Chapter 1

The first look at a genome

Sequence statistics

- Genomes and genomic sequences
- Probabilistic models of sequences
- Statistical properties of sequences
- Standard data formats and databases

1.1 Genomic era, year zero

In 1995 a group of scientists led by Craig Venter, at The Institute for Genomic Research (TIGR) in Maryland, published a landmark paper in the journal *Science*. This paper reported the complete DNA sequence (the genome) of a free-living organism, the bacterium *Haemophilus influenzae* (or *H. influenzae*, for short). Up until that moment, only small viral genomes or small parts of other genomes had been sequenced. The first viral genome sequence (that of phage *phiX174*) was produced by Fred Sanger's group in 1978, followed a few years later by the sequence of human mitochondrial DNA by the same group. Sanger – working in Cambridge, UK – was awarded two Nobel prizes, the first one in 1958 for developing protein sequencing techniques and the second one in 1980 for developing DNA sequencing techniques. A bacterial sequence, however, is enormously larger than a viral one, making the *H. influenzae* paper a true milestone. Given the order of magnitude increase in genome size that was sequenced by the group at TIGR, the genomic era can be said to have started in 1995.

A few months later the same group at TIGR published an analysis of the full genome of another bacterium, *Mycoplasma genitalium* – a microbe responsible for urethritis – and shortly thereafter the sequence of the first eukaryote, the fungus, *Saccharomyces cerevisiae* (or *S. cerevisiae*, baker's yeast) was published by other groups. The method created by the TIGR group to obtain and assemble genome sequences was itself a watershed; their method relied massively on computer technology, but is beyond the topics discussed here. In the years that followed, the number of complete genomes published grew enormously, and the pace is still increasing. Before the start of the genomic era the collection of publicly available DNA sequence data was distributed among scientists on magnetic tape, then on CD, and finally over the internet. Now whole genomes are available for fast download from a number of public databases.

After completing the sequencing of *H. influenzae*, Venter's group moved to what is the next phase of any genomic project: genome annotation. Annotation

Table 1.1 Some of the genomes discussed in this book, with their size and date of completion

Organism	Completion date	Size	Description
phage phiX174	1978	5,368 bp	1st viral genome
human mtDNA	1980	16,571 bp	1st organelle genome
lambda phage	1982	48,502 bp	important virus model
HIV	1985	9,193 bp	AIDS retrovirus
H. influenzae	1995	1,830 Kb	1st bacterial genome
M. genitalium	1995	580 Kb	smallest bacterial genome
S. cerevisiae	1996	12.5 Mb	1st eukaryotic genome
E. coli K12	1997	4.6 Mb	bacterial model organism
C. trachomatis	1998	1,042 Kb	internal parasite of eukaryotes
D. melanogaster	2000	180 Mb	fruit fly, model insect
A. thaliana	2000	125 Mb	thale cress, model plant
H. sapiens	2001	3,000 Mb	human
SARS	2003	29,751 bp	coronavirus

involves various phases, and is never really complete. However, most sequencing projects perform at least two steps: a first (usually simpler) analysis, aimed at identifying all of the main structures and characteristics of a genome; then a second (often more complicated) phase, aimed at predicting the biological function of these structures. The first chapters of this book present some of the basic tools that allow us to perform sequence annotation. We leave the more advanced topic of sequence assembly – the initial step of constructing the entire genome sequence that must occur before any analyses begin – to more advanced courses in bioinformatics.

Now that we have the complete genome sequences of various species, and of various individuals of the same species, scientists can begin to make whole-genome comparisons and analyze the differences between organisms. Of course the completion of the draft human genome sequence in 2001 attracted headlines, but this was just one of the many milestones of the genomic era, to be followed soon thereafter by mouse, rat, dog, chimp, mosquito, and others. Table 1.1 lists some important model organisms as well as all of the organisms used in examples throughout this book, with their completion dates and genome length (the units of length will be defined in the next section). We should stress here that there were many challenges in data storage, sharing, and management that had to be solved before many of the new analyses we discuss could even be considered.

In the rest of this chapter we begin our analysis of genomic data by reproducing some of the original analyses of the early genome papers. We continue this aim in the following chapters, providing the reader with the data, tools, and concepts necessary to repeat these landmark analyses. Before we start our first statistical examination of a complete genome; however, we will need to summarize some key biological facts about how DNA is organized in cells and the key statistical issues involved in the analysis of DNA.

It is also worth emphasizing at this point that genomic data include more than just DNA sequence data. In 1995 a group of scientists led by Pat Brown at Stanford University introduced a high-throughput technology that enabled them to measure the level of activity of all the genes in an organism in a single experiment. The analysis of the large datasets generated by these experiments will be addressed in Chapter 9 and, partly, Chapter 10.

1.2 The anatomy of a genome

As a first definition, we can say that a genome is the set of all DNA contained in a cell; shortly we will explain how some organisms actually have multiple genomes in a single cell. The genome is formed by one or more long stretches of DNA strung together into chromosomes. These chromosomes can be linear or circular, and are faithfully replicated by a cell when it divides. The entire complement of chromosomes in a cell contains the DNA necessary to synthesize the proteins and other molecules needed to survive, as well as much of the information necessary to finely regulate their synthesis. As we mentioned in the Prologue, each protein is coded for by a specific gene – a stretch of DNA containing the information necessary for that purpose.

DNA molecules consist of a chain of smaller molecules called *nucleotides* that are distinct from each other only in a chemical element called a base. For biochemical reasons, DNA sequences have an orientation: it is possible to distinguish a specific direction in which to read each chromosome or gene. The cell's enzymatic machinery reads the DNA from the 5′ to the 3′ end (these are chemical conventions of the nucleic acids that make up DNA), which are often represented as the left and right end of the sequence, respectively.

A DNA sequence can be either single-stranded or double-stranded. The double-stranded nature of DNA is caused by the pairing of bases. When it is double-stranded, the two strands have opposite direction and are complementary to one another. This complementarity means that for each A, C, G, T in one strand, there is a T, G, C, or A, respectively, in the other strand. Chromosomes are double-stranded – hence the "double helix" – and information about a gene can be contained in either strand. Importantly, this pairing introduces a complete redundancy in the encoding, which allows the cell to reconstitute the entire genome from just one strand, which in turn enables faithful replication. For simple convenience, however, we usually just write out the single strand of DNA sequence we are interested in from 5′ to 3′.

Example 1.1
Sequence orientation and complementarity. The sequence 5′ –ATGCATGC – 3′ is complementary to the sequence 3′ – TACGTACG – 5′, which would often be represented in print as simply ATGCATGC if no other directionality is provided.

We have seen that DNA molecules consist of chains of nucleotides, each characterized by the base it contains. As a result, the letters of the DNA alphabet

are variously called nucleotides (nt), bases, or base pairs (bp) for double stranded DNA. The length of a DNA sequence can be measured in bases, or in kilobases (1000 bp or Kb) or megabases (1 000 000 bp or Mb). The genomes present in different organisms range in size from kilobases to megabases, and often have very different biological attributes. Here we review some of the basic facts about the most common genomes we will come across.

Prokaryotic genomes. As of the writing of this book, the Comprehensive Microbial Resources website hosted at TIGR contains the sequences of 239 completed genomes: 217 from bacteria, and 21 from archaea (including the two bacterial genomes from 1995 discussed above). This number is sure to have risen since then. Eubacteria and archaea are the two major groups of prokaryotes: free-living organisms without nuclei, a structure within cells that is used by eukaryotes to house their genomes. Prokaryotic organisms generally have a single, circular genome between 0.5 and 13 megabases long. *M. genitalium* has the smallest prokaryotic genome known, with only 580 074 bases. In addition to having relatively small genomes, prokaryotes also have rather simple genes and genetic control sequences; in the next chapter we will explore these issues in depth and see how they affect the identification of genes. Because of this simplicity and their fundamental similarities to more complex genomes, we will use many prokaryotic genomes as first examples for the analyses performed in the book. We will focus in particular on *H. influenzae*, *M. genitalium*, and *Chlamydia trachomatis*.

Viral genomes. Although viruses are not free-living organisms, an examination of viral genomes can be very informative. At least a thousand viral genomes have been sequenced, starting from what is considered the "pre-genomic" era, dating back to the late 1970s. Although these genomes are usually very short – between 5 and 50 kilobases (Kb) – and contain very few genes, their sequencing was a milestone for biology, and they enabled scientists to develop conceptual tools that would become essential for the analysis of the genomes of larger, free-living organisms. As they are an excellent model for practicing many of the methods to be deployed later on larger genomes, we will use them to illustrate basic principles. Their analysis is also highly relevant for epidemiological and clinical applications, as has been demonstrated in cases involving HIV and SARS (see Chapters 6 and 7). Peculiarly, viral genomes can be either single- or double-stranded, and either DNA- or RNA-based (we will learn more about the molecule RNA in the next chapter). Because of their small size, we can analyze a large number of viral genomes simultaneously on a laptop, a task that would require a large cluster of machines in the case of longer genomic sequences.

Eukaryotic genomes. The nuclear genome of eukaryotes is usually considered *the* genome of such an organism (see the next paragraph for a description of the organellar genomes contained in many eukaryotes). These nuclear genomes can be much larger than prokaryotic genomes, ranging in size from 8 Mb for some fungi to 670 gigabases (billions of bases or Gb) for some species of the single-celled amoeba; humans come in at a middling 3.5 Gb. Because of the large size of eukaryotic genomes, their sequencing is still a large effort usually

undertaken by consortia of labs; these labs may divide the work up by each sequencing different linear chromosomes of the same genome. Currently we have sequences representing more than 50 different eukaryotic organisms, including various branches of the evolutionary tree: the fungus, *S. cerevisiae* (baker's yeast); the round worm, *Caenorhabditis elegans*; the zebrafish, *Danio rerio*; important insects like the fruitfly, *Drosophila melanogaster*, and mosquito, *Anopheles gambiae*; mammals such as humans, *Homo sapiens*, and mouse, *Mus musculus*; and plants such as rice, *Oryza sativa*. The large size of such genomes is generally due not to a larger number of genes, but rather to a huge amount of repetitive "junk" DNA. It is estimated that only 5% of the human genome is functional (e.g. codes for proteins), while at least 50% of the genome is known to be formed by repetitive elements and parasitic DNA. Added to the packaging problems associated with stuffing these large amounts of DNA into each cell of a multicellular organism, most eukaryotes carry two copies of their nuclear genome in each cell: one from each parent. We refer to the single complement as the *haploid* set, as opposed to the *diploid* set of both genomes.

Organellar genomes. In addition to these huge nuclear genomes, most eukaryotes also carry one or more smaller genomes in each cell. These are contained in cellular organelles, the most common of which are the mitochondrion and the chloroplast. These organellar genomes are likely the remnants of symbiotic prokaryotic organisms that lived within eukaryotic cells. We now have the genome sequence of the mitochondria and chloroplasts of at least 600 species, often with multiple whole genomes of different individuals within a species. These genomes are usually only tens of thousands of bases long, circular, and contain a few essential genes. There can be hundreds of each of these organelles in a cell, with more copies resulting in more expressed products. Mitochondrial DNA (*mtDNA*) is particularly important for anthropological analyses, and we will use it to discuss whole genome comparisons within humans in Chapter 5.

1.3 Probabilistic models of genome sequences

All models are wrong, but some are useful.
(G. E. P. Box)

When the first whole genome of a free-living organism was sequenced in 1995, much was already known about the general workings of cellular function, and many things were also known about DNA sequences. Complete genomes of small viruses and organelles were available in the early 1980s, as were the sequences of individual genes from a variety of organisms. Although nothing done before 1995 compared in scale with the problems that were to be addressed when analyzing even a simple bacterial genome, earlier experiences provided the basic statistical techniques that were to evolve into modern whole-genome analysis.

A large part of the study of computational genomics is comprised of statistical methods. While any undertaking involving millions or billions of data

points necessarily requires statistics, the problem is especially acute in the study of DNA sequences. One reason for this is that we often wish to find structures of interest (such as genes) in sequences of millions of bases, and in many important cases most of those sequences do not contain biologically relevant information. In other words, the signal-to-noise ratio in genome sequences may be very low. As we will see, interesting elements are often immersed in a random background – detecting them requires sophisticated statistical and algorithmic tools.

As a first step, we need to have probabilistic models of DNA sequences. These will set a foundation for all of the analyses that follow in this book. We first define some of the basic concepts in probabilistic models of DNA, and then present a simple statistical framework for the analysis of genome sequence data. To highlight the usefulness of probabilistic models we continue the chapter by carrying out simple analyses of genome sequences. Later chapters will introduce increasingly sophisticated biological questions and the commensurate statistical methods needed to answer them.

Alphabets, sequences, and sequence space. Although we should not forget that DNA is a complex molecule with three-dimensional properties, it is often convenient to model it as a one-dimensional object, a sequence of symbols from the alphabet {A, C, G, T}. This abstraction is extremely powerful, enabling us to deploy a large number of mathematical tools; it is also incorrect, in that it neglects all the information that might be contained in the three-dimensional structure of the molecule. In this book we make this approximation, without any further warning, and will develop statistical and computational methods of analysis based on it.

Definition 1.1
DNA sequences and genomes: formal model. A "DNA sequence," $\mathbf{s}$, is a finite string from the alphabet $\mathcal{N} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ of nucleotides. A "genome" is the set of all DNA sequences associated with an organism or organelle.

This representation of genomes as strings from an alphabet is very general, and enables us to develop statistical models of sequence evolution, sequence similarity, and various forms of sequence analysis. Some of them are discussed in this chapter.

Definition 1.2
Elements of a sequence. We denote the elements of a sequence as follows $\mathbf{s} = \mathbf{s}_1\mathbf{s}_2 \ldots \mathbf{s}_n$, where an individual nucleotide is represented by $\mathbf{s}_i$. Given a set of indices K, we can consider the sequence formed by concatenating together the corresponding elements of $\mathbf{s}$ in their original order: $\mathbf{s}(K) = \mathbf{s}_i\mathbf{s}_j\mathbf{s}_k$ if $K = \{i, j, k\}$. If the set is a closed interval of integers, $K = [i, j]$, we can denote it also as $K = (i : j)$; the corresponding subsequence is the substring $\mathbf{s}(i : j)$. If it is formed by just one element, $K = \{i\}$, then this indicates a single, specific symbol of the sequence: $\mathbf{s}_i = \mathbf{s}(i)$.

Example 1.2
Elements of a sequence. In the DNA sequence $\mathbf{s}$ =ATATGTCGTGCA we find $\mathbf{s}(\mathbf{3} : \mathbf{6})$ =ATGT and $\mathbf{s}(\mathbf{8}) = \mathbf{s}_8$ = G.

Remark 1.1
Strings and sequences. Note that what we call sequences in biology are usually called strings in standard computer science terminology. This distinction becomes relevant in computer science when defining subsequences and substrings, two very different objects. What we call subsequences in this book – contiguous, shorter sequences from a longer sequence – are called substrings in computer science, where subsequences refer to non-contiguous sets of symbols from a longer sequence.

Nearly all probabilistic sequence analysis methods assume one of two simple models, or variations thereof. They are the *multinomial* model and the *Markov* model, and will be described below. As is often the case in modeling, these do not need to mimic true DNA sequences in every respect. Their main feature is that they capture enough of the properties of sequences while still being efficiently computable. In other words, they are the result of a compromise between accuracy and efficiency.

Although in this chapter we deal with DNA sequences, in the rest of the book we will find various other types of biological sequences; sequences defined on different alphabets. All the algorithms we present will be valid for any type of sequence, and we will often define them in terms of a generic alphabet Σ, so as to preserve generality. The most common other types of biological sequences are RNA sequences (also defined over a 4 letter alphabet, $\mathcal{N}_{RNA} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$), and amino acid sequences (based on a 20 letter alphabet

$$\mathcal{A} = \{\mathtt{A}, \mathtt{R}, \mathtt{N}, \mathtt{D}, \mathtt{C}, \mathtt{Q}, \mathtt{E}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{L}, \mathtt{K}, \mathtt{M}, \mathtt{F}, \mathtt{P}, \mathtt{S}, \mathtt{T}, \mathtt{W}, \mathtt{Y}, \mathtt{V}\}$$

and discussed further in Chapter 2). It is often also useful to define a sequence of codons (see Chapter 2), where the alphabet is formed by all triplets from the nucleotide alphabet $\mathcal{N}$ and will be indicated by $\mathcal{C} = \{\mathtt{AAA}, \cdots, \mathtt{TTT}\}$.

Multinomial sequence models. The simplest model of DNA sequences assumes that the nucleotides are independent and identically distributed (the "i.i.d." assumption): the sequence has been generated by a stochastic process that produces any of the four symbols at each sequence-position i at random, independently drawing them from the same distribution over the alphabet $\mathcal{N}$. This is called a *multinomial sequence model*, and is simply specified by choosing a probability distribution over the alphabet, $p = (p_A, p_C, p_G, p_T)$, where the probability of observing any of the four nucleotides at position i of the sequence $\mathbf{s}$ is denoted by $p_x = p(\mathbf{s}(i) = x)$ and does not depend on the position i. This model can also be used to calculate the probability of observing the generic symbol $x \in \Sigma$ (i.e. x within any alphabet).

For this model we could assume that all four nucleotides are of equal frequency (p_A=p_C=p_G=p_T), or that they are the observed frequencies from some dataset. Our only requirement is that the distribution needs to satisfy the normalization constraint $p_A + p_C + p_G + p_T = 1$.

The multinomial model allows us to easily calculate the *probability* of a given sequence (denoted P), also called the *likelihood* of the data given the model (denoted $\mathcal{L}$). Given a sequence $\mathbf{s} = \mathbf{s}_1\mathbf{s}_2...\mathbf{s}_n$, its probability is

$$P(\mathbf{s}) = \prod_{i=1}^{n} p(\mathbf{s}(i)).$$

This is equivalent to multiplying together the probabilities of all of the individual nucleotides.

Of course we do not expect DNA sequences to be truly random, but having a model that describes the expectation for a randomly generated sequence can be very helpful. Furthermore, this simple model already captures enough of a sequence's behavior to be useful in certain applications, while remaining very easy to handle mathematically. Even finding violations of this simple model will point us towards interesting regions of the genome. We can easily test if this model is realistic by checking to see whether real data conform to its assumptions. We can do this either by estimating the frequencies of the symbols in various regions of the sequence – to check the assumption of stationarity of the independent and identically distributed (i.i.d). probability distribution across the sequence – or by testing for violations of the independence assumption by looking for correlations among neighboring nucleotides. Regions that both change in the frequency of `A`, `C`, `G`, and `T` and where there are strong correlations among nearby symbols can be quite interesting, as we will show later.

Markov sequence models. A more complex model of DNA sequences is provided by the theory of Markov chains. In Markov chains the probability of observing a symbol depends on the symbols preceding it in the sequence. In so doing, Markov chains are able to model local correlations among nucleotides. A Markov chain is said to be of order 1 if the probability of each symbol only depends on the one immediately preceding it, and of increasing order as the dependency on past symbols extends over greater distances. We can think of the multinomial model as a Markov chain of order 0 because there is no dependence on previous symbols. How do we know which model to pick, or what order Markov model to pick? This is a problem of hypothesis testing, a topic we address in Chapter 2. For now, we will simply discuss the basic aspects of Markov models. (Markov chains are central tools in bioinformatics, and we will encounter them again in Sections 5.4.1 and 5.4.2).

Briefly, a Markov chain is a process defined by a set of states (in this case the symbols of an alphabet) and by a transition probability from one state to the next. The transition probabilities are organized in the transition matrix T. The trajectory of the process through the state space defines a sequence. This is represented in Figure 1.1.

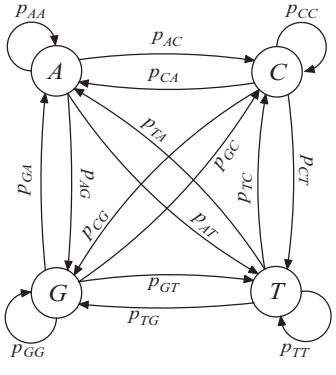

Fig. 1.1 The trajectory of a Markov chain process generates a sequence of symbols. Starting at any one of the four nucleotides, the probability of the next nucleotide in the sequence is determined by the current state.

The figure shows that, starting at any one of the four nucleotides, the probability of the next nucleotide in the sequence is determined by the current state. If we start at `G`, the probabilities of any of the four other nucleotides appearing next are given by p_{GA}, p_{GC}, p_{GG}, and p_{GT}. Moving to another nucleotide state means that there are new transition probabilities: if we moved to state `A`, these would be p_{AA}, p_{AC}, p_{AG}, and p_{AT}. In this manner, a Markov chain defines a DNA sequence. All of the transition probabilities are given by the matrix T,

and the probability for the start state is given by $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$, again with the obvious normalization constraint:

$$T = \begin{matrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CA} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{matrix}$$

$$\pi = \begin{matrix} \pi_A & \pi_C & \pi_G & \pi_{T.} \end{matrix}$$

The Markov model therefore no longer assumes that the symbols are independent, and short-range correlations can be captured (if the transition probabilities are all 0.25, we are again back at a multinomial model).

Example 1.3
Markov DNA sequence. Using the Markov chain defined by uniform starting probabilities (π) and the transition matrix

$T =$

	to A	*to* C	*to* G	*to* T
from A	0.6	0.2	0.1	0.1
from C	0.1	0.1	0.8	0
from G	0.2	0.2	0.3	0.3
from T	0.1	0.8	0	0.1

we generated the following sequence:

```
ACGCGTAATCAAAAAATCGGTCGTCGGAAAAAAAAAATCG
```

As you can see, many As are followed by As, Ts are followed by Cs, and Cs are followed by Gs, as described by our transition matrix.

The entries in the transition matrix are defined formally as follows:

$$p_{xy} = p(\mathbf{s}_{i+1} = y | \mathbf{s}_i = x).$$

This says that the probability of going from state x to state y is equivalent to the conditional probability of seeing state y given that it was preceded by state x. We generally define the probability of an entire sequence as a joint probability:

$$P(\mathbf{s}) = P(s_1 s_2 \cdots s_n).$$

The computation of this probability can be greatly simplified when we can factorize it. In the case of multinomial models, the factorization is obvious: $P(\mathbf{s}) = p(s_1)p(s_2)\ldots p(s_n)$; for Markov chains (of order 1) it is only slightly more complicated:

$$P(\mathbf{s}) = P(s_n|s_{n-1})P(s_{n-1}|s_{n-2})\cdots P(s_2|s_1)\pi(s_1)$$

or

$$P(\mathbf{s}) = \pi(s_1)\prod_{i=2}^{n} p(s_i|s_{i-1}) = \pi(s_1)\prod_{i=2}^{n} p_{s_{i-1}s_{i.}}$$

where $\pi(s_1 = x)$ represents the probability of seeing symbol x in position s_1. In other words, we exploit the fact that the probability of a symbol depends only on the previous one to simplify the computation of the joint probability of the entire sequence.

Table 1.2 Basic statistics of the *H. influenzae* genome: the count of each nucleotide and its relative frequency. The total length of the sequence is 1 830 138 bp.

Base	Number	Frequency
A	567,623	0.3102
C	350,723	0.1916
G	347,436	0.1898
T	564,241	0.3083

1.4 Annotating a genome: statistical sequence analysis

There are various elements of interest in a genome sequence, and many of them will be discussed in various chapters of this book. For example, Chapter 2 will discuss the structure of genes, how we find them, and the way in which they are regulated. Here we examine simpler statistical properties of the DNA sequences such as the frequency of nucleotides, dinucleotides (pairs of bases), and other short DNA words. We will see that this preliminary description is of fundamental importance for genome annotation, as well as a stepping stone for further analysis.

Base composition. One of the most fundamental properties of a genome sequence is its base composition, the proportion of A, G, C, and T nucleotides present. For *H. influenzae*, we can easily count the number of each type of base and divide by the total length of the genome (performing both operations on only one strand of the DNA sequence) to obtain the frequency of each base. Table 1.2 shows the number of times each base appears in the genome, and its relative frequency (the total length, L, of the *H. influenzae* genomes is 1 830 138 bp).

We can see that the four nucleotides are not used at equal frequency across the genome: A and T are much more common than G and C. In fact, it is fairly unusual for all of the bases to be used in equal frequencies in any genome. We should point out that, while we have only counted the bases on one strand of the DNA molecule, we know exactly what the frequency of all the bases are on the other strand because of the complementarity of the double helix. The frequencies in the complementary sequence will be T $= 0.3102$, G $= 0.1916$, C $= 0.1898$, and A $= 0.3083$.

In addition to the global base composition of a genome, it is of interest to consider local fluctuations in the frequencies of nucleotides across the sequence. We can measure local base composition by sliding a window of size k along a chromosome, measuring the frequency of each base in the window, and assigning these values to the central position of the window. This produces a vector of length $L - k + 1$ that can be plotted, as seen in Figures 1.2 and 1.3 (with window sizes 90 000 bp and 20 000 bp, respectively).