22
Bioinformatics

Russ B. Altman and Sean D. Mooney

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

- Why is 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 two changes should we anticipate in the medical record as a result of these new information sources?
- What are two computational challenges in bioinformatics for the future?

22.1 The Problem of Handling Biological Information

Bioinformatics is the study of how information is represented and analyzed in biological systems, starting 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 Chapter 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 them 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.

The discipline of bioinformatics is currently in a period of rapid growth, because the needs for information storage, retrieval, and analysis in biology—particularly in molec-

ular biology and **genomics**—have increased dramatically in the past decade. History has shown that scientific developments within the basic sciences tend to lag about a decade before their influence on clinical medicine is fully appreciated. The types of information being gathered by biologists today will drastically alter the types of information and technologies available to the health care workers of tomorrow.

22.1.1 Many Sources of Biological Data

There are three sources of information that are revolutionizing our understanding of human biology and that are creating significant challenges for computational processing. The most dominant new type of information is the **sequence information** produced 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 chromosomes.[1] 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.[2] Now efforts are under way to finish the sequence and to determine the variations that occur between the genomes of different individuals.[3] Essentially, the entire set of 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.

While we are studying the human genome, a second set of concurrent projects is studying the genomes of numerous other biological organisms, including important experimental animal systems (such as mouse, rat, and yeast) as well as important human pathogens (such as *Mycobacterium tuberculosis* or *Haemophilus influenzae*). Many of these genomes have recently been completely determined by sequencing experiments. 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 not only are proceeding at a furious pace, but also are accompanied in many cases by a new approach to biology, which produces a third new source of biomedical information: **proteomics**. In addition to small, relatively focused experimental studies aimed at particular molecules thought to be important for disease, large-scale experimental methodologies are used to collect data on thousands or millions of molecules simultaneously. Scientists apply these methodologies longitudinally over time and across a wide variety of organisms or (within an organism) organs to watch the evolution of various physiological phenomena. New technologies give us the abilities to follow the production and degradation of molecules on **DNA arrays**[4]

[1] http://www.genome.gov/page.cfm?pageID=10001694.
[2] http://www.genome.gov/10005139.
[3] http://www.genome.gov/page.cfm?pageID=10001688.
[4] These are small glass plates onto which specific DNA fragments can be affixed and then used to detect other DNA fragments present in a cell extract.

(Lashkari et al., 1997), to study the expression of large numbers of proteins with one another (Bai and Elledge, 1997), and to create multiple variations on a genetic theme to explore the implications of various mutations on biological function (Spee et al., 1993). 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.

22.1.2 Implications for Clinical Informatics

The effects of this new biological information on clinical medicine and clinical informatics are difficult to predict precisely. It is already clear, however, that some major changes to medicine will have to be accommodated.

1. *Sequence information in the medical record*. With the first set of human genomes now available, it will soon become cost-effective to consider sequencing or genotyping at least sections of many other genomes. The sequence of a gene involved in disease may provide the critical information that we need to select appropriate treatments. For example, the set of genes that produces essential hypertension may be understood at a level sufficient to allow us to target antihypertensive medications based on the precise configuration of these genes. It is possible that clinical trials may use information about genetic sequence to define precisely the population of patients who would benefit from a new therapeutic agent. Finally, clinicians may learn the sequences of infectious agents (such as of the *Escherichia coli* strain that causes recurrent urinary tract infections) and store them in a patient's record to record the precise pathogenicity and drug susceptibility observed during an episode of illness. In any case, it is likely that genetic information will need to be included in the medical record and will introduce special problems. Raw sequence information, whether from the patient or the pathogen, is meaningless without context and thus is not well suited to a printed medical record. Like images, it can come in high information density and must be presented to the clinician in novel ways. As there are for laboratory tests, there may be a set of nondisease (or normal) values to use as comparisons, and there may be difficulties in interpreting abnormal values. Fortunately, most of the human genome is shared and identical among individuals; less than 1 percent of the genome seems to be unique to individuals. Nonetheless, the effects of sequence information on clinical databases will be significant.
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. **Single nucleotide polymorphisms (SNPs)** and other genetic markers are used to identify how a patient's genome differs from the draft genome. 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. For

example, the severe acute respiratory syndrome (SARS) virus was determined to be a corona virus using a gene expression array containing the genetic information from several common pathogenic viruses.[5] In general, diagnostic tools based on the gene sequences within a patient are likely to increase greatly the number and variety of tests available to the physician. Physicians will not be able to manage these tests without significant computational assistance. Moreover, genetic information will be available to provide more accurate prognostic information to patients. What is the standard course for this disease? How does it respond to these medications? Over time, we will be able to answer these questions with increasing precision, and will develop computational systems to manage this information.

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. The *Human Gene Mutations Database* (HGMD) annotates mutations with disease phenotype.[6] 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.[7] As these data sets, and others like them, continue to improve, the first clinical benefits from the genome projects will be realized.

3. *Ethical considerations*. One of the critical questions facing the genome-sequencing projects is "Can genetic 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? 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 is disclosed (Durfy, 1993; see Chapter 10).

22.2 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 collec-

[5] http://www.cdc.gov/ncidod/sars/.
[6] http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html.
[7] http://pharmgkb.org.
[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.

tion 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 the sequence of DNA subunits in turn 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 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.

Understanding the basic building blocks of life requires understanding the function of genomic sequences, genes, and proteins. When are genes turned on? Once genes are transcribed and translated into proteins, into what cellular compartment are the proteins directed? How do the proteins function once there? Equally important, how are the proteins turned off? Experimentation and bioinformatics have divided the research into several areas, and the largest are: (1) genome and protein sequence analysis, (2) macromolecular structure–function analysis, (3) gene expression analysis, and (4) proteomics.

22.2.1 Roots of Modern Bioinformatics

Practitioners of bioinformatics have come from many backgrounds, including medicine, molecular biology, chemistry, physics, 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 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, there have been hosts of biological problems addressed with computational methods.

22.2.2 The Genomics Explosion

The Human Genome Project was completed and a nearly finished sequence was published in 2003.[9] 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

[9]http://www.genome.gov/10005139.

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 understated. 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. Where 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. 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.

22.3 Biology Is Now Data-Driven

Twenty 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 is because advances in research methods such as **microarray chips**, drug screening robots, **X-ray crystallography**, **nuclear magnetic resonance spectroscopy**, and DNA sequencing experiments have resulted in massive amounts of data. These data need to be properly stored, analyzed, and disseminated.

The volume of data being produced by genomics projects is staggering. There are now more than 22.3 million sequences in GenBank comprising more than 29 billion digits.[10] But these data do not stop with sequence data: PubMed contains over 15 million literature citations, the PDB contains three-dimensional structural data for over 40,000 protein sequences, and the Stanford Microarray Database (SMD) contains over 37,000 experiments (851 million data points). 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.

22.3.1 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.

One major difficulty within bioinformatics is that standard database models, such as relational database systems, are not well suited to sequence information. The basic problem is that sequences are important both as a set of elements grouped together and

[10] http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html.

treated in a uniform manner and as individual elements, with their relative locations and functions. Any given position in a sequence can be important because of its own identity, because it is part of a larger subsequence, or perhaps because it is part of a large set of overlapping subsequences, all of which have different significance. It is necessary to support queries such as, "What sequence motifs are present in this sequence?" It is often difficult to represent these multiple, nested relationships within standard relational database schema. In addition, the neighbors of a sequence element are also critical, and it is important to be able to perform queries such as, "What sequence elements are seen 20 elements to the left of this element?" For these reasons, researchers in bioinformatics are developing **object-oriented databases** (see Chapter 6) in which a sequence can be queried in different ways, depending on the needs of the user (Altman, 2003).

22.3.2 Structures in Biology

The sequence information mentioned in Section 22.3.1 is rapidly becoming inexpensive to obtain and easy to store. On the other hand, the **three-dimensional structure information** about the proteins that are produced from the DNA sequences is much more difficult and expensive to obtain, and presents a separate set of analysis challenges. Currently, only about 30,000 three-dimensional structures of biological macromolecules are known.[11] 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 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 vastly outsized 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. 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 grow the number of known structures considerably. In the end, it is the physical forces 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).

[11]For more information see http://www.rcsb.org/pdb/.

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),[12] which classifies proteins based on shape and function.

22.3.3 Expression Data in Biology

The development of DNA microarrays has led to a wealth of data and unprecedented insight into the fundamental biological machine. The premise is relatively simple; up to 40,000 gene sequences derived from genomic data are fixed onto a glass slide or filter. An experiment is performed where two groups of cells are grown in different conditions, 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 and converted back to DNA, using special enzymes. This leaves a pool of DNA molecules that are only from the genes that were turned on (expressed) in that group of cells. Using a chemical reaction, the experimental DNA sample is attached to a red fluorescent molecule and the control group is attached to a green fluorescent molecule. These two samples are mixed and then washed over the glass slide. The two samples contain only genes that were turned on in the cells, and they are labeled either red or green depending on whether they came from the experimental group or the control group. The labeled DNA in the pool sticks or hybridizes to the same gene on the glass slide. This leaves the glass slide with up to 40,000 spots and genes that were turned on in the cells are now bound with a label to the appropriate spot on the slide. Using a scanning confocal microscope and a laser to fluoresce the linkers, the amount of red and green fluorescence in each spot can be measured. The ratio of red to green determines whether that gene is being turned off (downregulated) in the experimental group or whether the gene is being turned on (upregulated). The experiment has now measured the activity of genes in an entire cell due to some experimental change. Figure 22.1 illustrates a typical gene expression experiment from SMD.[13]

Computers are critical for analyzing these data, because it is impossible for a researcher to comprehend the significance of those red and green spots. Currently scientists are using gene expression experiments to study how cells from different organ-

[12]See http://scop.mrc-lmb.cam.ac.uk/scop/.
[13]http://genome-www5.stanford.edu/MicroArray/SMD/.

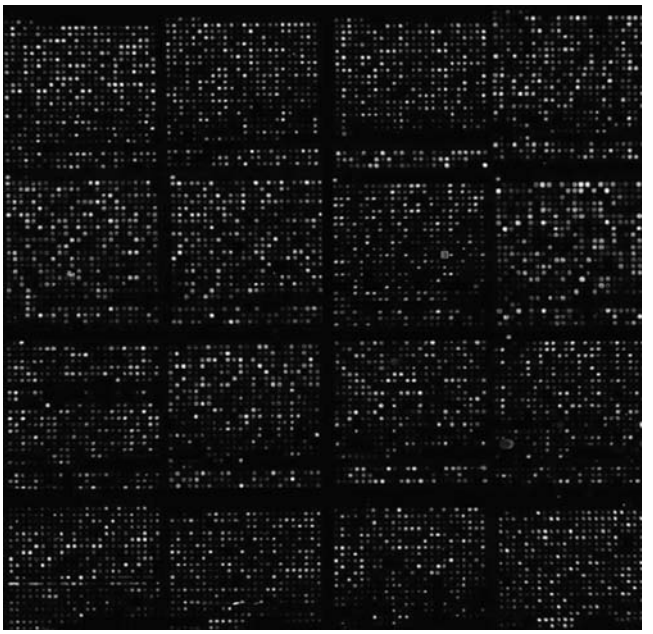

Figure 22.1. Measuring global levels of gene expression. Genomics has created a new need for bioinformatics tools. In this experiment (from the Stanford Microarray Database), stress-induced changes in the gene expression pattern for bakers yeast (*S. cerevisae*) are shown.

isms compensate for environmental changes, how pathogens fight antibiotics, and how cells grow uncontrollably (as is found in cancer). A new 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.

22.3.4 Systems Biology

With the completion of the human genome and the abundance of sequence, structural, and gene expression data, a new field of systems biology that tries to understand how proteins and genes interact at a cellular level is emerging. The basic algorithms for analyzing sequence and structure are now leading to opportunities for more integrated analysis of the pathways in which these molecules participate and ways in which molecules can be manipulated for the purpose of combating disease. A detailed understanding of the role of a particular molecule in the cell requires knowledge of the context—of the other molecules with which it interacts—and of the sequence of chemical transformations that take place in the cell. Thus, major research areas in bioinformatics are elucidating the key pathways by which chemicals are transformed, defining the molecules that catalyze these transformations, identifying the input compounds and the output compounds, and linking these pathways into

networks that we can then represent computationally and analyze to understand the significance of a particular molecule. The Alliance for Cell Signaling is generating large amounts of data related to how signal molecules interact and affect the concentration of small molecules within the cell.

22.4 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.

22.4.1 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 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 et al. (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) of 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:[14] the PDB now has over 30,000 detailed atomic structures and is the primary source of information about the relationship between protein sequence and protein structure. 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 2 million sequences and 100 billion bases. The GENBANK database of DNA sequence information supports the experimental reconstruction of genomes and acts as a focal point for experimental groups.[15] Numerous other databases store the sequences of protein molecules[16] and information about human genetic diseases.[17]

Included among the databases that have accelerated the development of bioinformatics is the Medline[18] database of the biomedical literature and its paper-based companion *Index Medicus* (see Chapter 19). Including articles as far back as 1953 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 Entrez and PubMed systems for integrating access to literature references and the associated databases.

22.4.2 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

[14]See http://www.rcsb.org/pdb/holdings.html.
[15]http://gdbwww.gdb.org/.
[16]The Protein Identification Resource: http://pir.georgetown.edu; Swiss-Prot at http://www.expasy.ch/sprot/.
[17]Online Mendelian Inheritance in Man: http://www3.ncbi.nlm.nih.gov/omim/.
[18]See http://www.ncbi.nlm.nih.gov/PubMed/.

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 22.2 shows an example of a Smith-Waterman matrix.

Unfortunately, the dynamic programming algorithms are computationally expensive to apply, 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 observations 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. Another tool that has found wide use in mining genome sequences is BLAT (Kent, 2003). BLAT is often used to search long genomic sequences with significant performance increases over BLAST. It achieves its 50-fold increase in speed over other tools by storing and indexing long sequences as nonoverlapping k-mers, allowing efficient storage, searching, and alignment on modest hardware.

22.4.3 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.[19] Recent results suggest that when two protein molecules have a high degree (more than 40 percent) of sequence similarity 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 percent, however, performance of these methods is much less reliable.

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 (such as those that SCOP maintains).

One of these algorithms is MinRMS (Jewett et al., 2003).[20] MinRMS works by finding the minimal root-mean-squared-distance (RMSD) alignments of two protein

[19] http://predictioncenter.org/.
[20] http://www.cgl.ucsf.edu/Research/minrms/.

a) Pairwise alignment between human chymotrypsin and human trypsin.

```
CTRB_HUMAN      MAFLWLLSCWALLGTTFGCGVPAIHPVLSGLSRIVNGEDAVPGSWPWQVSLQDKTGFHFC
TRY1_HUMAN      MNPLLILTFVA----------AALAAPFDDDDKIVGGYNCEENSVPYQVSLN--SGYHFC

CTRB_HUMAN      GGSLISEDWVVTAAHCGVRTSDVVVAGEFDQGSDEENIQVLKIAKVFKNPKFSILTVNND
TRY1_HUMAN      GGSLINEQWVVSAGHC-YKSRIQVRLGEHNIEVLEGNEQFINAAKIIRHPQYDRKTLNND

CTRB_HUMAN      ITLLKLATPARFSQTVSAVCLPSADDDFPAGTLCATTGWGKTKYNANKTPDKLQQAALPL
TRY1_HUMAN      IMLIKLSSRAVINARVSTISLPTAPP--ATGTKCLISGWGNTASSGADYPDELQCLDAPV

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

CTRB_HUMAN      STSSPGVYARVTKLIPWVQKILAAN-
TRY1_HUMAN      QKNKPGVYTKVYNYVKWIKNTIAANS
```

b) Smith Waterman matrix illustrating the aligned region in A, using the BLOSUM62 mutation matrix (Henikoff 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	-2	1
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
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

Figure 22.2. Example of sequence alignment using the Smith Waterman algorithm.

structures as a function of matching residue pairs. MinRMS generates a family of alignments, each with different number of residue position matches. This is useful for identifying local regions of similarity in a protein with multiple domains. MinRMS solves two problems. First, it determines which structural *superpositions*, or alignment, to evaluate. Then, given this superposition, it determines which residues should be

considered "aligned" or matched. Computationally, this is a very difficult problem. MinRMS reduces the search space by limiting superpositions to be the best superposition between four atoms. It then exhaustively determines all potential four-atom-matched superpositions and evaluates the alignment. Given this superposition, the number of aligned residues is determined, as any two residues with C-alpha carbons (the central atom in all amino acids) less than a certain threshold apart. The minimum average RMSD for all matched atoms is the overall score for the alignment. In Figure 22.3, an example of such a comparison is shown.

A related problem is that of using the structure of a large biomolecule and the structure of a small organic molecule (such as a 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

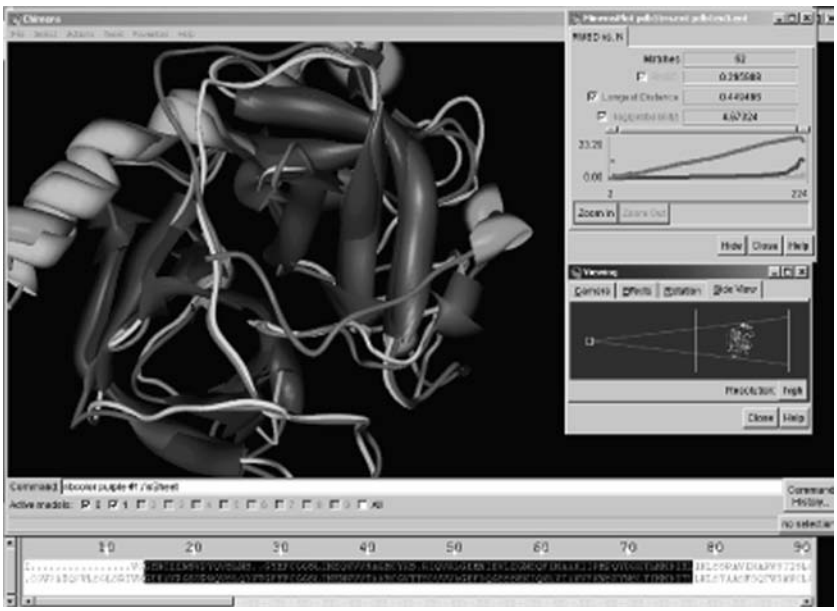

Figure 22.3. Example of structural comparison. Comparison of the chymotrypsin and trypsin protein structures using Chimera and MinRMS (http://www.cgl.ucsf.edu/chimera).

function. These methods can guess at general function for roughly 60 to 80 percent 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.

22.4.4 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. Within each cell is the red to green ratio of that gene's experimental results. 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 Guo, 2003, for 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. Clustering methods fall into two categories: supervised and unsupervised. Supervised learning methods require some preconceived knowledge of the data at hand. Usually, the method begins by selecting profiles that represent the different groups of data, 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 (Figure 22.4).

The future clinical usefulness of these algorithms cannot be understated. 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 can-

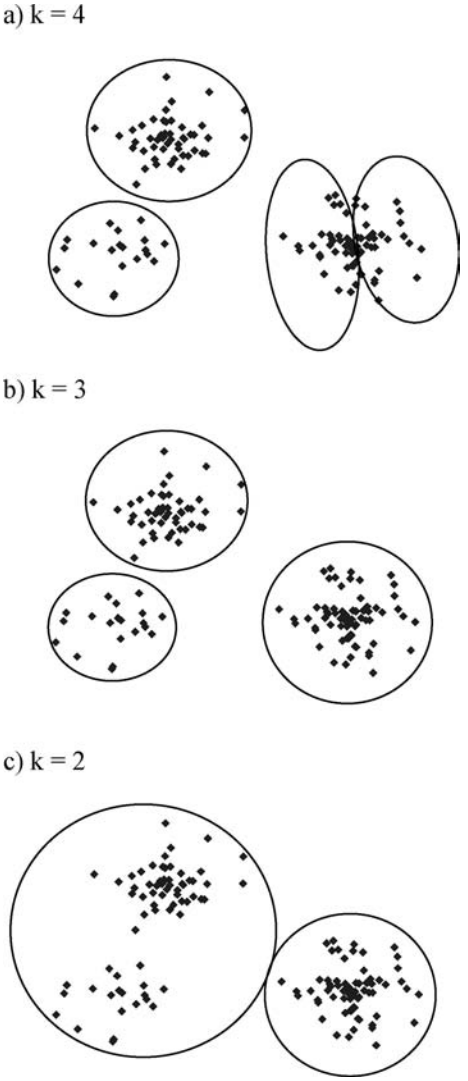

Figure 22.4. K-means clustering example with varying k. In this case, $k = 3$ is the most reasonable.

cers 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.

22.5 Current Application Successes from Bioinformatics

Biologists have embraced the Web in a remarkable way and have made Internet 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 *Entrez* system from the National Center for Biological Information (NCBI) 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.[21] The *Sequence Retrieval System* (SRS) from the European Molecular Biology Laboratory allows queries from one database to another to be linked and sequenced, thus allowing relatively complicated queries to be evaluated.[22] 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.

22.5.1 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 NCBI.[23] GENBANK is growing rapidly as genome-sequencing projects feed their data (often in an automated procedure) directly into the database. Figure 22.5 shows the logarithmic growth of data in GENBANK since 1982. Entrez Gene curates some of the many genes within GENBANK and presents the data in a way that is easy for the researcher to use (Figure 22.6).

[21]See http://www3.ncbi.nlm.nih.gov/Entrez/.
[22]See http://www.lionbioscience.com/solutions/products/srs/.
[23]http://www.ncbi.nlm.nih.gov/.

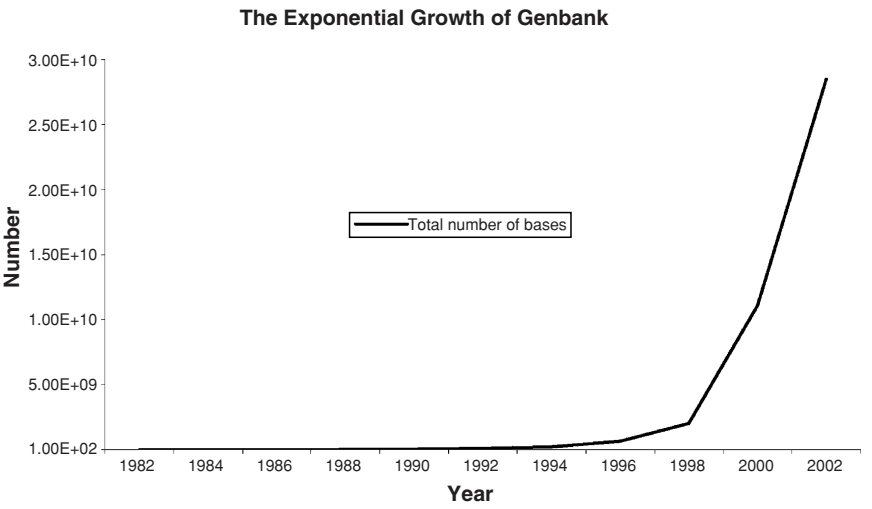

Figure 22.5. The exponential growth of GENBANK. This plot shows that since 1982 the number of bases in GENBANK has grown by five full orders of magnitude and continues to grow by a factor of 10 every 4 years.

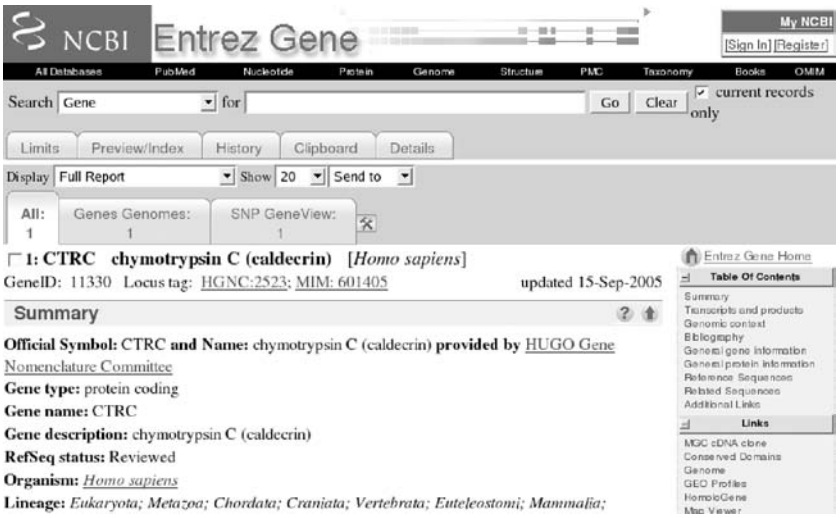

Figure 22.6. The Entrez Gene entry for the digestive enzyme chymotrypsin. Basic information about the original report is provided, as well as some annotations of the key regions in the sequence and the complete sequence of DNA bases (a, g, t, and c) is provided as a link. (Courtesy of NCBI)

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 assembly browser[24] (Figure 22.7). 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 variations (mutations and SNPs), genes, comparative maps with other organisms, and many other important data.

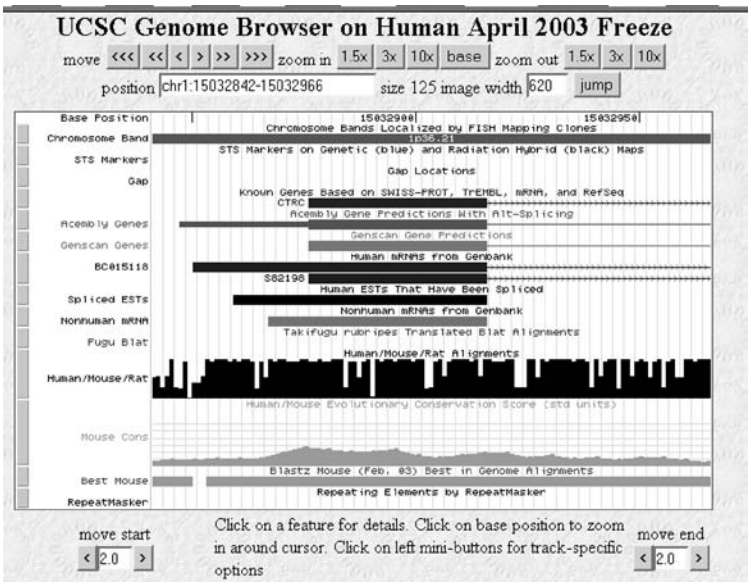

Figure 22.7. Screen from the UC Santa Cruz genome browser showing chymotrypsin. The rows in the browser show annotations on the gene sequence. The browser window here shows a small segment of human chromosome 15, as if the sequence of a, g, c and t are represented from left to right (5 to 3). The annotations 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.

[24] http://genome.ucsc.edu/.

22.5.2 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[25] for small molecules (usually less than 100 atoms) and the PDB[26] for macromolecules (see Section 22.3.2), including proteins and nucleic acids, and combinations of these macromolecules with small molecules (such as drugs, cofactors, and vitamins). The PDB has approximately 20,000 high-resolution structures, but this number is misleading because many of them are small variants on the same structural architecture (Figure 22.8). If an algorithm is applied to the database to filter out redundant structures, less than 2,000 structures remain.

There are approximately 100,000 proteins in humans; therefore many structures remain unsolved (e.g., Burley and Bonanno, 2002; Gerstein et al., 2003). In the PDB,

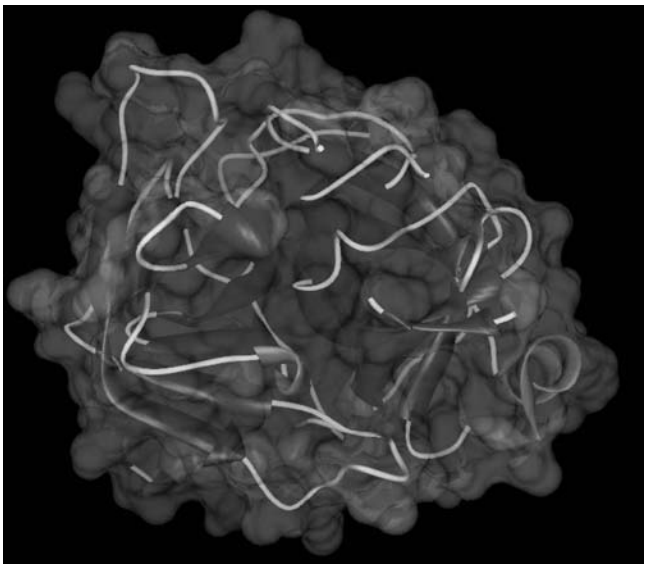

Figure 22.8. A stylized diagram of the structure of chymotrypsin, here shown with two identical subunits interacting. The red portion of the protein backbone shows α-helical regions, while the blue portion shows β-strands, and the white denotes connecting coils, while the molecular surface is overlaid in gray. The detailed rendering of all the atoms in chymotrypsin would make this view difficult to visualize because of the complexity of the spatial relationships between thousands of atoms.

[25] http://www.ccdc.cam.ac.uk/.
[26] http://www.rcsb.org/.

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 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. SCOP[27] provides a classification based on the overall structural features of proteins. It is a useful method for accessing the entries of the PDB.

22.5.3 *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.[28] 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 compounds on which they work. It also links these enzymes to their position on the genome to provide a useful interface into this information. The network of pathways within ECOCYC provides an excellent substrate on which useful applications can be built. For example, they could 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 datacets for organism genomes.[29]

22.5.4 *Postgenomic Databases*

A **postgenomic database** 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,[30] 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. Each entry contains links to special-purpose databases and thus provides links between clinical syndromes and basic molecular mechanisms (Figure 22.9).

---

[27]See http://scop.mrc-lmb.cam.ac.uk/scop/.
[28]http://www.ecocyc.org/.
[29]http://www.genome.ad.jp/kegg/.
[30]http://www3.ncbi.nlm.nih.gov/omim/.

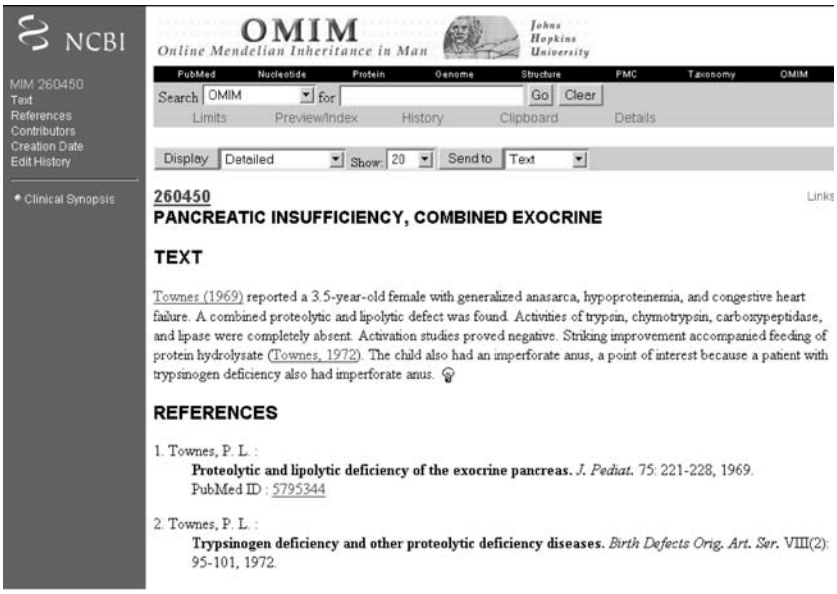

Figure 22.9. Screen from the Online Mendelian Inheritance in Man (OMIM) database showing an entry for pancreatic insufficiency, an autosomal recessive disease in which chymotrypsin (LocusLink entry shown in Figure 22.2) is totally absent (as are some other key digestive enzymes). (Courtesy of NCBI)

The SMD is another example of a postgenomic database that has proven extremely useful, but has also addressed some formidable challenges. As discussed previously in several sections, expression data are often represented as vectors of data values. In addition to the ratio values, the SMD stores images of individual chips, complete with annotated gene spots (see Figure 22.1). Further, the SMD must store experimental conditions, the type and protocol of the experiment, and other data associated with the experiment. Arbitrary analysis can be performed on different experiments stored in this unique resource.

A critical technical challenge within bioinformatics is the interconnection of databases. As biological databases have proliferated, researchers have been increasingly interested in linking them to support more complicated requests for information. Some of these links are natural because of the close connection of DNA sequence to protein structure (a straightforward translation). Other links are much more difficult because the semantics of the data items within the databases are fuzzy or because good methods for linking certain types of data simply do not exist. For example, in an ideal world, a protein sequence would be linked to a database containing information about that sequence's function. Unfortunately, although there are databases about protein function, it is not always easy to assign a function to a protein based on sequence information alone, and so the databases are limited by gaps in our understanding of biology. Some excellent recent work in the integration of diverse biological databases has been

done in connection with the NCBI Entrez/PubMed systems,[31] the SRS resource,[32] DiscoveryLink,[33] and the Biokleisli project.[34]

22.6 Future Challenges as Bioinformatics and Clinical Informatics Converge

The human genome sequencing projects will be complete within a decade, and if the only *raison d'etre* for bioinformatics is to support these projects, then the discipline is not well founded. If, on the other hand, we can identify a set of challenges for the next generations of investigators, then we can more comfortably claim disciplinary status for the field. Fortunately, there is a series of challenges for which the completion of the first human genome sequence is only the beginning.

22.6.1 Completion of Multiple Human Genome Sequences

With the first human genome in hand, the possibilities for studying the role of genetics in human disease multiply. A new challenge immediately emerges, however: collecting individual sequence data from patients who have disease. Researchers estimate that more than 99 percent of the DNA sequences within humans are identical, but the remaining sequences are different and account for our variability in susceptibility to and development of disease states. It is not unreasonable to expect that for particular disease syndromes, the detailed genetic information for individual patients will provide valuable information that will allow us to tailor treatment protocols and perhaps let us make more accurate prognoses. There are significant problems associated with obtaining, organizing, analyzing, and using this information.

22.6.2 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 which small groups of molecules interact, we are not able to fully explain 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 between 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.

---

[31] http://www.ncbi.nlm.nih.gov/PubMed/.
[32] http://srs.embl-heidelberg.de:8000/.
[33] http://www.research.ibm.com/journal/sj/402/haas.html.
[34] http://www.geneticxchange.com/.

22.6.3 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, because the data either are more intricate than those stored in GENBANK or PDB or 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 Chapter 20); 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.

The development of such schemes necessitates the creation of terminology standards, just as in clinical informatics. The RiboWeb project is undertaking this task in the domain of RNA biology (Chen et al., 1997). RiboWeb is a collaborative tool for ribosomal modeling that has at its center a knowledge base of the ribosomal structural literature. RiboWeb links standard bibliographic references to knowledge-based entries that summarize the key experimental findings reported in each paper. For each type of experiment that can be performed, the key attributes must be specified. Thus, for example, a *cross-linking experiment* is one in which a small molecule with two highly reactive chemical groups is added to an ensemble of other molecules. The reactive groups attach themselves to two vulnerable parts of the ensemble. Because the molecule is small, the two vulnerable areas cannot be any further from each other than the maximum stretched-out length of the small molecule. Thus, an analysis of the resulting reaction gives information that one part of the ensemble is "close" to another part. This experiment can be summarized formally with a few features—for example, target of experiment, cross-linked parts, and cross-linking agent.

The task of creating connections between published literature and basic data is a difficult one because of the need to create formal structures and then to create the necessary content for each published article. The most likely scenario is that biologists will write and submit their papers along with the entries that they propose to add to the

knowledge base. Thus, the knowledge base will become an ever-growing communal store of scientific knowledge. Reviewers of the work will examine the knowledge-based elements, perhaps will run a set of automated consistency checks, and will allow the knowledge base to be modified if they deem the paper to be of sufficient scientific merit. RiboWeb in prototype form can be accessed on the Web.[35]

22.6.4 A Complete Computational Model of Physiology

One of the most exciting goals for computational biology and bioinformatics is the creation of a unified computational model of physiology. Imagine a computer program that provides a comprehensive simulation of a human body. The simulation would be a complex mathematical model in which all the molecular details of each organ system would be represented in sufficient detail to allow complex "what if?" questions to be asked. For example, a new therapeutic agent could be introduced into the system, and its effects on each of the organ subsystems and on their cellular apparatus could be assessed. The side-effect profile, possible toxicities, and perhaps even the efficacy of the agent could be assessed computationally before trials are begun on laboratory animals or human subjects. The model could be linked to visualizations to allow the teaching of medicine at all grade levels to benefit from our detailed understanding of physiological processes—visualizations would be both anatomic (where things are) and functional (what things do). Finally, the model would provide an interface to human genetic and biological knowledge. What more natural user interface could there be for exploring physiology, anatomy, genetics, and biochemistry than the universally recognizable structure of a human that could be browsed at both macroscopic and microscopic levels of detail? As components of interest were found, they could be selected, and the available literature could be made available to the user.

The complete computational model of a human is not close to completion. First, all the participants in the system (the molecules and the ways in which they associate to form higher-level aggregates) must be identified. Second, the quantitative equations and symbolic relationships that summarize how the systems interact have not been elucidated fully. Third, the computational representations and computer power to run such a simulation are not in place. Researchers are, however, working in each of these areas. The genome projects will soon define all the molecules that constitute each organism. Research in simulation and the new experimental technologies being developed will give us an understanding of how these molecules associate and perform their functions. Finally, research in both clinical informatics and bioinformatics will provide the computational infrastructure required to deliver such technologies.

22.7 Conclusion

Bioinformatics is closely allied to clinical informatics. It differs in its emphasis on a reductionist view of biological systems, starting with sequence information and moving

[35] http://smi-web.stanford.edu/projects/helix/riboweb.html.

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 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 Medline.

Baldi P., Brunak S. (1998). *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.

Bishop M., Rawlings C. (Eds.) (1997). *DNA and Protein Sequence Analysis—A Practical Approach*. New York: IRL Press at Oxford University Press.
This book provides an introduction to sequence analysis for the interested biologist with limited computing experience.

Durbin R., Eddy 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.

Gribskov M., Devereux J. (1991). *Sequence Analysis Primer*. New York: Stockton Press.
This primer provides a good introduction to the basic algorithms used in sequence analysis, including dynamic programming for sequence alignment.

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.

Hunter L. (1993). *Artificial Intelligence and Molecular Biology*. Menlo Park, CA: AAAI Press/MIT Press.
This volume shows a variety of ways in which artificial intelligence techniques have been used to solve problems in biology.

Malcolm S., Goodship, J. (Eds.) (2001) *Genotype to Phenotype* (2nd ed.). Oxford: BIOS Scientific Publishers.
This volume illustrates the different efforts to understand how diseases are linked to genes

Salzberg S., Searls D., Kasif S. (Eds.) (1998). *Computational Methods in Molecular Biology*. New York: Elsevier Science.
This volume offers a useful collection of recent work in bioinformatics.

Setubal J., Medianis J. (1997). *Introduction to Computational Molecular Biology*. Boston: PWS Publishing.
Another introduction to bioinformatics, this text was written for computer scientists.

Stryer L. (1995). *Biochemistry*. New York: W.H. Freeman.
The textbook by Stryer is well written, and is illustrated and updated on a regular basis. It provides an excellent introduction to basic molecular biology and biochemistry.

Questions for Discussion

1. In what ways will bioinformatics and medical informatics interact in the future? Will the research agendas of the two fields merge, or will they always remain separable?
2. Will the introduction of DNA and protein sequence information change the way that medical records are managed in the future? Which types of systems will be most affected (laboratory, radiology, admission and discharge, financial, order entry)?
3. 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.
4. 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.
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. Will the same problems arise 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?