Bioinformatics and Molecular Evolution

Paul G. Higgs and *Teresa K. Attwood*

Blackwell
Publishing

BIOINFORMATICS AND MOLECULAR EVOLUTION

Bioinformatics and Molecular Evolution

Paul G. Higgs and *Teresa K. Attwood*

Blackwell
Publishing

BLACKWELL PUBLISHING
350 Main Street, Malden, MA 02148-5020, USA
108 Cowley Road, Oxford OX4 1JF, UK
550 Swanston Street, Carlton, Victoria 3053, Australia

First published 2005 by Blackwell Science Ltd

Library of Congress Cataloging-in-Publication Data

Higgs, Paul G.
Bioinformatics and molecular evolution / Paul G. Higgs and Teresa K. Attwood.
p. ; cm.
Includes bibliographical references and index.
ISBN 1–4051–0683–2 (pbk. : alk. paper)
1. Molecular evolution—Mathematical models. 2. Bioinformatics.
[DNLM: 1. Computational Biology. 2. Evolution, Molecular. 3. Genetics. QU 26.5 H637b 2005]
I. Attwood, Teresa K. II. Title.

QH371.3.M37H54 2005
572.8—dc22

2004021066

A catalogue record for this title is available from the British Library.

Set in 9½/12pt Photina
by Graphicraft Limited, Hong Kong
Printed and bound in the United Kingdom
by TJ International, Padstow, Cornwall

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

For further information on
Blackwell Publishing, visit our website:
www.blackwellpublishing.com

Short Contents

Full Contents

Preface

RATIONALE

Degree programs in bioinformatics at Masters or undergraduate level are becoming increasingly common and single courses in bioinformatics and computational biology are finding their way into many types of degrees in biological sciences. This book satisfies a need for a textbook that explains the scientific concepts and computational methods of bioinformatics and how they relate to current research in biological sciences. The content is based on material taught by both authors on the MSc program in bioinformatics at the University of Manchester, UK, and on an upper level undergraduate course in computational biology taught by P. Higgs at McMaster University, Ontario.

Many fundamental concepts in bioinformatics, such as sequence alignments, searching for homologous sequences, and building phylogenetic trees, are linked to evolution. Also, the availability of complete genome sequences provides a wealth of new data for evolutionary studies at the whole-genome level, and new bioinformatics methods are being developed that operate at this level. This book emphasizes the evolutionary aspects of bioinformatics, and includes a lot of material that will be of use in courses on molecular evolution, and which up to now has not been found in bioinformatics textbooks.

Bioinformatics chapters of this book explain the need for computational methods in the current era of complete genome sequences and high-throughput experiments, introduce the principal biological databases, and discuss methods used to create and search them. Algorithms for sequence alignment, identification of conserved motifs in protein families, and pattern-recognition methods using hidden Markov models and neural networks are discussed in detail. A full chapter on data analysis for microarrays and proteomics is included.

Evolutionary chapters of the book begin with a brief introduction to population genetics and the study of sequence variation within and between populations, and move on to the description of the evolution of DNA and protein sequences. Phylogenetic methods are covered in detail, and examples are given of application of these methods to biological questions. Factors influencing evolution at the level of whole genomes are also discussed, and methods for the comparison of gene content and gene order between species are presented.

The twin themes of bioinformatics and molecular evolution are woven throughout the book – see the Chapter Plan diagram below. We have considered several possible orders of presenting this material, and reviewers of this book have also suggested their own alternatives. There is no single right way to do it, and we found that no matter in which order we presented the chapters, there was occasionally need to forward-reference material in a later chapter. This order has been chosen so as to emphasize the links between the two themes, and to proceed from background material through standard methods to more advanced topics. Roughly speaking, we would consider everything up to the end of Chapter 9 as fundamental material, and Chapters 10–13 as more advanced methods or more recent applications.

Individual instructors are free to use any combination of chapters in whichever order suits them best.

This book is for people who want to understand bioinformatics methods and who may want to go on to develop methods for themselves. Intelligent use of bioinformatics software requires a proper understanding of the mathematical and statistical methods underlying the programs. We expect that many of our readers will be biological science students who are not confident of their mathematical ability. We therefore try to present mathematical material carefully at an accessible level. However, we do not avoid the use of equations, since we consider the theoretical parts of the book to be an essential aspect. More detailed mathematical sections are placed in boxes aside from the main text, where appropriate. The book contains an appendix summarizing the background mathematics that we would hope bioinformatics students should know. In our experience, students need to be encouraged to remember and practice mathematical techniques that they have been taught in their early undergraduate years but have not had occasion to use.

We discuss computational algorithms in detail but we do not cover programming languages or programming methods, since there are many other books on computing. Although we give some pointers to available software and useful Web sites, this book is not simply a list of programs and URLs. Such material becomes out of date extremely quickly, whereas the underlying methods and principles that we focus on here retain their relevance.

Features

• Comprehensive coverage of bioinformatics in a single text, including sequence analysis, biological databases, pattern recognition, and applications to genomics, microarrays, and proteomics.
• Places bioinformatics in the context of evolutionary biology, giving detailed treatments of molecular evolution and molecular phylogenetics and discussing evolution at the whole-genome level.
• Emphasizes the theoretical and statistical methods used in bioinformatics programs in a way that is accessible to biological science students.
• Extended problem questions provide reinforcement of concepts and application of chapter material.
• Periodic cumulative "self-tests" challenge the students to synthesize and apply their overall understanding of the material up to that point.
• Accompanied by a dedicated Web site at www.blackwellpublishing.com/higgs including the following:

– all art in downloadable JPEG format (also available to instructors on CD-ROM)
– all answers to self-tests
– downloadable sequences
– links to Web resources.

ACKNOWLEDGMENTS

We wish to thank Andy Brass for his tremendous input to the Manchester Bioinformatics MSc program over many years, and for his "nose" for which subjects are the important ones to teach. We thank Magnus Rattray for his help on the theory and algorithms module and for coming up with several of the problem questions in this book. Thanks are also due to several members of our research groups whose work has helped with some of the topics covered here: Vivek Gowri-Shankar, Daniel Jameson, Howsun Jow, Bin Tang, and Anna Gaulton.

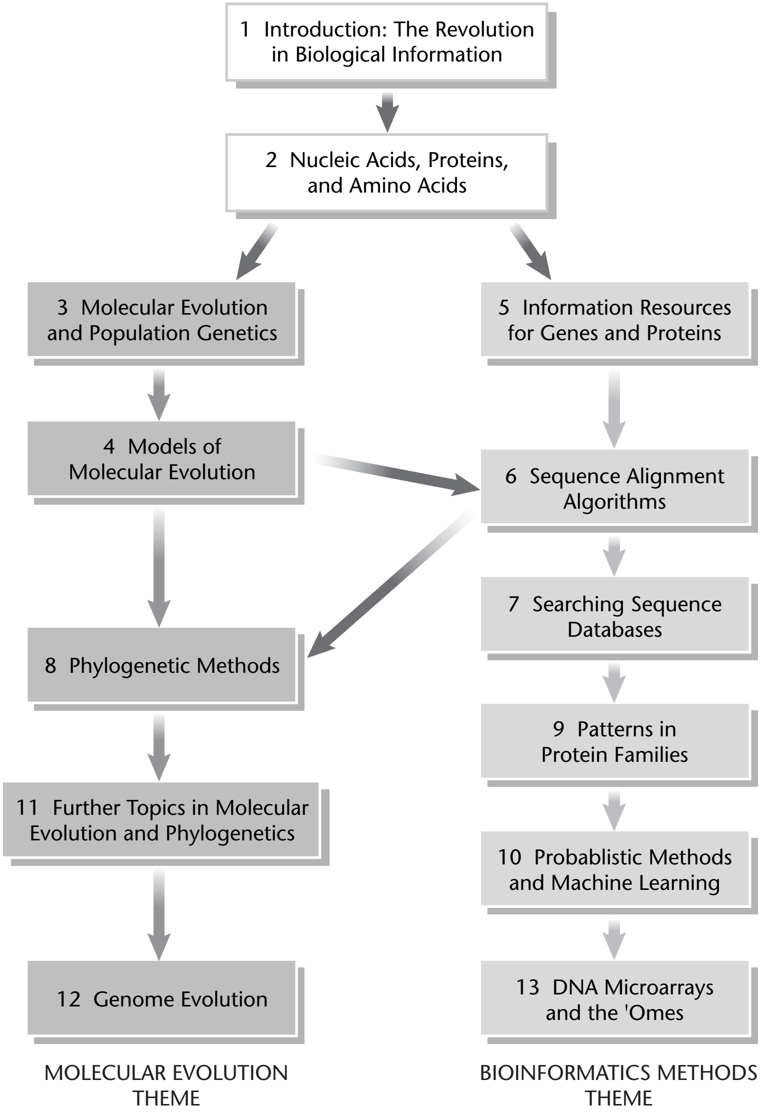
At Blackwell publishers, we would like to thank Liz Marchant, from the UK office, who originally commissioned this book, and Elizabeth Wald and Nancy Whilton at the US office, who have given advice and support in the later stages.

Paul Higgs would also like to thank the enlightened administrative authorities at the University of Manchester for employing a physicist in the biology department and the equally enlightened representatives of McMaster University for employing a bioinformatician in the physics department. The jump between disciplines and between institutions is not too much of a stretch:

```
M--CMASTER
MANCHESTER
```

Paul Higgs
Terri Attwood
May 2004

Bioinformatics and Molecular Evolution Chapter Plan


1 Introduction: The Revolution in Biological Information
2 Nucleic Acids, Proteins, and Amino Acids
3 Molecular Evolution and Population Genetics
5 Information Resources for Genes and Proteins
4 Models of Molecular Evolution
6 Sequence Alignment Algorithms
7 Searching Sequence Databases
8 Phylogenetic Methods
9 Patterns in Protein Families
11 Further Topics in Molecular Evolution and Phylogenetics
10 Probablistic Methods and Machine Learning
12 Genome Evolution
13 DNA Microarrays and the 'Omes
MOLECULAR EVOLUTION THEME
BIOINFORMATICS METHODS THEME

Introduction: The revolution in biological information

CHAPTER 1

CHAPTER PREVIEW

Here we consider the rapid expansion in the amount of biological sequence data available and compare this to the exponential growth in computer speed and memory size that has occurred in the same period. The reader should appreciate why bioinformatics is now essential for understanding the information contained in the sequences, and for efficient storage and retrieval of the information. We also consider some of the history of bioinformatics, and show that many of its foundations are related to molecular evolution and population genetics. Thus, the reader should understand what is meant by the term "bioinformatics" and the role of bioinformatics in relation to other disciplines.

1.1 DATA EXPLOSIONS

In the past decade there has been an explosion in the amount of DNA sequence data available, due to the very rapid progress of genome sequencing projects. There are three principal comprehensive databases of nucleic acid sequences in the world today.

- The EMBL (European Molecular Biology Laboratory) database is maintained at the European Bioinformatics Institute in Cambridge, UK (Stoesser *et al.* 2003).
- GenBank is maintained at the National Center for Biotechnology Information in Maryland, USA (Benson *et al.* 2003).
- The DDBJ (DNA Databank of Japan) is maintained at the National Institute of Genetics in Mishima, Japan (Miyazaki *et al.* 2003).

These three databases share information and hence contain almost identical sets of sequences. The objective of these databases is to ensure that DNA sequence information is stored in a way that is publicly, and freely, accessible and that it can be retrieved and used by other researchers in the future. Most scientific journals require submission of newly sequenced DNA to one of the public databases before a publication can be made that relies on the sequence. This policy has proved tremendously successful for the progress of science, and has led to a rapid increase in the size and usage of sequence databases.

As a measure of the rapid increase in the total available amount of sequence data, Fig. 1.1 and Table 1.1 show the total length of all sequences in GenBank, and the total number of sequences in GenBank as a function of time. Note that the vertical scale is logarithmic and the curves appear approximately as straight lines. This means that the size of GenBank is increasing exponentially with time (see Problem 1.1). The dotted line in the figure is a straight-line fit to the data for the total sequence length (the 1982 point seemed to be an outlier and was excluded). From this we can estimate that the yearly multiplication factor (i.e., the factor by which the amount of data goes up each year) is about 1.6, and that the database doubles in size every 1.4 years. All those sequencing machines are working hard! Interestingly, the curve for the number of sequences almost exactly parallels the curve for the total length. This means that the typical length of one sequence entry in GenBank has remained at

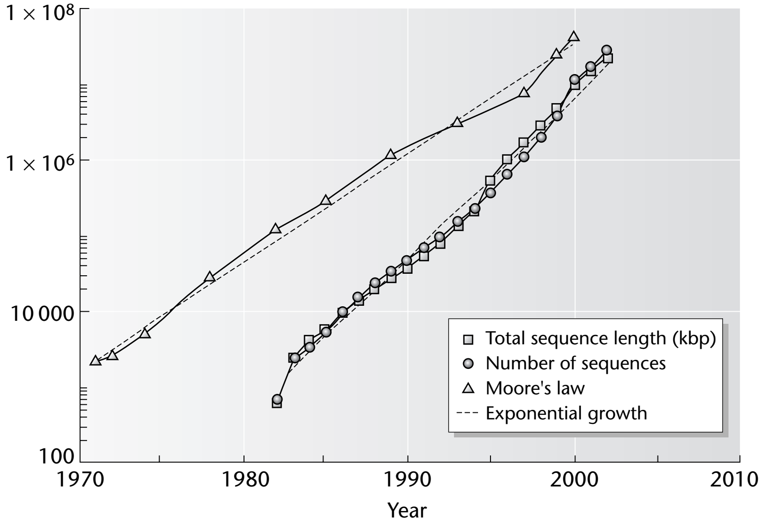


Fig. 1.1 Comparison of the rate of growth of the GenBank sequence (data from Table 1.1) with the rate of growth of the number of transistors in personal computer chips (Moore's law: data from Table 1.2). Dashed lines are fits to an exponential growth law.

Table 1.1 The growth of GenBank.

Year	Base pairs	Sequences
1982	680,338	606
1983	2,274,029	2,427
1984	3,368,765	4,175
1985	5,204,420	5,700
1986	9,615,371	9,978
1987	15,514,776	14,584
1988	23,800,000	20,579
1989	34,762,585	28,791
1990	49,179,285	39,533
1991	71,947,426	55,627
1992	101,008,486	78,608
1993	157,152,442	143,492
1994	217,102,462	215,273
1995	384,939,485	555,694
1996	651,972,984	1,021,211
1997	1,160,300,687	1,765,847
1998	2,008,761,784	2,837,897
1999	3,841,163,011	4,864,570
2000	11,101,066,288	10,106,023
2001	15,849,921,438	14,976,310
2002	28,507,990,166	22,318,883

Data obtained from http://www.ncbi.nih.gov/Genbank/genbankstats.html.

close to 1000. There are, of course, enormous variations in length between different sequence entries.

There is another famous exponentially increasing curve that goes by the name of Moore's law. Moore (1965) noticed that the number of transistors in integrated circuits appeared to be roughly doubling every year over the period 1959–65. Data on the size of Intel PC chips (Table 1.2) show that this exponential increase is still continuing. Looking at the data more carefully, however, we see that the estimate of doubling every year is rather overoptimistic. The chip size is actually doubling every **two** years and the yearly multiplication factor is 1.4. Although extremely impressive, this is substantially slower than the rate of increase of GenBank (see Fig. 1.1 and Table 1.3).

What about the world's fastest supercomputers? Jack Dongarra and colleagues from the University of Tennessee introduced the LINPACK benchmark, which measures the speed of computers at solving a complex set of linear equations. A list of the top 500 supercomputers according to this benchmark is published twice yearly (http://www.top500.org). Figure 1.2 shows the performance benchmark rate of the top computer at each release of the list. Once again, this is approximately an exponential (with large fluctuations). The best-fit straight line has a

Table 1.2 The growth of the number of transistors in personal computer processors.

Type of processor	Year of introduction	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386™ processor	1985	275,000
486™ DX processor	1989	1,180,000
Pentium® processor	1993	3,100,000
Pentium II processor	1997	7,500,000
Pentium III processor	1999	24,000,000
Pentium 4 processor	2000	42,000,000

Data obtained from Intel (http://www.intel.com/research/silicon/mooreslaw.htm).

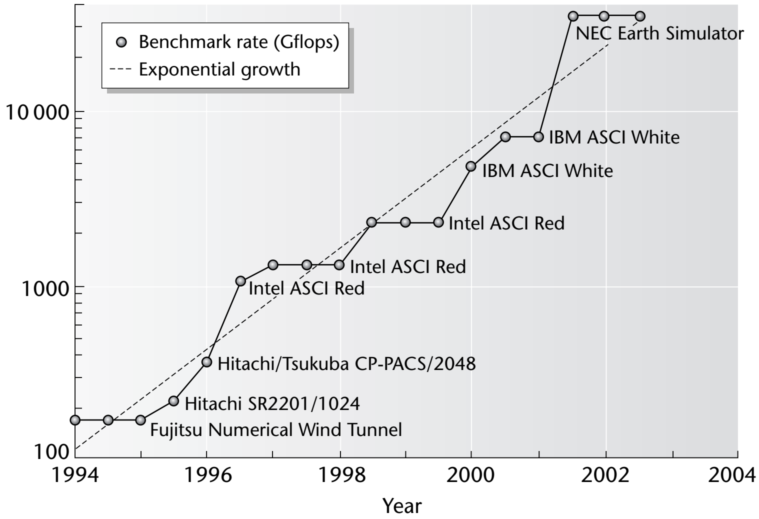


Fig. 1.2 The performance of the world's top supercomputers using the LINPACK benchmark (Gflops). Data from http://www.top500.org.

doubling time of 1.04 years. So supercomputers seem to be beating GenBank for the moment. However, most of us do not have access to a supercomputer. The PC chip size may be a better measure of the amount of computing power available to anyone using a desktop.

Clearly, we have reached a point where computers are essential for the storage, retrieval, and analysis of biological sequence data. However, we cannot simply rely on computers and stop thinking. If we stick with our same old computing methods, then we will be limited by the hardware. We still need people, because only people can think of better and faster algorithms for data analysis. That is what this book is about. We will discuss the methods and algorithms used in bioinformatics, so that hopefully you will understand enough to be able to improve those methods yourself.

Another important type of biological data that is exponentially increasing is protein structures. PDB is a database of protein structures obtained from X-ray crystallography and NMR experiments. From the number of entries in PDB in successive releases, we calculated that the doubling time for the number

Table 1.3 Comparison of rates of increase of several different data explosion curves.

Type of data	Growth rate, *r*	Doubling time, *T* (years)	Yearly multiplication factor, *R*
GenBank (total sequence length)	0.480	1.44	1.62
PC chips (number of transistors)	0.332	2.09	1.39
Supercomputer speed (LINPACK benchmark)	0.664	1.04	1.94
Protein structures (number of PDB entries)	0.209	3.31	1.23
Number of complete prokaryotic genomes	0.518	1.34	1.68
Abstracts: bioinformatics	0.587	1.18	1.80
Abstracts: genomics	0.569	1.22	1.77
Abstracts: proteomics	0.996	0.70	2.71
Abstracts: phylogenetic(s)	0.188	3.68	1.21
Abstracts: total	0.071	9.80	1.07

Table 1.4 The history of genome-sequencing projects.

Year	Archaea	Bacteria	Eukaryotes	Landmarks
1995	0	2	0	First bacterial genome: *Haemophilus influenzae*
1996	1	2	0	First archaeal genome: *Methanococcus jannaschii*
1997	2	4	1	First unicellular eukaryote: *Saccharomyces cerevisiae*
1998	1	5	1	First multicellular eukaryote: *Caenorhabditis elegans*
1999	1	4	1	—
2000	3	13	2	First plant genome: *Arabidopsis thaliana*
2001	2	24	3	First release of the human genome
2002	6	32	9	—
2003 (to July)	0	25	2	—
Total	16	111	19	—

Data from the Genomes OnLine Database (http://wit.integratedgenomics.com/GOLD/).

of available protein structures is 3.31 years (Table 1.3), which is considerably slower than the number of sequences. Since the number of experimentally determined structures is lagging further and further behind the number of sequences, computational methods for structure prediction are important. Many of these methods work by looking for similarities in sequence between a protein of unknown structure and a protein of known structure, and use this to make predictions about the unknown structure. These techniques will become increasingly useful as our knowledge of real examples increases.

In 1995, the bacterium *Haemophilus influenzae* entered history as the first organism to have its genome completely sequenced. Sequencing technology has advanced rapidly and has become increasingly automated. The sequencing of a new prokaryotic genome has now become almost commonplace. Table 1.4 shows the progress of complete genome projects with some historical landmarks. With the publication of the human genome in 2001, we can now truly say that we are in the "post-genome age". The number of complete prokaryotic genomes (total of archaea plus bacteria from Table 1.4) is going through its own data explosion. The doubling time is about 1.3 years and the yearly multiplication factor is about 1.7. For the present, complete eukaryotic genomes are still rather few, so that the publication of each individual genome still retains its status as a landmark event. It seems only a matter of time, however, before we shall be able to draw a data explosion curve for the number of eukaryotic genomes too.

This book emphasizes the relationship between bioinformatics and molecular evolution. The availability of complete genomes is tremendously important for evolutionary studies. For the first time we can begin to compare whole sets of genes between organisms, not just single genes. For the first time we can begin to study the processes that govern the evolution of whole genomes. This is therefore an exciting time to be in the bioinformatics area.

1.2 GENOMICS AND HIGH-THROUGHPUT TECHNIQUES

The availability of complete genomes has opened up a whole research discipline known as **genomics**. Genomics refers to scientific studies dealing with whole sets of genes rather than single genes. The advances made in sequencing technology have come at the same time as the appearance of new **high-throughput** experimental techniques. One of the most important of these is **microarray** technology, which allows measurement of the expression level (i.e., mRNA concentration) of thousands of genes in a cell simultaneously. For example, in the case of the yeast, *Saccharomyces cerevisiae*, where the complete genome is available, we can put probes for **all** the genes onto one microarray chip. We can then study the way the expression levels of all the genes respond to changes in external conditions or the way that they vary during the cell cycle. Complete genomes therefore change the way that experimental science is carried out, and allow us to address questions that were not possible before.

Another important field where high-throughput techniques are used is **proteomics**. Proteomics is the study of the proteome, i.e., the complete set of proteins in a cell. The experimental techniques used are principally two-dimensional gel electrophoresis for the separation of the many different proteins in a cell extract, and mass spectrometry for identifying proteins by their molecular masses. Once again, the availability of complete genomes is tremendously important, because the masses of the proteins determined by mass spectrometry can be compared directly to the masses of proteins expected from the predicted position of open reading frames in the genome.

High-throughput experiments produce large amounts of quantitative data. This poses challenges for bioinformaticians. How do we store information from a microarray experiment in such a way that it can be compared with results from other groups? How do we best extract meaningful information from the vast array of numbers generated? New statistical methods are needed to spot significant trends and patterns in the data. This is a new area of biological sciences where computational methods are essential for the progress of the experimental science, and where algorithms and experimental techniques are being developed side by side.

As a measure of the interest of the scientific community in genomics and related areas, let us look at the number of scientific papers published in these areas over the past few years. The ISI Science Citation Index allows searches for articles published in specific years that use specified words in their title, keywords, or abstract. Figure 1.3 shows the numbers of published articles (cumulative since 1981) for several important terms relevant to this book. Papers using the words "genomics" and "bioinformatics" increase at almost exactly the same rate, both having yearly multiplication factors of 1.8 and doubling times of 1.2 years. "Proteomics" is a very young field, with no articles found prior to 1998. The doubling time is 0.7 years: the fastest growth of any of the quantities considered in Table 1.3. References to "microarray" also increase rapidly. This curve appears significantly nonlinear because there are several different meanings for the term. Almost all the references prior to about 1996 refer to microarray electrodes, whereas in later years, almost all refer to DNA microarrays for gene expression. The rate of increase of the use of DNA microarrays is therefore steeper than it appears in the figure.

The number of papers using both "sequence" and "database" is much larger than those using any of the terms considered above (although it is increasing less rapidly). This shows how important biological databases and the algorithms for searching them have

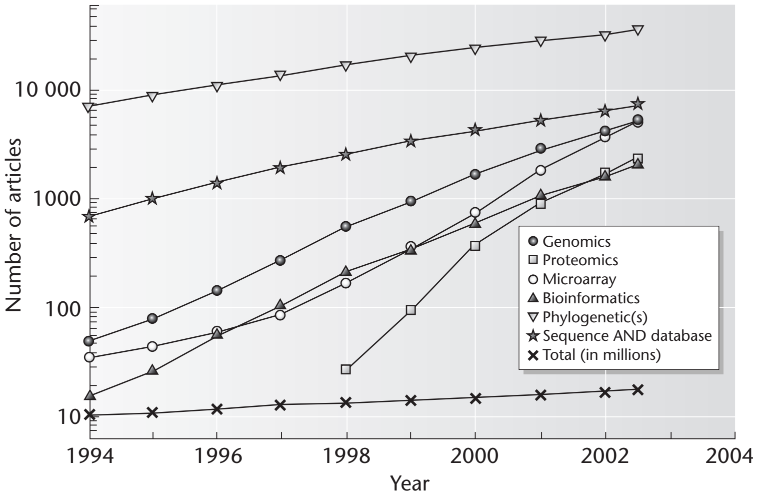


Fig. 1.3 Cumulative number of scientific articles published from 1981 to the date shown that use specific terms in the title, keywords, or abstract. Data from the Science Citation Index (SCI-EXPANDED) available at http://wos.mimas.ac.uk/ or http://isi6.isiknowledge.com.

become to the biological science community in the past decade. The number of papers using the term "phylogenetic" or "phylogenetics" dwarfs those using all the other terms considered here by at least an order of magnitude. This curve is a remarkably good exponential, although the doubling time is fairly long (3.7 years). Phylogenetics is a relatively old area, where morphological studies predate the availability of molecular sequences by several decades. The high level of interest in the field in recent years is largely a result of the availability of sequence data and of new methods for tree construction. Very large sequence data sets are now being used, and we are beginning to resolve some of the controversial aspects of evolutionary trees that have been argued over for decades.

As a comparison, for all the curves in Fig. 1.3 that refer to specific scientific terms, the figure also shows the total number of articles in the Science Citation Index (this curve is in millions of articles, whereas the others are in individual articles). The total level of scientific activity (or at least, scientific writing) has also been increasing exponentially, and hence we all have to read more and more in order to keep up. This curve is an almost perfect exponential, with a doubling time of 9.8 years. Thus, all the curves related to the individual subjects are increasing far more rapidly than the total accumulation of scientific knowledge.

At this point you will be suitably impressed by the importance of the subject matter of this book and will be eager to read the rest of it!

1.3 WHAT IS BIOINFORMATICS?

Since bioinformatics is still a fairly new field, people have a tendency to ask, "What is bioinformatics?" Often, people seem to worry that it is not very well defined, and tend to have a suspicious look in their eyes when they ask. These people would never trouble to ask "What is biology?" or "What is genetics?" In fact, bioinformatics is no more difficult or more easy to define than these other fields. Here is our short and simple definition.

> *Bioinformatics is the use of computational methods to study biological data.*

In case this is too short and too general for you, here is a longer one.

> *Bioinformatics is:*
> *(i) the development of computational methods for studying the structure, function, and evolution of genes, proteins, and whole genomes;*
> *(ii) the development of methods for the management and analysis of biological information arising from genomics and high-throughput experiments.*

If that is still too short, have another look at the contents list of this book to see what we think are the most important topics that make up the field of bioinformatics.

1.4 THE RELATIONSHIP BETWEEN POPULATION GENETICS, MOLECULAR EVOLUTION, AND BIOINFORMATICS

1.4.1 A little history . . .

The field of population genetics is concerned with the variation of genes within a population. The issues of natural selection, mutation, and random drift are fundamental to population genetics. Alternative versions of a gene are known as alleles. A large body of population genetics theory is used to interpret experimental data on allele frequency distributions and to ask questions about the behavior of the organisms being studied (e.g., effective population size, pattern of migration, degree of inbreeding). Population genetics is a well-established discipline with foundations dating back to Ronald Fisher and Sewall Wright in the first half of the twentieth century. These foundations predate the era of molecular sequences. It is possible to discuss the theory of the spread of a new allele, for example, without knowing anything about its sequence.

Molecular evolution is a more recent discipline that has arisen since DNA and protein sequence information has become available. Molecular techniques provide many types of data that are of great use to population geneticists, e.g., allozymes, microsatellites, restriction fragment length polymorphisms, single nucleotide polymorphisms, human mitochondrial haplotypes. Population geneticists are interested in what these molecular markers tell us about the organisms (see the many examples in the book by Avise 1994). In contrast, the focus of molecular evolution is on the molecules themselves, and understanding the processes of mutation and selection that act on the sequences. There are many genes that have now been sequenced in a large number of different species. This usually means that we have a representative example of a single gene sequence from each species. There are only a few species for which a significant amount of information about within-species sequence variation is available (e.g., humans and *Drosophila*). The emphasis in molecular evolution therefore tends to be on comparative molecular studies **between** species, while population genetics usually considers variation **within** a species.

The article by Zuckerkandl and Pauling (1965) is sometimes credited with inventing the field of molecular evolution. This was the first time that protein sequences were used to construct a molecular phylogeny and it set many people thinking about biological sequences in a **quantitative** way. 1965 was the same year in which Moore invented his law and in which computers were beginning to play a significant role in science. Indeed, molecular biology has risen to prominence in the biological sciences in the same time frame that computers have risen to prominence in society in general.

We might also argue that bioinformatics was beginning in 1965. The first edition of the *Atlas of Protein Sequence and Structure*, compiled by Margaret Dayhoff, appeared in printed form in 1965. The *Atlas* later became the basis for the PIR protein sequence database (Wu *et al.* 2002). However, this is stretching the point a little. The term bioinformatics was not around in 1965, and barring a few pioneers, bioinformatics was not an active research area at that time. As a discipline, bioinformatics is more recent. It arose from the recognition that efficient computational techniques were needed to study the huge amount of biological sequence information that was becoming available. If molecular biology arose at the same time as scientific computing, then we may also say that bioinformatics arose at the same time as the Internet. It is **possible** to imagine the existence of biological sequence databases without the Internet, but they would be a whole lot less useful. Database use would be restricted to those who subscribed to postal deliveries of database releases. Think of that cardboard box arriving each month and getting exponentially bigger each time. Amos Bairoch of the Swiss Institute of Bioinformatics comments (Bairoch 2000) that in 1988, the version of their PC/Gene database and software was shipped as 53 floppy disks! For that matter, think how difficult it would be to submit sequences to a database if it were not for email and the Internet.

At this point, the first author of this book starts to feel old. Coincidentally, I also first saw the light of day in 1965. Shortly afterwards, in 1985, I was happily writing programs with DO-loops in them for mainframes (students who are too young to know what mainframe computers are probably do not need to know). In 1989, someone first showed me how to use a mouse. I remember this clearly because I used the mouse for the first time when I began to write my Ph.D. thesis. It is scary to think almost all my education is pre-mouse. Possibly even more frightening is that I remember – it must have been in 1994 – someone explaining to our academic department how the World-Wide Web worked and what was meant by the terms URL and Netscape. A year or so after that, use of the Internet had become a daily affair for me. Now, of course, if the network is down for a day, it is impossible to do anything at all!

1.4.2 Evolutionary foundations for bioinformatics

Let's get back to the plot. Bioinformatics is a new discipline. Since this is a bioinformatics book, why do we need to know about the older subjects of molecular evolution and population genetics? There is a famous remark by the evolutionary biologist Theodosius Dobzhansky that, "Nothing in biology makes sense except in the light of evolution". You will find this quoted in almost every evolutionary textbook, but we will not apologize for quoting it once again. In fact, we would like to update it to, "Nothing in bioinformatics makes sense except in the light of evolution". Let's consider some examples to see why this is so.

The most fundamental and most frequently used procedure in bioinformatics is pairwise sequence alignment. When amino acid sequences are aligned, we use a scoring system, such as a PAM matrix, to determine the score for aligning any amino acid with any other. These scoring systems are based on evolutionary models. High scores are assigned to pairs of amino acids that frequently substitute for one another during protein sequence evolution. Low, or negative, scores are assigned to pairs of amino acids that interchange very rarely. When RNA sequences are aligned, we often use the fact that the secondary structure tends to be conserved, and that pairs of compensatory substitutions occur in the base-paired regions of the structure. Thus, creating accurate sequence alignments of both proteins and RNAs relies on an understanding of molecular evolution.

If we want to know something about a particular biological sequence, the first thing we do is search the database to find sequences that are similar to it. In particular, we are often interested in sequence motifs that are well conserved and that are present in a whole family of proteins. The logic is that important parts of a sequence will tend to be conserved during evolution. Protein family databases like PROSITE, PRINTS, and InterPro (see Chapter 5) identify important conserved motifs in protein alignments and use them to assign sequences to families. An important concept here is **homology**. Sequences are homologous if they descend from a common ancestor, i.e., if they are related by the evolutionary process of divergence. If a group of proteins all share a conserved motif, it will often be because all these proteins are homologous. If a motif is very short, however, there is some chance that it will have evolved more than once independently (**convergent evolution**). It is therefore important to try to distinguish chance similarities arising from convergent evolution from similarities arising from divergent evolution. The thrust of protein family databases is therefore to facilitate the identification of true homologs, by making the distinction between chance and real matches clearer.

Similar considerations apply in protein structural databases. It is often observed that distantly related proteins have relatively conserved structures. For example, the number and relative positions of α helices and β strands might be the same in two proteins that have very different sequences. Occasionally, the sequences are so different that it would be very difficult to establish a relationship between them if the structure were not known. When similar (or identical) structures are found in different proteins, it probably indicates homology, but the possibility of small structural motifs arising more than once still needs to be considered. Another important aspect of protein structure that is strongly linked to evolution is **domain shuffling**. Many large proteins are composed of smaller domains that are continuous sections of the sequence that fold into fairly well-defined three-dimensional structures; these assemble to form the overall protein structure. Particularly in eukaryotes, it is found that certain domains occur in many different proteins in different combinations and different orders. See the ProDom database (Corpet *et al.* 2000), for example. Although bioinformaticians will argue about what constitutes a domain and where the boundaries between domains lie, it is clear that the duplication and reshuffling of domains is a very useful way of evolving new complex proteins. The main message is that in order to create reliable information resources for protein sequences, structures, and domains, we need to have a good understanding of protein evolution.

In recent years, evolutionary studies have also become possible at the whole genome level. If we want to compare the genomes of two species, it is natural to ask which genes are shared by both species. This question can be surprisingly hard to answer. For each gene in the first species, we need to decide if there is a gene in the second species that is homologous to it. It may be difficult to detect similarity between sequences from different species simply because of the large amount of evolutionary change that has gone on since the divergence of the species. Most genomes contain many open reading frames that are thought to be genes, but for which no similar sequence can be found in other species. This is evidence for the limitations of our current methods as much as for the diversity generated by molecular evolution. In cases where we **are** able to detect similarity, then it can still be tough to decide which genes are homologous. Many genomes contain families of duplicated genes that often have slightly different functions or different sites of expression within the organism. Sequences from one species that are evolutionarily related and that diverged from one another due to a gene duplication event are called paralogous sequences, in contrast to orthologous sequences, which are sequences in different organisms that diverged from one another due to the split between the species. Duplications can occur in different lineages independently, so that a single gene in one species might be homologous to a whole family in the other species. Alternatively, if duplications occurred in a common ancestor, then both species should contain a copy of each member of the gene family – unless, of course, some genes have been deleted in one or other species. Another factor to consider, particularly for bacteria, is that genomes can acquire new genes by horizontal transfer of DNA from unrelated species. This sequence comparison can show up genes that are apparently homologous to sequences in organisms that are otherwise thought to be extremely distantly related. A major task for bioinformatics is to establish sets of homologous genes between groups of species, and to understand how those genes got to be where they are. The flip side of this is to be able to say which genes are **not** present in an organism, and how the organism manages to get by without them.

The above examples show that many of the questions addressed in bioinformatics have foundations in questions of molecular evolution. A fair amount of this book is therefore devoted to molecular evolution. What about population genetics? There are many other books on population genetics and hence this book does not try to be a textbook of this area. However, there are some key points that are usually considered in population genetics courses that we need to consider if we are to properly understand molecular evolution and bioinformatics. These questions concern the way in which sequence diversity is generated in populations and the way in which new variant sequences spread through populations. If we run a molecular phylogeny program, for example, we might be asking whether “the” sequence from humans is more similar to “the” sequence from chimpanzees or gorillas. It is important to remember that these sequences have diverged as a result of the fixation of new sequence variants in the populations. We should also not forget that the sequences we have are just samples from the variations that exist in each of the populations.

There are some bioinformatics areas that have a direct link to the genetics of human populations. We are accumulating large amounts of information about variant gene sequences in human populations, particularly where these are linked to hereditary diseases. Some of these can be major changes, like deletions of all or part of a gene or a chromosome region. Some are single nucleotide polymorphisms, or SNPs, where just a single base varies at a particular site in a gene. Databases of SNPs potentially contain information of great relevance to medicine and to the pharmaceutical industry. The area of **pharmacogenomics** attempts to understand the way that different patients respond more or less well to drug treatments according to which alleles they have for certain genes. The hope is that drug treatments can be tailored to suit the genetic profile of the patient. However, many important diseases are not caused by a single gene. Understanding the way that variations at many different loci combine to affect the susceptibility of individuals to different medical problems is an important goal, and developing computational techniques to handle data such as SNPs, and to extract information from the data, is an important application of bioinformatics.

SUMMARY

The amount of biological sequence information is increasing very rapidly and seems to be following an exponential growth law. Computational methods are playing an increasing role in biological sciences. New algorithms will be required to analyze this information and to understand what it means. Genome sequencing projects have been remarkably successful, and comparative analysis of whole genomes is now possible. This provides challenges and opportunities for new types of study in bioinformatics. At the same time, several types of experimental methods are being developed currently that may be classed as "high-throughput". These include microarrays, proteomics, and structural genomics. The philosophy behind these methods is to study large numbers of genes or proteins simultaneously, rather than to specialize in individual cases. Bioinformatics therefore has a role in developing statistical methods for analysis of large data sets, and in developing methods of information management for the new types of data being generated.

Evolutionary ideas underlie many of the methods used in bioinformatics, such as sequence alignments, identifying families of genes and proteins, and establishing homology between genes in different organisms. Evolutionary tree construction (i.e., molecular phylogenetics) is itself a very large field within computational biology. Since we now have many complete genomes, particularly in bacteria, we can also begin to look at evolutionary questions at the whole-genome level. This book will therefore pay particular attention to the evolutionary aspects of bioinformatics.

REFERENCES

Avise, J.C. 1994. *Molecular Markers, Natural History, and Evolution*. New York: Chapman and Hall.

Bairoch, A. 2000. Serendipity in bioinformatics: The tribulations of a Swiss bioinformatician through exciting times. *Bioinformatics*, **16**: 48–64.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2003. GenBank. *Nucleic Acids Research*, **31**: 23–7.

Corpet, F., Sernat, F., Gouzy, J., and Kahn, D. 2000. ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research*, **28**: 267–9. http://prodes.toulouse.inra.fr/prodom/2002.1/html/home.php

Miyazaki, S., Sugawara, H., Gojobori, T., and Tateno, Y. 2003. DNA Databank of Japan (DDBJ) in XML. *Nucleic Acids Research*, **31**: 13–16.

Moore, G.E. 1965. Cramming more components onto integrated circuits. *Electronics*, **38**(8): 114–17.

Stoesser, G., Baker, W., van den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., Nardone, R., Stoehr, P., Tuli, M.A., Tzouvara, K., and Vaughan, R. 2003. The EMBL Nucleotide Sequence Database: Major new developments. *Nucleic Acids Research*, **31**: 17–22.

Wu, C.H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z.Z., Ledley, R.S., Lewis, K.C., Mewes, H.W., Orcutt, B., Suzek, B.E., Tsugita, A., Vinayaka, C.R., Yeh, L.L., Zhang, J., and Barker, W.C. 2002. The Protein Information Resource: An integrated public resource of functional annotation of proteins. *Nucleic Acids Research*, **30**: 35–7. http://pir.georgetown.edu/.

Zuckerkandl, E. and Pauling, L. 1965. Evolutionary divergence and convergence in proteins. In V. Bryson, and H.J. Vogel (eds.), *Evolving Genes and Proteins*, pp. 97–166. New York: Academic Press.

PROBLEMS

1.1 The data explosion curves provide us with a good way of revising some fundamental points in mathematics that will come in handy later in the book. Now would also be a good time to check the Mathematical appendix and maybe have a go at the Self-test that goes with it.

For each type of data considered in this chapter, we have a quantity $N(t)$ that is increasing with time t, and we are assuming that it follows the law:

$$N(t) = N_0 \exp(rt)$$

Here, N_0 is the value at the initial time point, and r is the growth rate. In the cases we considered, time was measured in years. We defined the yearly multiplication factor as the factor by which N increases each year, i.e.

$$R = \frac{N(t+1)}{N(t)} = \exp(r)$$

If the data really follow an exponential law, then this ratio is the same at whichever time we measure it. Another way of writing the growth law is therefore:

$$N(t) = N_0 R^t$$

The other useful quantity that we measured was the doubling time T. To calculate T we require that the number after a time T is twice as large as its initial value.

$$\frac{N(T)}{N_0} = \exp(rT) = 2$$

Hence $rT = \ln(2)$ or $T = \ln(2)/r$.

If any of these steps is not obvious, then you should revise your knowledge of exponentials and logarithms. There are some helpful pointers in the Mathematical appendix of this book, Section M.1.

1.2 Use the data from Tables 1.1 and 1.2 and plot your own graphs. The figures in this chapter plot N directly against t and use a logarithmic scale on the vertical axis. This comes out to be a straight line because:

$$\ln(N) = \ln(N_0) + rt$$

so the slope of the line is r. The other way to do it is to calculate $\ln(N)$ at each time point with a calculator, and then to plot $\ln(N)$ against t using a linear scale on both axes. Plot the graphs both ways and make sure they look the same.

My graph-plotting package will do a best fit of an exponential growth law to a set of data points. This is how the values of r were obtained in Table 1.3. However, if your package will not do that, then you can also estimate r from the $\ln(N)$ versus t graph by using a straight-line fit. Try doing the fit to the data in both ways and make sure that you get the same answer.

1.3 The exponential growth law arises from the assumption that the rate of increase of N is proportional to its current value. Thus the growth law is the solution of the differential equation

$$\frac{dN}{dt} = rN$$

Now would be a good time to make sure you understand what this equation means (see Sections M.6 and M.8 for some help).

1.4 While the assumption in 1.3 might have some plausibility for the increase in the size of a rabbit population (if they have a limitless food supply), there does not seem to be a theoretical reason why the size of GenBank or the size of a PC chip should increase exponentially. It is just an empirical observation that it works that way. Presumably, sooner or later all these curves will hit a limit.

There are several other types of curve we might imagine to describe an increasing function of time.

Linear increase:	$N(t) = A + Bt$
Power law increase:	$N(t) = At^k$ (for some value of k not equal to 1)
Logarithmic increase:	$N(t) = A + B\ln(t)$

In each case, A, B, and k are arbitrary constants that could be obtained by fitting the curve to the data. Try to fit the data in Tables 1.1 and 1.2 to these other growth laws. Is it true that the exponential growth law fits better than the alternatives?

If you had some kind of measurements that you believed followed one of the other growth laws, how would you plot the graph so that the points would lie on a straight line?

Nucleic acids, proteins, and amino acids

CHAPTER 2

CHAPTER PREVIEW

This chapter begins with a basic introduction to the chemical and physical structure of nucleic acids and proteins for those who do not have a background in biochemistry or biology. We also give a reminder of the processes of transcription, RNA processing, translation and protein synthesis, and DNA replication. We then give a detailed discussion of the physico-chemical properties of the amino acids and their relevance in protein folding. We use the amino acid property data as an example when introducing two statistical techniques that are useful in bioinformatics: principal component analysis and clustering algorithms.

2.1 NUCLEIC ACID STRUCTURE

There are two types of nucleic acid that are of key importance in cells: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The chemical structure of a single strand of RNA is shown in Fig. 2.1. The backbone of the molecule is composed of ribose units (five-carbon sugars) linked by phosphate groups in a repeating polymer chain. Two repeat units are shown in the figure. The carbons in the ribose are conventionally numbered from 1 to 5, and the phosphate groups are linked to carbons 3 and 5. At one end, called the 5′ ("five prime") end, the last carbon in the chain is a number 5 carbon, whereas at the other end, called the 3′ end, the last carbon is a number 3. We often think of a strand as beginning at the 5′ end and ending at the 3′ end, because this is the direction in which genetic information is read. The backbone of DNA differs in that deoxyribose sugars are used instead of ribose. The OH group on carbon number 2 in ribose is simply an H in deoxyribose, but the molecules are otherwise the same.

Each sugar is linked to a molecule known as a base. In DNA, there are four types of base, called adenine, thymine, guanine, and cytosine, usually referred to simply as A, T, G, and C. The structures of these molecules are shown in Fig. 2.2. In RNA, the base uracil (U; Fig. 2.2) occurs instead of T. The structure of U is similar to that of T but lacks the CH_3 group linked to the ring of the T molecule. In addition, a variety of bases of slightly different structures, called modified bases, can also be found in some types of RNA molecule. A and G are known as purines. They both have a double ring in their chemical structure. C, T, and U are known as pyrimidines. They have a single ring in their chemical structure. The fundamental building block of nucleic acid chains is called a nucleotide: this is a unit of one base plus one sugar plus one phosphate. We usually think of the "length" of a nucleic acid sequence as the number of nucleotides in the chain. Nucleotides are also found as separate molecules in the cell, as well as being part of nucleic acid polymers. In this case, there are usually two or three phosphate groups attached to the same nucleotide. For example, ATP (adenosine triphosphate) is an important molecule in cellular metabolism, and it has three phosphates attached in a chain.

DNA is usually found as a double strand. The two strands are held together by hydrogen bonding

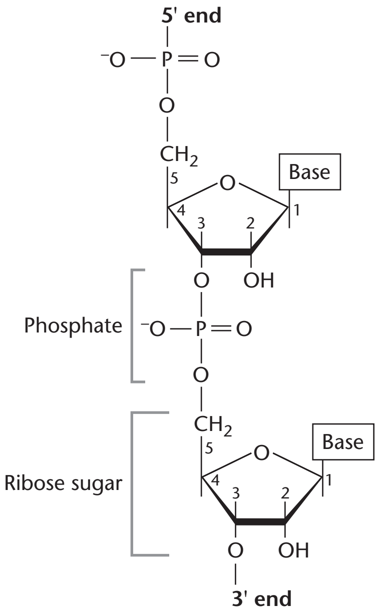


Fig. 2.1 Chemical structure of the RNA backbone showing ribose units linked by phosphate groups.

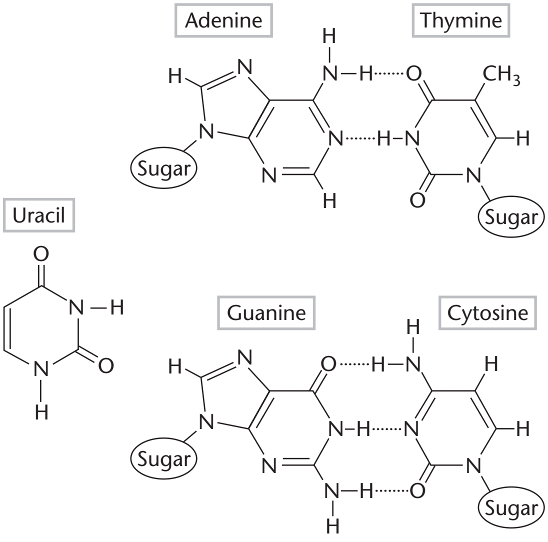


Fig. 2.2 The chemical structure of the four bases of DNA showing the formation of hydrogen-bonded AT and GC base pairs. Uracil is also shown.

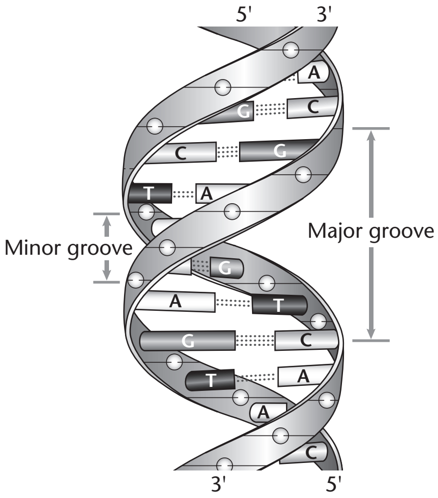


Fig. 2.3 Schematic diagram of the DNA double helical structure.

between A and T and between C and G bases (Fig. 2.2). The two strands run in opposite directions and are exactly complementary in sequence, so that where one has A, the other has T and where one has C the other has G. For example:

5′–C–C–T–G–A–T–3′
3′–G–G–A–C–T–A–5′

The two strands are coiled around one another in the famous double helical structure elucidated by Watson and Crick 50 years ago. This is shown schematically in Fig. 2.3.

In contrast, RNA molecules are usually single stranded, and can form a variety of structures by base pairing between short regions of complementary sequences within the same strand. An example of this is the cloverleaf structure of transfer RNA (tRNA), which has four base-paired regions (stems) and three hairpin loops (Fig. 2.4). The base-pairing rules in RNA are more flexible than DNA. The principal pairs are GC and AU (which is equivalent to AT in DNA), but GU pairs are also relatively frequent, and a variety of unusual, so-called “non-canonical”, pairs are also found in some RNA structures (e.g., GA pairs). A two-dimensional drawing of the base-pairing pattern is called a secondary structure

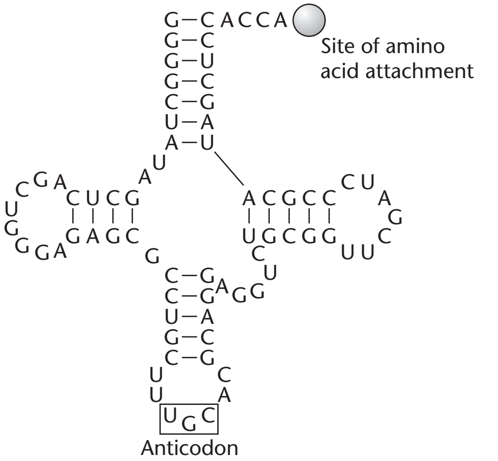


Fig. 2.4 Secondary structure of tRNA-Ala from *Escherichia coli* showing the anticodon position and the site of amino acid attachment.

diagram. In Chapter 11, we will discuss RNA secondary structure in more detail.

2.2 PROTEIN STRUCTURE

The fundamental building block of proteins is the amino acid. An amino acid has the chemical structure shown in Fig. 2.5(a), with an amine group on one side and a carboxylic acid group on the other. In solution, these groups are often ionized to give NH_3^+ and COO^-. There are 20 types of amino acid found in proteins. These are distinguished by the nature of the side-chain group, labeled R in Fig. 2.5(a). The central carbon to which the R group is attached is known as the α carbon. Proteins are linear polymers composed of chains of amino acids. The links are formed by removal of an OH from one amino acid and an H from the next to give a water molecule. The resultant linkage is called a peptide bond. These are shown in boxes in Fig. 2.5(b), which illustrates a tripeptide, i.e., a chain composed of three amino acids. Proteins, or “polypeptides”, are typically composed of several hundred amino acids.

The chemical structures of the side chains are given in Fig. 2.6. Each amino acid has a standard three-letter abbreviation and a standard one-letter code, as shown in the figure. A protein can be represented simply by a sequence of these letters, which is very convenient for storage on a computer, for example:

```
MADIQLSKYHVSKDIGFLLEPLQDVLPDYFAPWNR
LAKSLPDLVASHKFRDAVKEMPLLDSSKLAGYRQK
```

is the first part of a real protein. The two ends of a protein are called the N terminus and the C terminus because one has an unlinked NH_3^+ group and the other has an unlinked COO^- group. Protein sequences are traditionally written from the N to the C terminus, which corresponds to the direction in which they are synthesized.

The four atoms involved in the peptide bond lie in a plane and are not free to rotate with respect to one another. This is due to the electrons in the chemical bonds, which are partly delocalized. The flexibility of the protein backbone comes mostly from rotation about the two bonds on either side of each α carbon. Many proteins form globular three-dimensional structures due to this flexibility of the backbone. Each protein has a structure that is specific to its sequence. The formation of this three-dimensional structure is called “protein folding”. The amino acids vary considerably in their properties. The combination of repulsive and attractive interactions between the different amino acids, and between the amino acids and water, determines the way in which a protein folds. An important role of proteins is as catalysts of

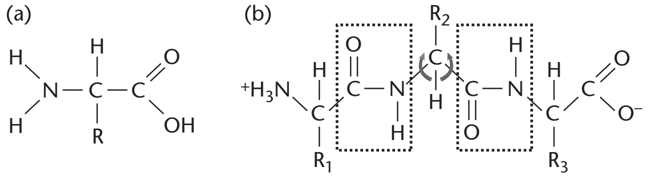


Fig. 2.5 Chemical structure of an amino acid (a) and the protein backbone (b). The peptide bond units (boxed) are planar and inflexible. Flexibility of the backbone comes from rotation about the bonds next to the α carbons (indicated by arrows).