

1

CHAPTER

An Introduction to Bioinformatics for Computer Scientists

David W. Corne University of Reading

Gary B. Fogel Natural Selection, Inc.

1.1 INTRODUCTION

In June 2000, the Sanger Center in Cambridge, England, announced one of the definitive achievements of modern times: The release of the first draft of the human genome, the 3,000,000,000-letter code that distinguishes *Homo sapiens* from other species. The achievement, a culmination of worldwide efforts involving 16 major biotechnology research centers, was unquestionably epochal, and the contribution to science (from this and the wider context of worldwide sequencing and similar efforts) was no less monumental. However, the contribution to *knowledge* was much less clear. If we treat knowledge in this context as a collection of answers to such challenging scientific questions as “Which genes are involved in the human immune system?” “Are we more closely related to chimpanzees or to gorillas?” “How can we slow the aging process?” then the publication of a DNA sequence provides no immediate help at all. There is a considerable gap between the availability of new sequence data and a scientific understanding of that information. This gap can be filled through the use of *bioinformatics*. Bioinformatics is an interdisciplinary field bringing together biology, computer science, mathematics, statistics, and information theory to analyze biological data for interpretation and prediction.

For instance, a *gene* is a sequence of DNA (typically 100–5000 symbols in length) that codes for the manufacture of a particular molecule called a *protein*, and proteins carry out actions in cells. Given a newly sequenced genome, one might ask the question “How many genes are in this genome?” In reference to the human genome, such an answer might help us identify differences relative to other animals, how we evolved, and perhaps supply a window of opportunity to

appreciate how *little* we understand about our own genome. But the identification of genes in new sequence information is not a trivial problem. One popular approach is to develop a predictive computer model from a database of known gene sequences and use the resulting model to predict where genes are likely to be in newly generated sequence information. Currently, bioinformatic approaches to this problem range from statistical modeling to machine learning techniques such as artificial neural networks, hidden Markov models, and support vector machines. Indeed, the explosive growth in biological data demands that the most advanced and powerful ideas in machine learning be brought to bear on such problems. Discovery of coding regions in DNA sequences can therefore be viewed as a pattern recognition problem that can be addressed with such techniques as evolutionary computation (see Chapter 9).

However, computer scientists cannot expect biologists to hand over their datasets with an expectation of easily gaining worthwhile results. Knowledge and insight into the application domain can be incorporated into computational analysis in a variety of ways to help develop a more successful approach. The added value of domain-specific knowledge cannot be overestimated, especially in a field that is at the intersection of both biology and computer science. An ideal combination brings together domain experts from biology and computer science with someone capable of bridging both domains. Rarely does one individual have the requisite expertise in both domains. This unification has already paid dividends in terms of our rapidly growing understanding of the human genome, but there are many problems still awaiting the field and many new techniques to be applied.

The development of predictive computer models can be accomplished in many ways. A technique that has generated significant attention for its flexibility, ease of parallelization, and useful performance is *evolutionary computation* (EC). Broadly speaking, EC can be viewed as a paradigm for optimization. For pattern recognition, EC can be used to optimize the parameters or structure (or both) of any type of classifier or predictive model. EC can also be applied to problems in bioinformatics that do not necessarily involve pattern recognition. For example, the protein-folding problem is that of determining the most likely three-dimensional structure of a protein, given only its primary amino-acid sequence. Given a primary sequence that specifies the amino acids of concern, constraints that limit the number of degrees of freedom, and a suitable method for evaluating the quality of a candidate structure, EC can be used to predict with significant accuracy how this amino acid string folds in 3D. Chapters 6 through 8 highlight this particular application of EC.

To summarize, bioinformatics is the interdisciplinary science that seeks to uncover knowledge from a vast quantity of biological data by using computational and informational approaches. The complexity and amount of this data necessi-

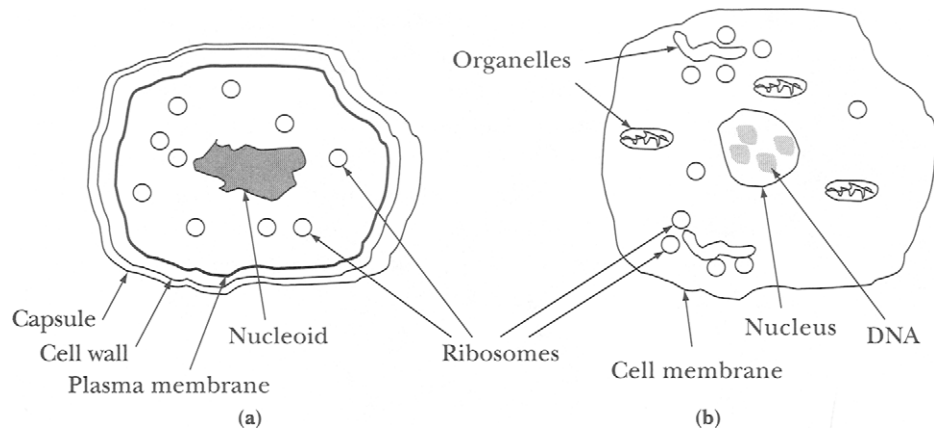
tates close collaboration between biologists and computer scientists. Increased value is added to this equation if the collaborators have at least an elementary knowledge of each other's field. The first two chapters of this book are provided with this goal in mind. Chapter 1 is intended for the computer scientist who requires some additional background material for the biological problems addressed in this book. Chapter 2 introduces the technique of EC to the biologist with only a limited knowledge of this particular field of computer science. Whereas Chapter 1 focuses on the problems, Chapter 2 focuses on the methods of EC that are used to generate useful solutions.

1.2 BIOLOGY—THE SCIENCE OF LIFE

The science of biology attempts to provide an understanding of the nature of all living things at a variety of levels—from molecules to cells, individuals, groups, populations, and ecosystems. The cell, however, is almost universally accepted as the primary *unit* of life: All living things are either composed of a single cell or a collection of them. From this perspective, adult humans can be considered as a collection of 100 trillion cells. At the moment, the major focus of bioinformatics is on problems at the level of molecules and cells, and this book reflects this trend. This is not to say that problems at higher levels are any less important; however, significant amounts of scientific research are currently focused on problems at the lower levels of the hierarchy. As a result, tremendous volumes of data continue to be generated at the lower levels and the use of bioinformatics has followed this trend.

Each cell contains a dynamic environment consisting of molecules, chemical reactions, and a copy of the genome for that organism. Although the DNA is the same in all of our cells, the expression of that code can be very different in different cells and leads to dramatic cellular specialization during development of the individual organism. The precise mechanisms controlling this specialization have only recently begun to be understood. The many chemical reactions in a cell are mainly the result of proteins (*enzymes*). The network of reactions is fed by nutrients absorbed by the cell from its surrounding environment. However certain changes in the environment can cause considerable changes in this network, as discussed later in this chapter.

Historically, life has been divided into *Kingdoms* or *domains* based on similarity of cell morphology. Biologists commonly refer to *prokaryotes* and *eukaryotes*. Prokaryotes are generally classified as single-celled organisms such as bacteria (e.g., *Escherichia coli*). Prokaryotes do not shelter the DNA in their cells within a special membrane, whereas eukaryotes (e.g., fish, amphibians, reptiles, insects,



1.1
 FIGURE 1.1 (a) Structure of a prokaryotic cell. (b) Structure of a eukaryotic cell. In prokaryotes, the DNA is a component of the *nucleoid*, which is a tightly packed giant composite DNA wrapped around certain protein molecules. This DNA/protein complex is not separated from the remainder of the cell. In a eukaryotic cell, the DNA is enclosed within a compartment called the *nucleus*.

birds, mammals, fungi) protect their DNA within the nuclear membrane of each cell (Figure 1.1) (Lewin, 2001).

Comparison of biological sequence information for a wide variety of organisms through bioinformatics has led to the appreciation that there are three major domains of life on Earth: the Eukarya (eukaryotes), the Eubacteria, and the Archaea (Marshall and Schopf, 1996). Archaea and Eubacteria are both prokaryotes and are morphologically similar under a microscope. However, Archaea are known to live in extreme environments (e.g., highly acidic or basic conditions, near deep-sea hydrothermal vents at extreme temperatures and pressures), in which other living things are unable to survive. This generally sets them quite apart from the rest of living things. Our knowledge of this domain of life is reliant on our ability to use bioinformatics to understand how Archaea survive in their extreme environments.

1.3 THE CENTRAL DOGMA OF MOLECULAR BIOLOGY

The DNA in each organism controls the activities in each cell by specifying the synthesis of enzymes and other proteins. To do this, a gene does not build a protein directly, but instead generates a template in the form of a strand of RNA (ribo-

nucleic acid), which in turn codes for protein production. This flow of information in cells was termed the “central dogma of molecular biology” by Francis Crick (Lewin, 2001).

DNA (deoxyribonucleic acid) consists of two long strands, each strand being made of units called phosphates, deoxyribose sugars, and nucleotides (adenine [A], guanine [G], cytosine [C], and thymine [T]) linked in series. For ease of understanding, biologists commonly represent DNA molecules simply by their different nucleotides using the symbols {A, G, C, T}. The DNA in each cell provides the full genetic blueprint for that cell (and in the case of the multicellular eukaryotes, all other cells in the organism). The DNA molecule is a combination of two of these strands running in an antiparallel orientation to form a double helix following the base pairing rules (A pairs with T) and (C pairs with G). Because of these pairing rules (also known as *Chargaff's rules*), the two strands of DNA are complementary: each strand is the structural complement of the other.

The DNA molecules in the cell provide the blueprint for the production of RNA and ultimately for that of proteins. The transfer of information from DNA to specific protein (via RNA) takes place according to a *genetic code*. This code is not universal for all organisms, but there is a standard code used by the majority of organisms, given in Table 1.1. Using this code, the DNA sequence:

AGTCTCGTTACTTCTTCAAAT

is first *transcribed* into an RNA sequence using the nucleotides adenine (A), guanine (G), cytosine (C) and uracil (U), which is used in place of thymine in the RNA strand:

AGUCUCGUUACUUCUCAAU .

In a eukaryote, this RNA sequence is typically exported out of the nucleus to the cytoplasm for *translation* into a protein primary sequence. The RNA is then deciphered as a series of three-letter *codons*:

AGU CUC GUU ACU UCU UCA AAU ,

where each codon corresponds to a particular amino acid specified in Tables 1.1 and 1.2. A special codon called the *start codon* signals the beginning of the translation process; the process ends when one of three *stop codons* is reached. In this example, the protein sequence SLVTFLN would be generated (the amino acid abbreviations used in this sequence are defined in Table 1.2). Note that during this process, information has been transferred from the DNA (the information-storage molecule) to RNA (information-transfer molecule) to a specific protein (a functional, noncoding product). Each of these levels has a maze of regulatory elements and interactions that have only been understood for 50 years. Our

First base is A

Second base	Third base			
	A	C	G	U
A	AAA—lys	AAC—ans	AAG—lys	AAU—asn
C	ACA—thr	ACC—thr	ACG—thr	ACU—thr
G	AGA—arg	AGC—ser	AGG—arg	AGU—ser
U	AUA—ile	AUC—ile	AUG—met	AUU—ile

First base is C

Second base	Third base			
	A	C	G	U
A	CAA—gln	CAC—his	CAG—gln	CAU—his
C	CCA—pro	CCC—pro	CCG—pro	CCU—pro
G	CGA—arg	CGC—arg	CGG—arg	CGU—arg
U	CUA—leu	CUC—leu	CUG—leu	CUU—leu

First base is G

Second base	Third base			
	A	C	G	U
A	GAA—glu	GAC—asp	GAG—glu	GAU—asp
C	GCA—ala	GCC—ala	GCG—ala	GCU—ala
G	GGA—gly	GGC—gly	GGG—gly	GGU—gly
U	GUA—val	GUC—val	GUG—val	GUU—val

First base is U


Third base	Third base			
	A	C	G	U
A	UAA—STOP	UAC—tyr	UAG—STOP	UAU—tyr
C	UCA—leu	UCC—phe	UCG—leu	UCU—phe
G	UGA—STOP	UGC—cys	UGG—trp	UGU—cys
U	UUA—leu	UUC—phe	UUG—leu	UUU—phe

1.1

TABLE

The genetic code: how each RNA codon corresponds to an amino acid. For each possible codon, the three-letter code of the corresponding amino acid is given.

Amino acid	Three-letter	One-letter	Size	Charge
Alanine	ala	A	Small	Neutral
Arginine	arg	R	Large	+
Asparagine	asn	N	Medium	Neutral
Aspartic acid	asp	D	Medium	-
Cysteine	cys	C	Small	Neutral
Glutamic acid	glu	E	Large	-
Glutamine	gln	Q	Large	Neutral
Glycine	gly	G	Small	Neutral
Histidine	his	H	Large	+
Isoleucine	ile	I	Medium	Neutral
Leucine	leu	L	Medium	Neutral
Lysine	lys	K	Large	+
Methionine	met	M	Large	Neutral
Phenylalanine	phe	F	Medium	Neutral
Proline	pro	P	Medium	Neutral
Serine	ser	S	Small	Neutral
Threonine	thr	T	Small	Neutral
Tryptophan	trp	W	Large	Neutral
Tyrosine	tyr	Y	Medium	Neutral
Valine	val	V	Small	Neutral

1.2

TABLE The 20 amino acids. Full names, standard three-letter abbreviations, and standard one-letter abbreviations (as appear in primary sequence data) are given, along with size (large, medium, or small) and electrostatic charge (+, -, or neutral).

knowledge of these interactions continues to expand, raising hopes in medicine that the interactions can be controlled at the cellular level in individual patients through molecular manipulation.

To gain better insight into the nature of this information flow, we must first produce the *genomes* (the full complement of genetic material) for many organisms. The genes must then be mapped and our knowledge of the genetic code used to predict RNA and protein products, and perhaps even to infer a putative function for the newly discovered gene based on the similarity of its products to those of known genes. For this effort, many computational tools are required, including sequence and structure alignment, pattern recognition, feature selection, methods to correctly predict the folding of RNA and proteins, and even inference of properties about the amazing web of interactions within a cell. The chapters of this book are arranged to focus on particular applications of EC to each of these

problem domains. The reader is directed to the many other books on bioinformatics to gain a better understanding of other methods that can be used (Durbin et al., 1998; Baldi and Brunak, 2001; Mount, 2001). We will continue below with a closer examination of the main flow of information in a cell so that the reader can form a better appreciation for the scope of the problems being investigated.

1.3.1 Anatomy of a DNA Sequence

As mentioned previously, DNA can be represented as a series of symbols from the set {A, G, C, T}. Biologists refer to each symbol on one strand as a *base* or *nucleotide* (nt) position and refer to complementary positions across both strands of the DNA as a *base pair* (bp). When attempting to discover genes in a previously unannotated DNA sequence, the researcher looks for any segment that appears between a start and stop codon and so could potentially encode a protein: such a segment is a candidate gene. But we cannot be certain of the primary protein sequence that the putative gene encodes. There may be uncertainty regarding the position of the correct starting nucleotide and consequent uncertainty regarding the appropriate *reading frame*. This concept is illustrated in Figure 1.2, which shows three possible reading frames in a gene sequence whose starting position is unknown. Indeed, there are several known cases, especially in the often compact genomes of viruses or bacteria, in which overlapping genes result in multiple reading frames within the same segment of DNA.

A number of regulatory elements (e.g., promoters, enhancers) typically flank actual gene sequences. The nucleotides that make up these elements can control the rate of gene expression; the presence of these elements can also be used by researchers to help identify actual gene sequences. In eukaryotes, the coding regions—that is, those regions that code for proteins—make up only a small fraction of the total DNA; this fraction varies considerably across organisms. Even within a single gene, there may be regions of DNA that give rise to a functional product (*exons*) and regions that do not give rise to any functional product (*introns*). Prokaryotes have much less noncoding DNA than do eukaryotes and have their own system of gene regulation and morphology. The many differences at this level between and within eukaryotes and prokaryotes are well beyond the scope of this chapter, but the reader should be aware that models of eukaryotic systems may not transfer readily to prokaryotes and vice versa.

Within many eukaryotes, the remaining noncoding DNA is typified by an assortment of multiple contiguous repeats of simple short sequences (*microsatellites*), former genes (*pseudogenes*), and segments of DNA that are capable of self-splicing and/or copying to other locations in the genome (*transposons*). For instance, the Alu sequence, a transposon of 280 bp in length, has been actively

DNA	A	G	T	C	T	C	G	T	T	A	C	T	T	C	T	C	A	A	A	T
Frame 1	S			L			V			T			F			E				
Frame 2	V			L			L			L			L			K				
Frame 3	F			R			Y			F			L			N				

1.2

 FIGURE

Three possible reading frames of a DNA sequence. The top row shows a DNA sequence. Its interpretation as a sequence of three-letter codons depends upon the starting point. Frames 1, 2, and 3 begin at the first, second, and third nucleotide positions, respectively. For each frame, the translation into an amino acid sequence is given, using Table 1.2 to determine the amino acid (where T is replaced by U) and using Table 1.1 to recover the standard single-letter code for that amino acid.

spreading in human genomes for much of our evolutionary history. These and similar transposons make up the category called SINES (short interspersed nuclear elements). There are also longer transposon elements called LINES. A type of LINE called L1, for example, is a sequence of 7 kilobases (kb), which comprises almost 15% of the entire human genome. The various repetitive elements clearly represent a substantial part of our genome. There is great interest and debate about their roles and effects in our evolutionary history and our current cellular processes. The remaining noncoding or *intergenic* DNA is largely uncharacterized; however, there are many reasons to refrain from dismissing it as “junk.”

1.3.2 Transcription and Translation

As mentioned previously, genes in DNA are first transcribed into RNA, and then translated into protein. The nucleic acids DNA and RNA share a very similar language of symbols (nucleotides) and therefore the change of DNA to RNA is analogous to transcribing a book from one copy to the next using a very similar language. But the transfer of the information from nucleic acids to proteins requires translation across two very different languages through the use of the genetic code.

To transcribe DNA into RNA, a *pre-initiation complex* is assembled around the promoter region just upstream of the gene that is to be expressed. This complex of proteins attached to the DNA attracts a very special protein called RNA polymerase, which causes the strands of the DNA to separate in the region near the start of the gene; the RNA polymerase enzyme then binds to one of the strands. The RNA polymerase subsequently moves along the gene from left to right (biologists refer to this as the *5' to 3' direction* because of the biochemistry of nucleic acids), leading to the production of an RNA transcript. The RNA is made by incor-

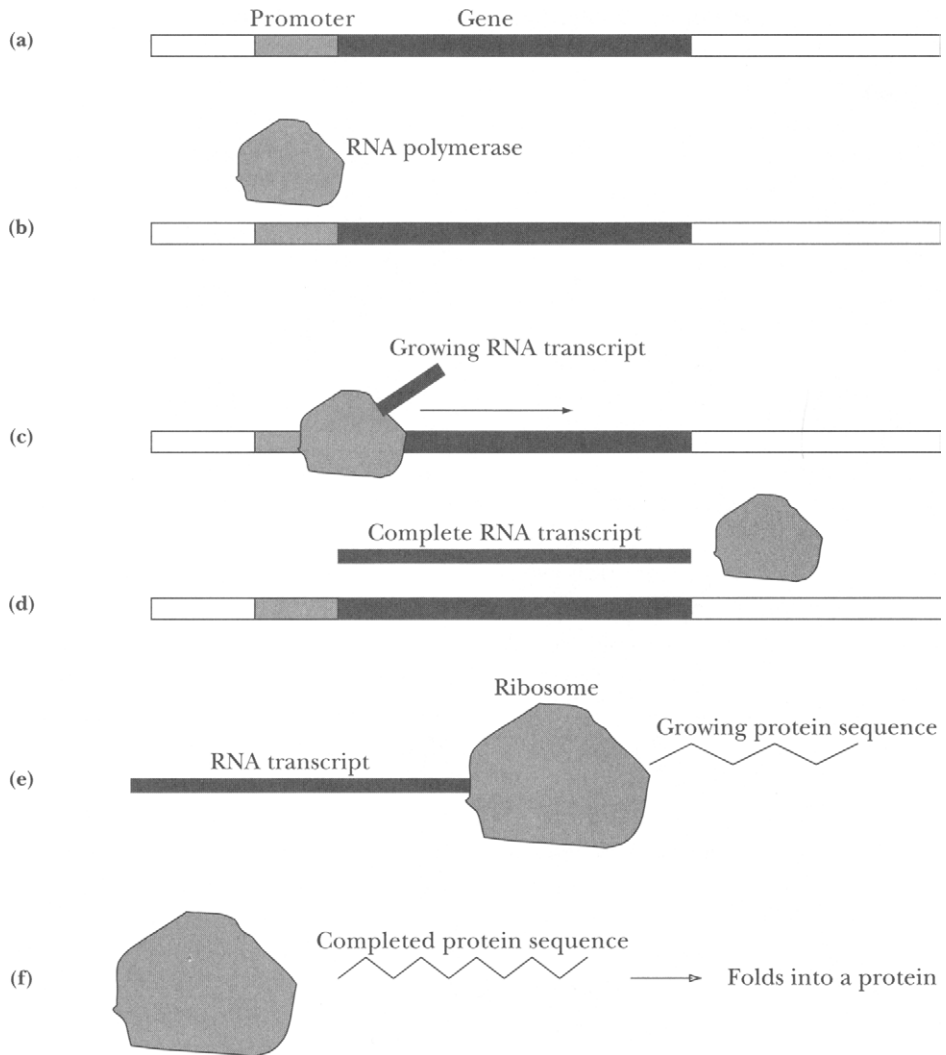
porating the correct complementary nucleotide from the set {A, T, G, U} for each nucleotide represented in the gene using the previously mentioned pairing rules: A pairs with U; G with C. This process finishes when the RNA polymerase encounters a termination signal and the resulting *messenger RNA* (mRNA) is left free-floating in the cell for export or further processing. In eukaryotic cells, the mRNA is exported out of the nucleus and into the cell's cytoplasm. There it encounters a structure in the cell called a *ribosome*, which begins the process of decoding the information in the mRNA and translating this information into the language of amino acids and proteins (Figures 1.2 and 1.3). A very large assortment of RNA and protein sequences assist in this process, whose description is beyond the scope of this chapter.

The resulting protein is quickly able to assume its native three-dimensional form (called its *conformation*) and is then ready for immediate action in the cell. The process of gene expression outlined above occurs constantly in all living cells, including those of your own body, where, depending on the cellular environment, particular genes may be switched on or off and thus expressed at varying rates over time. Next we focus on the protein products of these genes and return at the end of the chapter to discuss gene networks and their expression.

1.3.3 Proteins

Proteins make up most of an organism's biomass, and play a key role in its metabolic and other cellular and bodily processes. Each protein has a distinct three-dimensional shape and is typically composed of between 1000 and 50,000 atoms. The amino-acid sequence of the protein is the functional portion of information flow in a cell. For example, the protein collagen makes up about 25% of all of the protein in a human body. A single collagen molecule is much like a strong cable, and groups of them self-organize to provide support for our skin, internal organs, bones, teeth, and muscles. Of the tens of thousands of proteins that exist in human cells, relatively few have readily discernible effects at the level of the individual: Many proteins determine our characteristics in subtle ways. For example, a large collection of proteins called *immunoglobins* are key to the functioning of our immune systems, and hence control how well we fight infections of various types. The *histone* proteins exist in every eukaryotic cell and are crucial to DNA structural organization, and therefore crucial to how the DNA blueprint is interpreted. The interplay of the many proteins in a cell is an amazing web of multiple interactions that affect the behavior of every cell.

Despite the huge variation in structure and function, all proteins share a common structural vocabulary. The essence of a protein's structure is illustrated in Figure 1.4. All proteins have a backbone of carbon (C) and nitrogen (N) atoms



1.3
 FIGURE

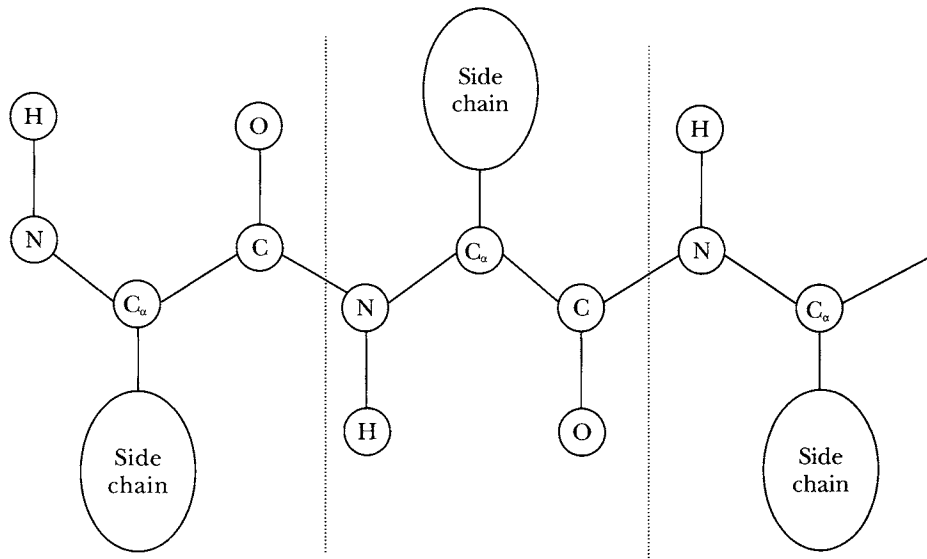
How a protein is made from genetic information. (a) Section of DNA containing a gene, preceded by a promoter region. (b) An RNA polymerase complex (a collection of specialized proteins) is attracted to the DNA around the promoter and begins to interact with the DNA at the start of the gene. (c) The RNA polymerase moves along from left to right, gradually building an RNA transcript of the gene sequence. (d) The transcript is complete, and it now separates from both the RNA polymerase and the DNA. It floats around the cell until it encounters a ribosome. (e) The ribosome uses it to automatically manufacture a protein molecule. (f) The process is complete, and the protein folds into its native form and goes about its business in the cell.

linked in sequence. The N–C–C motif is referred to as a *peptide unit* (the bond between the carbon and nitrogen atoms is the *peptide bond*); sometimes proteins are referred to as *polypeptides*. Oxygen (O) and hydrogen (H) atoms are attached to the nitrogen and second carbon atoms in each peptide unit as shown in the figure, but the central carbon atom in each unit (referred to as C_α) is a hook for any of 20 possible amino acid attachments called *side chains* or *residues* (Figure 1.4).

A protein can therefore be specified by its sequence of amino acids (also known in this context as side chains or residues). Biologists represent each of the 20 amino acids by a one-letter symbol, as shown in Table 1.2. The sequence PNAYYA, for example, specifies a protein whose backbone is made up of six peptide units; the first amino acid is proline, the second is asparagine, and so forth. In this way, any sequence of amino acids fully specifies the so-called *primary structure* of a protein molecule. Given the backbone structure and the complete chemical structure of each amino acid, the protein is then characterized—at a very basic level—by the chemical bonds between its constituent atoms.

The structural details of a protein molecule are intimately related to its functional role in an organism and, in pathological cases, determine how the protein fails. For example, the protein hemoglobin is directly involved in oxygen transport within the bloodstream. The genes that encode the two main constituents of hemoglobin (α -hemoglobin and β -hemoglobin) specify particular primary sequences each of some 100 amino-acid residues in length. Some humans have a variant with a single change in the 15th bp of the β -hemoglobin gene. This alteration at the level of the gene results in a change of the hemoglobin primary amino acid sequence. The amino acid change results in an altered protein structure, which, in this case, is responsible for the condition known as sickle-cell anemia. Molecules of this variant form, when oxygen-free, are prone to link together to form long, rigid chains. These chains distort the cells containing them into a relatively rigid sickle-like shape, which can then get lodged within small blood vessels, causing considerable pain. However, individuals with this particular affliction are less likely to contract malaria. The three-dimensional structure of a protein is thus critical to its proper function and changes to this shape can have deleterious, beneficial, or neutral effects on the individual.

Knowledge of the primary structure alone is not sufficient to determine the three-dimensional shape of a protein (Branden and Tooze, 1998; see Figure 1.5). The side chains of amino acids jut out at regular intervals, supporting and stabilizing the roughly globular structure. It is immensely difficult to infer the precise three-dimensional structure of an arbitrary protein from its primary sequence. This is in fact one of the more pressing issues in bioinformatics, and three chapters in this volume (Chapters 6, 7, and 8) address this problem.



1.4
 FIGURE

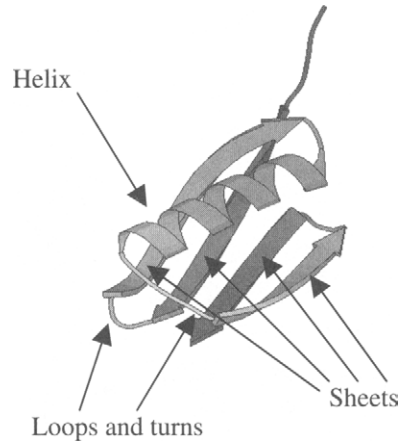
The essentials of protein structure. A snippet of the polypeptide backbone of a protein molecule is shown. It consists of repeated triplets of the motif N–C–C_α, which assume alternating orientations. One of 20 side-chain residues is attached to each central carbon atom (the C_α atom) of the repeated motif.


1.4 GENE NETWORKS

1.4

Genes are expressed at varying rates throughout the life of a cell. The expression rates of different genes in the same genome may vary continuously from 0 to about 100,000 proteins per second. Several factors influence the expression level of each gene, including the expression patterns of other genes. It is useful to visualize the expression process in any particular cell as a dynamic network, the nodes of which are genes, and the links between genes having real-valued (positive or negative) weights that model the degree to which the expression of one gene affects the expression of another. The field of systems biology attempts to provide a better understanding of these relationships between different gene expressions (Kitano, 2001).

The links in this dynamic network tend to model indirect effects. For example, gene 1 might encode a protein that binds tightly to the DNA sequence CTACTG. This will interfere with the expression of any gene that contains this substring. Promoters in close proximity to this substring might also be affected. Another gene (call it gene 2), however, may find its expression increased by the product of gene 1. The target binding sequence may occur somewhere near gene 2, and



1.5

 FIGURE

Streptococcal protein G (adapted from Figure 6.7), depicted in the way structural biologists commonly view protein structures, as constituted from regions of distinct secondary structure. This protein's backbone folds into four distinct β -sheet regions and an α -helix. The β -sheet regions are, as is typical, largely parallel.

its influence on the DNA in that region may affect the DNA folding around gene 2 such that there is an increase in its level of transcription. The vector of gene expression levels in a given cell is, in some sense, the fingerprint of that cell. This fingerprint varies over time. Each cell contains the same genetic material as all other cells in the same organism, but the cell's history and its current environment profoundly affect its gene expression networks. *Microarrays* (or *DNA chips*) can be used to interrogate the expression levels of many thousands of genes at a given time in a given environmental situation. The environment can then be altered such that a new pattern of expression emerges. This technology has only recently become available and promises to help with our understanding of the many interactions in a genome. Part V of this book discusses methods to analyze these and other interactions.

1.5 SEQUENCE ALIGNMENT



Following the identification of a new gene and its protein product, the biologist typically looks for similar sequences in previously collected data. Similarities might lead to clues regarding the evolutionary history of the gene or the function of the protein. To look for similarities, biologists use several search engines available online (see the Appendix) that use sequence and structural information to find sim-

ilar sequences in sequence databases. These search engines compare the query and database sequences through a series of *alignments*. Consider two sequences, one newly discovered and the other previously discovered in a database:

ATCTCTGGCA

TACTCGCA

An alignment of the two sequences might yield the following result:

ATCTCTGGCA

-T-ACTCGCA

In this case, gaps (-) were inserted in locations that are missing between the two sequences. Also, the two sequences do not exactly correspond at all the places which do not involve gaps. Such an alignment is meant to convey a hypothesis about the evolutionary relationship between the two sequences and is based on the assumption that they both evolved from a common ancestor sequence. In this case the gaps indicate either that the longer sequence has arisen from insertions during its evolutionary descent from the common ancestor, or that the shorter sequence has arisen from deletions, or that perhaps there was some mixture of the two. Aligned loci that do not correspond simply indicate different mutations at the same position along the two lines of descent.

Alignments are made with pairs (or larger sets) of sequences, which may provide additional information for important nucleotide positions that have remained invariant over time or positions that have been free to vary throughout evolutionary history. When the length and number of sequences increases, a far greater collection of possible alignments is available, and we must settle for what seems to be the most *likely* best solution, without even having a convincing measure of likelihood. The standard approach is to calculate the *cost* of an alignment, which reflects the presumed chances of the collection of mutation events expressed. Chapters 4 and 5 review the application of sequence alignment to obtain information about both DNA and protein sequences.

1.6 CONCLUDING REMARKS

The field of bioinformatics is full of exciting new challenges for the computer scientist. The various applications in this book all fall within the scope of processes that occur at the molecular and cellular levels. Although modern molecular biology appears to be focused on a variety of genome projects, generating ever-increasing amounts of DNA-sequence information, by itself this information is

only a stepping stone to the answers to challenging questions. Bioinformatics provides the means to obtain these answers.

However, many of the problems faced in bioinformatics are daunting in their size and scope. The number of potential solutions to a given problem (such as protein folding) can be so large that it precludes exhaustive search. Given this dilemma, the biologist generally tends to simplify the problem to generate a space of possible solutions that can be searched with exhaustive or gradient-descent methods. Such a simplification generally leads to the right answer to the wrong problem. What is required is a method to search large computational spaces in an efficient manner without simplification. One such method, EC, has already been proven on a wide range of engineering problems as an efficient technique for searching large spaces while examining only a fraction of the possible solutions. Such techniques are currently being brought to bear on problems in bioinformatics.

REFERENCES

- Baldi, P., and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, Mass.
- Branden, C., and Tooze, J. (1998). *Introduction to Protein Structure*. Second edition. Garland Publishing Inc., New York.
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, U.K.
- Kitano, H. (2001). *Foundations of Systems Biology*. MIT Press, Cambridge, Mass.
- Lewin, B. (2001). *Genes VII*. Oxford University Press, New York.
- Marshall, C. R., and Schopf, J. W. (1996). *Evolution and the Molecular Revolution*. Jones and Bartlett Publishers International, London.
- Mount, D. W. (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.