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

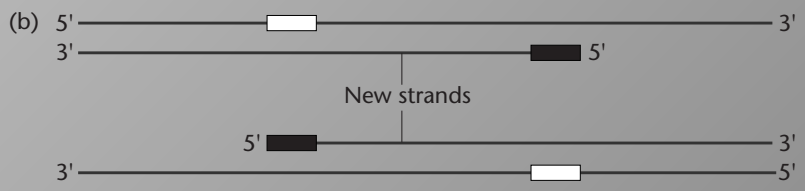
As we mentioned in Section 1.1, we have many protein sequences for which experimentally determined three-dimensional structures are unavailable. A long-standing goal of bioinformatics has been to predict protein structure from sequence. Some methods for doing this will be discussed in Chapter 10 on pattern 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

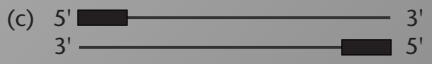
BOX 2.1
Polymerase chain reaction (PCR)

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 sequencing. To carry out PCR, it is not necessary to know the sequence to be amplified, but it is necessary to know the sequence of two short sequences at either end of the region to be amplified. These will be used as primers and are indicated by white boxes below.

The long sequence of DNA containing the region of interest is denatured by heating, and mixed with oligonucleotides of the two primer sequences, indicated by black boxes below. The complementary strands are synthesized by a DNA polymerase called *Taq* polymerase from the thermophilic bacterium *Thermus aquaticus*. This enzyme is able to withstand the high temperatures used in the denaturing cycles used in PCR. The primers determine the position where the polymerase begins. The situation now looks like this:



These molecules are again denatured and the complementary strands are synthesized, and the cycle of denaturation and DNA synthesis is carried out many times. Included in the mixture of products are some pieces of DNA that are bounded by the primers, like this:



These strands can multiply exponentially because the products of the DNA synthesis can be used as templates at the next cycle. After many cycles, the specified sequence dominates the population of DNA sequences, with only negligible fractions of the longer DNA sequences being present.

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	Fract. area
Alanine	Ala	A	67	11.50	0.00	6.00	1.8	1.6	113	0.74
Arginine	Arg	R	148	14.28	52.00	10.76	−4.5	−12.3	241	0.64
Asparagine	Asn	N	96	12.28	3.38	5.41	−3.5	−4.8	158	0.63
Aspartic acid	Asp	D	91	11.68	49.70	2.77	−3.5	−9.2	151	0.62
Cysteine	Cys	C	86	13.46	1.48	5.05	2.5	2.0	140	0.91
Glutamine	Gln	Q	114	14.45	3.53	5.65	−3.5	−4.1	189	0.62
Glutamic acid	Glu	E	109	13.57	49.90	3.22	−3.5	−8.2	183	0.62
Glycine	Gly	G	48	3.40	0.00	5.97	−0.4	1.0	85	0.72
Histidine	His	H	118	13.69	51.60	7.59	−3.2	−3.0	194	0.78
Isoleucine	Ile	I	124	21.40	0.13	6.02	4.5	3.1	182	0.88
Leucine	Leu	L	124	21.40	0.13	5.98	3.8	2.8	180	0.85
Lysine	Lys	K	135	15.71	49.50	9.74	−3.9	−8.8	211	0.52
Methionine	Met	M	124	16.25	1.43	5.74	1.9	3.4	204	0.85
Phenylalanine	Phe	F	135	19.80	0.35	5.48	2.8	3.7	218	0.88
Proline	Pro	P	90	17.43	1.58	6.30	−1.6	−0.2	143	0.64
Serine	Ser	S	73	9.47	1.67	5.68	−0.8	0.6	122	0.66
Threonine	Thr	T	93	15.77	1.66	5.66	−0.7	1.2	146	0.70
Tryptophan	Trp	W	163	21.67	2.10	5.89	−0.9	1.9	259	0.85
Tyrosine	Tyr	Y	141	18.03	1.61	5.66	−1.3	−0.7	229	0.76
Valine	Val	V	105	21.57	0.13	5.96	4.2	2.6	160	0.86
Mean			109	15.35	13.59	6.03	−0.5	−1.4	175	0.74
Std. dev.			28	4.53	21.36	1.72	2.9	4.8	44	0.11

Vol., volume calculated from van der Waals radii (Creighton 1993); Bulk., bulkiness index (Zimmerman, Eliezer, and Simha 1968); Polarity, polarity index (Zimmerman, Eliezer, and Simha 1968); pI, pH of the isoelectric point (Zimmerman, Eliezer, and Simha 1968); Hyd.1, hydrophobicity scale (Kyte and Doolittle 1982); Hyd.2, hydrophobicity scale (Engelman, Steitz, and Goldman 1986); Surface area, surface area accessible to water in unfolded peptide (Miller *et al.* 1987); Fract. area, fraction of accessible area lost when a protein folds (Rose *et al.* 1985).

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

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

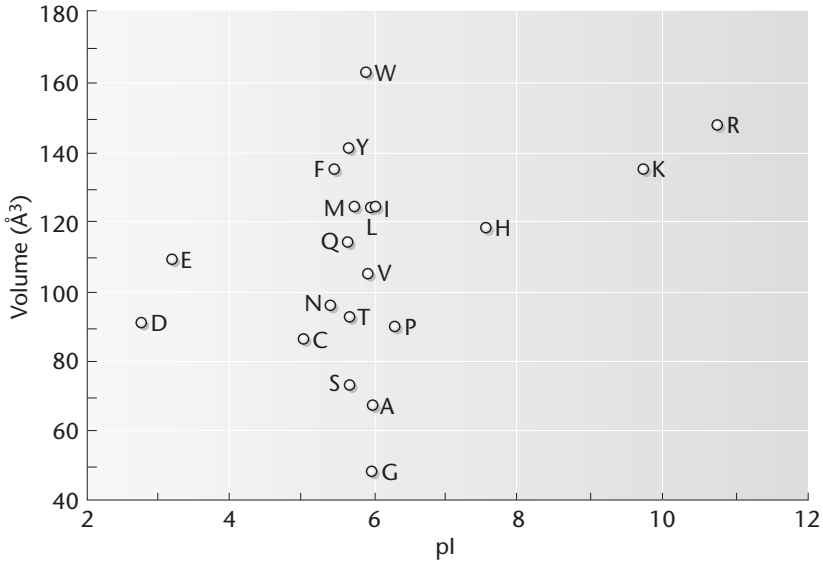


Fig. 2.8 Plot of amino acid volume against pI – two properties thought to be important in protein folding.

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

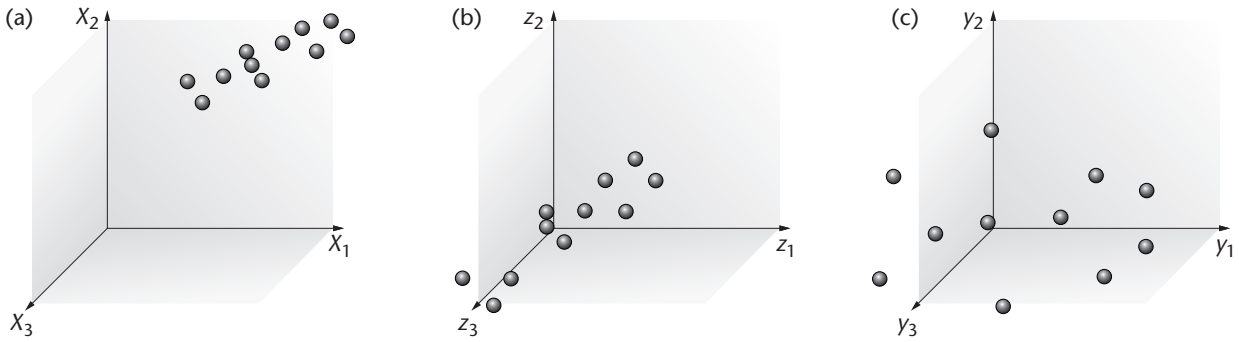


Fig. 2.9 Schematic illustration of principal component analysis. (a) Original data. (b) Scaled and centered on the origin. (c) Rotated onto principal components.

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/).

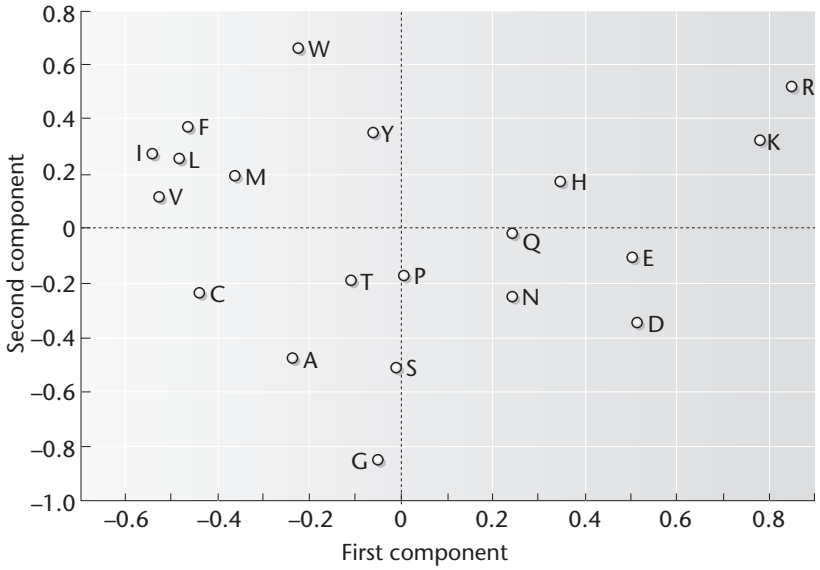


Fig. 2.10 Plot of the amino acids on the first two components of the principal component analysis.

	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.

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., they both tend to be large at the same time and small at the same time. If $C_{jk} < 0$, the properties are negatively correlated, i.e., 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 illustrate features of the data that are not easy to see in the original matrix. For example, volume has a strong positive correlation with surface area and bulkiness, and a fairly weak correlation with the other properties. The two hydrophobicity scales have strong positive correlation with each other and also with the fractional area property, but they have a significant negative correlation with the polarity scale.

It can be shown that the vectors $\mathbf{v}_j$ that define the principal component axes are the eigenvectors of the correlation 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) vector is the eigenvector with the largest eigenvalue. Subsequent PCs can be listed in order of decreasing size of eigenvalue. The first two eigenvalues 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:

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

We know that the variance of each of the z coordinates is 1, hence the total variance of all the coordinates is P. When we change the coordinates to the principal components, we just rotate the points in space, so the total variance in the PC space is still P. The fraction of the total variance represented by the first two PCs is therefore $(\lambda_1 + \lambda_2)/P$, which in our case is $(3.57 + 2.81)/8 = 0.797$. This is why it is useful to look at the data on the PC plot (as in Fig. 2.10). Roughly 80% of the variation in the positioning of the points in the original coordinates can be seen with just two PCs. When points appear close in the two-dimensional plot of the first two PCs, they really are close in the eight-dimensional space, because the remaining six dimensions that we can't see do not contribute much to the distance between the points. This means that if we spot patterns in the data in the PC plot, such as clusters of closely spaced points, then these are likely to give a true impression of the patterns in the full data.

	Vol	Bulk.	Pol.	pI	Hyd.1	Hyd.2	S.A.	Fr.A.
Vol.	1.00	0.73	0.24	0.37	–0.08	–0.16	0.99	0.18
Bulk.	0.73	1.00	–0.20	0.08	0.44	0.32	0.64	0.49
Pol.	0.24	–0.20	1.00	0.27	–0.69	–0.85	0.29	–0.53
pI	0.37	0.08	0.27	1.00	–0.20	–0.27	0.36	–0.18
Hyd.1	–0.08	0.44	–0.67	–0.20	1.00	0.85	–0.18	0.84
Hyd.2	–0.16	0.32	–0.85	–0.27	0.85	1.00	–0.23	0.79
S.A.	0.99	0.64	0.29	0.36	–0.18	–0.23	1.00	0.12
Fr.A.	0.18	0.49	–0.53	–0.18	0.84	0.79	0.12	1.00

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:

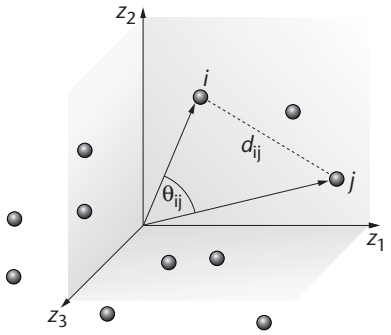


Fig. 2.11 Illustration of the data points as vectors in multidimensional space.

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

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:

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

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 ($\cos\theta$ 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 $\cos\theta$, 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 $\cos\theta$ 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 $\cos\theta$, 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 $\cos\theta$ 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:

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

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 two exactly complementary strands. RNA is also composed of four nucleotide building blocks, but U is used instead of T. RNA molecules are usually single stranded and fold to form complex stem-loop secondary structures by base pairing between short sections of the same strand. Proteins are polymers composed of sequences of 20 types of amino acid linked by peptide bonds.

The process of synthesis of RNA using a DNA strand as a template is called transcription. It is carried out by RNA polymerase. The process of protein synthesis using mRNA as a template is called translation. It is carried out by the ribosome. Protein-coding DNA sequences store information in the form of groups of three bases called codons. Each codon codes for either an amino acid or a stop signal. The mapping from codons to amino acids is known as the genetic code. During translation, the anticodon sequences in tRNA molecules pair with the codon sequences in the mRNA. Each tRNA is charged with a specific amino acid, and this amino acid gets transferred from the tRNA to the growing protein chain due to the catalytic activity of the ribosome.

Amino acids vary greatly in size, charge, hydrophobicity, and other physical properties. Principal component analysis (PCA) is a way of visualizing the important features of multidimensional data sets, such as tables of

amino acid properties. PCA chooses coordinates that are linear combinations of the original variables in such a way that the maximum variability between the data points is explained by the first few coordinates. Clustering analysis is another way of looking for patterns in complex data sets. A large variety of clustering methods is possible, including hierarchical and direct clustering. These methods give slightly different answers; hence some thought is required in order to interpret the resulting clusters and to decide which method is most appropriate for a given data set. Clustering and PCA are applied to the amino acid data in this chapter because they reveal interesting properties of the amino acids and also because the methods are general and are useful in many areas, such as microarray data analysis, as we shall discuss in Chapter 13.

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.