Nucleic acids, proteins, and amino acids

CHAPTER 2

CHAPTER PREVIEW

This chapter begins with a basic introdu[illegible] structure of nucleic acids and proteins [illegible] in biochemistry or biology. We also give a [illegible] tion, RNA processing, translation [illegible] We then give a detailed discussion [illegible] amino acids and then [illegible] property data as an example [illegible] are useful in bioinformatics [illegible] 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

PTER

2

:he molecules are
he same.
ʒar is linked to a
aown as a base. In
: are four types of
adenine, thymine,
nd cytosine, usu-
l to simply as A, T,
The structures of
:ules are shown in
n RNA, the base
of T. The structure
cks the CH_3 group
:ule. In addition, a
t structures, called
d in some types of
n as purines. They
hemical structure.
dines. They have a
ıcture. The funda-
ıcid chains is called
ɔase plus one sugar
iink of the "length"
: number of nucle-
are also found as
vell as being part of
e, there are usually
tached to the same
adenosine triphos-
n cellular metabol-
ttached in a chain.
louble strand. The
hydrogen bonding

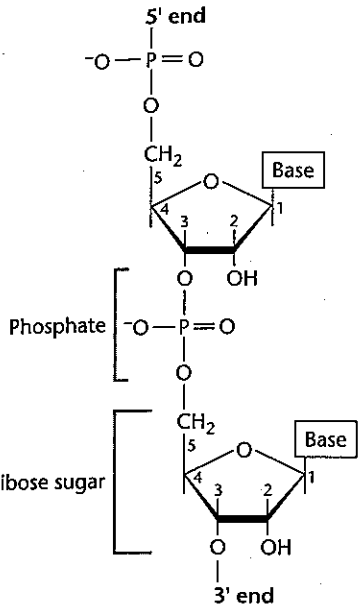


Fig. 2.1 Chemical structure of the RNA backbone showing ribose units linked by phosphate groups.

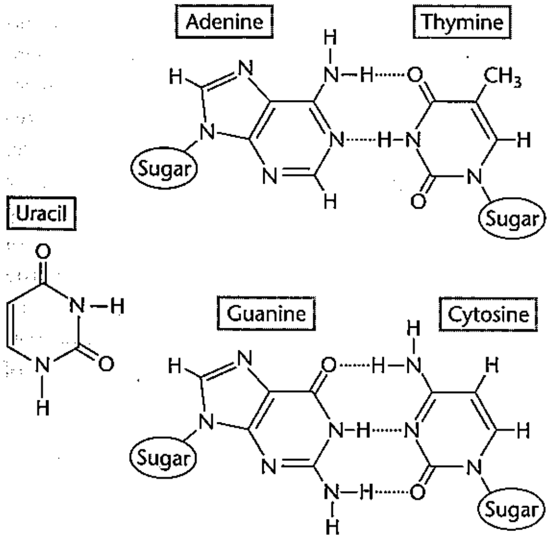


Fig. 2.2 The chemical structure of the four bases of DNA showing the formation of hydrogen-bonded AT and GC base pairs. Uracil is also shown.

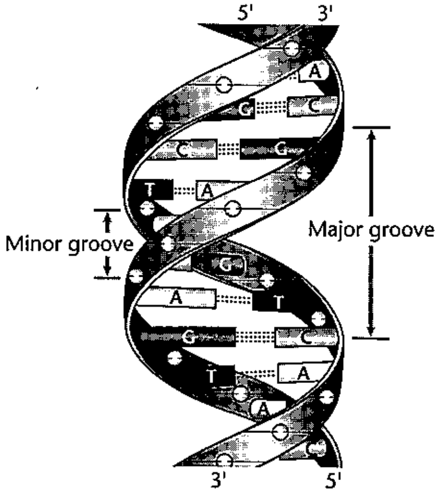


Fig. 2.3 Schematic diagram of the DNA double helical structure.

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

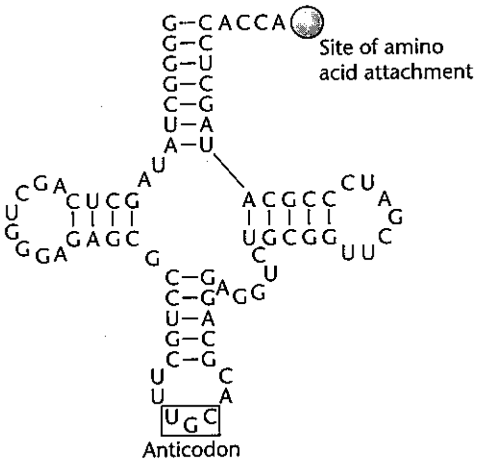


Fig. 2.4 Secondary structure of tRNA-Ala from *Escherichia coli* showing the anticodon position and the site of amino acid attachment.

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

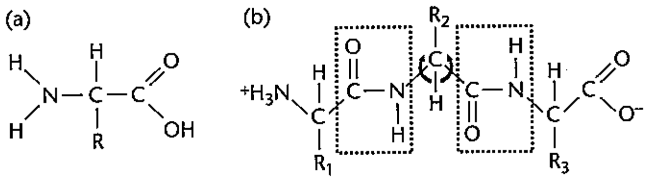


Fig. 2.5 Chemical structure of an amino acid (a) and the protein backbone (b). The peptide bond units (boxed) are planar and inflexible. Flexibility of the backbone comes from rotation about the bonds next to the α carbons (indicated by arrows).

;ptide bond. These are
(b), which illustrates
posed of three amino
es", are typically com-
.o acids.
ie side chains are given
has a standard three-
ıdard one-letter code,
ein can be represented
letters, which is very
mputer, for example:

'LQDVLPDYFAPWNR
IPLLDSSKLAGYRQK

in. The two ends of a
us and the C terminus
l NH_3^+ group and the
group. Protein sequ-
n from the N to the C
s to the direction in

the peptide bond lie in
ate with respect to one
ctrons in the chemical
lized. The flexibility of
mostly from rotation
side of each α carbon.
ree-dimensional struc-
ie backbone. Each pro-
pecific to its sequence.
limensional structure
he amino acids vary
s. The combination of
ractions between the
ween the amino acids
ay in which a protein
oteins is as catalysts of

mical structure of an
) and the protein
The peptide bond units
anar and inflexible.
he backbone comes from
t the bonds next to the α
cated by arrows).

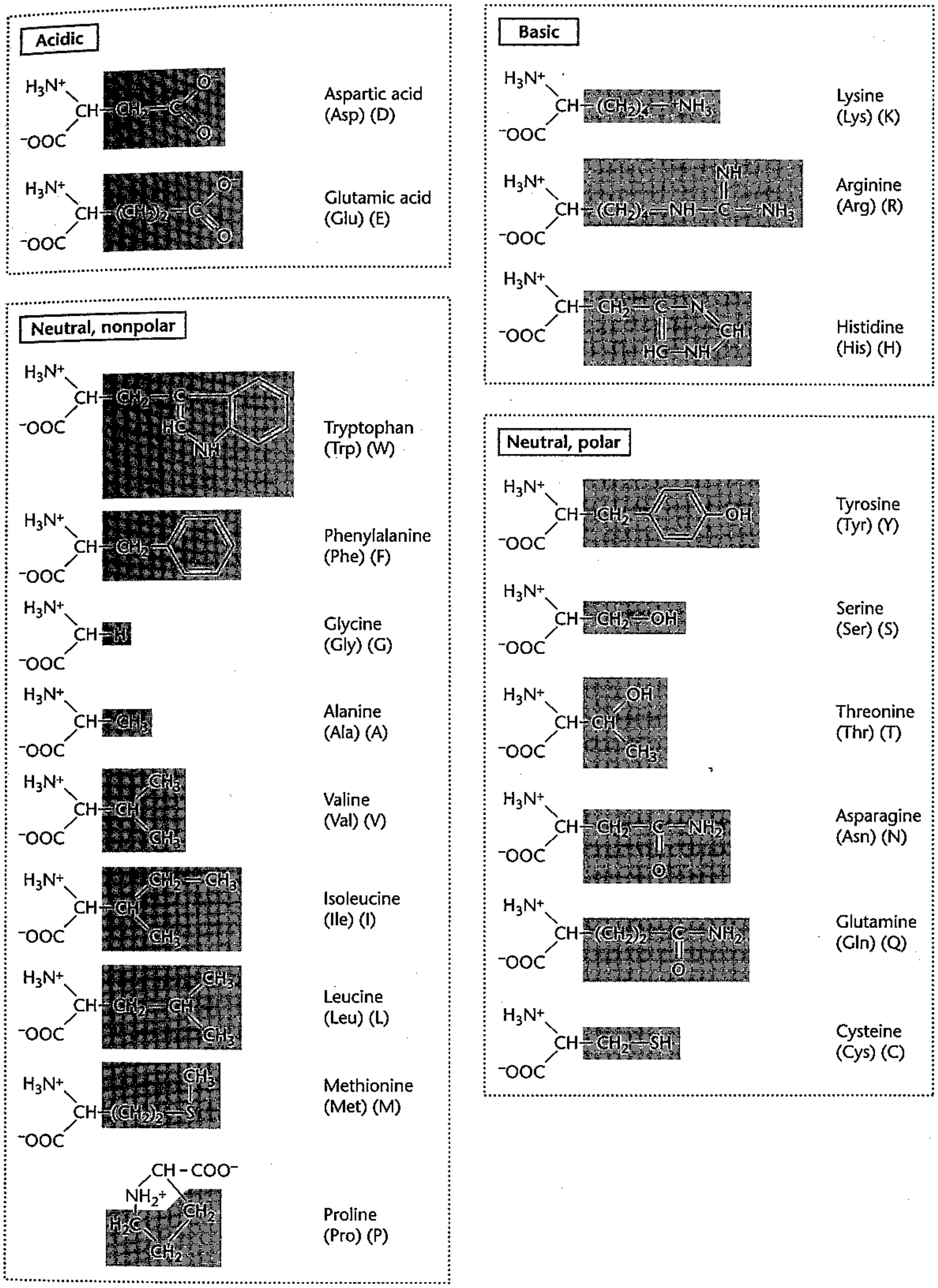


Fig. 2.6 Chemical structures of the 20 amino acid side chains.

biochemical reactions – protein catalysts are called enzymes. Proteins are able to catalyze a huge variety of chemical reactions due to the wide range of different side groups of the amino acids and the vast number of sequences that can be assembled by combining the 20 building blocks in specific orders.

Many protein structures are known from X-ray crystallography. Plate 2.1(a) illustrates the interaction between a glutaminyl-tRNA molecule and a protein called glutaminyl-tRNA synthetase. We will discuss the functions of these molecules below when we describe the process of translation and protein synthesis. At this point, the figure is a good example of the three-dimensional structure of both RNAs and proteins. Rather than drawing all the atomic positions, the figure is drawn at a coarse-grained level so that the most important features of the structure are visible. The backbone of the RNA is shown in yellow, and the bases in the RNA are shown as colored rectangular blocks. The tRNA has the cloverleaf structure with the four stems shown in Fig. 2.4. In three dimensions, the molecule is folded into an L-shaped globule. The anticodon loop (see Section 2.3.4) at the bottom of Fig. 2.4 is the top loop in Plate 2.1(a). The stem regions in the secondary structure diagram are the short sections of the double helix in the three-dimensional structure.

The backbone of the protein structure is shown as a purple ribbon in Plate 2.1(a). Two characteristic aspects of protein structures are visible in this example. In many places, the protein backbone forms a helical structure called an α helix. The helix is stabilized by hydrogen bonds between the CO group in the peptide link and the NH in another peptide link further down the chain. These hydrogen bonds are roughly parallel to the axis of the helix. The side groups of the amino acids are pointing out perpendicular to the helix axis. For more details, see a textbook on protein structure, e.g., Creighton (1993). The other features visible in the purple ribbon diagram of the protein are β sheets. The strands composing the sheets are indicated by arrows in the ribbon diagram that point along the backbone from the N to the C terminus. A β sheet consists of two or more strands that are folded to run more or less side by side with one another in either parallel or antiparallel orientations. The strands are held together by hydrogen bonds between CO groups on one strand and NH groups on the neighboring strand. The α helices and β sheets in a protein are called elements of secondary structure (whereas the secondary structure of an RNA sequence refers to the base-paired stem regions).

Plate 2.1(b) gives a second example of a molecular structure. This shows a dimer of the lac repressor protein bound to DNA. The *lac* operon in *Escherichia coli* is a set of genes whose products are responsible for lactose metabolism. When the repressor is bound to the DNA, the genes in the operon are turned off (i.e., not transcribed). This happens when there is no lactose in the growth medium. This is an example of a protein that can recognize and bind to a specific sequence of DNA bases. It does this without separating the two strands of the DNA double helix, as it is able to bind to the "sides" of the DNA bases that are accessible in the grooves of the double helix. The binding of proteins to DNA is an important way of controlling the expression of many genes, allowing genes to be turned on in some cells and not others.

Plates 2.1(a) and (b) give an idea of the relative sizes of proteins and nucleic acids. The protein in Plate 2.1(a) has 548 amino acids, which is probably slightly larger than the average globular protein. The tRNA has 72 nucleotides. Most RNAs are much longer than this. The diameter of an α helix is roughly 0.5 nm, whereas the diameter of a nucleic acid double helix is roughly 2 nm.

2.3 THE CENTRAL DOGMA

2.3.1 *Transcription*

There is a principle, known as the central dogma of molecular biology, that information passes from DNA to RNA to proteins. The process of going from DNA to RNA is called transcription, while the process of going from RNA to protein is called translation. Synthesis of RNA involves simply rewriting (or transcribing) the DNA sequence in the same language of nucleotides, whereas synthesis of proteins involves translating from the language of nucleotides to the language of amino acids.

ıre held together
groups on one
ghboring strand.
ırotein are called
whereas the sec-
ence refers to the

ple of a molecular
the lac repressor
ron in *Escherichia*
ts are responsible
epressor is bound
on are turned off
; when there is no
s is an example of
bind to a specific
; without separat-
uble helix, as it is
NA bases that are
louble helix. The
important way of
y genes, allowing
and not others.
lea of the relative
s. The protein in
which is probably
globular protein.
ıt RNAs are much
of an α helix is
neter of a nucleic

IA

central dogma of
tion passes from
cess of going from
on, while the pro-
ı is called transla-
nply rewriting (or
in the same lan-
ıthesis of proteins
ıage of nucleotides

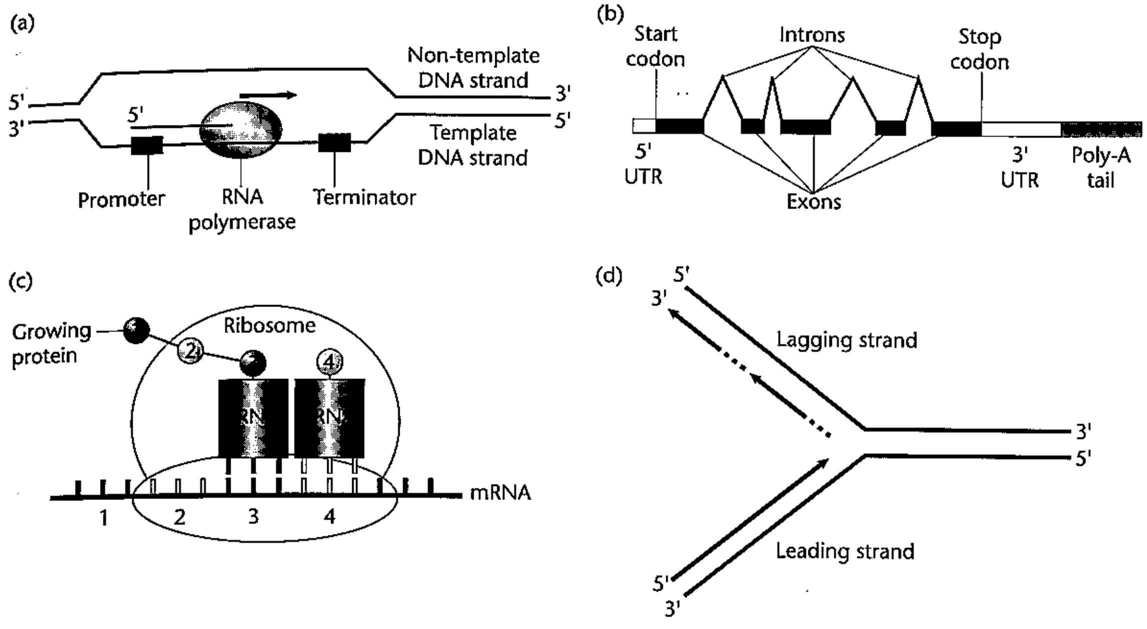


Fig. 2.7 Four important mechanisms. (a) Transcription. (b) Structure and processing of prokaryotic mRNA. (c) Translation. (d) DNA replication.

DNA is usually found in cells in very long pieces. For example, the genomes of most bacteria are circular loops of DNA a few million base pairs (Mbp) long, while humans have 23 pairs of linear chromosomes, with lengths varying from 19 to 240 Mbp. Genes are regions of DNA that contain the information necessary to make RNA and protein molecules. Transcription is the process of synthesis of RNA using DNA as a template. Typically sections of DNA a few thousand base pairs long are transcribed that correspond to single genes (or sometimes a small number of sequential genes). Transcription is carried out by an enzyme called RNA polymerase. The RNA polymerase binds to one of the two strands of DNA that are temporarily separated from one another during the transcription process (see Fig. 2.7(a)). This strand is called the template strand. The polymerase catalyzes the assembly of individual ribonucleotides into an RNA strand that is complementary to the template DNA strand. Base pairing occurs between the template strand and the growing RNA strand initially, but as the polymerase moves along, the RNA separates from the template and the two DNA strands close up again. When the template is a C, G, or T, the base added to RNA is a G, C, or A, as usual. If the template has an A, then a U base is added to the RNA rather than a T. Since the RNA is complementary to the template strand, it is actually **the same** as the **non-template** strand, with the exception that Ts are converted to Us. When people talk about the DNA sequence of a gene, they usually mean the non-template strand, because it is this sequence that is the same as the RNA, and this sequence that is subsequently translated into the protein sequence.

RNA polymerase moves along the template from the 3′ to the 5′ end, hence the RNA is synthesized from its 5′ end to its 3′ end. The polymerase needs to know where to stop and start. This information is contained in the DNA sequence. A promoter is a short sequence of DNA bases that is recognized as a start signal by RNA polymerase. For example, in *E. coli*, most promoters have a sequence TATAAT about 10 nucleotides before the start point of transcription and a sequence TTGACA about 35 nucleotides before the start. However, these sequences are not fixed, and there is considerable variation between genes. Since promoters are relatively short and relatively variable in sequence, it is actually quite a hard problem to write a computer program to reliably locate them.

The stop signal, or terminator, for RNA polymerase is often in the form of a specific sequence that has the ability to form a hairpin loop structure in the RNA sequence. This structure delays the progression of the polymerase along the template and causes it to dissociate. We still have a lot more to learn about these signals.

2.3.2 *RNA processing*

An RNA strand that is transcribed from a protein-coding region of DNA is called a messenger RNA (mRNA). The mRNA is used as a template for protein synthesis in the translation process discussed below. In prokaryotes, mRNAs consist of a central coding sequence that contains the information for making the protein and short untranslated regions (UTRs) at the 5′ and 3′ ends. The UTRs are parts of the sequence that were transcribed but will not be translated.

In eukaryotes, the RNA transcript has a more complicated structure (Fig. 2.7(b)). When the RNA is newly synthesized, it is called a pre-mRNA. It must be processed in several ways before it becomes a functional mRNA. At the 5′ end, a structure known as a cap is added, which consists of a modified G nucleotide and a protein complex called the cap-binding complex. At the 3′ end, a poly-A tail is added, i.e., a string of roughly 200 A nucleotides. Proteins called poly-A binding proteins bind to the poly-A tail. Many mRNAs have a rather short lifetime (a few minutes) in the cell because they are broken down by nuclease enzymes. These are proteins that break down RNA strands into individual nucleotides, either by chopping them in the middle (endonucleases), or by eating them up from the end one nucleotide at a time (exonucleases). Having proteins associated with the mRNA, particularly at the ends, slows down the nucleases. Variation in the types of binding protein on different mRNAs is an important way of controlling mRNA lifetimes, and hence controlling the amount of protein synthesized by limiting the number of times an mRNA can be used in translation.

Probably the most important type of RNA processing occurs in the middle of the pre-mRNA rather than at the ends. Eukaryotic gene sequences are broken up into alternating sections called exons and introns. Exons are the pieces of the sequence that contain the information for protein coding. These pieces will be translated. Introns do not contain protein-coding information. The introns, indicated by the inverted Vs in Figure 2.7(b), are cut out of the pre-mRNA and are not present in the mRNA after processing. When an intron is removed, the ends of the exons on either side of it are linked together to form a continuous strand. This is known as splicing.

Splicing is carried out by the spliceosome, a complex of several types of RNA and proteins bound together and acting as a molecular machine. The spliceosome is able to recognize signals in the pre-mRNA sequence that tell it where the intron–exon boundaries are and hence which bits of the sequence to remove. As with promoter sequences, the signals for the splice sites are fairly short and somewhat variable, so that reliable identification of the intron–exon structure of a gene is a difficult problem in bioinformatics. Nevertheless, the spliceosome manages to do it.

Introns that are spliced out by the spliceosome are called spliceosomal introns. This is the majority of introns in most organisms. In addition, there are some interesting, but fairly rare, self-splicing introns, which have the ability to cut themselves out of an RNA strand without the action of the spliceosome. There are surprisingly large numbers of introns in many eukaryotic genes – 10 or 20 in one gene is not uncommon. In contrast, most prokaryotic genes do not contain introns. It is still rather controversial where and when introns appeared, and what is the use, if any, of having them.

In eukaryotes, the DNA is contained in the nucleus, and transcription and RNA processing occur in the nucleus. The mRNA is then transported out of the nucleus through pores in the nuclear membrane, and translation occurs in the cytoplasm.

ng mRNA lifetimes, and
ınt of protein synthesized
times an mRNA can be

tant type of RNA process-
ɔf the pre-mRNA rather
otic gene sequences are
ıg sections called exons
e pieces of the sequence
tion for protein coding.
ıslated. Introns do not
formation. The introns,
Vs in Figure 2.7(b), are
nd are not present in the
ıen an intron is removed,
ither side of it are linked
ous strand. This is known

the spliceosome, a com-
NA and proteins bound
molecular machine. The
ɔgnize signals in the pre-
t where the intron–exon
vhich bits of the sequence
r sequences, the signals for
:t and somewhat variable,
ı of the intron–exon struc-
roblem in bioinformatics.
me manages to do it.
ut by the spliceosome are
s. This is the majority of
s. In addition, there are
rare, self-splicing introns,
cut themselves out of an
ction of the spliceosome.
ge numbers of introns in
0 or 20 in one gene is not
ıost prokaryotic genes do
still rather controversial
ppeared, and what is the

; contained in the nucleus,
\ processing occur in the
:hen transported out of
in the nuclear membrane,
ıe cytoplasm.

Table 2.1 The standard genetic code. This is used in most prokaryotic genomes and in the nuclear genomes of most eukaryotes.

[illegible]	Second position				[illegible]
	U	C	A	[illegible]	
[illegible]	Phe (F) Phe (F) Leu (L) Leu (L)	Ser (S) Ser (S) Ser (S) Ser (S)	Tyr (Y) Tyr (Y) STOP [illegible]	[illegible]	[illegible]
[illegible]	Leu (L) Leu (L) Leu (L) [illegible]	Pro (P) [illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

2.3.3 *The genetic code*

We now need to consider the way information in the form of sequences of four types of base is turned into information in the form of sequences of 20 types of amino acid. The mRNA sequence is read in groups of three bases called codons. There are $4^3 = 64$ codons that can be made with four bases. Each of these codons codes for one type of amino acid, and since 64 is greater than 20, most amino acids have more than one codon that codes for them. The set of assignments of codons to amino acids is known as the genetic code, and is given in Table 2.1.

The table is divided into blocks that have the same bases in the first two positions. For example, codons of the form UCN (where N is any of the four bases) all code for Ser. There are many groups of four codons where all four code for the same amino acid and the base at the third position does not make any difference. There are several groups where there are two pairs of two amino acids in a block, e.g., CAY codes for His and CAR codes for Gln (Y indicates a pyrimidine, C or U; and R indicates a purine, A or G). There are only two amino acids that have a single codon: UGG = Trp, and AUG = Met. Ile is unusual in having three codons, while Leu, Ser, and Arg all have six codons, consisting of a block of four and a block of two. There are three codons that act as stop signals rather than coding for amino acids. These denote the end of the coding region of a gene.

When the genetic code was first worked out in the 1960s, it was thought to be universal, i.e., identical in all species. Now we realize that it is extremely widespread but not completely universal. The standard code shown in Table 2.1 applies to almost all prokaryotic genomes (including both bacteria and archaea) and to the nuclear genomes of almost all eukaryotes. In mitochondrial genomes, there are several different genetic codes, all differing from the standard code in small respects (e.g., the reassignment of the stop codon UGA to Trp, or the reassignment of the Ile codon AUA to Met). There are also some changes in the nuclear genome codes for specific groups of organisms, such as the ciliates (a group of unicellular eukaryotes including *Tetrahymena* and *Paramecium*), and *Mycoplasma* bacteria use a slightly different code from most bacteria. These changes are all quite small, and presumably they occurred at a relatively late stage in evolution. The main message is that the code is shared between all three domains of life (archaea, bacteria, and eukaryotes) and hence must have evolved before the divergence of these groups. Thus the last universal common ancestor of all current life must have used this genetic code.

Here we will pause to write a letter of complaint to the BBC. When the release of the human genome sequence was announced in 2001, there were many current affairs broadcasters who commented on how exciting it is that we now know the complete "human genetic code". We have known the genetic code for 40 years! What is new is that we now have the complete genome sequence. Please do not confuse the genetic code with the genome. We now have the complete book, whereas 40 years ago we only knew the words in which the book is written. It will probably take us another 40 years to understand what the book means.

2.3.4 *Translation and protein synthesis*

Translation is the process of synthesis of a protein sequence using mRNA as a template. A key molecule in the process is transfer RNA (tRNA). The structure of tRNA was already shown in Fig. 2.4 and Plate 2.1(a). The three bases in the middle of the central hairpin loop in the cloverleaf are called the anticodon. The sequence shown in Fig. 2.4 is a tRNA-Ala, i.e., a tRNA for the amino acid alanine. The anticodon of this molecule is UGC (reading from 5′ to 3′) in the tRNA. This can form complementary base pairs with the codon sequence GCA (reading from 5′ to 3′ in the mRNA) like this:

```
 |       |
 C – G – U     tRNA
–G – C – A–    mRNA
```

Note that GCA is an alanine codon in the genetic code. Organisms possess sets of tRNAs capable of base pairing with all 61 codons that denote amino acids. These tRNAs have different anticodons, and are also different from one another in many other parts of the sequence, but they all have the same cloverleaf secondary structure.

It is not true, however, that there is one tRNA that exactly matches every codon. Many tRNAs can pair with more than one codon due to the flexibility of the pairing rules that occurs at the third position in the codon – this is known as wobble. For example, most bacteria have two types of tRNA-Ala. One type, with anticodon UGC, decodes the codons GCA and GCG, while the other type, with anticodon GGC, decodes the codons GCU and GCC. The actual number of tRNAs varies considerably between organisms. For example, the *E. coli* K12 genome has 86 tRNA genes, of which three have UGC and two have GGC anticodons. In contrast *Rickettsia prowazeckii*, another member of the proteobacteria group, has a much smaller genome with only 32 tRNAs, and only one of each type of tRNA-Ala. These figures are all taken from the genomic tRNA database (Lowe and Eddy, 1997). In eukaryotes, the wobble rules tend to be less flexible, so that a greater number of distinct tRNA types are required. Also the total number of tRNA genes can be much larger, due to the presence of duplicate copies. Thus, in humans, there are about 496 tRNAs in total, and for tRNA-Ala there are 10 with UGC anticodon, five with CGC, and 25 with AGC. In contrast,, in most mitochondrial genomes, there are only 22 tRNAs capable of decoding the complete set of codons. In this case, whenever there is a box of four codons, only one tRNA is required. Pairing at the third position is extremely flexible (sometimes known as hyperwobble). For example the tRNA-Ala, with anticodon UGC, decodes all codons of the form GCN.

Transfer RNA acts as an adaptor molecule. The anticodon end connects to the mRNA, and the other end connects to the growing protein chain. Each tRNA has an associated enzyme, known as an amino acyl-tRNA synthetase, whose function is to attach an amino acid of the correct type to the 3′ end of the tRNA. The enzyme and the tRNA recognize one another specifically, due to their particular shape and intermolecular interactions. The interaction between glutaminyl-tRNA synthetase and tRNA-glutamine is shown in Plate 2.1(a).

Protein synthesis is carried out by another molecular machine called a ribosome. The ribosome is composed of a large and a small subunit (represented by the two large ellipses in the cartoon in Fig. 2.7(c)). In bacteria, the small subunit contains the small subunit ribosomal RNA (SSU rRNA), which is typically 1500 nucleotides long, together with about 20 ribosomal proteins. The large subunit contains large subunit ribosomal RNA (LSU rRNA), which is typically 3000 nucleotides long, together with about 30 proteins and another smaller ribosomal RNA known as 5S rRNA. The ribosomes of eukaryotes are larger – the two major rRNA molecules are significantly longer and the number of proteins in each subunit is greater.

Figure 2.7(c) illustrates the mechanism of protein synthesis. The ribosome binds to the mRNA and moves along it one codon at a time. tRNAs, charged with their appropriate amino acid, are able to bind to the mRNA at a site inside the ribosome. The amino acid is then removed from the tRNA and attached to the end of a growing protein chain. The old tRNA then leaves and can be recharged with another molecule of the same type of amino acid and used again. The tRNA corresponding to the next codon then binds to the mRNA and the ribosome moves along one codon.

Just as with transcription, translation also requires signals to tell it where to start and stop. We already mentioned stop codons. These are codons that do not have a matching tRNA. When the ribosome reaches a stop codon, a protein known as a release factor enters the appropriate site in the ribosome instead of a tRNA. The release factor triggers the release of the completed protein from the ribosome.

There is also a specific start codon, AUG, which codes for methionine. The ribosome begins protein synthesis at the first AUG codon it finds, which will be slightly downstream of the place where it initially binds to the mRNA. In bacteria, mRNAs contain a conserved sequence of about eight nucleotides, called the Shine–Dalgarno sequence, close to their 5′ end. This sequence is complementary to part of SSU rRNA in the small subunit of the ribosome. This interaction triggers the binding of the ribosome to the mRNA. The first tRNA involved is known as an fMet initiator tRNA. This is a special type of tRNA-Met, where a formyl group has been added to the methionine on the charged tRNA. The fMet is only used when an AUG is a start codon. Other AUG codons occurring in the middle of a gene sequence lead to the usual form of Met being added to the protein sequence.

In the last few years, we have been able to obtain three-dimensional crystal structures of the ribosome (e.g., Yusupov *et al.* 2001), and we are getting closer to understanding the mechanism by which the ribosome actually works. The ribosome is acting as a catalyst for the process of peptide bond formation. "Ribozyme" is the term used for a catalytic RNA molecule, by analogy with "enzyme", which is a catalytic protein. It had previously been thought that rRNA was simply a scaffold onto which the ribosomal proteins attached themselves, and that it was the catalytic action of the proteins that achieved protein synthesis. Recent experiments are making it clear that rRNA plays an essential role in the catalysis, and hence that rRNA is a type of ribozyme.

2.3.5 *Closing the loop: DNA replication*

As stated above, the central dogma is the principle that information is stored in DNA, is transferred from DNA to RNA, and then from RNA to proteins. We have now briefly explained the mechanisms by which this occurs. In order to close the loop in our explanation of the synthesis of nucleic acids and proteins, we still need to explain how DNA is formed.

DNA needs to be replicated every time a cell divides. In a multicellular organism, each cell contains a full copy of the genome of the organism to which it belongs (with the exception of certain cells without nuclei, such as red blood cells). The DNA is needed in every cell in order that transcription and translation can proceed in those cells. DNA replication is also essential for reproduction, because DNA contains the genetic information that ensures heredity.

DNA replication is semi-conservative. This means that the original double strand is replicated to give two double strands, each of which contains one of the original strands and one newly synthesized strand that is complementary to it. Clearly, both strands of DNA contain the full information necessary to recreate the other strand. The key processes of DNA replication occur at a replication fork (Fig. 2.7(d)). At this point, the two old strands are separated from one another and the new strands are synthesized. The main enzyme that does this job is DNA polymerase III. This enzyme catalyzes the addition of nucleotides to the 3′ ends of the growing strands (at the heads of the arrows in Fig. 2.7(d)). The new strand is therefore synthesized in the 5′ to 3′ direction (as with mRNA synthesis during transcription). On one strand, called the leading strand, synthesis is possible in a continuous unbroken fashion. However, on the lagging strand on the opposite side, continuous synthesis is not possible and it is necessary to initiate synthesis independently many times. The new strand is therefore formed in pieces, which are known as Okazaki fragments.

DNA polymerase III is able to carry out the addition of new nucleotides to a strand but it cannot initiate a new strand. This is in contrast to RNA polymerase, which is able to perform both initiation and addition. DNA polymerase therefore needs a short sequence, called a primer, from which to begin. Primers are short sequences of RNA (indicated by dotted lines in Fig. 2.7(d)) that are synthesized by a form of RNA polymerase called primase. The processes of DNA synthesis initiated by primers has been harnessed to become an important laboratory tool, the polymerase chain reaction or PCR (see Box 2.1).

Once the fragments on the lagging strand have been synthesized, it is necessary to connect them together. This is done by two more enzymes. DNA polymerase I removes the RNA nucleotides of the primers and replaces them with DNA nucleotides. DNA ligase makes the final connection between the fragments. Both DNA polymerase I and III have the ability to excise nucleotides from the 3′ end if they do not match the template strand. This process of error correction is called proof-reading. This means that the fidelity of replication of DNA polymerase is increased by several orders of magnitude with respect to RNA polymerases. Errors in DNA replication cause heritable point mutations, whereas errors in RNA replication merely lead to mistakes in a single short-lived mRNA. Hence accurate DNA replication is very important.

We called this section "closing the loop" because, in the order that we presented things here, DNA replication is the last link in the cycle of mechanisms for synthesis of the major biological macromolecules. There is, however, a more fundamental sense in which this whole process is a loop. Clearly proteins cannot be synthesized without DNA because proteins do not store genetic information. DNA **can** store this information, but it cannot carry out the catalytic roles necessary for metabolism in a cell, and it cannot replicate itself without the aid of proteins. There is thus a chicken and egg situation: "Which came first, DNA or proteins?" Many people now believe that RNA preceded both DNA and proteins, and that there was a period in the Earth's history when RNA played both the genetic and catalytic roles. This is a tempting hypothesis, because several types of catalytic RNA are known (both naturally occurring and artificially synthesized sequences), and because many viruses use RNA as their genetic material today. As with all conjectures related to the origin of life and very early evolution, however, it is difficult to prove that an RNA world once existed.

2.4 PHYSICO-CHEMICAL PROPERTIES OF THE AMINO ACIDS AND THEIR IMPORTANCE IN PROTEIN FOLDING

As we mentioned in Section 1.1, we have many protein sequences for which experimentally determined three-dimensional structures are unavailable. A long-standing goal of bioinformatics has been to predict protein structure from sequence. Some methods for doing this will be discussed in Chapter 10 on pattern recognition. In this section, we will introduce some of the physico-chemical properties that are thought to be important for determining the way a protein folds.

One property that obviously matters for amino acids is size. Proteins are quite compact in structure, and the different residues pack together in a way

mistakes in a sin-
ırate DNA replica-

the loop" because,
things here, DNA
cycle of mechan-
biological macro-
nore fundamental
s is a loop. Clearly
without DNA be-
ıetic information.
ɔut it cannot carry
for metabolism in
ːlf without the aid
ken and egg situ-
r proteins?" Many
eceded both DNA
ıs a period in the
l both the genetic
ıpting hypothesis,
: RNA are known
rtificially synthes-
nany viruses use
ıy. As with all con-
ife and very early
to prove that an

PROPERTIES
ID THEIR
N FOLDING

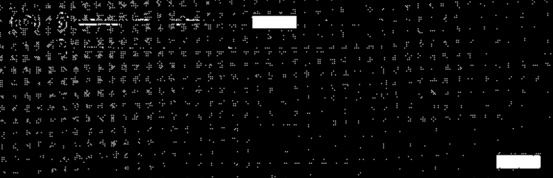
/e have many pro-
ːntally determined
e unavailable. A
cs has been to pre-
ce. Some methods
Chapter 10 on pat-
we will introduce
roperties that are
mining the way a

natters for amino
ıpact in structure,
ogether in a way

BOX 2.1 Polymerase chain reaction (PCR)

The object of PCR is to create many copies of a specified sequence of DNA that is initially present in a very small number of copies. The amplified section can then be used in further experiments or for DNA [illegible]. To carry out PCR, it is not [illegible] to be amplified [illegible] [illegible]

The long sequence of DNA con[illegible] is denatured by heating, and [illegible] nucleotides of the two primer [illegible] boxes below. The [illegible]



that is almost space filling. The volume occupied by the side groups is important for protein folding, and also for molecular evolution. It would be difficult to substitute a very large amino acid for a small one because this would disrupt the structure. It is more difficult than we might think at first to define the volume of an amino acid. We have a tendency to think of molecules as "balls and sticks", but really molecules contain atomic nuclei held together by electrons in molecular orbitals. However, if you push atoms together too much, they repel and hence it is possible to define a radius of an atom, known as a van der Waals radius, on the basis of these repulsions. A useful measure of amino acid volume is to sum the volumes of the spheres defined by the van der Waals radii of its constituent atoms. These figures are given in Table 2.2 (in units of $Å^3$). There is a significant variation in volume between the amino acids. The largest amino acid, tryptophan, has roughly 3.4 times the volume of the smallest amino acid, glycine. Creighton (1993) gives more information on van der Waals interactions and on amino acid volumes. Since protein folding occurs in water, another way to define the amino acid volume

Table 2.2 Physico-chemical properties of the amino acids.

			Vol.	Bulk.	Polarity	pI	Hyd.1	Hyd.2	Surface area	[illegible]
Alanine	Ala	A	67	11.50	0.00	6.00	1.8	1.6	[illegible]	[illegible]
Arginine	Arg	R	148	14.28	52.00	10.76	−4.5	−12.3	[illegible]	[illegible]
Asparagine	Asn	N	96	12.28	3.38	5.41	−3.5	[illegible]	[illegible]	[illegible]
Aspartic acid	Asp	D	91	11.68	49.70	2.77	−3.5	[illegible]	[illegible]	[illegible]
Cysteine	Cys	C	86	13.46	1.48	5.05	2.5	2.0	[illegible]	[illegible]
Glutamine	Gln	Q	114	14.45	3.53	5.65	[illegible]	[illegible]	[illegible]	[illegible]
Glutamic acid	Glu	E	109	13.57	49.90	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Glycine	Gly	G	48	3.40	0.00	5.97	[illegible]	[illegible]	[illegible]	[illegible]
Histidine	His	H	118	13.69	51.60	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Isoleucine	Ile	I	124	21.40	0.13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Leucine	Leu	L	124	21.40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Lysine	Lys	K	135	15.71	49.50	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Methionine	Met	M	124	16.25	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Phenylalanine	Phe	F	135	19.80	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Proline	Pro	P	90	17.43	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Serine	Ser	S	73	9.47	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Threonine	Thr	T	93	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Tryptophan	Trp	W	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Tyrosine	Tyr	Y	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Valine	Val	V	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Mean			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Std. dev.			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Vol, volume calculated [illegible] and Simha 1968) [illegible] point (Zimmerman, Eliezer, and [illegible] hydrophobicity scale [illegible] in unfolded [illegible] *et al.* 1986).

is to consider the increase in volume of a solution when an amino acid is dissolved in it. This is known as the partial volume. Partial volumes are closely correlated with the volumes calculated from the van der Waals radii, and we do not show them in the table.

Zimmerman, Eliezer, and Simha (1968) presented data on several amino acid properties that are relevant in the context of protein folding. Rather than simply considering the volume, they defined the "bulkiness" of an amino acid as the ratio of the side chain volume to its length, which provides a measure of the average cross-sectional area of the side chain. These figures are shown in Table 2.2 (in Å^2). Zimmerman, Eliezer, and Simha (1968) also introduced a measure of the polarity of the amino acids. They calculated the electrostatic force of the amino acid acting on its surroundings at a distance of 10 Å. This is composed of the force from the electric charge (for the amino acids that have a charged side group) plus the force from the dipole moment (due to the non-uniformity of electronic charge across the amino acid). The total force (in units scaled for convenience) was used as a polarity index, and this is shown in Table 2.2. The electrostatic charge term, where it exists, is much larger than the dipole term. Hence, this

Table 2.2 (in Å^2).
(1968) also intro-
f the amino acids.
orce of the amino
ı distance of 10 Å.
he electric charge
arged side group)
ment (due to the
: across the amino
l for convenience)
this is shown in
term, where it ex-
term. Hence, this

measure clearly distinguishes between the charged and uncharged amino acids.

The polarity index does not distinguish between the positively and negatively charged amino acids, however, since both have high polarity. A quantity that does this is the pI, which is defined as the pH of the isoelectric point of the amino acid. Acidic amino acids (Asp and Glu) have pI in the range 2–3. This means that these amino acids would be negatively charged at neutral pH due to ionization of the COOH group to COO$^-$. We need to put them in an acid solution in order to shift the equilibrium and balance this charge. The basic amino acids (Arg, Lys, and His) have pI greater than 7. All the others usually have uncharged side chains in real proteins. They have pI in the range 5–6. Thus, pI is a useful measure of acidity of amino acids that distinguishes clearly between positive, negative, and uncharged side chains.

A key factor in protein folding is the "hydrophobic effect", which arises as a result of the unusual characteristics of water as a solvent. Liquid water has quite a lot of structure due to the formation of chains and networks of molecules interacting via hydrogen bonds. When other molecules are dissolved in water, the hydrogen-bonded structure is disrupted. Polar amino acid residues are also able to form hydrogen bonds with water. They therefore disrupt the structure less than non-polar amino acids that are unable to form hydrogen bonds. We say that the non-polar amino acids are hydrophobic, because they do not "want" to be in contact with water, whereas the polar amino acids are hydrophilic, because they "like" water. It is generally observed that hydrophobic residues in a protein are in the interior of the structure and are not in contact with water, whereas hydrophilic residues are on the surface and are in contact with water. In this way the free energy of the folded molecule is minimized.

Kyte and Doolittle (1982) defined a hydrophobicity (or hydropathy) scale that is an estimate of the difference in free energy (in kcal/mol) of the amino acid when it is buried in the hydrophobic environment of the interior of a protein and when it is in solution in water. Positive values on the scale mean that the residue is hydrophobic: it costs free energy to take the residue out of the protein and put it in water. Another version of the hydrophobicity scale was developed by Engelman, Steitz, and Goldman (1986), who were particularly interested in membrane proteins. The interior of a lipid bilayer is hydrophobic, because it mostly consists of the hydrocarbon tails of the lipids. They estimated the free energy cost for removal of an amino acid from the bilayer to water. These two scales are similar but not identical; therefore both scales are shown in the table.

Another property that is thought to be relevant for protein folding is the surface area of the amino acid that is exposed (accessible) to water in an unfolded peptide chain and that becomes buried when the chain folds. Table 2.2 shows the accessible surface areas of the residues when they occur in a Gly–X–Gly tripeptide (Miller *et al.* 1987, Creighton 1993). Rose *et al.* (1985) calculated the average fraction of the accessible surface area that is buried in the interior in a set of known crystal structures. They showed that hydrophobic residues have a larger fraction of the surface area buried, which supports the argument that the "hydrophobic effect" is important in determining protein structure.

2.5 VISUALIZATION OF AMINO ACID PROPERTIES USING PRINCIPAL COMPONENT ANALYSIS

So far, this chapter has summarized some of the fundamental aspects of molecular biology that we think every bioinformatician should know. In the rest of the chapter, we want to introduce some simple methods for data analysis that are useful in bioinformatics. We will use the data on amino acid properties.

Table 2.2 shows eight properties of each amino acid (and we could easily have included several more columns using data from additional sources). It would be useful to plot some kind of diagram that lets us visualize the information in this table. It is straightforward to take any two of the properties and use these as the coordinates for the points in a two-dimensional graph. Figure 2.8 shows a plot of volume against pI. This clearly shows the acidic amino acids at low pI, the basic amino acids at high pI, and all the rest in the middle. It also shows the

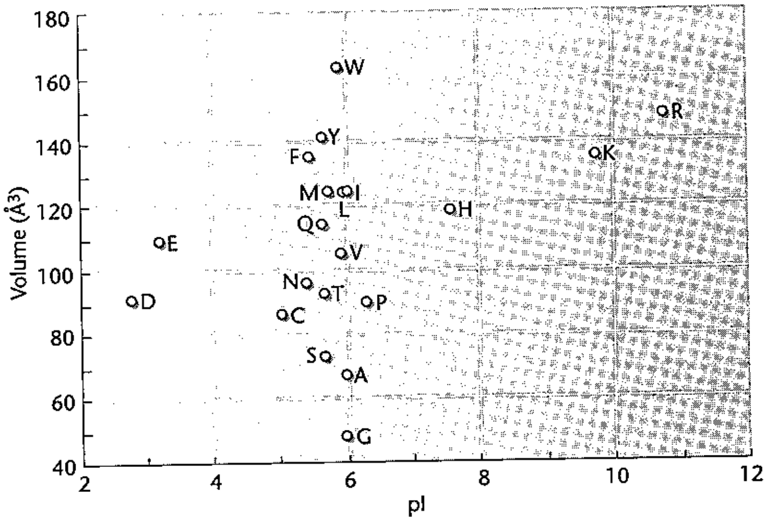


Fig. 2.8 Plot of amino acid volume against pI – two properties thought to be important in protein folding.

large spread of the middle group along the volume axis. However, the figure does not distinguish between the hydrophilic and hydrophobic amino acids in the middle group: N and Q appear very close to M and V, for example. We could separate these by using one of the hydrophobicity scales on the axis instead of pI, but then the acidic and basic groups would appear close together because both are hydrophilic (negative on the hydrophobicity scale). What we need is a way of combining the information from all eight properties into a two-dimensional graph. This can be done with principal component analysis (PCA).

In general with PCA, we begin with the data in the form of an $N \times P$ matrix, like Table 2.2. The number of rows, N, is the number of objects in our data set (in this case $N = 20$ amino acids), and the number of columns, P, is the number of properties of those objects (in this case $P = 8$). Each row in the data matrix can be thought of as the coordinates of a point in P-dimensional space. The whole data set is a cloud of these points. The PCA method transforms this cloud of points first by scaling them and shifting them to the origin, and then by rotating them in such a way that the points are spread out as much as possible, and the structure in the data is made easier to see.

Let the original data matrix be X_{ij}(i.e., X_{ij} is the value of the j^{th} property of object i). The mean and standard deviation of the properties are

$$\mu_j = \frac{1}{N} \sum_i X_{ij}$$

and

$$\sigma_j = \left(\frac{1}{N} \sum_i (X_{ij} - \mu_j)^2 \right)^{1/2}$$

The mean and standard deviation are listed at the foot of Table 2.2. Since the properties all have different scales and different mean values, the first step of PCA is to define scaled data values by

$$z_{ij} = (X_{ij} - \mu_j)/\sigma_j$$

The z_{ij} matrix measures the deviation of the values from the mean values for each property. By definition, the mean value of each column in the z_{ij} matrix is 0 and the standard deviation is 1. Scaling the data in this way means that all the input properties are placed on an equal footing, and all the properties will contribute equally to the data analysis.

We now choose a set of vectors $v_j = (v_{j1}, v_{j2}, v_{j3}, \ldots v_{jP})$ that define the directions of the principal

ιino acid volume
perties thought to
tein folding.

X_{ij}(i.e., X_{ij} is the
i). The mean and
s are

ı are listed at the
ies all have differ-
es, the first step of
by

tion of the values
roperty. By defin-
ın in the z_{ij} matrix
. Scaling the data
out properties are
the properties will
ysis.
's $v_j = (v_{j1}, v_{j2}, v_{j3},$
of the principal

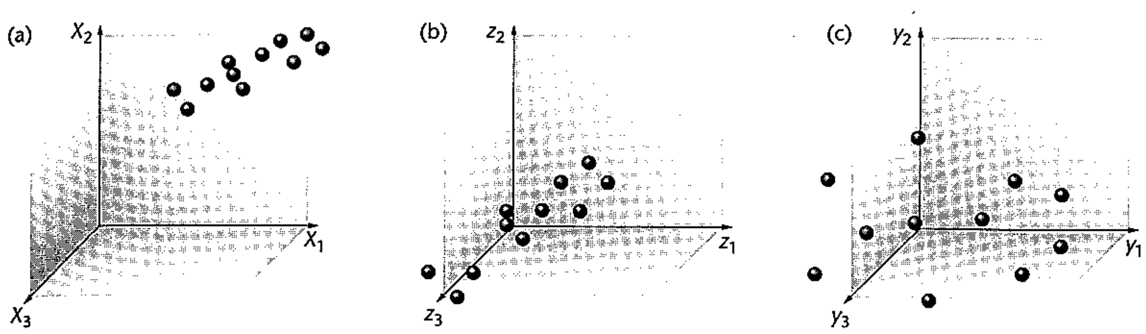


Fig. 2.9 Schematic illustration of principal component analysis. (a) Original data. (b) Scaled and centered on the origin. (c) Rotated onto principal components.

components. These vectors are of unit length, i.e., $\sum_k v_{jk}^2 = 1$ for each vector, and they are all orthogonal to one another, i.e., $\sum_k v_{ik}v_{jk} = 0$, when i and j are not equal. Each vector represents a new coordinate axis that is a linear combination of the old coordinates. The positions of the points in the new coordinate system are given by

$$y_{ij} = \sum_k v_{jk}z_{ik}$$

The new y coordinate system is a rotation of the z coordinate system – see Fig. 2.9.

There are still P coordinates, so we can only use two of them if we plot a two-dimensional graph. However, we can define the y coordinates so that as much of the variation between the points as possible is visible in the first few coordinates. We therefore choose the v_{1k} values so that the variance of the points along the first principal component axis, $\frac{1}{N}\sum_i y_{i1}^2$ is as large as possible. (Note that the means of the y's are all zero because the means of the z's were zero.) We then choose the v_{2k} for the second component by maximizing the variance $\frac{1}{N}\sum_i y_{i2}^2$, with the constraint that the second axis is orthogonal to the first, i.e., $\sum_k v_{1k}v_{2k} = 0$. If we wish, we can define further components by maximizing the variance with the constraint that each component is orthogonal to the previous ones. Calculation of the v_{jk} is discussed in more detail in Box 2.2.

The results of PCA for the amino acid data in Table 2.2 are shown in Fig. 2.10. The first two principal component vectors are shown in the matrix on p. 28. For component 1, the largest contributions in the vector are the negative contributions from the hydrophobicity scales. Thus hydrophobic amino acids appear on the left side and hydrophilic ones on the right. For component 2, the largest contributions are positive ones from volume, bulkiness, and surface area. Thus large amino acids appear near the top of the figure and small ones near the bottom. However, all the properties contribute to some extent to each of the components; therefore, the resulting figure is not the same as we would have got by simply plotting hydrophobicity against volume.

Figure 2.10 illustrates several points about the data that seem intuitive. There is a cluster of medium-sized hydrophobic residues, I, L, V, M, and F. The two acids, D and E, are close, and so are the two amides, Q and N. Two of the basic residues, R and K, are very close, and H is fairly close to these. The two largest residues, W and Y, are quite close to one another. The PCA diagram manages to do a fairly good job at illustrating all these similarities at the same time.

The PCA calculation in this section was done using the program pca.c by F. Murtagh (http://astro.u-strasbg.fr/~fmurtagh/mda-sw/).

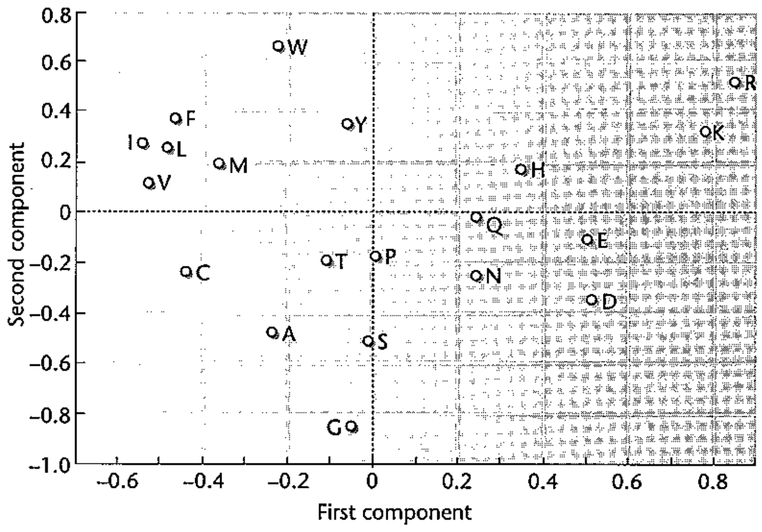


Fig. 2.10 Plot of the amino acids on the first two components of the principal component analysis.

	Vol	Bulk.	Pol.	pI	Hyd.1	Hyd.2	S.A.	Fr.A.
Comp. 1	(0.06,	−0.22,	0.44,	0.19,	−0.49,	−0.51,	0.10,	−0.45)
Comp. 2	(0.58,	0.48,	0.10,	0.25,	0.03,	−0.03,	0.56,	0.17)

2.6 CLUSTERING AMINO ACIDS ACCORDING TO THEIR PROPERTIES

2.6.1 Handmade clusters

When we look at a figure like 2.10, it is natural to try to group the points into "clusters" of similar objects. We already remarked above that I, L, V, M, and F look like a cluster. So, where would you put clusters? Before going any further, make a few photocopies of Fig. 2.10. Now take one of the copies and draw rings around the groups of points that you think should be clustered. You can decide how many clusters you think there should be – somewhere between four and seven is probably about right. You can also decide how big the clusters should be – you can put lots of points together in one cluster if you like, or you can leave single points on their own in a cluster of size one. OK, go ahead!

When we presented the chemical structures of the amino acids in Fig. 2.6, we chose four groups:

Neutral, nonpolar	W, F, G, A, V, I, L, M, P
Neutral, polar	Y, S, T, N, Q, C
Acidic	D, E
Basic	K, R, H

This is one possible clustering. We chose these four clusters because this is the way the amino acids are presented in most molecular biology textbooks. Try drawing rings round these four clusters on another copy of Fig. 2.10. The acidic and basic groups work quite well. The neutral polar group forms a rather spread-out cluster in the middle of the figure, but unfortunately it has P in the middle of it. The nonpolar group can hardly be called a cluster, as it takes up about half the diagram, and contains points that are very far from one another, like G and W. You probably think that you did a better job when you made up your own clusters a few minutes ago.

We now want to consider ways of clustering data that are more systematic than drawing rings on paper.

the amino acids on
[illegible]nents of the
[illegible]nt analysis.

[illegible]	Fr.A.
0,	−0.45)
6,	0.17)

A, V, I, L, M, P
N, Q, C

[illegible] chose these four [illegible]e amino acids are [illegible]gy textbooks. Try [illegible]usters on another [illegible]basic groups work [illegible]up forms a rather [illegible] of the figure, but [illegible]e of it. The nonpo- [illegible]ster, as it takes up [illegible]ins points that are [illegible]and W. You prob- [illegible] when you made [illegible]s ago.

[illegible] of clustering data [illegible]ing rings on paper.

BOX 2.2
Principal component analysis in more detail

From the $N \times P$ data matrix, we can define a $P \times P$ matrix of correlation coefficients, C_{jk}, between the properties:

$$C_{jk} = \frac{1}{N\sigma_j\sigma_k}\sum_i (X_{ij} - \mu_j)(X_{ik} - \mu_k) = \frac{1}{N}\sum_i z_{ij}z_{ik}$$

The coefficients are always in the range −1 to 1. If $C_{jk} > 0$, the two properties are positively correlated, i.e. [illegible] tend to be large at the same time and small at the same time. If $C_{jk} < 0$, the properties are negatively [illegible] one tends to be large when the other is small. The correlation matrix for the amino acid data looks like this.

The matrix is symmetric ($C_{jk} = C_{kj}$) and all the diagonal elements are 1.00 by definition. The values [illegible] tures of the data that are not easy to see in the original matrix. For example, volume has a strong positive [illegible] with surface area and bulkiness, and a fairly weak correlation with the other properties. The [illegible] phobicity scales have strong positive correlation with each other and also with the fractional area property [illegible] have a significant negative correlation with the polarity scale.

It can be shown that the vectors $\mathbf{v}_n$ that define the principal component axes are the eigenvectors [illegible] matrix, i.e., they satisfy the equation:

$$\sum_j v_{nj} C_{jk} = \lambda_n v_{nk}$$

where the λ_n are constants called eigenvalues. The first principal component (PC) [illegible] largest eigenvalue. Subsequent PCs can be listed in order of decreasing size of [illegible] in this case are $\lambda_1 = 3.57$ and $\lambda_2 = 2.81$.

The variance along the n^{th} PC axis is equal to the corresponding eigenvalue:

$$\frac{1}{N}\sum_i y_{in}^2 = \frac{1}{N}\sum_i\sum_j\sum_k v_{nj}z_{ij}v_{nk}z_{ik} = \sum_j\sum_k v_{nj}C_{jk}v_{nk} = \sum_j \lambda_n [illegible] = \lambda_n$$

We know that the variance of each of the z coordinates is 1, [illegible] we change the coordinates to the principal components, [illegible] in the PC space is still P. The fraction of the total variance [illegible] $(\lambda_1 + \lambda_2)/P$, which in our case is $(3.57 + 2.81)/8 = 0.797$ [illegible] in Fig. 2.10). Roughly 80% of the variation in the [illegible] with just two PCs. When points appear close in the [illegible] the eight-dimensional space, because [illegible] distance between the points. [illegible] spaced points, then these are likely [illegible]

	Vol	Bulk	[illegible]
Vol.	1.00	0.7[illegible]	[illegible]
Bulk	0.7[illegible]	1.00	[illegible]
[illegible]	0.2[illegible]	−0.2[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
H1	[illegible]	[illegible]	[illegible]
H2	[illegible]	[illegible]	[illegible]
SA	[illegible]	[illegible]	[illegible]
Fr.A.	0.1[illegible]	[illegible]	[illegible]

In fact, there is a **huge** number of different clustering methods. This testifies to the fact that there are a lot of different people from a lot of different disciplines who find clustering useful for describing the patterns in their data. Unfortunately, it also means that there is not one single clustering method that everyone agrees is best. Different methods will give different answers when applied to the same data; therefore, there has to be some degree of subjectivity in deciding which method to use for any particular data set.

In the context of the amino acids, clustering according to physico-chemical properties is actually quite helpful when we come to do protein sequence alignments. We usually want to align residues with similar properties with one another, even if the residues are not identical. There are several sequence alignment editors that ascribe colors to residues, assigning the same color to clusters of similar amino acids. In well-aligned parts of protein sequences, we often find that all the residues in a column have the same color. The coloring scheme can thus help with constructing alignments and spotting important conserved motifs. When we look at protein sequence evolution (Chapter 4) it turns out that substitutions are more frequent between amino acids with similar properties. So, clustering according to properties is also relevant for evolution. In the broader context, however, clustering algorithms are very general and can be used for almost any type of data. In this book, they will come up again in two places: in Chapter 8 we discuss distance matrix methods for molecular phylogenetics, which are a form of hierarchical clustering; and in Chapter 13 we discuss applications of clustering algorithms on microarray data. It is therefore worth spending some time on these methods now, even if you are getting a bit bored with amino acid properties.

2.6.2 Hierarchical clustering methods

In a hierarchical clustering method, we need to choose a measure of similarity between the data points, then we need to choose a rule for measuring the similarity of clusters.

We will use the scaled coordinates z as in the previous section. There is a vector $\mathbf{z}_i$ from the origin to each point i in the data set (see Fig. 2.11). The length of the vector is:

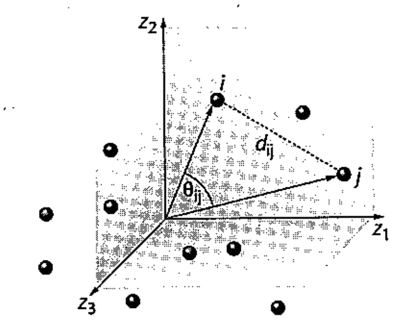


Fig. 2.11 Illustration of the data points as vectors in multidimensional space.

$$|\mathbf{z}_i| = \left(\sum_k \mathbf{z}_{ik}^2\right)^{1/2}$$

We want to measure how similar the vectors are for two points i and j. A simple way to do this is to use the cosine of the angle θ_{ij} between the vectors. If the two vectors are pointing in almost the same direction, θ_{ij} will be small and $\cos\theta_{ij}$ will be close to 1. Vectors with no correlation will have θ_{ij} close to 90° and $\cos\theta_{ij}$ close to 0. Vectors with negative correlation will have $\theta_{ij} > 90°$ and $\cos\theta_{ij} < 0$.

From standard geometry,

$$\cos\theta_{ij} = \frac{\sum_k z_{ik} z_{jk}}{|\mathbf{z}_i||\mathbf{z}_j|}$$

Another possible similarity measure is the correlation coefficient between the $\mathbf{z}$ vectors:

$$R_{ij} = \frac{1}{Ps_i s_j} \sum_k (z_{ik} - m_i)(z_{jk} - m_j)$$

where m_i and s_i are the mean and standard deviation of the elements in the i^{th} row (see also Box 2.2, where we define the correlation between the columns). R_{ij} is in the range -1 to 1.

In what follows, we shall assume that we have calculated an $N \times N$ matrix of similarities between the data points that could be $\cos\theta_{ij}$ or R_{ij}, or any other measure of similarity that appears appropriate for the data in question. We will call the similarity matrix S_{ij} from now on, to emphasize that the method is general and works the same way, whichever measure we use for similarity.

During the process of hierarchical clustering, points are combined into clusters, and small clusters are combined to give progressively larger clusters. To decide in what order these clusters will be connected, we will need a definition of similarity between clusters. Suppose we already have two clusters A and B. We want to define the similarity S_{AB} of these clusters. There are (at least) three ways of doing this:

- Group average. S_{AB} = the mean of the similarities S_{ij} between the individual data points, averaged over all pairs of points, where i is in cluster A and j is in cluster B.
- Single-link rule. S_{AB} = maximum similarity S_{ij} for any i in A and j in B.
- Complete-link rule. S_{AB} = minimum similarity S_{ij} for any i in A and j in B.

The reasons for the terms "single link" and "complete link" will be made more clear in Section 2.6.3.

An algorithm is a computational recipe that specifies how to solve a problem. Algorithms come up throughout this book, and we will discuss some general points about algorithms in Chapter 6. For the moment, we will present a very simple algorithm for hierarchical clustering. This works in the same way, whatever the definitions of similarity between data points and between clusters. We begin with each point in a separate cluster of its own.

1 Join the two clusters with the highest similarity to form a single larger cluster.

2 Recalculate similarities between all the clusters using one of the three definitions above.

3 Repeat steps 1 and 2 until all points have been connected to a single cluster.

This procedure is called "hierarchical" because it generates a set of clusters within clusters within clusters. For this reason, the results of a hierarchical clustering procedure can be represented as a tree. Each branching point on the tree is a point where two smaller clusters were joined to form a larger one. Reading backwards from the twigs of the tree to the root tells us the order in which the clusters were connected.

Plate 2.2(a) shows a hierarchical clustering of the amino acid data. This was performed using the CLUTO package (Karypis 2002). The similarity measure used was cos θ and the group-average rule was used for the similarity between clusters. The tree on the left of Plate 2.2(a) shows the order in which the amino acids were clustered. For example, L and I are very similar, and are clustered at the beginning. The LI cluster is later combined with V. In the meantime M and F are clustered, and then the MF cluster is combined with VLI, and so on. The tree indicates what happens if the clustering is continued to the point where there is only one cluster left. In practice, we want to stop the clustering at some stage where there is a moderate number of clusters left. The right side of Plate 2.2(a) shows the clusters we get if we stop when there are six clusters. These can be summarized as follows.

Cluster 1:	Basic residues	K, R, H
Cluster 2:	Acid and amide residues	E, D, Q, N
Cluster 3:	Small residues	P, T, S, G, A
Cluster 4:	Cysteine	C
Cluster 5:	Hydrophobic residues	V, L, I, M, F
Cluster 6:	Large, aromatic residues	W, Y

The central part of Plate 2.2(a) is a representation of the scaled data matrix z_{ij}. Red/green squares indicate that the value is significantly higher/lower than average; dark colors indicate values close to the average. This color scheme makes sense in the context of microarrays, as we shall see in Chapter 13. We have named the clusters above according to what seemed to be the most important feature linking members of the cluster. The basic cluster contains all the residues that are red on both the pI and polarity scales. The acid and amide cluster contains all the residues that are green on the hydrophobicity scales and also on the pI scale. Note that if we had stopped the clustering with a larger number of clusters, the acids and the amides would have been in separate clusters. We called cluster 3 "small" because the most noticeable thing is that these residues are all green on the volume and surface area scales. These residues are quite mixed in terms of hydrophobicities. Cluster 4 contains only cysteine. Cysteine has an unusual role in protein structure because of its potential to form disulfide bonds between pairs of cysteine residues. For this reason, cysteines tend to be important when they occur and it is difficult to interchange them for other residues. Cysteine does not appear to be particularly extreme in any of the eight properties used here, and none of the eight properties captures the important factor of disulfide bonding. Nevertheless, it is interesting that this cluster analysis manages to spot some of its uniqueness. Cluster 5 is clearly hydrophobic, and cluster 6 contains the two largest amino acids, which both happen to be aromatic. It is worth noting, however, that the other aromatic residue, phenylalanine (F), is in cluster 5. Phenylalanine has a simple hydrocarbon ring as a side group and therefore is hydrophobic. In contrast, tryptophan and tyrosine are only moderate on the hydrophobicity scales used here.

At the top of Plate 2.2(a), there is another tree indicating a clustering of the eight properties. This is done so that the properties can be ordered in a way that illustrates groups of properties that are correlated. The tree shows very similar information to the correlation matrix given in Box 2.2, i.e., volume and surface area are correlated, the two hydrophobicity scales are correlated with the fractional area scale, etc.

2.6.3 *Variants on hierarchical clustering*

Take another copy of Fig. 2.10 and draw rings around the six clusters specified by the hierarchical method. These clusters seem to make sense, and they are probably as good as we are likely to get with these data as input. They are not the only sensible set of clusters, however, and the details of the clusters we get depend on the details of the method.

First, the decision to stop at six clusters is subjective. If we use the same method (cos θ and group average) and stop at seven, the difference is that the acids are separated from the amides. If we stop at five, cysteine is joined with the hydrophobic cluster.

A second point to consider is the rule for similarity between clusters. In hierarchical clustering methods, we could in principle plot the similarity of the pair of clusters that we connect at each step of the process as a function of the number of steps made. This level begins at one, and gradually descends and the clusters get bigger and the similarity between the clusters gets lower. In the group-average method, the similarity of the clusters is the mean of the similarities of the pairs of points in the cluster. Therefore, roughly half of the pairs of points will have similarities greater than or equal to the similarity level at which the connection is made. When the single-link rule is used, the level at which the connection is made is the similarity of the most similar pair of points in the two clusters connected. This means that clusters can be very spread out. Two points in the same cluster may be very different from one another as long as there is a chain of points between them, such that each link in the chain corresponds to a high similarity pair. In contrast, the complete-link rule will only connect a pair of clusters when all the pairs of points in the two clusters have similarity greater than the current connection level. Thus each point is completely linked to all other points in the cluster. In our case, using cos θ, the single-link rule and stopping at six clusters yields the same six clusters as with the group-average rule, except that WY is linked with VLIMF and QN is split from DE. Using cos θ with the complete-link rule gives the same as the group-average method, with the exception that C is linked with VLIMF and TP is split from SGA.

These are relatively minor changes. We also tried using the correlation coefficient as the similarity measure instead of cos θ, and this gave a more significant change in the result. With the group-average rule we obtained: EDH; QNKR; YW; VLIMF; PT; SGAC. These clusters seem less intuitive than those obtained with the cos θ measure, and also appear less well defined in the PCA plot. The correlation coefficient therefore seems to work less well on this particular data set. The general message is that it is worth considering several different methods on any real data, because differences will arise.

So far we have been treating the data in terms of similarities. It is also possible to measure distances between data points that measure how "far apart" the points are, rather than how similar they are. We already have points in our P-dimensional space defined by the z coordinates (Fig. 2.11). Therefore we can straightforwardly measure the Euclidean distance between these points:

reen the clusters
nethod, the sim-
the similarities of
ıerefore, roughly
have similarities
ity level at which
single-link rule is
tion is made is the
points in the two
at clusters can be
same cluster may
as long as there is
ıch that each link
similarity pair. In
ill only connect a
points in the two
han the current
nt is completely
ıster. In our case,
ıd stopping at six
ters as with the
VY is linked with
Using cos θ with
ıme as the group-
n that C is linked
A.
ges. We also tried
as the similarity
his gave a more
With the group-
NKR; YW; VLIMF;
ess intuitive than
ıeasure, and also
plot. The correla-
work less well on
al message is that
erent methods on
will arise.
e data in terms of
neasure distances
e how "far apart"
nilar they are. We
imensional space
2.11). Therefore
ıre the Euclidean

$$d_{ij} = \left(\sum_k (z_{ik} - z_{jk})^2 \right)^{1/2}$$

We can use the matrix of distances between points instead of the matrix of similarities. The only difference in the hierarchical clustering procedure is always to connect the pair of clusters with the smallest distance, rather than the pair with the highest similarity. Group-average, single-link, and complete-link methods can still be used with distances. Even though the clustering rule is basically the same, clustering based on distances and similarities will give different results because the data are input to the method in a different way – the distances are not simple linear transformations of the similarities.

One of the first applications of clustering techniques, including the ideas of single-link, complete-link, and group-average clusters, was for construction of phylogenetic trees using morphological characters (Sokal and Sneath 1963). Distance-matrix clustering methods are still important in molecular phylogenetics. In that case, the data consist of sequences, rather than points in Euclidean space. There are many ways of defining distances between sequences (Chapter 4), but once a distance matrix has been defined, the clustering procedure is the same. In the phylogenetic context, the group-average method starting with a distance matrix is usually called UPGMA (see Section 8.3).

2.6.4 *Non-hierarchical clustering methods*

All the variants discussed above give rise to a nested set of clusters within clusters that can be represented by a tree. There are other types of clustering method, sometimes called "direct" clustering methods, where we simply specify the number, K, of clusters required and we try to separate the objects into K groups without any notion of a hierarchy between the groups. Direct clustering methods require us to define a function that measures how good a set of clusters is. One function that does this is

$$I_2 = \sum_A \sqrt{\sum_{i,j \in A} S_{ij}}$$

Here, A labels the cluster, and we are summing over all clusters $A = 1, 2 \ldots K$. The notation $i,j \in A$ means that we are summing over all pairs of objects i and j that are in cluster A. We called this function I_2, following the notation in the manual for the CLUTO software (Karypis 2002). Given any proposed division of the objects into clusters, we can evaluate I_2. We can then choose the set of clusters that maximizes I_2.

There are many other optimization functions that we might think of to evaluate the clusters. Basically, we want to maximize some function of the similarities of objects within clusters or minimize some function of the similarities of objects in different clusters. I_2 is the default option in CLUTO, but several other functions can be specified as alternatives. Note that if a cluster has n objects, there are n^2 pairs of points in the cluster. The square root in I_2 provides a way of balancing the contributions of large and small clusters to the optimization function. Using the I_2 optimization function on the amino acid data with $K = 6$ gives the clusters: KRH; EDQN; PT; CAGS; VLIMF; WY. This is another slight variant on the one shown in Plate 2.2(a), but one that also seems to make sense intuitively and when drawn on the principal components plot.

Another well-known form of direct clustering, known as K-means (Hartigan 1975), treats the data in the form of distances instead of similarities. In this case, we define an error function E and choose the set of clusters that minimizes E. Let μ_{Aj} be the mean value of z_{ij} for all objects i assigned to cluster A. The square of the distance of object i from the mean point of the cluster to which it belongs is

$$d_{iA}^2 = \sum_j (z_{ij} - \mu_{Aj})^2$$

and the error function is

$$E = \sum_A \sum_{i \in A} d_{iA}^2$$

In direct clustering methods, we have a well-defined function that is being optimized. However, we do not have a well-defined algorithm for finding the set of clusters. It is necessary to write a computer program that tries out very many possible solutions and saves the best one that it finds. Typically, we might begin with some random partition of the data into K clusters and then try moving one object at a time into a different cluster in such a way as to make the best possible improvement in the optimization function. If there is no movement of an object that would improve the optimization function, then we have found at least a local optimum solution. If the process is repeated several times from different starting positions, we have a good chance of finding the global optimum solution.

For the hierarchical methods in the previous section, the algorithm tells us exactly how to form the clusters, so there is no trial and error involved. However, there is no function that is being optimized. Exactly the same distinction will be made when we discuss phylogenetic methods in Chapter 8: distance matrix methods have a straightforward algorithm but no optimization criterion, whereas maximum-parsimony and maximum-likelihood methods have well-defined optimization criteria, but require a trial-and-error search procedure to locate the optimal solution.

There are many issues related to clustering that we have not covered here. Some methods do not fit into either the hierarchical or the direct clustering categories. For example, we can also do top-down clustering where we make successive partitions of the data, rather than successive amalgamations, as in hierarchical methods. It is worth stating an obvious point about all the clustering methods discussed in this chapter: clusters are defined to be non-overlapping. An object cannot be in more than one cluster at once. When we run a clustering algorithm, we are forcing the data into non-overlapping groups. Sometimes the structure of the data may not warrant this, in which case we should be wary of using clustering methods or of reading too much into the clusters produced. Statistical tests for the significance of clusters are available, and these would be important if we were in doubt whether a clustering method was appropriate for our data.

To illustrate the limitations of non-overlapping clusters, we tried to plot a Venn diagram illustrating as many relevant properties of amino acids as possible: see Plate 2.2(b). These properties **do** overlap. For example, several amino acids are not strongly polar or nonpolar, and are positioned in the overlap area. There are aromatic amino acids on both the polar and nonpolar sides, so the aromatic ring overlaps the others. This diagram is surprisingly hard to draw (this is at least the fourth version we tried!). There were some things in earlier versions that got left out of this one, for example tyrosine (Y) is sometimes weakly acidic (so should it be in a ring with D and E?) and histidine is only weakly basic (so should we move it into the polar neutral area?). The general message is that clusters are useful, but they have limitations, and we should keep this in mind when clustering more complex data sets, such as the microarray data discussed in Chapter 13.

SUMMARY

DNA is composed of sequences of four types of nucleotide building blocks known as A, C, G and T. It is the molecule that stores the genetic information of the cell. It usually exists as a double helix composed of [illegible] exactly complementary strands. RNA is also [illegible] of four nucleotide building blocks, [illegible] of T. RNA molecules are usually single [illegible] to form complex stem-loop [illegible] base pairing between short [illegible] Proteins are polymers [illegible] types of amino acid [illegible]

The process of [illegible] a template is called [illegible] polymerase. The [illegible] e in more than one a clustering algo- to non-overlapping of the data may not should be wary of reading too much istical tests for the ailable, and these n doubt whether a te for our data.

of non-overlapping iagram illustrating amino acids as pos- perties **do** overlap. ls are not strongly oned in the overlap acids on both the aromatic ring over- urprisingly hard to version we tried!). er versions that got yrosine (Y) is some- be in a ring with D kly basic (so should area?). The general eful, but they have this in mind when sets, such as the apter 13.

amino acid properties. PCA chooses coordinates that ar [illegible] linear combinations of the original variables in such a [illegible] way that the maximum variability between the [illegible] points is explained by the first fe [illegible] Clustering analysis is another way of looking [illegible] in complex data sets. A large variety of [illegible] ods is possible, including hierarchical [illegible] ing. These methods give [illegible]

REFERENCES

Bell, C.E. and Lewis, M. 2001. Crystallographic analysis of lac repressor bound to natural operator O1. *Journal of Molecular Biology*, **312**: 921–6.

Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R., and Schneider, B. 1992. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Journal of Biophysics*, **63**: 751–9. (http://ndbserver.rutgers.edu/index.html)

Creighton, T.E. 1993. *Proteins: Structures and Molecular Properties*. New York: W.H. Freeman.

Engelman, D.A., Steitz, T.A., and Goldman, A. 1986. Identifying non-polar transbilayer helices in amino acid sequences of membrane proteins. *Annual Review of Biophysics and Biophysical Chemistry*, **15**: 321–53.

Hartigan, J.A. 1975. *Clustering Algorithms*. New York: Wiley.

Karypis, G. 2002. CLUTO – a clustering toolkit. University of Minnesota technical report #02–017 (http://www-users.cs.umn.edu/~karypis/cluto/)

Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**: 105–32.

Lowe, T.M. and Eddy, S.R. 1997. tRNA-scan-SE: A program for improved detection of transfer RNA genes in genomic sequences. *Nucleic Acids Research*, **25**: 955–64. (http://rna.wustl.edu/tRNAdb/)

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. 1987. Interior and surface of monomeric proteins. *Journal of Molecular Biology*, **196**: 641–57.

Parry-Smith, D.J., Payne, A.W.R., Michie, A.D., and Attwood, T.K. 1998. CINEMA: A novel colour interactive editor for multiple alignments. *Gene*, **221**: GC57–GC63.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science*, **228**: 834–8.

Sherlin, L.D., Bullock, T.L., Newberry, K.J., Lipman, R.S.A., Hou, Y.M., Beijer, B., Sproat, B.S., and Perona, J.J. 2000. Influence of transfer RNA tertiary structure on aminoacylation efficiency by glutaminyl- and cysteinyl-tRNA synthetases. *Journal of Molecular Biology*, **299**: 431–46.

Sokal, R.R. and Sneath, P.H.A. 1963. *Principles of Numerical Taxonomy*. San Francisco: W.H. Freeman.

Yusupov, M.M., Yusupova, G.Z., Baucom, A., Lieberman, K., Earnest, T.N., Cate, J.H.D., and Noller, H.F. 2001. Crystal structure of the ribosome at 5.5 angstrom resolution. *Science*, **292**: 883–96.

Zimmerman, J.M., Eliezer, N., and Simha, R. 1968. The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology*, **21**: 170–201.