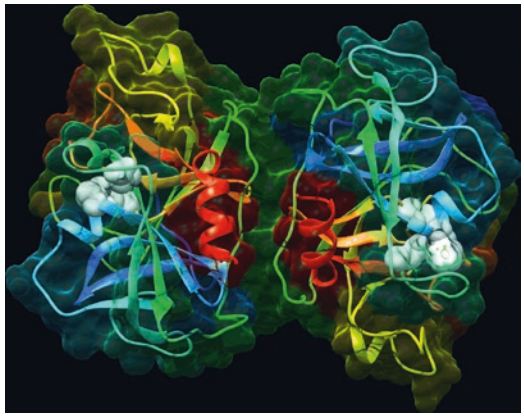


Fig. 9.2 Example of structural visualization and comparison. Comparison of the serine protease protein structures and catalytic amino acids using Chimera (► http://www.cgl.ucsf.edu/chimera; accessed December 15, 2018)

as PyMol[18] and UCSF Chimera[19] provide sophisticated and extensible applications for relatively easy visualization of 3D structures. Tools for 3D alignment of the structures are provided with these applications.

9.4.3 Prediction of Structure and Function from Sequence

One of the primary challenges in bioinformatics is taking a newly determined DNA sequence (as well as its translation into a protein sequence) and predicting the structure of the associated molecules, as well as their function. Both problems are difficult, being fraught with all the dangers associated with making predictions without hard experimental data. Nonetheless, the available sequence data are starting to be sufficient to allow good predictions in a few cases. For example, there is a Web site devoted to the assessment of biological macromolecular structure prediction methods.[20] Results suggest that when two protein molecules have a high degree (more than 40%) of sequence identity and one of the structures is known, a reliable model of the other can be built by analogy. In the case that sequence similarity is less than 25%, however, performance of these methods is much less reliable.

With the advent of deep learning, there has been an acceleration of progress in many machine learning tasks, including structure prediction. Recently, the use of convolutional neural networks by DeepMind Inc. called AlphaFold (Senior, et al. 2020) has lead to a quantum leap in the quality of predicted structures—so much so that some experts in protein structure prediction have said that parts of this challenge can now be considered "solved[21]." They make this claim because on multiple prediction tasks, the accuracy of the predicted structure is similar to those determined experimentally. Of course, it is likely that there are classes of proteins that may not perform as well, but for a large fraction of protein sequences, the structure seems to be predictable by these methods. An important caveat is that these methods must be carefully reviewed by the community, reproduced and made generally available before they will have their full impact

When scientists investigate biological structure, they commonly perform a task analogous to sequence alignment, called **structural alignment**. Given two sets of three-dimensional coordinates for a set of atoms, what is the best way to superimpose them so that the similarities and differences between the two structures are clear? Such computations are useful for determining whether two structures share a common ancestry and for understanding how the structures' functions have subsequently been refined during evolution. There are numerous published algorithms for finding good structural alignments. We can apply these algorithms in an automated fashion whenever a new structure is determined, thereby classifying the new structure into one of the protein families.

There are also algorithms for using the structure of a large biomolecule and the structure of a small organic molecule (such as a

18 ► https://pymol.org/ (accessed December 1, 2018).
19 ► http://www.cgl.ucsf.edu/chimera/ (accessed December 1, 2018).
20 ► http://predictioncenter.org/ (accessed December 1, 2018).
21 ► https://www.nature.com/articles/d41586-020-03348-4.

drug or cofactor) to try to predict the ways in which the molecules will interact. An understanding of the structural interaction between a drug and its target molecule often provides critical insight into the drug's mechanism of action. The most reliable way to assess this interaction is to use experimental methods to solve the structure of a drug–target complex. Once again, these experimental approaches are expensive, so computational methods play an important role. Typically, we can assess the physical and chemical features of the drug molecule and can use them to find complementary regions of the target. For example, a highly electronegative drug molecule will be most likely to bind in a pocket of the target that has electropositive features.

Prediction of function often relies on use of sequential or structural similarity metrics and subsequent assignment of function based on similarities to molecules of known function. These methods can guess at general function for roughly 60–80% of all genes, but leave considerable uncertainty about the precise functional details even for those genes for which there are predictions, and have little to say about the remaining genes.

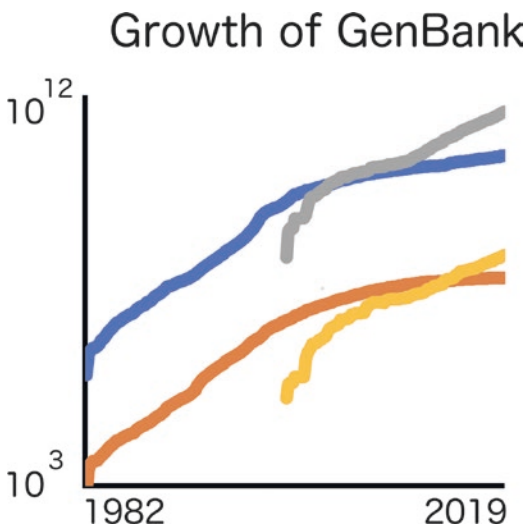
9.4.4 Clustering of Gene Expression Data

Analysis of gene expression data often begins by clustering the expression data. A typical experiment is represented as a large table, where the rows are the genes on each chip and the columns represent the different experiments, whether they be time points or different experimental conditions. Each row is then a vector of values that represent the results of the experiment with respect to a specific gene. Clustering can then be performed to determine which genes are being expressed similarly. Genes that are associated with similar expression profiles are often functionally associated. For example, when a cell is subjected to starvation (fasting), ribosomal genes are often downregulated in anticipation of lower protein production by the cell. It has similarly been shown that genes associated with neoplastic progression could be identified relatively easily with this method, making gene expression experiments a powerful assay in cancer research (see Yan and Gu 2009, for a review). In order to cluster expression data, a distance metric must be determined to compare a gene's profile with another gene's profile. If the vector data are a list of values, Euclidian distance or correlation distances can be used. If the data are more complicated, more sophisticated distance metrics may be employed. These methods fall into two categories: supervised and unsupervised. Supervised learning methods require some preconceived knowledge of the data at hand (discussed below). Usually, the method begins by selecting profiles that represent the different groups of data, e.g., genes that represent certain pathways, and then the clustering method associates each of the genes with the representative profile to which they are most similar. Unsupervised methods are more commonly applied because these methods require no knowledge of the data, and can be performed automatically.

Two such unsupervised learning methods are the hierarchical and K-means clustering methods. Hierarchical methods build a dendrogram, or a tree, of the genes based on their expression profiles. These methods are agglomerative and work by iteratively joining close neighbors into a cluster. The first step often involves connecting the closest profiles, building an average profile of the joined profiles, and repeating until the entire tree is built. K-means clustering builds k clusters or groups automatically. The algorithm begins by picking k representative profiles randomly. Then each gene is associated with the representative to which it is closest, as defined by the distance metric being employed. Then the center of mass of each cluster is determined using all of the member gene's profiles. Depending on the implementation, either the center of mass or the nearest member to it becomes the new representative for that cluster. The algorithm then iterates until the new center of mass and the previous center of mass are within some threshold. The result is k groups of genes that are regulated similarly. One drawback of K-means is that one must chose the value for k. If k is too large, logical "true" clusters may be split into pieces and if k is too small, there

will be clusters that are merged. One way to determine whether the chosen k is correct is to estimate the average distance from any member profile to the center of mass. By varying k, it is best to choose the lowest k where this average is minimized for each cluster. Another drawback of K-means is that different initial conditions can give different results, therefore it is often prudent to test the robustness of the results by running multiple runs with different starting configurations (◘ Fig. 9.3).

The future clinical usefulness of these algorithms cannot be overstated. In 2002, van't Veer et al. (2002) found that a gene expression profile could predict the clinical outcome of breast cancer. The global analysis of gene expression showed that some cancers were associated with different prognosis, not detectable using traditional means. Another exciting advancement in this field is the potential use of microarray expression data to profile the molecular effects of known and potential therapeutic agents. This molecular understanding of a disease and its treatment will soon help clinicians make more informed and accurate treatment choices (for more, see ► Chap. 26).



◘ **Fig. 9.3** The exponential growth of GENBANK. This plot shows that since 1982 the number of bases in GENBANK has grown by five full orders of magnitude and continues to grow by a factor of 10 every 4 years

9.4.4.1 Classification and Prediction

A high level description of some common approaches to classification or supervised learning are described below, but note that entire courses could be, and are, taught on each of these methods. For further details we refer readers to the suggested texts at the end of this chapter.

One of the simplest methods for classification is that of *k*-nearest-neighbor, or KNN. Essentially, KNN uses the classification of the *k* closest instances to a given input as a set of votes regarding how that instance should be classified. Unfortunately, KNN tends not to be useful for omics-based classification because it tends to break down in high-dimensional space. For high-dimensional data, KNN has difficulty in finding enough neighbors to make prediction, which will lead to large variation in the classification. This breakdown is one aspect of the "curse of dimensionality," described in more detail below (Hastie et al. 2009).

A more general statistical approach to supervised learning, and one which encompasses a number of popular methods, is that of function approximation. In this approach, one attempts to find a useful approximation of the function $f(x)$ that underlies the actual relation between the inputs and outputs. In this case, one chooses a metric by which to judge the accuracy of the approximation, for example the residual sum of squares, and uses this metric to optimize the model to fit the training data. Bayesian modeling, logistic regression, and Support Vector Machines all use variations on this approach.

Finally, there is the class of rule-based classifiers. This type of classifier may be thought of as a series of rules, each of which splits the set of instances based on a given characteristic. Details such as what criteria are used to choose the feature on which to base a rule, and whether the algorithm uses enhancements such ensemble learning (i.e., multiple models together) determine the specifics of the classifier type, for example decision trees, random forests, or covering rules.

Which approach to use depends both on the nature of the data and the question being

asked. The question might prioritize sensitivity over specificity or vice versa. For example, for a test to detect a life-threatening infection that is easily treatable by readily available antibiotics, one might want to err on the side of sensitivity. In addition, data may be numeric or categorical or have differing degrees of noise, missing values, correlated features or non-linear interactions among features. These different qualities are better handled by different methods. In many cases the best approach is actually to try a number of different methods and to compare the results. Such comparative analysis is facilitated through freely available software packages such as R/Bioconductor[22] and Weka.[23]

9.4.5 The Curse of Dimensionality

In the post-genomic era, there is no shortage of data to analyze. Rather, many researchers have more data than they know what to do with. However this overabundance tends to be a factor of the dimensionality of the data, rather than the number of subjects. This mismatch can lead to challenges for experimental design and statistical analysis. Type 1 error, or the tendency to incorrectly reject the null hypothesis and say that indeed there is statistical significance to a pattern (see ▶ Chap. 13), is amplified by looking at high-dimensional data. This is one aspect of what is known as the "curse of dimensionality" (Hastie et al. 2009). Consider analysis of gene expression data for 20,000 genes, trying to detect a pattern that can predict outcome. In a sample of, say, 30 subjects—a reasonable number when testing a single hypothesis—by random chance, some number of genes will correlate with outcome. Essentially one is testing not one but 20,000 hypotheses simultaneously. One must therefore correct for multiple hypothesis testing. The Bonferroni method is a common and straightforward approach to correct for multiple hypothesis testing.[24] It entails dividing the threshold p-value one would use, traditionally 0.05, by the number of hypotheses. So, for a test of 20,000 genes, one would require a p-value of 2.5×10^{-6} to call a gene significant. Typically, analyses using high dimensional data such as gene expression are not sufficiently powered to pass this stringent test. One would need thousands of samples to be sufficiently powered. Another approach is to use q-value, or false discovery rate (Storey and Tibshirani 2003), rather than p-value. This approach relies on empirical permutation to determine the expected number of false positives if indeed the null hypothesis is correct, which enables approximation of the proportion of false positives among all reported positives. Consider again the microarray experiment above in which each array includes 20,000 genes. We want to know whether gene X was differentially expressed in cases versus controls. Choosing a threshold p-value, or false *positive* rate, of 0.05 means that 1 time in 20 we will erroneously reject the null hypothesis and predict a false positive. If a statistical test returns 2000 positives, i.e. 2000 genes appear to be significantly differentially expressed, we expect 1 in 20 of the genes being analyzed ($20{,}000 \times (1/20) = 1000$) or approximately half of them to be false positives. A false *discovery* rate of 0.05, on the other hand, would mean that 5% *of those called positive*, in this case 100 out of 2000, are false positives. Q-value is thus less stringent than p-value, but may be of greater utility in a high-dimensional omics context than a traditional p-value or correction for multiple hypotheses.

Another approach to analysis of high dimensional data sets is to use dimensionality reduction methods such as feature selection or feature extraction. Feature selection entails extracting only a subset of the features at hand, in this case genes. This may be done in a number of ways, based on which genes vary the most, or on which genes seem to best predict the categorization at hand. In contrast,

22 ▶ http://bioconductor.org/ (accessed December 1, 2018).

23 ▶ http://www.cs.waikato.ac.nz/ml/weka/ (accessed December 1, 2018).

24 ▶ http://en.wikipedia.org/wiki/Bonferroni_correction (accessed December 1, 2018).

feature extraction creates a new smaller set of features that captures the essence of the original variation. As an example, imagine a plane flight from Seattle, WA to Key West, FL. One could use a 3-dimensional vector consisting of latitude and longitude to describe the plane's position at any given point along the way. In this case, one value would describe how far the plane had gone in the north/south direction, and one would indicate how far the plane had gone in the east/west direction. However, if we change the axis along which we are measuring to instead be the direct route along which the plane is flying, then we only need 1 dimension to describe where the plane is located. The distance flown tells us where the plane is located at any given time. This approach of changing the axes is the basis for principle components analysis (PCA), a common method for feature extraction. Instead of going from two dimensions to one, PCA on gene expression data typically goes from tens of thousands of features to just a few. Both for feature selection and feature extraction, it is important to replicate the findings in an independently generated data set in order to be sure the model is not over fitting the data on which it was trained.

9.5 Current Application Successes from Bioinformatics

Biologists have embraced the Internet in a remarkable way and have made access to data a normal and expected mode for doing business. Hundreds of databases curated by individual biologists create a valuable resource for the developers of computational methods who can use these data to test and refine their analysis algorithms. With standard Internet search engines, most biological databases can be found and accessed within moments. The large number of databases has led to the development of meta-databases that combine information from individual databases to shield the user from the complex array that exists. There are various approaches to this task.

The National Center for Biotechnology Information (NCBI) suite of databases and software (previously known as the 'Entrez' gives integrated access to the biomedical literature, protein, and nucleic acid sequences, macromolecular and small molecular structures, and genome project links (including both the Human Genome Project and sequencing projects that are attempting to determine the genome sequences for organisms that are either human pathogens or important experimental model organisms) in a manner that takes advantages of either explicit or computed links between these data resources.[25] Newer technologies are being developed that will allow multiple heterogeneous databases to be accessed by search engines that can combine information automatically, thereby processing even more intricate queries requiring knowledge from numerous data sources. One example is the Bioconductor project, a toolbox for bioinformatics in the R programming language.[26]

9.5.1 Data Sharing

In 1996, the First International Strategy Meeting on Human Genome Sequencing was held in Bermuda. In this meeting, a set of principles was agreed upon regarding sharing of human genome sequencing data. These principles came to be known as the Bermuda principles. They stipulated that (1) all sequence assemblies larger than 1 kb should be released as soon as possible, ideally within 24 h; (2) finished annotated sequences should be published immediately to public databases; and (3) that all human sequence data generated in large-scale sequencing centers should be made available in the public domain.[27]

Increasingly, journals and funders require that researchers deposit all types of research data in publicly available repositories (Fischer and Zigmond 2010). In 2009, President Obama announced an Open Government

25 ► https://www.ncbi.nlm.nih.gov/search/ (accessed December 7th, 2020).

26 ► http://bioconductor.org/ (accessed December 1, 2018).

27 ► http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml (accessed December 1, 2018).

Directive that included plans to make federally funded research data available to the public.[28] This announcement describes the NIH's policy regarding published manuscripts in particular, but also notes that the results of vgovernment-funded research can take many forms, including data sets. Currently the NIH requires that proposals for funding of over $500,000 include a data sharing plan.[29]

To that end, a significant advancement in bioinformatics is in making research datasets more available and reusable. From the community of researchers who are enabling this effort the concept of FAIR data has emerged. FAIR datasets are Findable, Accessible, Interoperable and Reusable. FAIR data principles lay out a framework to encourage increased sharing and use of scientific datasets. Findable data includes the use of global persistent identifiers and metadata standards. Accessible data is available on the Internet and searchable through metadata usage. Interoperable data use a "formal, accessible, shared and broadly applicable language for knowledge representation". Finally, reusable data have clear attribution and license that enables reuse. The webportal FAIRsharing provides curated resources on datasets, standards and collections that are more FAIR.[30] Resources such as BioCaddie DataMed enable discovery of datasets through a Data Discovery Index.[31]

9.5.2 Data Standards, Metadata and Biomedical Ontologies

▶ Chapter 7 on standards in biomedical informatics addresses standardized terminologies as well as standards for data exchange, and terminologies for translational research are discussed in ▶ Chap. 27. The development of such schemes necessitates the creation of terminology standards, just as in clinical informatics. There are now many controlled vocabularies (or ontologies) and metadata standards for annotation of genomic or proteomic data. Metadata standards help define information which should be collected and annotated upon various types of datasets. Furthermore, a great many tools have been developed to help researchers access and analyze this data. For example, the previously mentioned Bioconductor project provides bioinformatic tools in the R language for solving common problems. Other commonly used tools include BioPerl, BioPython and MATLAB.[32]

Biomedical ontologies have become a key component in the development of metadata standards for the management and exchange of bioinformatic datasets and in making data more FAIR (see ▶ Sect. 9.5.1). The open biomedical ontologies consortium (OBO) has developed a number of reference ontologies that are in wide use in bioinformatics including Gene Ontology, Human Phenotype Ontology and the UBERON anatomy ontology (Smith et al. 2007). For example, **Gene Ontology (GO)** is an ontology used for annotation of gene function, and arguably the most widely used ontology in basic research. Ontologies enable indexing, exchange and computing with biomedical datasets and metadata.

Metadata standards for bioinformatics datasets are an intellectual challenge for researchers to enable the sharing and interoperability of data and to make data more FAIR. There are a number of tools and web portals such as the Center for Expanded Data Annotation and Retrieval (CEDAR) provide tools for creation and sharing of metadata about datasets.[33] Metadata can include information about an experiment such as the protocol, the time the experiment was performed, who performed the experiment and technology used to generate or analyze the experiment, but

28 ▶ http://edocket.access.gpo.gov/2009/E9-29322.htm (accessed December 1, 2018).
29 ▶ http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed December 1, 2018).
30 ▶ https://fairsharing.org/ (accessed December 1, 2018).
31 ▶ https://datamed.org/ (accessed April 20, 2019).
32 ▶ http://www.open-bio.org/ (accessed December 1, 2018).
33 ▶ https://metadatacenter.org/ (accessed December 1, 2018).

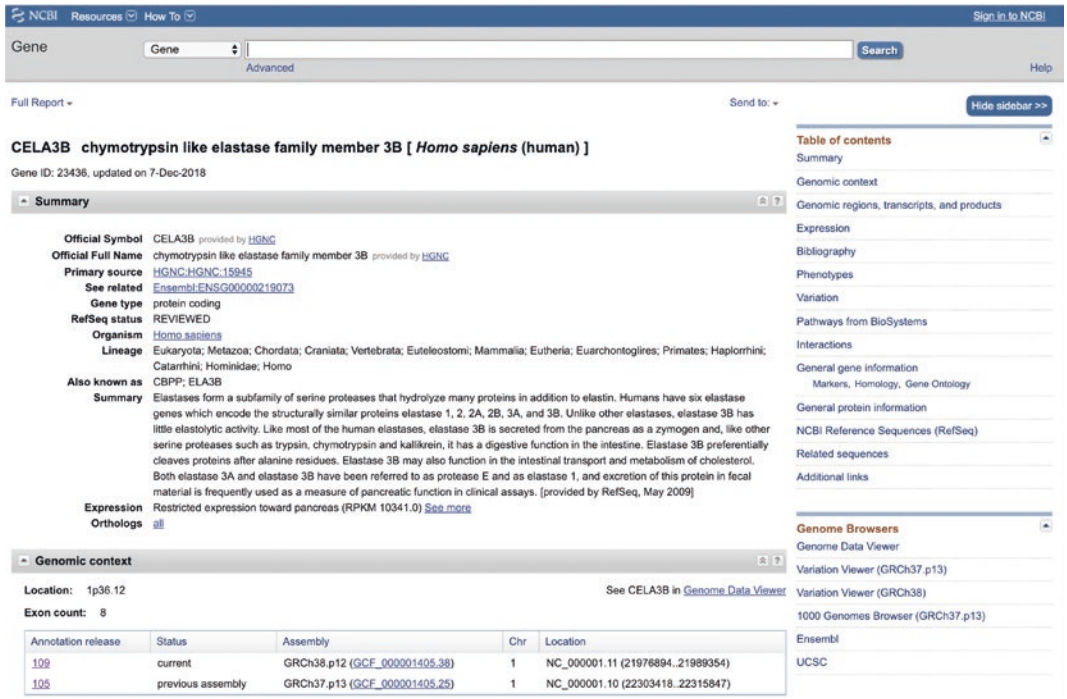


Fig. 9.4 The NCBI Gene entry for the digestive enzyme chymotrypsin. Basic information about the original report is provided, as well as some annotations of the key regions in the sequence and the complete sequence of DNA bases (a, g, t, and c) is provided as a link. (Courtesy of NCBI)

can also include information such as organism, disease model, tissue, conditions, etc.

9.5.2.1 Sequence and Genome Databases

The main types of sequence information that must be stored are DNA and protein. One of the largest DNA **sequence databases** is GENBANK, which is managed by the NCBI.[23] GENBANK is growing rapidly as genome-sequencing projects feed their data (often in an automated procedure) directly into the database. Figure 9.3 shows the logarithmic growth of data in GENBANK since 1982. NCBI Gene curates some of the many genes within GENBANK and presents the data in a way that is easy for the researcher to use (Fig. 9.4).

In addition to GENBANK, there are numerous special-purpose DNA databases for which the curators have taken special care to clean, validate, and annotate the data. The work required of such curators indicates the degree to which raw sequence data must be interpreted cautiously. GENBANK can be searched efficiently with a number of algorithms and is usually the first stop for a scientist with a new sequence who wonders "Has a sequence like this ever been observed before? If one has, what is known about it?" There are increasing numbers of stories about scientists using GENBANK to discover unanticipated relationships between DNA sequences, allowing their research programs to leap ahead while taking advantage of information collected on similar sequences.

A database that has become very useful recently is the University of California Santa Cruz Genome Browser[34] (Fig. 9.5). This data set allows users to search for specific sequences in the UCSC version of the human genome. Powered by the similarity search tool BLAT, users can quickly find annotations on the human genome that contain their sequence of interest. These annotations include known

34 ▶ http://genome.ucsc.edu/ (accessed December 1, 2018).

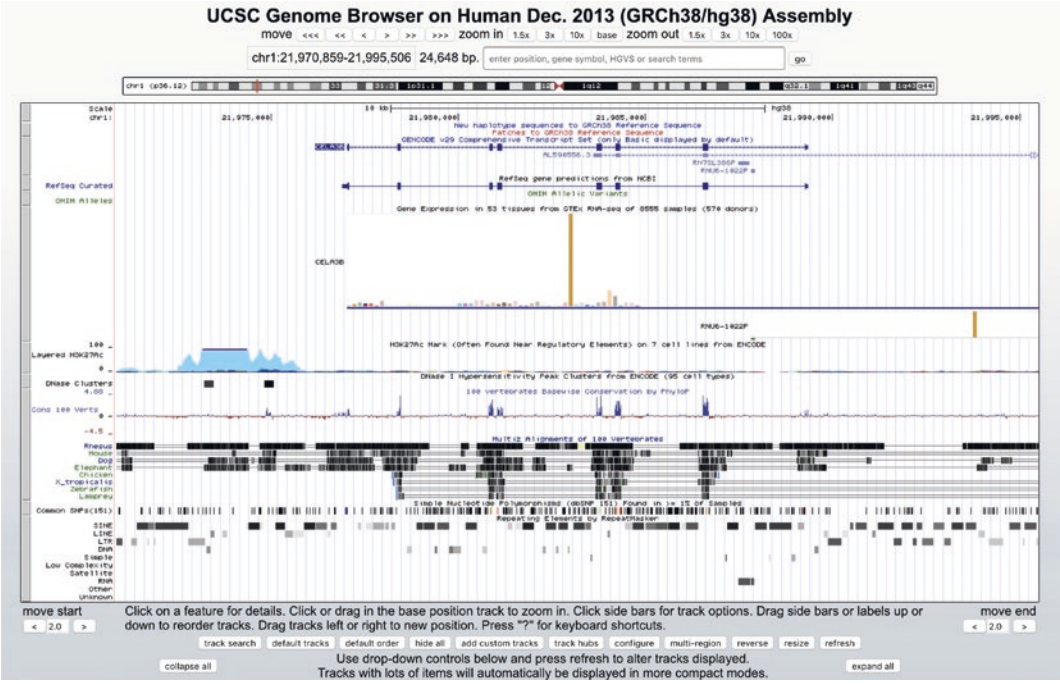


Fig. 9.5 Screen from the UC Santa Cruz genome browser showing the chymotrypsin C gene. The rows in the browser show annotations on the gene sequence. The browser window here shows a small segment of human chromosome 15, as if the sequence of a, g, c and t are represented from left to right (5–3). The annotations include gene predictions and annotations as well as an alignment of the similarity of this region of the genome when compared with the mouse genome

variations (mutations and SNPs), genes, comparative maps with other organisms, and many other important data.

9.5.3 Structure Databases

Although sequence information is obtained relatively easily, structural information remains expensive on a per-entry basis. The experimental protocols used to determine precise molecular structural coordinates are expensive in time, materials, and human power. Therefore, we have only a small number of structures for all the molecules characterized in the sequence databases. The two main sources of structural information are the Cambridge Structural Database[35] for small molecules (usually less than 100 atoms) and the PDB[36] for macromolecules (see ▶ Sect. 9.3.2), including proteins and nucleic acids, and combinations of these macromolecules with small molecules (such as drugs, cofactors, and vitamins). The PDB has approximately 75,000 high-resolution structures, but this number is misleading because many of them are small variants on the same structural architecture. There are approximately 100,000 proteins in humans; therefore, many structures remain unsolved (e.g., Burley and Bonanno 2002). In the PDB, each structure is reported with its biological source, reference information, manual annotations of interesting features, and the Cartesian coordinates of each atom within the molecule. Given knowledge of the three-dimensional structure of

35 ▶ https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/ (accessed December 15, 2018).

36 ▶ https://www.rcsb.org/ (accessed December 18, 2018).

molecules, the function sometimes becomes clear. For example, the ways in which the medication methotrexate interacts with its biological target have been studied in detail for two decades. Methotrexate is used to treat cancer and rheumatologic diseases, and it is an inhibitor of the protein dihydrofolate reductase, an important molecule for cellular reproduction. The three-dimensional structure of dihydrofolate reductase has been known for many years and has thus allowed detailed studies of the ways in which small molecules, such as methotrexate, interact at an atomic level. As the PDB increases in size, it becomes important to have organizing principles for thinking about biological structure. SCOP2[37] provides a classification based on the overall structural features of proteins. It is a useful method for accessing the entries of the PDB.

9.5.4 Analysis of Biological Pathways and Understanding of Disease Processes

The ECOCYC project is an example of a computational resource that has comprehensive information about biochemical pathways. ECOCYC is a knowledge base of the metabolic capabilities of *E. coli*; it has a representation of all the enzymes in the *E. coli* genome and of the chemical compounds those enzymes transform.[38] It also links these enzymes to their genes, and genes are mapped to the genome sequence.

EcoCyc also encodes the genetic regulatory network of *E. coli*, describing all protein and RNA regulators of *E. coli* genes. The network of pathways within ECOCYC provides an excellent substrate on which useful applications can be built. For example, they provide: (1) the ability to guess the function of a new protein by assessing its similarity to *E. coli* genes with a similar sequence, (2) the ability to ask what the effect on an organism would be if a critical component of a pathway were removed (would other pathways be used to create the desired function, or would the organism lose a vital function and die?), and (3) the ability to provide a rich user interface to the literature on *E. coli* metabolism. Similarly, the Kyoto Encyclopedia of Genes and Genomes (KEGG) provides pathway datasets for organism genomes.[39]

9.5.5 Integrative Databases

A **integrative database** is a postgenomic database that bridges the gap between molecular biological databases with those of clinical importance. One excellent example of a postgenomic database is the Online Mendelian Inheritance in Man (OMIM) database, which is a compilation of known human genes and genetic diseases, along with manual annotations describing the state of our understanding of individual genetic disorders.[40] Each entry contains links to special-purpose databases and thus provides links between clinical syndromes and basic molecular mechanisms (Fig. 9.6).

9.6 Future Challenges as Bioinformatics and Clinical Informatics Converge

Bioinformatics didn't solve all of its problems with the sequencing of the human genome. There is a series of challenges for which the completion of the first human genome sequence is only the beginning.

9.6.1 Linkage of Molecular Information with Symptoms, Signs, and Patients

There is currently a gap in our understanding of disease processes. Although we have a good understanding of the principles by

37 ► http://scop2.mrc-lmb.cam.ac.uk/ (accessed December 15, 2018).

38 ► http://ecocyc.org/ (accessed December 15, 2018).

39 ► http://www.genome.jp/kegg/pathway.html (accessed December 1, 2018).

40 ► http://www.ncbi.nlm.nih.gov/omim/ (accessed December 1, 2018).

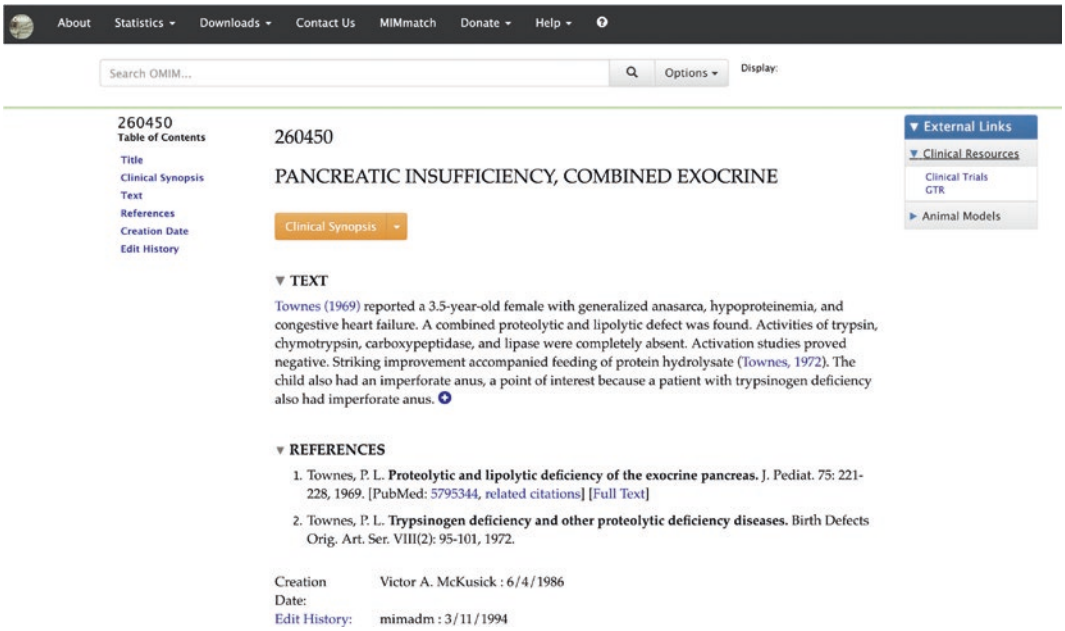


Fig. 9.6 Screen from the Online Mendelian Inheritance in Man (*OMIM*) database showing an entry for pancreatic insufficiency, an autosomal recessive disease in which chymotrypsin (NCBI Gene entry shown in Fig. 9.2) is totally absent (as are some other key digestive enzymes). (Courtesy of NCBI)

which small groups of molecules interact, we are not able to explain fully how thousands of molecules interact within a cell to create both normal and abnormal physiological states. As the databases continue to accumulate information ranging from patient-specific data to fundamental genetic information, a major challenge is creating the conceptual links among these databases to create an audit trail from molecular-level information to macroscopic phenomena, as manifested in disease. The availability of these links will facilitate the identification of important targets for future research and will provide a scaffold for biomedical knowledge, ensuring that important literature is not lost within the increasing volume of published data.

9.6.2 Computational Representations of the Biomedical Literature

An important opportunity within bioinformatics is the linkage of biological experimental data with the published papers that report them. Electronic publication of the biological literature provides exciting opportunities for making data easily available to scientists. Already, certain types of simple data that are produced in large volumes are expected to be included in manuscripts submitted for publication, including new sequences that are required to be deposited in GENBANK and new structure coordinates that are deposited in the PDB. However, there are many other experimental data sources that are currently difficult to provide in a standardized way, either because the data are more intricate than those stored in GENBANK or PDB or they are not produced in a volume sufficient to fill a database devoted entirely to the relevant area. Knowledge base technology can be used, however, to represent multiple types of highly interrelated data.

Knowledge bases can be defined in many ways (see ▶ Chap. 24); for our purposes, we can think of them as databases in which (1) the ratio of the number of tables to the number of entries per table is high compared with usual databases, (2) the individual entries (or records) have unique names, and (3) the values of many fields for one record in the database are the names of other records, thus creating

a highly interlinked network of concepts. The structure of knowledge bases often leads to unique strategies for storage and retrieval of their content. To build a knowledge base for storing information from biological experiments, there are some requirements. First, the set of experiments to be modeled must be defined. Second, the key attributes of each experiment that should be recorded in the knowledge base must be specified. Third, the set of legal values for each attribute must be specified, usually by creating a controlled terminology for basic data or by specifying the types of knowledge-based entries that can serve as values within the knowledge base.

9.6.3 Computational Challenges with an Increasing Deluge of Biomedical Data

An increasing challenge in biomedicine is storing, interpreting and integrating the massive amount of datasets the biomedical community is generating, largely from modern technologies in high throughput experimentation. The amount of DNA sequence data being generated over time has dwarfed Moore's Law, for example. This issue is important for all areas of biomedical informatics, and is discussed in more detail in the on Translational Bioinformatics (► Chap. 26).

9.7 Conclusion

Bioinformatics is closely allied to translational and clinical informatics. It differs in its emphasis on a reductionist view of biological systems, starting with sequence information and moving to structural and functional information. The emergence of the genome sequencing projects and the new technologies for measuring metabolic processes within cells is beginning to allow bioinformaticians to construct a more synthetic view of biological processes, which will complement the whole-organism, top-down approach of clinical informatics. More importantly, there are technologies that can be shared between bioinformatics and clinical informatics because they both focus on representing, storing, and analyzing biological or biomedical data. These technologies include the creation and management of standard terminologies and data representations, the integration of heterogeneous databases, the organization and searching of the biomedical literature, the use of machine learning techniques to extract new knowledge, the simulation of biological processes, and the creation of knowledge-based systems to support advanced practitioners in the two fields.

Suggested Readings

Altman, R. B., Dunker, A. K., Hunter, L., & Klein, T. E. (2003). Pacific symposium on Biocomputing'03. Singapore: World Scientific Publishing. The proceedings of one of the principal meetings in bioinformatics, this is an excellent source for up-to-date research reports. Other important meetings include those sponsored by the International Society for Computational Biology (ISCB, http://www.iscb.org/), Intelligent Systems for Molecular Biology (ISMB, http://iscb.org/conferences.shtml.35), and the RECOMB meetings on computational biology (http://www.ctw-congress.de/recomb/). ISMB and PSB have their proceedings indexed in PubMed.

Baldi, P., & Brunak, S. (2001). Bioinformatics: The machine learning approach. Cambridge, MA: MIT Press. This introduction to the field of bioinformatics focuses on the use of statistical and artificial intelligence techniques in machine learning.

Baldi, P., & Hatfield, G. W. (2002). DNA microarrays and gene expression. Cambridge: Cambridge University Press. Introduces the different microarray technologies and how they are analyzed.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2010). Biochemistry. New York: W.H. Freeman. The textbook by Stryer and colleagues is well written, and is illustrated and updated on a regular basis. It provides an excellent introduction to basic molecular biology and biochemistry.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.

This edited volume provides an excellent introduction to the use of probabilistic representations of sequences for the purposes of alignment, multiple alignment, and analysis.

Gusfield, D. (1997). Algorithms on strings, trees and sequences: Computer science and computational biology. Cambridge: Cambridge University Press. Gusfield's text provides an excellent introduction to the algorithmics of sequence and string analysis, with special attention paid to biological sequence analysis problems.

Malcolm, S., & Goodship, J. (Eds.). (2007). Genotype to phenotype (2nd ed.). Oxford: BIOS Scientific Publishers. This volume illustrates the different efforts to understand how diseases are linked to genes.

Pevsner, P. (2009). Bioinformatics and functional genomics. Hoboken: Wiley. A widely used excellent introduction to bioinformatics algorithms.

Questions for Discussion

1. How are DNA and protein sequence information changing the way that medical records are managed? Which types of systems are or will be most affected (laboratory, radiology, admission and discharge, financial, order entry)?
2. It has been postulated that clinical informatics and bioinformatics are working on the same problems, but in some areas one field has made more progress than the other. Identify three common themes. Describe how the issues are approached by each subdiscipline.
3. Why should an awareness of bioinformatics be expected of clinical informatics professionals? Should a chapter on bioinformatics appear in a clinical informatics textbook? Explain your answers.
4. Why should an awareness of clinical informatics be expected of bioinformatics professionals? Should a chapter on clinical informatics appear in a bioinformatics textbook? Explain your answers.
5. One major problem with introducing computers into clinical medicine is the extreme time and resource pressure placed on physicians and other health care workers. Do you think that the same problems are arising in basic biomedical research?
6. Why have biologists and bioinformaticians embraced the Web as a vehicle for disseminating data so quickly, whereas clinicians and clinical informaticians have been more hesitant to put their primary data online?
7. If a patient's entire genome were present in their medical record how would one go about interpreting it clinically? Similarly, if we had an entire electronic health record database that included human genomes, how would a researcher go about finding new or novel genetic associations?
8. With the many high throughput experiments that are used in biomedical research, how are some ways to integrate those datasets using systems biology? For example, if you had a microarray dataset that annotated gene expression levels and a proteomics dataset that identified protein interactions, how could you jointly use both datasets to identify markers for a disease?

References

Altschul, S. F., Gish, W., Mille, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403–410.

Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., et al. (2010). Clinical assessment incorporating a personal genome. *The Lancet, 375*, 1525–1535.

Babior, B. M., & Matzner, Y. (1997). The familial Mediterranean fever gene—cloned at last. *The New England Journal of Medicine, 337*(21), 1548–1549.

Bai, C., & Elledge, S. J. (1997). Gene identification using the yeast two-hybrid system. *Methods in Enzymology, 283*, 141–156.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development, 16*(1), 6–21.

Brown, D. G., Rao, S., Weir, T. L., O'Malia, J., Bazan, M., Brown, R. J., & Ryan, E. P. (2016). Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool. *Cancer and Metabolism, 4*, 11.

Burley, S. K., & Bonanno, J. B. (2002). Structuring the universe of proteins. *Annual Review of Genomics and Human Genetics, 3*, 243–262.

Cech, T. R. (2000). Structural biology. The ribosome is a ribozyme. *Science, 289*(5481), 878–879.

Davies, K. (2010). Physicians and their use of information: A survey comparison between the United States, Canada, and the United Kingdom. *Journal of the Medical Library Association, 99*, 88–91.

Dayhoff, M. O. (1974). Computer analysis of protein sequences. *Federation Proceedings, 33*(12), 2314–2316.

Durfy, S. J. (1993). Ethics and the human genome project. *Archives of Pathology & Laboratory Medicine, 117*(5), 466–469.

Fischer, B. A., & Zigmond, M. J. (2010). The essential nature of sharing in science. *Science and Engineering Ethics, 16*(4), 783–799.

Gibson, K., & Scheraga, H. (1967). Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease S-peptide. *Proceedings of the National Academy of Sciences, 58*(2), 420–427.

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A landscape takes shape. *Cell, 128*(4), 635–638.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences: Computer science and computational biology*. Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Karplus, M., & Weaver, D. L. (1976). Protein-folding dynamics. *Nature, 260*(5550), 404–406.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell, 150*(2), 389–401.

Kent, W. J. (2003). BLAT—the BLAST-like alignment tool. *Genome Research, 12*(4), 656–664.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*, 860–921.

Langridge, R. (1974). Interactive three-dimensional computer graphics in molecular biology. *Federation Proceedings, 33*(12), 2332–2335.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., et al. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences, 94*(24), 13057–13062.

Levitt, M. (1983). Molecular dynamics of native protein. I. Computer simulation of trajectories. *Journal of Molecular Biology, 168*(3), 595–617.

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: Short oligonucleotide alignment program. *Bioinformatics, 24*(5), 713–714.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*(3), 443–453.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics, 42*, 30–35.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*(7285), 59–65.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry, 34*, 167–339.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020) Improved protein structure prediction using potentials from deep learning. Nature, 577, 706–710.

Shapiro, E., Biezunner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics, 14*, 618–630.

Smith, T., & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147*(1), 195–197.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology, 25*(11), 1251–1255.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9440–9445.

Van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 484–485.

Wei, L., & Altman, R. B. (1998). Recognizing protein binding sites using statistical descriptions of their 3D environments. In *Proceedings of the pacific symposium on Biocomputing '98* (pp. 497–508), Singapore.

Yan, J., & Gu, W. (2009). Gene expression microarrays. In Y. Lu & R. I. Mahato (Eds.), *Cancer research pharmaceutical perspectives of cancer therapeutics* (pp. 645–672). New York: Springer.