1 Introduction
Working with the molecules of life in the computer

[Jim Kent] embarked on a four-week programming marathon, icing his wrists at night to prevent them from seizing up as he churned out computer code by day. His deadline was June 26, when the completion of the rough draft was to be announced.

(*DNA: The Secret of Life*, describing Jim Kent's efforts in the human genome sequencing project in 2000; Watson and Berry, 2003)

It was a somewhat historic event when President Bill Clinton announced, on 26 June 2000, the completion of the first survey of the entire human genome. We were able for the first time to read all three billion letters of the human genetic make-up. This information was the ground-breaking result of the Human Genome Project. The success of this project relied on advanced technology, such as a number of experimental molecular biology methods. However, it also required a significant contribution from more theoretical disciplines such as computer science. Thus, in the final phase of the project, numerous pieces of information like those in a giant jigsaw puzzle needed to be appropriately combined. This step was critically dependent on programming efforts. Adding further tension to the programming exercises was the fact that a private company, Celera, was competing with the academic Human Genome Project. This competition was sometimes referred to as 'the Genome War' (Shreeve, 2004). While computationally talented people like Jim Kent 'churned out computer code', other gifted bioinformaticians, such as Gene Myers at Celera, worked on related jigsaw-puzzle problems. Ideally, scientists should not war against each other; however, there was an important conclusion from these projects in which important genetic information was generated: *computing is an essential part of biological research*.

This book will present a number of examples from the world of biology and biomedicine, where programming can significantly help us out. This first chapter will discuss very basic biology concepts and will introduce fundamentals of genetic information at the molecular level. We will then go on to examine some simple programming code to represent and process that genetic information. The code

will not be the advanced work of the programmers mentioned above, but it is a start.

Life on Earth and evolution

What are the characteristics of life on Earth? We observe a wealth of living organisms in a large number of ecological niches; on land and in fresh- and seawater; in the soil and deep beneath the Earth's surface. There is a variety of organism types, all the way from small viruses and bacteria to plants and animals.[1] As we compare many of the life forms, they seem very different. For example, an oak tree, a cow and a slime mould do not look that similar to the human eye. But in a sense these differences are somewhat superficial. In fact, a closer look at organisms at a microscopic or molecular level reveals that they are strikingly similar. For instance, all living forms (with the exception of viruses) are built from cells. All cells have certain characteristics in common. For instance, each cell is surrounded by a membrane. Passage of molecules across the membrane is rigorously controlled, and as a result the environment within a cell is very different from that of the outside world. All kinds of cells have a general biochemical machinery in common. Thus, the basic metabolism of living cells is very similar in all kinds of organisms. As one example, a large majority of cells have the capacity to extract energy by oxidation of carbohydrates. Yet another principle shared by all organisms is a basic system for storage and processing of genetic information. That information is stored in the form of a DNA molecule, and the genetic message is a long sequence of chemical units referred to as *bases* in DNA. Furthermore, there is a universal system for production of RNA and protein molecules as specified by this genetic information, as will be explained in some detail later in this chapter.

Why are species similar at the molecular level? The simple answer is one major principle of life: species are related by *evolution*. Evolution is a major key to the understanding of life. Or, as expressed by Theodosius Dobzhansky in his famous quotation: 'Nothing in biology makes sense except in the light of evolution', where he argues against creationism and intelligent design (Dobzhansky, 1964).

What are the basic concepts of evolution? One key element is that all species on our planet are related and have a common ancestor. Our planet is now about 4500 million years old; the first living forms appeared when it was about

[1] To be more strict about this, there are three major kingdoms of life: eubacteria, archaea and eukaryotes. The eubacteria and archea are single-cell organisms, whereas eukaryotes typically are multicellular organisms (e.g. plants and animals). Eukaryotes have a more organized intracellular structure, with a membrane-surrounded nucleus and organelles such as mitochondria. In addition to the organisms of the three kingdoms of life, there are viruses. These are not free-living organisms as they are dependent on their host organism for propagation.

1000 million years old. It is believed that life came about as a result of favourable conditions, such as the presence of water and other critical compounds, together with a suitable temperature range. Early life forms were very primitive compared to many of the living organisms that we observe today. Eventually, further evolution gave rise to diversification and new, more 'complex' species. The total number of species on our planet today is not known exactly, but exceeds two million. Although a systematic inventory of species was started by Carl von Linneaus and others in the eighteenth century, new species are still being identified today.

For evolution to proceed there must be a certain degree of change in the genetic message as it is being transmitted between generations. Such changes are referred to as *mutations*. Another key element to understanding evolution is the process of *selection*, or 'survival of the fittest', as was one early wording. This selection determines what mutations will stay in a population. To put it simply, the evolution of species is determined by factors such as environmental conditions and by competition between different individuals and species. For instance, bacteria that are able to adapt rapidly to a change in their environment may be more likely to survive compared to bacteria that do not have this property. As an example from more recent evolution, humans with enhanced brain functions presumably had a selective advantage compared to primates that did not.

What is the scientific evidence of evolution? What is, for instance, the evidence that species such as man developed from more 'primitive' species? We will not go into detail about this here, but Charles Darwin accumulated an overwhelming amount of data in support of his theory of evolution. Today, we also have detailed information about the genetic information of many species and are able to monitor changes during the course of evolution in remarkable detail. With this recent information there is even more overwhelming support of Darwinian theory.

There is no doubt that the concept of evolution is fundamental in genomics and bioinformatics, and many of the issues brought up in this book are directly or indirectly related to this concept. For instance, consider a common situation where biologists have identified a specific sequence of bases in a DNA molecule present in humans and they want to attribute this sequence to a biological function. This sequence, or a similar one, might be present in another species – let's say mice. This is because sequences of bases, just as species, are related by evolution. We may also find that the function of the sequence in mice has been elucidated. In such a case we may infer a function of the human sequence, based on what we have collected by studying mice. We will see a few cases of this sort later in the book (Chapters 11–13). These are examples of situations where we are making use of the concept of evolution without necessarily being

interested in evolution as such. Then again, we may actually be interested in evolutionary events; topics like these are dealt with in Chapters 8–10.

In summary, evolution is fundamental in bioinformatics.[2] If we are to understand more about the actual mechanisms of evolution, we need to know more about how mutations occur and what they mean in terms of biological information. For this, in turn, we need to know more about the molecular genetic machinery.

The machinery of genetic information: more about DNA

We have already seen that the DNA (deoxyribonucleic acid) molecule carries genetic information, i.e. the information that will be transmitted from one generation to the next. DNA is copied in a process called *replication*. The information in DNA ultimately guides the production of different proteins. First in this process, an RNA (ribonucleic acid) copy is made in a process known as *transcription*. Non-informational portions of the primary product of transcription (introns) are removed in a process called *splicing*. The RNA then directs the *synthesis of protein* (a process also called *translation*). See Fig. 1.1 for an overview of transcription, splicing and translation.

Each protein has a specific function in the cell. In a human cell there are in the order of 21 000 different proteins. These come in a variety of functional classes. One important category is *enzymes*. These are biocatalysts responsible for catalysing thousands of chemical reactions within the cell, reactions that would not be possible without enzymes. Other important functions of proteins are to act in the transport of molecules and to act as a 'skeleton', maintaining a specific cell architecture. Proteins are the molecules that directly determine all properties of a living cell. DNA does not do anything like that; it simply carries the genetic information that specifies which proteins are to be made.

We will now examine how the flow of information occurs from the DNA to the protein in more detail. First, we have to understand the basic chemistry of the DNA molecule. It is a very long polymer with repeating sugar and phosphate units. Attached to the sugars are bases. The bases are the variable units in DNA, and it is the sequence of bases that constitutes the actual genetic information.

A DNA molecule has a distinct chemical polarity. One end of the DNA is referred to as the 5′ end, because the hydroxyl group in that end is attached to

[2] If, for some reason, you prefer not to believe in this basic Darwinian concept (and there are in fact people that tend to support other ideas about the development of species) genomics and bioinformatics will not make sense to you at all and there is no point in reading further.

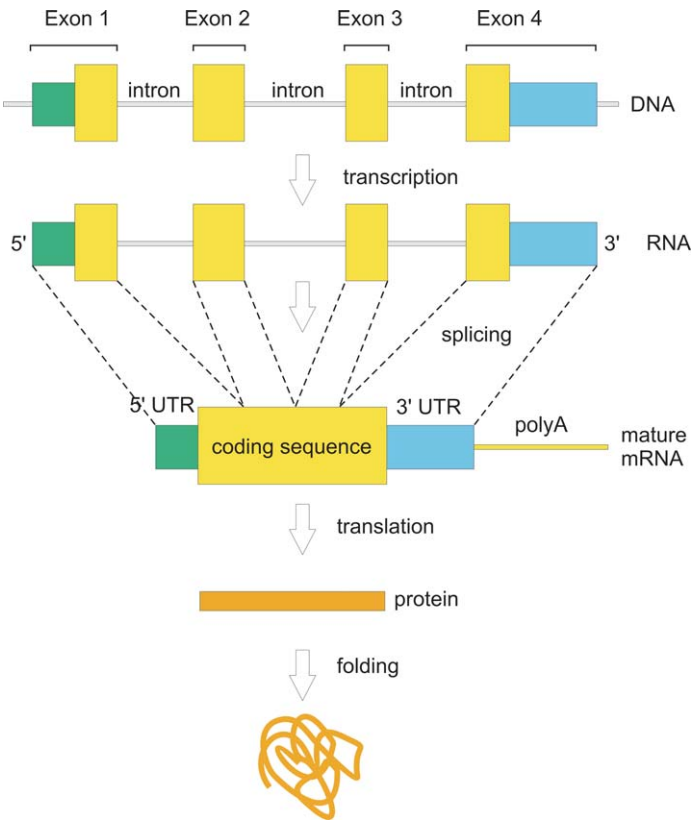

Fig. 1.1 *Flow of genetic information in a eukaryotic cell*. A primary RNA transcript is first produced in the process of transcription. This transcript is subject to splicing, where the exons are joined. The mature mRNA contains the exons as well as a polyA tail at its 3′ end. This RNA has a 5′ untranslated region (5′ UTR, green), a coding sequence (yellow) and a 3′ UTR (blue). The coding sequence determines the sequence of amino acids that are incorporated into the protein during translation. Finally, the amino acid sequence of the protein determines the folding of the protein into a three-dimensional structure. This structure has specific biological properties. Cells from all kingdoms of life are characterized by this flow of genetic information, although bacterial genes do not have introns and are therefore not subject to splicing.

a carbon of the sugar and that carbon has the number five, according to the numbering scheme for the sugar structure. Similarly, the other end of the DNA molecule is referred to as 3′. The polarity of DNA means that the sequence of bases read from one end of the molecule is not equivalent to the sequence of bases read from the other end.

The bases are referred to as adenine, guanine, cytosine and thymine, abbreviated as A, G, C and T (Fig. 1.2). The A and G bases have a double ring structure and are referred to as *purines*, whereas C and T have a simple ring structure

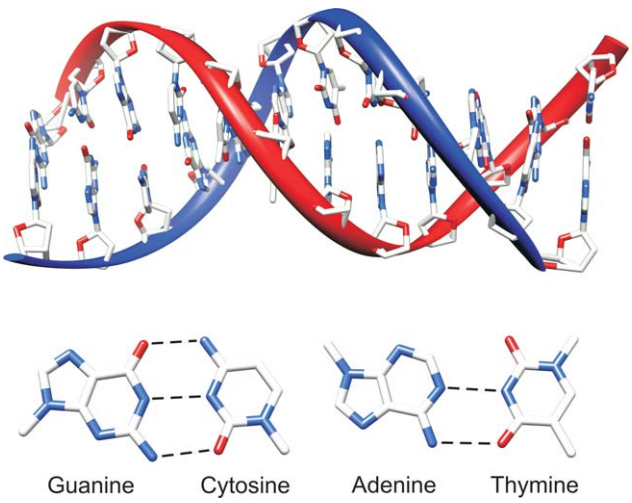

Fig. 1.2 *Three-dimensional structure of the DNA double helix*. The sugar–phosphate backbones of the two strands are shown as red and blue ribbons. Base-pair interactions between guanine and cytosine – with three hydrogen bonds – and between adenine and thymine – with two hydrogen bonds – are shown in the lower panel. Structures are from the Protein Data Bank entry 7BNA. The sequence of both DNA strands is CGCGAATTCGCG. The figure was produced with the UCSF Chimera software (Pettersen *et al.*, 2004).

and are called *pyrimidines*. The sugar–phosphate–base unit in DNA is referred to as a *nucleotide*. From an information perspective a nucleotide is equivalent to a base because the base is the only variable unit in DNA.

We now arrive at one basic concept of sequence bioinformatics. The genetic message in DNA can be represented as a simple string of the letters A, G, C and T. For example:

```
5′ - AGGACACGACGACTATTGG - 3′
```

Normally you see the sequence written in the direction shown here, i.e. with the 5′ end to the left and the 3′ end to the right.

DNA may sometimes occur as a *single-stranded* molecule (as in certain viruses), but normally we find DNA as a *double-stranded* unit. It has a double helical structure with two antiparallel strands. In this case the word 'antiparallel' means that the two strands run in 'parallel', but have opposite polarity.

The structure of double-stranded DNA was elucidated by Francis Crick and James Watson and was presented in a famous paper in *Nature* in 1953, an *annus mirabilis* for science[3] (Watson and Crick, 1953) (Fig. 1.2). The paper was a landmark in molecular biology, and in 1962 the two authors received the Nobel Prize in Physiology or Medicine for their discovery. The double-stranded DNA is held together by pairing between bases. Chemical bonds known as hydrogen bonds are formed between bases and the pairing is always such that adenine pairs with thymine and cytosine pairs with guanine. This distinct pairing is referred to as *complementarity*, and one strand of DNA is said to be the complement of the other. The double-stranded DNA may thus be represented as two strings, like this:

[3] http://www.nature.com/nature/dna50/archive.html.

```
5′ - AGGACACGACGACTATTGG - 3′
3′ - TCCTGTGCTGCTGATAACC - 5′
```

Note that the two strands have opposite polarity. As realized early by Watson and Crick, the base complementarity is the basis for replication of DNA. In this process one strand of DNA serves as the template for the synthesis of the other.

When sequences of DNA are deposited in databases, there is no need to put both strands there, as one of them may easily be inferred from the other. Similarly, only one strand is provided as input to many computer programs for sequence analysis, although the other strand is implicit. There is a somewhat confusing terminology as to the two strands of DNA (the reference strand and its complement), where you can encounter expressions such as plus/minus, forward/reverse, top/bottom, sense/antisense or even Watson/Crick.

The fact that the information in DNA may be represented as a long string of characters is of fundamental importance in sequence bioinformatics. It is actually quite surprising how useful the string representation is, as the DNA molecule has a specific three-dimensional structure, and as such is far from one-dimensional.

Genes and genomes

The example above shows a relatively short piece of DNA sequence. In reality, DNA molecules are very long. For instance, the genetic material of humans is distributed among *chromosomes*. There are 22 human chromosomes in addition to the X and Y chromosomes. Each chromosome is, in essence, a very long DNA molecule. The typical length of a human chromosome is about 100 million base pairs. We refer to the base or nucleotide sequence of all chromosomes of an organism as the *genome* of that species. Thus, when we say 'the human genome' we should think of the genetic information in *all* the human chromosomes. Genomes and the biological signals they contain are studied in the science referred to as *genomics*.

More than ten years ago the complete base sequence of all human chromosomes, more than three billion bases, was determined. This was indeed a significant advance in biological research (Lander *et al.*, 2001; Venter *et al.*, 2001). Among many different important medical applications, it helps us to better understand all diseases with a genetic background. Knowing the complete sequence of the human genome, we come to the obvious question: what exactly is the information carried within the genome? What biological signals are found as we examine a human chromosome from one end to the other? To begin with, one important category of elements are *genes*. A gene is a portion of the DNA molecule that contains all the information for production of a protein. The length of human genes is highly variable; they can be as small as a few

thousand and as large as one million bases. Some genes are specified by one of the strands of DNA, and others are specified by the other (complementary) strand.

Genes at work: transcription and translation

What matters to the physiology of a cell are the processes in which DNA serves as a template for production of RNA and protein products. The first step in the flow of genetic information is transcription, in which the information in DNA is copied into an RNA molecule. RNA is a nucleic acid, in the same way as DNA. However, RNA has a sugar unit that is ribose, instead of deoxyribose. In addition, RNA contains the base uracil (abbreviated as U) instead of thymine (T). Like DNA, RNA has 5′ to 3′ polarity.

One of the strands in DNA serves as the template for the synthesis of RNA. The RNA produced is, unlike DNA, single-stranded. Note that the sequence of the RNA produced is complementary to one of the DNA strands and it is equivalent in sequence (although U replaces T) to the other DNA strand (Fig. 1.3).

Transcription takes place with the help of the enzyme *RNA polymerase*. The biochemical machinery of transcription is to some extent similar to DNA replication, because in both of these processes a new strand of nucleic acid is formed by copying information from a template strand of DNA.

The different RNAs produced by transcription have different functions in the cell. However, a very important functional class of RNAs is *messenger RNA* (*mRNA*). These RNAs contain information for the production of proteins. Proteins are large polymers like DNA and RNA, but the building blocks are amino acids rather than nucleotides. There are 20 different amino acids. These are presented along with some of their properties in Table 1.1.

During protein synthesis the sequence of bases in RNA guides the incorporation of amino acids into proteins. Protein synthesis is also named *translation* because a 'translation' occurs from the language of nucleic acids (DNA/RNA) to the language of proteins. Information contained in a protein may be represented as a string of letters, the same as it can be for RNA and DNA. As there are 20 different amino acids, the protein alphabet contains 20 characters (Table 1.1), compared to four in the nucleic acid alphabet (A, T, C and G in the case of DNA).

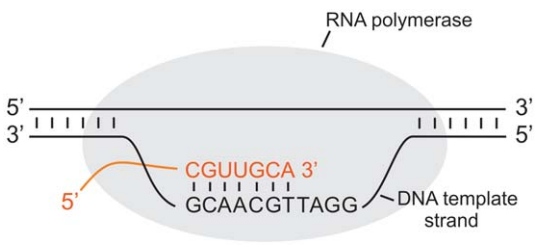

Fig. 1.3 *Transcription*. Double-helical DNA is unwound, allowing RNA polymerase to synthesize RNA in 5′ to 3′ direction (in orange) using one of the strands of DNA as a template. The RNA polymerase moves from left to right.

A set of well-defined rules determine the relationship between the RNA sequence and the sequence of amino acids in proteins. A sequence of three bases, known as a *codon*, specifies a distinct amino acid. The *genetic code* is a table showing

Table 1.1 *The amino acids*

Amino acid	Physicochemical property	Three-letter abbreviation	One-letter abbreviation
Alanine	Non-polar	Ala	A
Glycine		Gly	G
Cysteine	Non-polar, sulfhydryl group	Cys	C
Isoleucine	Non-polar, aliphatic	Ile	I
Leucine		Leu	L
Methionine		Met	M
Valine		Val	V
Proline	Aliphatic, side chain joined to both α carbon and amino group	Pro	P
Serine	Aliphatic – hydroxyl group	Ser	S
Threonine		Thr	T
Phenylalanine	Aromatic	Phe	F
Tryptophan		Trp	W
Tyrosine		Tyr	Y
Asparagine	Polar	Asn	N
Glutamine		Gln	Q
Aspartic acid	Acidic	Asp	D
Glutamic acid		Glu	E
Arginine	Basic	Arg	R
Histidine		His	H
Lysine		Lys	K

what codons correspond to what amino acids (Fig. 1.4). The genetic code was elucidated in the 1960s. Later, we will see how this code may be represented in a Perl program.

There are 64 codons in the genetic code and 61 of these specify an amino acid. The other three, UAA, UAG and UGA, are termination signals during protein synthesis. Because there are 61 codons that specify an amino acid and there are only 20 amino acids, for most amino acids there is more than one codon; because of this, the genetic code is said to be *degenerate*. The codon AUG, coding for methionine, is also used as a start codon, i.e. the start site of translation will (almost) always be the codon AUG.[4] Fig. 1.5 shows a region of bacterial mRNA

[4] The codon AUG is also used to encode internal methionines in a protein. Therefore, there are additional criteria to distinguish a start AUG from an internal AUG. For instance, in bacteria there is a specific sequence upstream of the start codon that helps in initiating protein synthesis.

UUU Phe F	UCU Ser S	UAU Tyr Y	UGU Cys C
UUC Phe F	UCC Ser S	UAC Tyr Y	UGU Cys C
UUA Leu L	UCA Ser S	UAA Stop	UGA Stop
UUG Leu L	UCG Ser S	UAG Stop	UGG Trp W
CUU Leu L	CCU Pro P	CAU His H	CGU Arg R
CUC Leu L	CCC Pro P	CAC His H	CGC Arg R
CUA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R
CUG Leu L	CCG Pro P	CAG Gln Q	CGG Arg R
AUU Ile I	ACU Thr T	AAU Asn N	AGU Ser S
AUC Ile I	ACC Thr T	AAC Asn N	AGC Ser S
AUA Ile I	ACA Thr T	AAA Lys K	AGC Arg R
AUG Met M	ACG Thr T	AAG Lys K	AGG Arg R
GUU Val V	GCU Ala A	GAU Asp D	GGU Gly G
GUC Val V	GCC Ala A	GAC Asp D	GGC Gly G
GUA Val V	GCA Ala A	GAA Glu E	GGA Gly G
GUG Val V	GCG Ala A	GAG Glu E	GGG Gly G

Fig. 1.4 *The universal genetic code*. Every three-base word (codon) corresponds to an amino acid (in three- and one-letter abbreviations), except UAA, UAG and UGA, which are termination signals during protein synthesis.

encoding a protein; a region starting with an AUG codon and ending with a stop codon.

What is the machinery responsible for carrying out the instructions as presented in the genetic code? Critical molecules are *transfer RNAs* (tRNAs), which act as adaptors between the languages of nucleic acids and proteins. This is possible because every tRNA carries a sequence, the *anticodon*, which is complementary to the codon on the mRNA, and to the amino acid corresponding to the codon.

The protein formed in the process of translation is a linear sequence of amino acids. That sequence will govern folding of the protein molecule into a distinct three-dimensional shape. Even more importantly, that shape is associated with one or more specific biological functions.

In summary, all cells are characterized by a flow of genetic information where DNA is first copied to RNA, which in turn is used to guide the production

```
AAAAUUCCUAUGAAGGUGAAUUUAGAGUGGAUAAUUAAACAGUUACAAAUGAUAGUUAAA
            M  K  V  N  L  E  W  I  I  K  Q  L  Q  M  I  V  K
AGAGCAUAUACUCCCUUUUCUAACUUUAAAGUUGCAUGUAUGAUUAUUGCUAACAACCAA
 R  A  Y  T  P  F  S  N  F  K  V  A  C  M  I  I  A  N  N  Q
ACUUUUUUUGGAGUUAACAUUGAAAAUUCUUCCUUUCCAGUAACUUUGUGUGCUGAAAGA
 T  F  F  G  V  N  I  E  N  S  S  F  P  V  T  L  C  A  E  R
AGCGCCAUUGCUAGCAUGGUUACAAGUGGUCAUAGGAAAAUUGAUUAUGUUUUUGUUUAC
 S  A  I  A  S  M  V  T  S  G  H  R  K  I  D  Y  V  F  V  Y
UUCAAUACUAAAAAUAAGAGUAACUCACCCUGUGGAAUGUGCAGACAAAACUUACUGGAA
 F  N  T  K  N  K  S  N  S  P  C  G  M  C  R  Q  N  L  L  E
UUUUCCCAUCAAAAAACAAAGCUUUUUUGUAUUGAUAAUGAUAGUAGUUAUAAACAAUUU
 F  S  H  Q  K  T  K  L  F  C  I  D  N  D  S  S  Y  K  Q  F
UCCAUUGAUGAAUUAUUAAUGAAUGGUUUUAAAAAGAGCUAAUGGAUAAACUUAGAUUA
 S  I  D  E  L  L  M  N  G  F  K  K  S STOP
```

Fig. 1.5 *Open reading frame (ORF)*. A region of the genome of the bacterium *Mycoplasma genitalium* that encodes the protein cytidine deaminase is depicted. The start site of translation, the codon AUG, is indicated as well as the stop codon UAA. The sequence in orange is the amino acid sequence of the deaminase protein, with one-letter symbols for the amino acids as listed in the genetic code in Fig. 1.4 and Table 1.1.