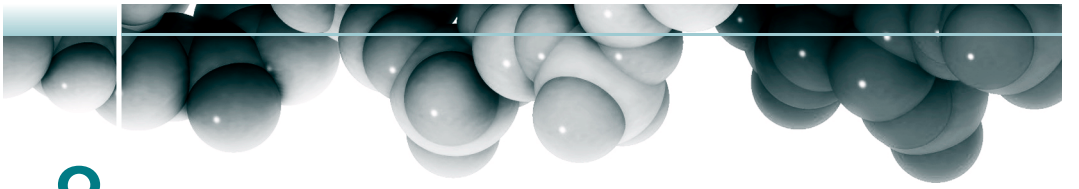


8 Protein structure, purification, characterisation and function analysis

J. WALKER

8.1 IONIC PROPERTIES OF AMINO ACIDS AND PROTEINS

Twenty amino acids varying in size, shape, charge and chemical reactivity are found in proteins and each has at least one codon in the genetic code (Section 5.3.5). Nineteen of the amino acids are α-amino acids (i.e. the amino and carboxyl groups are attached to the carbon atom that is adjacent to the carboxyl group) with the general formula $RCH(NH_2)COOH$, where R is an aliphatic, aromatic or heterocyclic group. The only exception to this general formula is proline, which is an imino acid in which the $-NH_2$ group is incorporated into a five-membered ring. With the exception of the simplest amino acid glycine (R = H), all the amino acids found in proteins contain one asymmetric carbon atom and hence are optically active and have been found to have the L configuration.

For convenience, each amino acid found in proteins is designated by either a three-letter abbreviation, generally based on the first three letters of their name, or a one-letter symbol, some of which are the first letter of the name. Details are given in Table 8.1.

Since they possess both an amino group and a carboxyl group, amino acids are ionised at all pH values, i.e. a neutral species represented by the general formula does not exist in solution irrespective of the pH. This can be seen as follows:

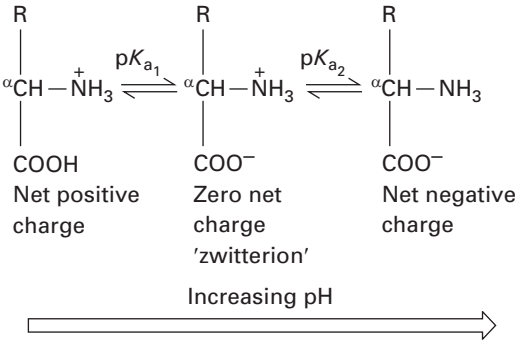


Table 8.1 **Abbreviations for amino acids**

Amino acid	Three-letter code	One-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Asparagine or aspartic acid	Asx	B
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glutamine or glutamic acid	Glx	Z
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W

Table 8.1 (*cont.*)

Amino acid	Three-letter code	One-letter code
Tyrosine	Tyr	Y
Valine	Val	V

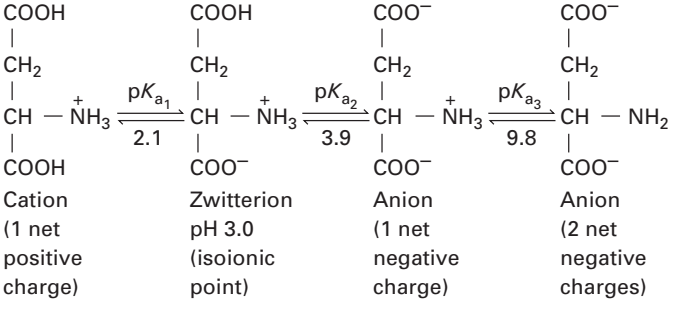
Thus at low pH values an amino acid exists as a cation and at high pH values as an anion. At a particular intermediate pH the amino acid carries no net charge, although it is still ionised, and is called a zwitterion. It has been shown that, in the crystalline state and in solution in water, amino acids exist predominantly as this zwitterionic form. This confers upon them physical properties characteristic of ionic compounds, i.e. high melting point and boiling point, water solubility and low solubility in organic solvents such as ether and chloroform. The pH at which the zwitterion predominates in aqueous solution is referred to as the isoionic point, because it is the pH at which the number of negative charges on the molecule produced by ionisation of the carboxyl group is equal to the number of positive charges acquired by proton acceptance by the amino group. In the case of amino acids this is equal to the isoelectric point (pI), since the molecule carries no net charge and is therefore electrophoretically immobile. The numerical value of this pH for a given amino acid is related to its acid strength (pK_a values) by the equation:

$$\text{pI} = \frac{\text{p}K_{a1} + \text{p}K_{a2}}{2} \tag{8.1}$$

where pK_{a1} and pK_{a2} are equal to the negative logarithm of the acid dissociation constants, K_{a1} and K_{a2} (Section 1.3.2).

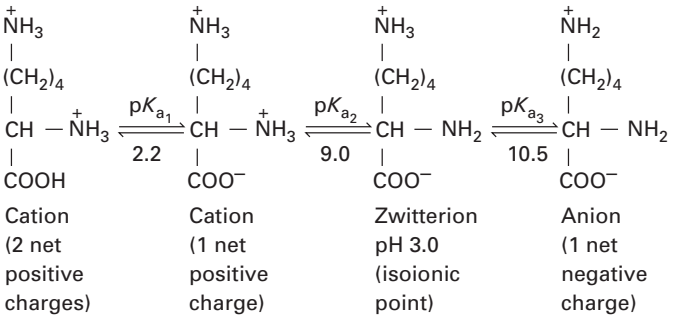
In the case of glycine, pK_{a1} and pK_{a2} are 2, 3 and 9.6, respectively, so that the isoionic point is 6.0. At pH values below this, the cation and zwitterion will coexist in equilibrium in a ratio determined by the Henderson–Hasselbalch equation (Section 1.3.3), whereas at higher pH values the zwitterion and anion will coexist in equilibrium.

For acidic amino acids such as aspartic acid, the ionisation pattern is different owing to the presence of a second carboxyl group:



In this case, the zwitterion will predominate in aqueous solution at a pH determined by pK_{a_1} and pK_{a_2}, and the isoelectric point is the mean of pK_{a_1} and pK_{a_2}.

In the case of lysine, which is a basic amino acid, the ionisation pattern is different again and its isoionic point is the mean of pK_{a_2} and pK_{a_3}:

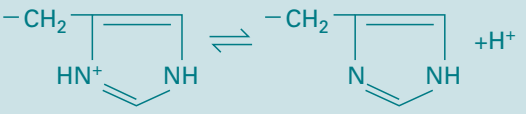
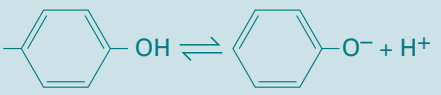


As an alternative to possessing a second amino or carboxyl group, an amino acid side chain may contain in the R of the general formula a quite different chemical group that is also capable of ionising at a characteristic pH. Such groups include a phenolic group (tyrosine), guanidino group (arginine), imidazolyl group (histidine) and sulphydryl group (cysteine) (Table 8.2). It is clear that the state of ionisation of the main groups of amino acids (acidic, basic, neutral) will be grossly different at a particular pH. Moreover, even within a given group there will be minor differences due to the precise nature of the R group. These differences are exploited in the electrophoretic and ion-exchange chromatographic separation of mixtures of amino acids such as those present in a protein hydrolysate (Section 8.4.2).

Proteins are formed by the condensation of the α-amino group of one amino acid with the α-carboxyl of the adjacent amino acid (Section 8.2). With the exception of the two terminal amino acids, therefore, the α-amino and carboxyl groups are all involved in peptide bonds and are no longer ionisable in the protein. Amino, carboxyl, imidazolyl, guanidino, phenolic and sulphydryl groups in the side chains are, however, free to ionise and of course there will be many of these. Proteins fold in such a manner that the majority of these ionisable groups are on the outside of the molecule, where they can interact with the surrounding aqueous medium. Some of these groups are located within the structure and may be involved in electrostatic attractions that help to stabilise the three-dimensional structure of the protein molecule. The relative numbers of positive and negative groups in a protein molecule influence aspects of its physical behaviour, such as solubility and electrophoretic mobility.

The isoionic point of a protein and its isoelectric point, unlike that of an amino acid, are generally not identical. This is because, by definition, the isoionic point is the pH at which the protein molecule possesses an equal number of positive and negative groups formed by the association of basic groups with protons and dissociation of acidic groups, respectively. In contrast, the isoelectric point is the pH at which the protein is electrophoretically immobile. In order to determine electrophoretic mobility experimentally, the protein must be dissolved in a buffered medium containing anions and cations, of low relative molecular mass, that are capable of binding to the multi-ionised protein. Hence the observed balance of charges at the isoelectric point could be due in

Table 8.2 **Ionisable groups found in proteins**

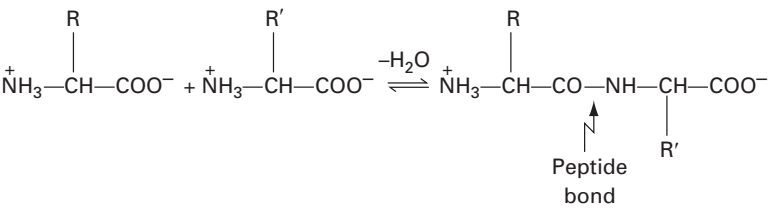
Amino acid group	pH-dependent ionisation	Approx. pK_a
N-terminal α-amino	$-NH_3 \rightleftharpoons NH_2 + H^+$	8.0
C-terminal α-carboxyl	$-COOH \rightleftharpoons COO^- + H^+$	3.0
Asp-β-carboxyl	$-CH_2COOH \rightleftharpoons CH_2COO^- + H^+$	3.9
Glu-γ-carboxyl	$-(CH_2)_2COOH \rightleftharpoons (CH_2)_2COO^- + H^+$	4.1
His-imidazolyl		6.0
Cys-sulphydryl	$-CH_2SH \rightleftharpoons -CH_2S^- + H^+$	8.4
Tyr-phenolic		10.1
Lys-ε-amino	$-(CH_2)_4\overset{+}{N}H_3 \rightleftharpoons -(CH_2)_4NH_2 + H^+$	10.3
Arg-guanidino	$-NH-C(=NH_2^+)-NH_2 \rightleftharpoons -NH-C(=NH)-NH_2 + H^+$	12.5

part to there being more bound mobile anions (or cations) than bound cations (anions) at this pH. This could mask an imbalance of charges on the actual protein.

In practice, protein molecules are always studied in buffered solutions, so it is the isoelectric point that is important. It is the pH at which, for example, the protein has minimum solubility, since it is the point at which there is the greatest opportunity for attraction between oppositely charged groups of neighbouring molecules and consequent aggregation and easy precipitation.

8.2 PROTEIN STRUCTURE

Proteins are formed by condensing the α-amino group of one amino acid or the imino group of proline with the α-carboxyl group of another, with the concomitant loss of a molecule of water and the formation of a peptide bond.



The progressive condensation of many molecules of amino acids gives rise to an unbranched polypeptide chain. By convention, the N-terminal amino acid is taken as the beginning of the chain and the C-terminal amino acid as the end of the chain (proteins are biosynthesised in this direction). Polypeptide chains contain between 20 and 2 000 amino acid residues and hence have a relative molecular mass ranging between about 2 000 and 2 00 000. Many proteins have a relative molecular mass in the range 20 000 to 1 00 000. The distinction between a large peptide and a small protein is not clear. Generally, chains of amino acids containing fewer than 50 residues are referred to as peptides, and those with more than 50 are referred to as proteins. Most proteins contain many hundreds of amino acids (ribonuclease is an extremely small protein with only 103 amino acid residues) and many biologically active peptides contain 20 or fewer amino acids, for example oxytocin (9 amino acid residues), vasopressin (9), enkephalins (5), gastrin (17), somatostatin (14) and lutenising hormone (10).

The primary structure of a protein defines the sequence of the amino acid residues and is dictated by the base sequence of the corresponding gene(s). Indirectly, the primary structure also defines the amino acid composition (which of the possible 20 amino acids are actually present) and content (the relative proportions of the amino acids present).

The peptide bonds linking the individual amino acid residues in a protein are both rigid and planar, with no opportunity for rotation about the carbon–nitrogen bond, as it has considerable double bond character due to the delocalisation of the lone pair of electrons on the nitrogen atom; this, coupled with the tetrahedral geometry around each α-carbon atom, profoundly influences the three-dimensional arrangement which the polypeptide chain adopts.

Secondary structure defines the localised folding of a polypeptide chain due to hydrogen bonding. It includes structures such as the α-helix and β-pleated sheet. Certain of the 20 amino acids found in proteins, including proline, isoleucine, tryptophan and asparagine, disrupt α-helical structures. Some proteins have up to 70% secondary structure but others have none.

Tertiary structure defines the overall folding of a polypeptide chain. It is stabilised by electrostatic attractions between oppositely charged ionic groups ($-\overset{+}{\mathrm{N}}\mathrm{H}_3, \mathrm{COO}^-$), by weak van der Waals forces, by hydrogen bonding, hydrophobic interactions and, in some proteins, by disulphide (-S – S-) bridges formed by the oxidation of spatially adjacent sulphydryl groups (-SH) of cysteine residues (Fig. 8.1). The three-dimensional folding of polypeptide chains is such that the interior consists predominantly of non-polar, hydrophobic amino acid residues such as valine, leucine and phenylalanine. The polar, ionised, hydrophilic residues are found on the outside of the molecule, where they are compatible with the aqueous environment. However, some proteins also have hydrophobic residues on their outside and the presence of these residues is important in the processes of ammonium sulphate fractionation and hydrophobic interaction chromatography (Section 8.3.4).

Quaternary structure is restricted to oligomeric proteins, which consist of the association of two or more polypeptide chains held together by electrostatic attractions, hydrogen bonding, van der Waals forces and occasionally disulphide bridges. Thus disulphide bridges may exist within a given polypeptide chain (intra-chain) or

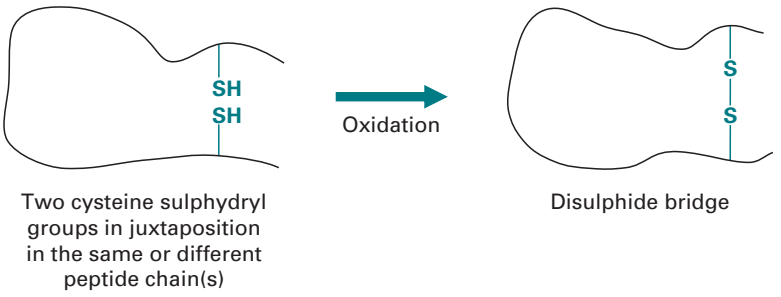


Fig. 8.1 The formation of a disulphide bridge.

linking different chains (inter-chain). An individual polypeptide chain in an oligomeric protein is referred to as a subunit. The subunits in a protein may be identical or different: for example, haemoglobin consists of two α- and two β-chains, and lactate dehydrogenase of four (virtually) identical chains.

Traditionally, proteins are classified into two groups – globular and fibrous. The former are approximately spherical in shape, are generally water soluble and may contain a mixture of α-helix, β-pleated sheet and random structures. Globular proteins include enzymes, transport proteins and immunoglobulins. Fibrous proteins are structural proteins, generally insoluble in water, consisting of long cable-like structures built entirely of either helical or sheet arrangements. Examples include hair keratin, silk fibroin and collagen. The native state of a protein is its biologically active form.

The process of protein denaturation results in the loss of biological activity, decreased aqueous solubility and increased susceptibility to proteolytic degradation. It can be brought about by heat and by treatment with reagents such as acids and alkalis, detergents, organic solvents and heavy-metal cations such as mercury and lead. It is associated with the loss of organised (tertiary) three-dimensional structure and exposure to the aqueous environment of numerous hydrophobic groups previously located within the folded structure.

In enzymes, the specific three-dimensional folding of the polypeptide chain(s) results in the juxtaposition of certain amino acid residues that constitute the active site or catalytic site. Oligomeric enzymes may possess several such sites. Many enzymes also possess one or more regulatory site(s). X-ray crystallography studies have revealed that the active site is often located in a cleft that is lined with hydrophobic amino acid residues but which contains some polar residues. The binding of the substrate at the catalytic site and the subsequent conversion of substrate to product involves different amino acid residues.

Some oligomeric enzymes exist in multiple forms called isoenzymes or isozymes (Section 15.1.2). Their existence relies on the presence of two genes that give similar but not identical subunits. One of the best-known examples of isoenzymes is lactate dehydrogenase, which reversibly interconverts pyruvate and lactate. It is a tetramer and exists in five forms (LDH1 to 5) corresponding to the five permutations of arranging the two types of subunits (H and M), which differ only in a single amino acid substitution, into a tetramer:

H_4	LDH1
H_3M	LDH2
H_2M_2	LDH3
HM_3	LDH4
M_4	LDH5

Each isoenzyme promotes the same reaction but has different kinetic constants (K_m, V_{max}), thermal stability and electrophoretic mobility. The tissue distribution of isoenzymes within an organism is frequently different, for example, in humans LDH1 is the dominant isoenzyme in heart muscle but LDH5 is the most abundant form in liver and muscle. These differences are exploited in diagnostic enzymology to identify specific organ damage, for example following myocardial infarction, and thereby aiding clinical diagnosis and prognosis.

8.2.1 Post-translational modifications

Proteins are synthesised at the ribosome and as the growing polypeptide chain emerges from the ribosome it folds up into its native three-dimensional structure. However, this is often not the final active form of the protein. Many proteins undergo modifications once they leave the ribosome, where one or more amino acid side chains are modified by the addition of a further chemical group; this is referred to as post-translational modification. Such changes include extensive modifications of the protein structure, for example the addition of chains of carbohydrates to form glycoproteins (see Section 8.4.4), where in some cases the final protein consists of as much as over 40% carbohydrate. Less dramatic, but equally important modifications include the addition of a hydroxyl group to proline to produce hydroxyproline (found in the structure of collagen), or the phosphorylation of one or more amino acids (tyrosine, serine and threonine residues are all capable of being phosphorylated). Many cases are known, for example, where the addition of a single phosphate group (by enzymes known as kinases) can activate a protein molecule, and the subsequent removal of the phosphate group (by a phosphatase) can inactivate the molecule; protein phosphorylation reactions are a central part of intracellular signalling. Another example can be found in the post-translational modification of proline residues in the transcription factor HIF (the α subunit of the hypoxia-inducible factor), which is a key oxygen-sensing mechanism in cells. Many proteins therefore are not in their final active, biological form until post-translational modifications have taken place. Over 200 different post-translational modifications have been reported for proteins from microbial, plant and animal sources. Mass spectrometry is used to determine such modifications (see Section 9.5.5).

8.3 PROTEIN PURIFICATION

8.3.1 Introduction

At first sight, the purification of *one* protein from a cell or tissue homogenate that will typically contain 10 000–20 000 different proteins, seems a daunting task. However,

in practice, on average, only four different fractionation steps are needed to purify a given protein. Indeed, in exceptional circumstances proteins have been purified in a single chromatographic step. Since the reason for purifying a protein is normally to provide material for structural or functional studies, the final degree of purity required depends on the purposes for which the protein will be used, i.e. you may not need a protein sample that is 100% pure for your studies. Indeed, to define what is meant by a 'a pure protein' is not easy. Theoretically, a protein is pure when a sample contains only a single protein species, although in practice it is more or less impossible to achieve 100% purity. Fortunately, many studies on proteins can be carried out on samples that contain as much as 5–10% or more contamination with other proteins. This is an important point, since each purification step necessarily involves loss of some of the protein you are trying to purify. An extra (and unnecessary) purification step that increases the purity of your sample from, say, 90% to 98% may mean that you now have a more pure protein, but insufficient protein for your studies. Better to have studied the sample that was 90% pure and have enough to work on!

For example, a 90% pure protein is sufficient for amino acid sequence determination studies as long as the sequence is analysed quantitatively to ensure that the deduced sequence does not arise from a contaminant protein. Similarly, immunisation of a rodent to provide spleen cells for monoclonal antibody production (Section 7.2.2) can be carried out with a sample that is considerably less than 50% pure. As long as your protein of interest raises an immune response it matters not at all that antibodies are also produced against the contaminating proteins. For kinetic studies on an enzyme, a relatively impure sample can be used provided it does not contain any competing activities. On the other hand, if you are raising a monospecific polyclonal antibody in an animal (see Section 7.2.1), it is necessary to have a highly purified protein as antigen, otherwise immunogenic contaminating proteins will give rise to additional antibodies. Equally, proteins that are to have a therapeutic use must be extremely pure to satisfy regulatory (safety) requirements. Clearly, therefore, the degree of purity required depends on the purpose for which the protein is needed.

8.3.2 The determination of protein concentration

The need to determine protein concentration in solution is a routine requirement during protein purification. The only truly accurate method for determining protein concentration is to acid hydrolyse a portion of the sample and then carry out amino acid analysis on the hydrolysate (see Section 8.4.2). However, this is relatively time-consuming, particularly if multiple samples are to be analysed. Fortunately, in practice, one rarely needs decimal place accuracy and other, quicker methods that give a reasonably accurate assessment of protein concentrations of a solution are acceptable. Most of these (see below) are colorimetric methods, where a portion of the protein solution is reacted with a reagent that produces a coloured product. The amount of this coloured product is then measured spectrophotometrically and the amount of colour related to the amount of protein present by appropriate calibration. However, none of these methods is absolute,

since, as will be seen below, the development of colour is often at least partly dependent on the amino acid composition of the protein(s). The presence of prosthetic groups (e.g. carbohydrate) also influences colorimetric assays. Many workers prepare a standard calibration curve using bovine serum albumin (BSA), chosen because of its low cost, high purity and ready availability. However, it should be understood that, since the amino acid composition of BSA will differ from the composition of the sample being tested, any concentration values deduced from the calibration graph can only be approximate.

Ultraviolet absorption

The aromatic amino acid residues tyrosine and tryptophan in a protein exhibit an absorption maximum at a wavelength of 280 nm. Since the proportions of these aromatic amino acids in proteins vary, so too do extinction coefficients for individual proteins. However, for most proteins the extinction coefficient lies in the range 0.4–1.5; so for a complex mixture of proteins it is a fair approximation to say that a solution with an absorbance at 280 nm (A_{280}) of 1.0, using a 1 cm pathlength, has a protein concentration of approximately 1 mg cm^{-3}. The method is relatively sensitive, being able to measure protein concentrations as low as 10 μg cm^{-3}, and, unlike colorimetric methods, is non-destructive, i.e. having made the measurement, the sample in the cuvette can be recovered and used further. This is particularly useful when one is working with small amounts of protein and cannot afford to waste any. However, the method is subject to interference by the presence of other compounds that absorb at 280 nm. Nucleic acids fall into this category having an absorbance as much as 10 times that of protein at this wavelength. Hence the presence of only a small percentage of nucleic acid can greatly influence the absorbance at this wavelength. However, if the absorbances (A) at 280 and 260 nm wavelengths are measured it is possible to apply a correction factor:

$$\text{Protein (mg cm}^{-3}) = 1.55\ A_{280} - 0.76A_{260}$$

The great advantage of this protein assay is that it is non-destructive and can be measured continuously, for example in chromatographic column effluents.

Even greater sensitivity can be obtained by measuring the absorbance of ultraviolet light by peptide bonds. The peptide bond absorbs strongly in the far ultraviolet, with a maximum at about 190 nm. However, because of the difficulties caused by the absorption by oxygen and the low output of conventional spectro-photometers at this wavelength, measurements are usually made at 205 or 210 nm. Most proteins have an extinction coefficient for a 1 μg cm^{-3} solution of about 30 at 205 nm and about 20 at 210 nm. Clearly therefore measuring at these wavelengths is 20 to 30 times more sensitive than measuring at 280 nm, and protein concentration can be measured to less than 1 μg cm^{-3}. However, one disadvantage of working at these lower wavelengths is that a number of buffers and other buffer components commonly used in protein studies also absorb strongly at this wavelength, so it is not always practical to work at this lower wavelength.

Nowadays all purpose-built column chromatography systems (e.g. fast protein liquid chromatography and high-performance liquid chromatography (HPLC)) have

in-line variable wavelength ultraviolet light detectors that monitor protein elution from columns.

Lowry (Folin–Ciocalteau) method

In the past this has been the most commonly used method for determining protein concentration, although it is tending to be replaced by the more sensitive methods described below. The Lowry method is reasonably sensitive, detecting down to 10 μg cm^{-3} of protein, and the sensitivity is moderately constant from one protein to another. When the Folin reagent (a mixture of sodium tungstate, molybdate and phosphate), together with a copper sulphate solution, is mixed with a protein solution, a blue-purple colour is produced which can be quantified by its absorbance at 660 nm. As with most colorimetric assays, care must be taken that other compounds that interfere with the assay are not present. For the Lowry method this includes Tris, zwitterionic buffers such as Pipes and Hepes, and EDTA. The method is based on both the Biuret reaction, where the peptide bonds of proteins react with Cu^{2+} under alkaline conditions producing Cu^{+}, which reacts with the Folin reagent, and the Folin–Ciocalteau reaction, which is poorly understood but essentially involves the reduction of phosphomolybdotungstate to hetero-polymolybdenum blue by the copper-catalysed oxidation of aromatic amino acids. The resultant strong blue colour is therefore partly dependent on the tyrosine and tryptophan content of the protein sample.

The bicinchoninic acid method

This method is similar to the Lowry method in that it also depends on the conversion of Cu^{2+} to Cu^{+} under alkaline conditions. The Cu^{+} is then detected by reaction with bicinchoninic acid (BCA) to give an intense purple colour with an absorbance maximum at 562 nm. The method is more sensitive than the Lowry method, being able to detect down to 0.5 μg protein cm^{-3}, but perhaps more importantly it is generally more tolerant of the presence of compounds that interfere with the Lowry assay, hence the increasing popularity of the method.

The Bradford method

This method relies on the binding of the dye Coomassie Brilliant Blue to protein. At low pH the free dye has absorption maxima at 470 and 650 nm, but when bound to protein has an absorption maximum at 595 nm. The practical advantages of the method are that the reagent is simple to prepare and that the colour develops rapidly and is stable. Although it is sensitive down to 20 μg protein cm^{-3}, it is only a relative method, as the amount of dye binding appears to vary with the content of the basic amino acids arginine and lysine in the protein. This makes the choice of a standard difficult. In addition, many proteins will not dissolve properly in the acidic reaction medium.

Kjeldahl analysis

This is a general chemical method for determining the nitrogen content of any compound. It is not normally used for the analysis of purified proteins or for monitoring column fractions but is frequently used for analysing complex solid samples and microbiological samples for protein content. The sample is digested by boiling

Example 1 **PROTEIN ASSAY**

Question A series of dilutions of bovine serum albumin (BSA) was prepared and 0.1 cm^3 of each solution subjected to a Bradford assay. The increase in absorbance at 595 nm relative to an appropriate blank was determined in each case, and the results are shown in the table.

Concentration of BSA ($mg\,cm^{-3}$)	A_{595}
1.5	1.40
1.0	0.97
0.8	0.79
0.6	0.59
0.4	0.37
0.2	0.17

A sample (0.1 cm^3) of a protein extract from *E. coli* gave an A_{595} of 0.84 in the same assay. What was the concentration of protein in the *E. coli* extract?

Answer If a graph of BSA concentration against A_{595} is plotted it is seen to be linear. From the graph, at an A_{595} of 0.84 it can be seen that the protein concentration of the *E. coli* extracted is 0.85 $mg\,cm^{-3}$.

with concentrated sulphuric acid in the presence of sodium sulphate (to raise the boiling point) and a copper and/or selenium catalyst. The digestion converts all the organic nitrogen to ammonia, which is trapped as ammonium sulphate. Completion of the digestion stage is generally recognised by the formation of a clear solution. The ammonia is released by the addition of excess sodium hydroxide and removed by steam distillation in a Markham still. It is collected in boric acid and titrated with standard hydrochloric acid using methyl red–methylene blue as indicator. It is possible to carry out the analysis automatically in an autokjeldahl apparatus. Alternatively, a selective ammonium ion electrode may be used to directly determine the content of ammonium ion in the digest. Although Kjeldahl analysis is a precise and reproducible method for the determination of nitrogen, the determination of the protein content of the original sample is complicated by the variation of the nitrogen content of individual proteins and by the presence of nitrogen in contaminants such as DNA. In practice, the nitrogen content of proteins is generally assumed to be 16% by weight.

8.3.3 Cell disruption and production of initial crude extract

The initial step of any purification procedure must, of course, be to disrupt the starting tissue to release proteins from within the cell. The means of disrupting the tissue will depend on the cell type (see Cell disruption, below), but thought must first be given to the composition of the buffer used to extract the proteins.

Extraction buffer

Normally extraction buffers are at an ionic strength (0.1–0.2 M) and pH (7.0–8.0) that is considered to be compatible with that found inside the cell. Tris or phosphate buffers are most commonly used. However, in addition a range of other reagents may be included in the buffer for specific purposes. These include:

- *An anti-oxidant*: Within the cell the protein is in a highly reducing environment, but when released into the buffer it is exposed to a more oxidising environment. Since most proteins contain a number of free sulphydryl groups (from the amino acid cysteine) these can undergo oxidation to give inter- and intramolecular disulphide bridges. To prevent this, reducing agents such as dithiothreitol, β-mercaptoethanol, cysteine or reduced glutathione are often included in the buffer.
- *Enzyme inhibitors*: Once the cell is disrupted the organisational integrity of the cell is lost, and proteolytic enzymes that were carefully packaged and controlled within the intact cells are released, for example from lysosomes. Such enzymes will of course start to degrade proteins in the extract, including the protein of interest. To slow down unwanted proteolysis, all extraction and purification steps are carried out at 4 °C, and in addition a range of protease inhibitors is included in the buffer. Each inhibitor is specific for a particular type of protease, for example serine proteases, thiol proteases, aspartic proteases and metalloproteases. Common examples of inhibitors include: di-isopropylphosphofluoridate (DFP), phenylmethyl sulphonylfluoride (PMSF) and tosylphenylalanyl-chloromethylketone (TPCK) (all serine protease inhibitors); iodoacetate and cystatin (thiol protease inhibitors); pepstatin (aspartic protease inhibitor); EDTA and 1,10-phenanthroline (metalloprotease inhibitors); and amastatin and bestatin (exopeptidase inhibitors).
- *Enzyme substrate and cofactors*: Low levels of substrate are often included in extraction buffers when an enzyme is purified, since binding of substrate to the enzyme active site can stabilise the enzyme during purification processes. Where relevant, cofactors that otherwise might be lost during purification are also included to maintain enzyme activity so that activity can be detected when column fractions, etc. are screened.
- *EDTA*: This can be present to remove divalent metal ions that can react with thiol groups in proteins giving *mercaptids*.

$$R - SH + Me^{2+} \rightarrow R - S - Me^{+} + H^{+}$$

- *Polyvinylpyrrolidone (PVP)*: This is often added to extraction buffers for plant tissue. Plant tissues contain considerable amounts of phenolic compounds (both monomeric, such as *p*-hydroxybenzoic acid, and polymeric, such as tannins) that can bind to enzymes and other proteins by non-covalent forces, including hydrophobic, ionic and hydrogen bonds, causing protein precipitation. These phenolic compounds are also easily oxidised, predominantly by endogenous phenol oxidases, to form quinones, which are highly reactive and can combine with reactive groups in proteins causing cross-linking, and further aggregation and precipitation. Insoluble PVP (which mimics the polypeptide backbone) is therefore added to adsorb the phenolic compounds which

can then be removed by centrifugation. Thiol compounds (reducing agents) are also added to minimise the activity of phenol oxidases, and thus prevent the formation of quinones.

- *Sodium azide*: For buffers that are going to be stored for long periods of time, antibacterial and/or antifungal agents are sometimes added at low concentrations. Sodium azide is frequently used as a bacteriostatic agent.

Membrane proteins

Membrane-bound proteins (normally glycoproteins) require special conditions for extraction as they are not released by simple cell disruption procedures alone. Two classes of membrane proteins are identified. Extrinsic (or peripheral) membrane proteins are bound only to the surface of the cell, normally via electrostatic and hydrogen bonds. These proteins are predominantly hydrophilic in nature and are relatively easily extracted either by raising the ionic concentration of the extraction buffer (e.g. to 1 M NaCl) or by changes of pH (e.g. to pH 3–5 or pH 9–12). Once extracted, they can be purified by conventional chromatographic procedures. Intrinsic membrane proteins are those that are embedded in the membrane (integrated membrane proteins). These invariably have significant regions of hydrophobic amino acids (those regions of the protein that are embedded in the membrane, and associated with lipids) and have low solubility in aqueous buffer systems. Hence, once extracted into an aqueous polar environment, appropriate conditions must be used to retain their solubility. Intrinsic proteins are usually extracted with buffer containing detergents. The choice of detergent is mainly one of trial and error but can include ionic detergents such as sodium dodecyl sulphate (SDS), sodium deoxycholate, cetyl trimethylammonium bromide (CTAB) and CHAPS, and non-ionic detergents such as Triton X-100 and Nonidet P-40.

Once extracted, intrinsic membrane proteins can be purified using conventional chromatographic techniques such as gel filtration, ion-exchange chromatography or affinity chromatography (using lectins). However, in each case it is necessary to include detergent in all buffers to maintain protein solubility. The level of detergent used is normally 10- to 100-fold less than that used to extract the protein, in order to minimise any interference of the detergent with the chromatographic process.

Cell disruption

Unless one is isolating proteins from extracellular fluids such as blood, protein purification procedures necessarily start with the disruption of cells or tissue to release the protein content of the cells into an appropriate buffer. This initial extract is therefore the starting point for protein purification. Clearly one chooses, where possible, a starting material that has a high level of the protein of interest. Depending on the protein being isolated one might therefore start with a microbial culture, plant tissue, or mammalian tissue. The last of these has generally been the tissue of choice where possible, owing to the relatively large amounts of starting material available. However, the ability to clone and overexpress genes for proteins from any source, in both bacteria and yeast, means that nowadays more and more protein purification protocols are starting with a microbial lysate. The different methods available for

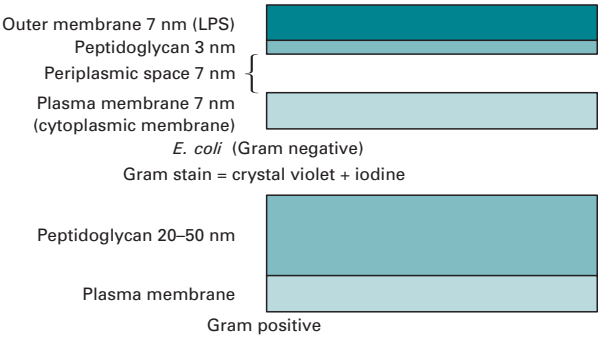


Fig. 8.2 The structure of the cell wall of Gram-positive and of Gram-negative bacteria. LPS, lipopolysaccharide.

disrupting cells are described below. Which method one uses depends on the nature of the cell wall/membrane being disrupted.

Mammalian cells

Mammalian cells are of the order of 10 μm in diameter and enclosed by a plasma membrane, weakly supported by a cytoskeleton. These cells therefore lack any great rigidity and are easy to disrupt by shear forces.

Plant cells

Plant cells are of the order of 100 μm in diameter and have a fairly rigid cell wall, comprising carbohydrate complexes and lignin or wax that surround the plasma membrane. Although the plasma membrane is protected by this outer layer, the large size of the cell still makes it susceptible to shear forces.

Bacteria

Bacteria have cell diameters of the order of 1 to 4 μm and generally have extremely rigid cell walls. Bacteria can be classified as either Gram positive or Gram negative depending on whether or not they are stained by the Gram stain (crystal violet and iodine). In Gram-positive bacteria (Fig. 8.2) the plasma membrane is surrounded by a thick shell of peptidoglycan (20–50 nm), which stains with the Gram stain. In Gram-negative bacteria (e.g. *Escherichia coli*) the plasma membrane is surrounded by a thin (2–3 nm) layer of peptidoglycan but this is compensated for by having a second outer membrane of lipopolysaccharide. The negatively charged lipopolysaccharide polymers interact laterally, being linked by divalent cations such as Mg^{2+}. A number of Gram-negative bacteria secrete proteins into the periplasmic space.

Fungi and yeast

Filamentous fungi and yeasts have a rigid cell wall that is composed mainly of polysaccharide (80–90%). In lower fungi and yeast the polysaccharides are mannan and glucan. In filamentous fungi it is chitin cross-linked with glucans. Yeasts also have a small percentage of glycoprotein in the cell wall, and there is a periplasmic space between the cell wall and cell membrane. If the cell wall is removed the cell content, surrounded by a membrane, is referred to as a spheroplast.

Cell disruption methods

Blenders

These are commercially available, although a typical domestic kitchen blender will suffice. This method is ideal for disrupting mammalian or plant tissue by shear force. Tissue is cut into small pieces and blended, in the presence of buffer, for about 1 min to disrupt the tissue, and then centrifuged to remove debris. This method is inappropriate for bacteria and yeast, but a blender can be used for these microorganisms if small glass beads are introduced to produce a bead mill. Cells are trapped between colliding beads and physically disrupted by shear forces.

Grinding with abrasives

Grinding in a pestle and mortar, in the presence of sand or alumina and a small amount of buffer, is a useful method for disrupting bacterial or plant cells; cell walls are physically ripped off by the abrasive. However, the method is appropriate for handling only relatively small samples. The Dynomill is a large-scale mechanical version of this approach. The Dynomill comprises a chamber containing glass beads and a number of rotating impeller discs. Cells are ruptured when caught between colliding beads. A 600 cm^3 laboratory scale model can process 5 kg of bacteria per hour.

Presses

The use of a press such as a French Press, or the Manton–Gaulin Press, which is a larger-scale version, is an excellent means for disrupting microbial cells. A cell suspension ($\sim$50 cm^3) is forced by a piston-type pump, under high pressure (10 000 PSI = $lbf in.^{-2} \approx 1450$ kPa) through a small orifice. Breakage occurs due to shear forces as the cells are forced through the small orifice, and also by the rapid drop in pressure as the cells emerge from the orifice, which allows the previously compressed cells to expand rapidly and effectively burst. Multiple passes are usually needed to lyse all the cells, but under carefully controlled conditions it can be possible to selectively release proteins from the periplasmic space. The X-Press and Hughes Press are variations on this method; the cells are forced through the orifice as a frozen paste, often mixed with an abrasive. Both the ice crystal and abrasive aid in disrupting the cell walls.

Enzymatic methods

The enzyme lysozyme, isolated from hen egg whites, cleaves peptidoglycan. The peptidoglycan cell wall can therefore be removed from Gram-positive bacteria (see Fig. 8.2) by treatment with lysozyme, and if carried out in a suitable buffer, once the cell wall has been digested the cell membrane will rupture owing to the osmotic effect of the suspending buffer.

Gram-negative bacteria can similarly be disrupted by lysozyme but treatment with EDTA (to remove Ca^{2+}, thus destabilising the outer lipopolysaccharide layer) and the inclusion of a non-ionic detergent to solubilise the cell membrane are also needed. This effectively permeabilises the outer membrane, allowing access of the lysozyme to the peptidoglycan layer. If carried out in an isotonic medium so that the cell membrane is not ruptured, it is possible to selectively release proteins from the periplasmic space.

Yeast can be similarly disrupted using enzymes to degrade the cell wall and either osmotic shock or mild physical force to disrupt the cell membrane. Enzyme digestion alone allows the selective release of proteins from the periplasmic space. The two most commonly used enzyme preparations for yeast are zymolyase or lyticase, both of which have β-1, 3-glucanase activity as their major activity, together with a proteolytic activity specific for the yeast cell wall. Chitinase is commonly used to disrupt filamentous fungi. Enzymic methods tend to be used for laboratory-scale work, since for large-scale work their use is limited by cost.

Sonication

This method is ideal for a suspension of cultured cells or microbial cells. A sonicator probe is lowered into the suspension of cells and high frequency sound waves (<20 kHz) generated for 30–60 s. These sound waves cause disruption of cells by shear force and cavitation. Cavitation refers to areas where there is alternate compression and rarefaction, which rapidly interchange. The gas bubbles in the buffer are initially under pressure but, as they decompress, shock waves are released and disrupt the cells. This method is suitable for relatively small volumes (50–100 cm^3). Since considerable heat is generated by this method, samples must be kept on ice during treatment.

8.3.4 Fractionation methods

Monitoring protein purification

As will be seen below, the purification of a protein invariably involves the application of one or more column chromatographic steps, each of which generates a relatively large number of test tubes (fractions) containing buffer and protein eluted from the column. It is necessary to determine how much protein is present in each tube so that an elution profile (a plot of protein concentration versus tube number) can be produced. Appropriate methods for detecting and quantifying protein in solution are described in Section 8.3.2. A method is also required for determining which tubes contain the protein of interest so that their contents can be pooled and the pooled sample progressed to the next purification step. If one is purifying an enzyme, this is relatively easy as each tube simply has to be assayed for the presence of enzyme activity.

For proteins that have no easily measured biological activity, other approaches have to be used. If an antibody to the protein of interest is available then samples from each tube can be dried onto nitrocellulose and the antibody used to detect the protein-containing fractions using the dot blot method (Section 5.9.2). Alternatively, an immunoassay such as ELISA or radioimmunoassay (Section 7.3.1) can be used to detect the protein. If an antibody is not available, then portions from each fraction can be run on a sodium dodecyl sulphate–polyacrylamide gel and the protein-containing fraction identified from the appearance of the protein band of interest on the gel (Section 10.3.1).

An alternative approach that can be used for cloned genes that are expressed in cells is to express the protein as a fusion protein, i.e. one that is linked via a short peptide sequence to a second protein. This can have advantages for protein purification (see Section 8.3.5). However, it can also prove extremely useful for monitoring the purification of a protein that has no easily measurable activity. If the second protein is an enzyme that can be easily assayed (e.g. using a simple colorimetric

assay), such as β-galactosidase, then the presence of the protein of interest can be detected by the presence of the linked β-galactosidase activity.

A successful fractionation step is recognised by an increase in the specific activity of the sample, where the specific activity of the enzyme relates its total activity to the total amount of protein present in the preparation:

$$\text{specific activity} = \frac{\text{total units of enzyme in fraction}}{\text{total amount of protein in fraction}}$$

The measurement of units of an enzyme relies on an appreciation of certain basic kinetic concepts and upon the availability of a suitable analytical procedure. These are discussed in Section 15.2.2.

The amount of enzyme present in a particular fraction is expressed conventionally not in terms of units of mass or moles but in terms of units based upon the rate of the reaction that the enzyme promotes. The international unit (IU) of an enzyme is defined as the amount of enzyme that will convert 1 μmole of substrate to product in 1 minute under defined conditions (generally 25 or 30 °C at the optimum pH). The SI unit of enzyme activity is defined as the amount of enzyme that will convert 1 mole of substrate to product in 1 second. It has units of katal (kat) such that $1\ \text{kat} = 6 \times 10^7$ IU and $1\ \text{IU} = 1.7 \times 10^{-8}$ kat. For some enzymes, especially those where the substrate is a macromolecule of unknown relative molecular mass (e.g. amylase, pepsin, RNase, DNase), it is not possible to define either of these units. In such cases arbitrary units are used generally that are based upon some observable change in a chemical or physical property of the substrate.

For a purification step to be successful, therefore, the specific activity of the protein must be greater after the purification step than it was before. This increase is best represented as the fold purification:

$$\text{fold purification} = \frac{\text{specific activity of fraction}}{\text{original specific activity}}$$

A significant increase in specific activity is clearly necessary for a successful purification step. However, another important factor is the yield of the step. It is no use having an increased specific activity if you lose 95% of the protein you are trying to purify. Yield is defined as follows:

$$\text{yield} = \frac{\text{units of enzyme in fraction}}{\text{units of enzyme in original preparation}}$$

A yield of 70% or more in any purification step would normally be considered as acceptable. Table 8.3 shows how yield and specific activity vary during a purification schedule.

Preliminary purification steps

The initial extract, produced by the disruption of cells and tissue, and referred to at this stage as a homogenate, will invariably contain insoluble matter. For example, for mammalian tissue there will be incompletely homogenised connective and/or vascular tissue, and small fragments of non-homogenised tissue. This is most easily removed by filtering through a double layer of cheesecloth or by low speed (5 000 *g*)

Example 2 ENZYME FRACTIONATION

Question A tissue homogenate was prepared from pig heart tissue as the first step in the preparation of the enzyme aspartate aminotransferase (AAT). Cell debris was removed by filtration and nucleic acids removed by treatment with polyethyleneimine, leaving a total extract (solution A) of 2 dm^3. A sample of this extract (50 mm^3) was added to 3 cm^3 of buffer in a 1 cm pathlength cuvette and the absorbance at 280 nm shown to be 1.7.

(i) Determine the approximate protein concentration in the extract, and hence the total protein content of the extract.

(ii) One unit of AAT enzyme activity is defined as the amount of enzyme in 3 cm^3 of substrate solution that causes an absorbance change at 260 nm of 0.1 min^{-1}. To determine enzyme activity, 100 mm^3 of extract was added to 3 cm^3 of substrate solution and an absorbance change of 0.08 min^{-1} was recorded. Determine the number of units of AAT actively present per cm^3 of extract A, and hence the total number of enzyme units in the extract.

(iii) The initial extract (solution A) was then subjected to ammonium sulphate fractionation. The fraction precipitating betweeen 50% and 70% saturation was collected and redissolved in 120 cm^3 of buffer (solution B). Solution B (5 mm^3 (0.005 cm^3)) was added to 3 cm^3 of buffer and the absorbance at 280 nm determined to be 0.89 using a 1 cm pathlength cuvette. Determine the protein concentration, and hence total protein content, of solution B.

(iv) Solution B 20 mm^3 was used to assay for AAT activities and an absorbance change of 0.21 per min at 260 nm was recorded. Determine the number of AAT units cm^{-3} in solution B and hence the total number of enzyme units in solution B.

(v) From your answers to (i) to (iv), determine the specific activity of AAT in both solutions A and B.

(vi) From your answers to question (v), determine the fold purification achieved by the ammonium sulphate fractionation step.

(vii) Finally, determine the yield of AAT following the ammonium sulphate fractionation step.

Answer

(i) Assuming the approximation that a 1 mg protein cm^{-3} solution has an absorbance of 1.0 at 280 nm using a 1 cm pathlength cell, then we can deduce that the protein concentration *in the cuvette* is approximately 1.7 mg cm^{-3}. Since 50 µl (0.05 cm^3) of the solution A was added to 3.0 cm^3 then the solution A sample had been diluted by a factor of 3.05/0.05 = 61.

Therefore the protein concentration of solution A is 61 × 1.7 mg cm^{-3} = ~104 mg cm^{-3}. Since there is 2 dm^3 (2000 cm^3) of solution A, the *total* amount of protein in solution A is 2000 × 104 = 208 000 mg or 208 g.

(ii) Since one enzyme unit causes an absorbance change of 0.1 per minute, there was 0.08/0.1 = 0.8 enzyme units in the cuvette. These 0.8 enzyme units came from the 100 mm^3 of solution A that was added to the cuvette.

Example 2 (*cont.*)

Therefore in 100 mm^3 of solution A there is 0.8 enzyme unit.
Therefore in 1 cm^3 of solution A there are 8.0 enzyme units.
Since we have 2000 cm^3 of solution A there is a total of $2000 \times 8.0 = 16\,000$ enzyme units in solution A.

(iii) Using the same approach as in Example 2(i), the protein concentration of solution B is $3.005/0.005 \times 0.89 = 601 \times 0.89 = 535$ mg cm^{-3}.
Therefore the total protein present in solution B $= 120 \times 535 = 64\,200$ mg.

(iv) Using the same approach as in Example 2(ii), there are $0.21/0.1 = 2.1$ units of enzyme activity in the cuvette. These units came from the 20 mm^3 that was added to the cell.
Therefore, 20 mm^3 (0.020 cm^3) of solution B contains 2.1 enzyme units.
Thus, 1 cm^3 of solution B contains $1.0/0.02 \times 2.1 = 105$ units. Therefore, solution B has 105 units cm^{-3}.
Since there are 120 cm^3 of solution B, total units in solution B $= 120 \times 105 = 12\,600$.

(v) For solution A, specific activity $= 16\,000/208\,000 = 0.077$ units mg^{-1}.
For solution B, specific activity $= 12\,600/64\,200 = 0.197$ units mg^{-1}.

(vi) Fold purification $= 0.197/0.077 = 2.6$ (approx.).

(vii) Yield $= (12\,600/16\,000) \times 100\% = 79\%$.

centrifugation. Any fat floating on the surface can be removed by coarse filtration through glass wool or cheesecloth. However, the solution will still be cloudy with organelles and membrane fragments that are too small to be conveniently removed by filtration or low speed centrifugation. These may not be much of a problem as they will often be lost in the preliminary stages of protein purification, for example during salt fractionation. However, if necessary they can be removed first by precipitation using materials such as Celite (a diatomaceous earth that provides a large surface area to trap the particles), Cell Debris Remover (CDR, a cellulose-based absorber), or any number of flocculants such as starch, gums, tannins or polyamines, the resultant precipitate being removed by centrifugation or filtration.

It is tempting to assume that the cell extract contains only protein, but of course a range of other molecules is present such as DNA, RNA, carbohydrate and lipid as well as any number of small molecular weight metabolites. Small molecules tend to be removed later on during dialysis steps or steps that involve fractionation based on size (e.g. gel filtration) and therefore are of little concern. However, specific attention has to be paid at this stage to macromolecules such as nucleic acids and polysaccharides. This is particularly true for bacterial extracts, which are particularly viscous owing to the presence of chromosomal DNA. Indeed microbial extracts can be extremely difficult to centrifuge to produce a supernatant extract. Some workers include DNase I in the extraction buffer to reduce viscosity, the small DNA fragments generated being removed at later dialysis/gel filtration steps. Likewise RNA can be removed by treatment with RNase. DNA and RNA can also be removed by precipitation with protamine

Table 8.3 **Example of a protein purification schedule**

Fraction	Volume (cm^3)	Protein concentration ($mg\ U\ cm^{-3}$)	Total protein (mg)	Activity[a] ($mg\ U\ cm^{-3}$)	Total activity (U)	Specific activity ($U\ mg^{-1}$)	Purification factor[b]	Overall yield[c] (%)
Homogenate	8 500	40	340 000	1.8	15 300	0.045	1	100
45%–70%$(NH_4)_2SO_4$	530	194	103 000	23.3	12 350	0.12	2.7	81
CM-cellulose	420	19.5	8 190	25	10 500	1.28	28.4	69
Affinity chromotography	48	2.2	105.6	198	9 500	88.4	1 964	62
DEAE-Sepharose	12	2.3	27.6	633	7 600	275	6 110	50

Notes: [a]The unit of enzyme activity (U) is defined as that amount which produces 1 μmole of product per minute under standard assay conditions.
[b]Defined as: purification factor = (specific activity of fraction/specific activity of homogenate).
[c]Defined as: overall yield = (total activity of fraction/total activity of homogenate).
Reproduced with permission from *Methods in Molecular Biology*, **59**, *Protein Purification Protocols*, ed. S. Doonan (1996), Humana Press Inc., Totowa, NJ.

sulphate. Protamine sulphate is a mixture of small, highly basic (i.e. positively charged) proteins, whose natural role is to bind to DNA in the sperm head. (Protamines are usually extracted from fish organs, which are obtained as a waste product at canning factories.) These positively charged proteins bind to negatively charged phosphate groups on nucleic acids, thus masking the charged groups on the nucleic acids and rendering them insoluble. The addition of a solution of protamine sulphate to the extract therefore precipitates most of the DNA and RNA, which can subsequently be removed by centrifugation. An alternative is to use polyethyleneimine, a synthetic long chain cationic (i.e. positively charged) polymer (molecular mass 24 kDa). This also binds to the phosphate groups in nucleic acids, and is very effective, precipitating DNA and RNA almost instantly. For bacterial extracts, carbohydrate capsular gum can also be a problem as this can interfere with protein precipitation methods. This is best removed by totally precipitating the protein with ammonium sulphate (see below) leaving the gum in solution. The protein can then be recovered by centrifugation and redissolved in buffer. However, if lysozyme (plus detergent) is used to lyse the cells (see Section 8.3.3) capsular gum will not be a problem as it is digested by the lysozyme.

The clarified extract is now ready for protein fractionation steps to be carried out. The concentration of the protein in this initial extract is normally quite low, and in fact the major contaminant at this stage is water! The initial purification step is frequently based on solubility methods. These methods have a high capacity, can therefore be easily applied to large volumes of initial extracts and also have the advantage of concentrating the protein sample. Essentially, proteins that differ considerably in their physical characteristics from the protein of interest are removed at this stage, leaving a more concentrated solution of proteins that have more closely similar physical characteristics. The next stages, therefore, involve higher resolution techniques that can separate proteins with similar characteristics. Invariably these high resolution techniques are chromatographic. Which technique to use, and in which order, is more often than not a matter of trial and error. The final research paper that describes in four pages a three-step, four-day protein purification procedure invariably belies the months of hard work that went into developing the final 'simple' purification protocol!

All purification techniques are based on exploiting those properties by which proteins differ from one another. These different properties, and the techniques that exploit these differences, are as follows.

Stability

Denaturation fractionation exploits differences in the heat sensitivity of proteins. The three-dimensional (tertiary) structure of proteins is maintained by a number of forces, mainly hydrophobic interactions, hydrogen bonds and sometimes disulphide bridges. When we say that a protein is denatured we mean that these bonds have by some means been disrupted and that the protein chain has unfolded to give the insoluble, 'denatured' protein. One of the easiest ways to denature proteins in solution is to heat them. However, different proteins will denature at different temperatures, depending on their different thermal stabilities; this, in turn, is a measure of the number of bonds holding the tertiary structure together. If the protein of interest is particularly heat stable, then heating the extract to a temperature at which the protein is stable yet other

proteins denature can be a very useful preliminary step. The temperature at which the protein being purified is denatured is first determined by a small-scale experiment. Once this temperature is known, it is possible to remove more thermolabile contaminating proteins by heating the mixture to a temperature 5–10 °C below this critical temperature for a period of 15–30 min. The denatured, unwanted protein is then removed by centrifugation. The presence of the substrate, product or a competitive inhibitor of an enzyme often stabilises it and allows an even higher heat denaturation temperature to be employed. In a similar way, proteins differ in the ease with which they are denatured by extremes of pH (< 3 and > 10). The sensitivity of the protein under investigation to extreme pH is determined by a small-scale trial. The whole protein extract is then adjusted to a pH not less than 1 pH unit within that at which the test protein is precipitated. More sensitive proteins will precipitate and are removed by centrifugation.

Solubility

Proteins differ in the balance of charged, polar and hydrophobic amino acids that they display on their surfaces. Charged and polar groups on the surface are solvated by water molecules, thus making the protein molecule soluble, whereas hydrophobic residues are masked by water molecules that are necessarily found adjacent to these regions. Since solubility is a consequence of solvation of charged and polar groups on the surfaces of the protein, it follows that, under a particular set of conditions, proteins will differ in their solubilities. In particular, one exploits the fact that proteins precipitate differentially from solution on the addition of species such as neutral salts or organic solvents. It should be stressed here that these methods precipitate native (i.e. active) protein that has become insoluble by aggregation; we have not denatured the protein.

Salt fractionation is frequently carried out using ammonium sulphate. As increasing salt is added to a protein solution, so the salt ions are solvated by water molecules in the solution. As the salt concentration increases, freely available water molecules that can solvate the ions become scarce. At this stage those water molecules that have been forced into contact with hydrophobic groups on the surface of the protein are the next most freely available water molecules (rather than those involved in solvating polar groups on the protein surface, which are bound by electrostatic interactions and are far less easily given up) and these are therefore removed to solvate the salt molecules, thus leaving the hydrophobic patches exposed. As the ammonium sulphate concentration increases, the hydrophobic surfaces on the protein are progressively exposed. Thus revealed, these hydrophobic patches cause proteins to aggregate by hydrophobic interaction, resulting in precipitation. The first proteins to aggregate are therefore those with the most hydrophobic residues on the surface, followed by those with less hydrophobic residues. Clearly the aggregates formed are made of mixtures of more than one protein. Individual identical molecules do not seek out each other, but simply bind to another adjacent molecule with an exposed hydrophobic patch. However, many proteins are precipitated from solution over a narrow range of salt concentrations, making this a suitably simple procedure for enriching the proteins of interest.

Organic solvent fractionation is based on differences in the solubility of proteins in aqueous solutions containing water-miscible organic solvents such as ethanol, acetone and butanol. The addition of organic solvent effectively ‘dilutes out’ the water present

(reduces the dielectric constant) and at the same time water molecules are used up in hydrating the organic solvent molecules. Water of solvation is therefore removed from the charged and polar groups on the surface of proteins, thus exposing their charged groups. Aggregation of proteins therefore occurs by charge (ionic) interactions between molecules. Proteins consequently precipitate in decreasing order of the number of charged groups on their surface as the organic solvent concentration is increased.

Organic polymers can also be used for the fractional precipitation of proteins. This method resembles organic solvent fractionation in its mechanism of action but requires lower concentrations to cause protein precipitation and is less likely to cause protein denaturation. The most commonly used polymer is polyethylene glycol (PEG), with a relative molecular mass in the range 6000–20 000.

The fractionation of a protein mixture using ammonium sulphate is given here as a practical example of fractional precipitation. As explained above, as increasing amounts of ammonium sulphate are dissolved in a protein solution, certain proteins start to aggregate and precipitate out of solution. Increasing the salt strength results in further, different proteins precipitating out. By carrying out a controlled pilot experiment where the percentage of ammonium sulphate is increased stepwise say from 10% to 20% to 30% etc., the resultant precipitate at each step being recovered by centrifugation, redissolved in buffer and analysed for the protein of interest, it is possible to determine a fractionation procedure that will give a significantly purified sample. In the example shown in Table 8.3, the original homogenate was made in 45% ammonium sulphate and the precipitate recovered and discarded. The supernatant was then made in 70% ammonium sulphate, the precipitate collected, redissolved in buffer, and kept, with the supernatant being discarded. This produced a purification factor of 2.7. As can be seen, a significant amount of protein has been removed at this step (237 000 mg of protein) while 81% of the total enzyme present was recovered, i.e. the yield was good. This step has clearly produced an enrichment of the protein of interest from a large volume of extract and at the same time has concentrated the sample.

Isoelectric precipitation fractionation is based upon the observations that proteins have their minimum solubility at their isoelectric point. At this pH there are equal numbers of positive and negative charges on the protein molecule; intermolecular repulsions are therefore minimised and protein molecules can approach each other. This therefore allows opposite charges on different molecules to interact, resulting in the formation of insoluble aggregates. The principle can be exploited either to remove unwanted protein, by adjusting the pH of the protein extract so as to cause the precipitation of these proteins but not that of the test protein, or to remove the test protein, by adjusting the pH of the extract to its pI. In practice, the former alternative is preferable, since some denaturation of the precipitation protein inevitably occurs.

Finally, an unusual solubility phenomenon can be utilised in some cases for protein purification from *E. coli*. Early workers who were overexpressing heterologous proteins in *E. coli* at high levels were alarmed to discover that, although their protein was expressed in high yield (up to 40% of the total cell protein), the protein aggregated to form insoluble particles that became known as inclusion bodies. Initially this was seen as a major impediment to the production of proteins in *E. coli*, the inclusion bodies effectively being a mixture of monomeric and polymeric denatured proteins formed

by partial or incorrect folding, probably due to the reducing environment of the *E. coli* cytoplasm. However, it was soon realised that this phenomenon could be used to advantage in protein purification. The inclusion bodies can be separated from a large proportion of the bacterial cytoplasmic protein by centrifugation, giving an effective purification step. The recovered inclusion bodies must then be solubilised and denatured and subsequently allowed to refold slowly to their active, native configuration. This is normally achieved by heating in 6 M guanidinium hydrochloride (to denature the protein) in the presence of a reducing agent (to disrupt any disulphide bridges). The denatured protein is then either diluted in buffer or dialysed against buffer, at which time the protein slowly refolds. Although the refolding method is not always 100% successful, this approach can often produce protein that is 50% or more pure.

Having carried out an initial fractionation step such as that described above, one would then move towards using higher resolution chromatographic methods. Chromatographic techniques for purifying proteins are summarised in Table 8.4, and some of the more commonly used methods are outlined below. The precise practical details of each technique are discussed in Chapter 11.

Charge

Proteins differ from one another in the proportions of the charged amino acids (aspartic and glutamic acids, lysine, arginine and histidine) that they contain. Hence proteins will differ in net charge at a particular pH. This difference is exploited in ion-exchange chromatography (Section 11.6), where the protein of interest is bound onto a solid support material bearing charged groups of the opposite sign (ion-exchange resin). Proteins with the same charge as the resin pass through the column to waste, after which bound proteins, containing the protein of interest, are selectively released from the column by gradually increasing the strength of salt ions in the buffer passing through the column or by gradually changing the pH of the eluting buffer. These ions compete with the protein for binding to the resin, the more weakly charged protein being eluted at the lower salt strength and the more strongly charged protein being eluted at higher salt strengths.

Size

Differences between proteins can be exploited in molecular exclusion (also known as gel filtration) chromatography. The gel filtration medium consists of a range of beads with slighly differing amounts of cross-linking and therefore slightly different pore sizes. The separation process depends on the different abilities of the various proteins to enter some, all or none of the beads, which in turn relates to the size of this protein (Section 11.7). The method has limited resolving power, but can be used to obtain a separation between large and small protein molecules and therefore be useful when the protein of interest is either particularly large or particularly small. This method can also be used to determine the relative molecular mass of a protein (Section 11.7.2) and for concentrating or desalting a protein solution (Section 11.7.2).

Affinity

Certain proteins bind strongly to specific small molecules. One can take advantage of this by developing an affinity chromatography system where the small molecule

Table 8.4 **Summary of chromatographic techniques commonly used in protein purification**

Technique	Property exploited	Capacity	Resolution	Practical points	Further details
Hydrophobic interaction	Hydrophobicity	High	Medium	Can cope with high ionic strength samples, e.g. ammonium sulphate precipitates. Fractions are of varying pH and/or ionic strength. Medium yield. Commonly used in early stages of purification protocol. Unpredictable	Section 11.4.3
Ion exchange	Charge	High	Medium	Sample ionic strength must be low. Fractions are of varying pH and/or ionic strength. Medium yield. Commonly used in early stages of purification protocol	Section 11.6
Affinity	Biological function	Medium (cost limited)	High	Limited by availability of immobilised ligand. Elution may denature protein. Yield medium–low. Commonly used towards end of purification protocol	Section 11.8
Dye affinity	Structure and hydrophobicity		High	Necessary to carry out initial screening of a wide range of dye–ligand supports	Section 11.8.5
Covalent	Thiol groups	Medium–low	High	Specific for thiol-containing proteins. Limited by high cost and long (3 h) regeneration time	Section 11.8.6
Metal chelate	Imidazole, thiol, tryptophan groups	Medium–low	High	Expensive	Section 11.8.4
Exclusion	Molecular size	Medium	Low	Can give information about protein molecular weight. Good for desalting protein samples	Section 11.7

(ligand) is bound to an insoluble support. When a crude mixture of proteins containing the protein of interest is passed through the column, the ligand binds the protein to the matrix whilst all other proteins pass through the column. The bound protein can then be eluted from the column by changing the pH, increasing salt strength or passing through a high concentration of unbound free ligand. For example, the protein concanavalin A (con A) binds strongly to glucose. An affinity column using glucose as the ligand can therefore be used to bind con A to the matrix, and the con A can be recovered by passing a high concentration of glucose through the column. Affinity chromatography is covered in detail in Section 11.8.

Hydrophobicity

Proteins differ in the amount of hydrophobic amino acids that are present on their surface. This difference can be exploited in salt fractionation (see above) but can also be used in a higher resolution method using hydrophobic interaction chromatography (HIC) (Section 11.4.3). A typical column material would be phenyl-Sepharose, where phenyl groups are bonded to the insoluble support Sepharose. The protein mixture is loaded on the column in high salt (to ensure hydrophobic patches are exposed) where hydrophobic interaction will occur between the phenyl groups on the resin and hydrophobic regions on the proteins. Proteins are then eluted by applying a decreasing salt gradient to the column and should emerge from the column in order of increasing hydrophobicity. However, some highly hydrophobic proteins may not even be eluted in the total absence of salt. In this case it is necessary to add a small amount of water-miscible organic solvent such as propanol or ethylene glycol to the column buffer solution. This will compete with the proteins for binding to the hydrophobic matrix and will elute any remaining proteins.

8.3.5 Engineering proteins for purification

With the ability to clone and overexpress genes for proteins using genetic engineering methodology has also come the ability to aid considerably the purification process by manipulation of the gene of interest prior to expression. These manipulations are carried out either to ensure secretion of the proteins from the cell or to aid protein purification.

Ensuring secretion from the cell

For cloned genes that are being expressed in microbial or eukaryotic cells, there are a number of advantages in manipulating the gene to ensure that the protein product is secreted from the cell:

- *To facilitate purification*: Clearly if the protein is secreted into the growth medium, there will be far fewer contaminating proteins present than if the cells had to be ruptured to release the protein, when all the other intracellular proteins would also be present.
- *Prevention of intracellular degradation of the cloned protein*: Many cloned proteins are recognised as 'foreign' by the cell in which they are produced and are therefore degraded by intracellular proteases. Secretion of the protein into the culture medium should minimise this degradation.

- *Reduction of the intracellular concentration of toxic proteins*: Some cloned proteins are toxic to the cell in which they are produced and there is therefore a limit to the amount of protein the cell will produce before it dies. Protein secretion should prevent cell death and result in continued production of protein.
- *To allow post-translational modification of proteins*: Most post-translational modifications of proteins occur as part of the secretory pathway, and these modifications, for example glycosylation (see Section 8.4.4), are a necessary process in producing the final protein structure. Since prokaryotic cells do not glycosylate their proteins, this explains why many proteins have to be expressed in eukaryotic cells (e.g. yeast) rather than in bacteria. The entry of a protein into a secretory pathway and its ultimate destination is determined by a short amino acid sequence (signal sequence) that is usually at the N terminus of the protein. For proteins going to the membrane or outside the cell the route is via the endoplasmic reticulum and Golgi apparatus, the signal sequence being cleaved-off by a protease prior to secretion. For example, human γ-interferon has been secreted from the yeast *Pichia pastoris* using the protein's native signal sequence. Also there are a number of well-characterised yeast signal sequences (e.g. the α-factor signal sequence) that can be used to ensure secretion of proteins cloned into yeast.

Fusion proteins to aid protein purification

This approach requires an additional gene to be joined to the gene of the protein of interest such that the protein is produced as a fusion protein (i.e. linked to this second protein, or tag). As will be seen below, the purpose of this tag is to provide a means whereby the fusion protein can be selectively removed from the cell extract. The fusion protein can then be cleaved to release the protein of interest from the tag protein. Clearly the amino acid sequence of the peptide linkage between tag and protein has to be carefully designed to allow chemical or enzymatic cleavage of this sequence. The following are just a few examples of many different types of fusion proteins that have been used to aid protein purification.

Flag™

This is a short hydrophilic amino acid sequence that is attached to the N-terminal end of the protein, and is designed for purification by immunoaffinity chromatography.

Asp-Tyr-Lys-Asp-Asp-Asp-Asp-Lys-Protein

A monoclonal antibody against this Flag sequence is available on an immobilised support for use in affinity chromatography. The cell extract, which includes the Flag-labelled protein, is passed through the column where the antibody binds to the Flag-labelled protein, allowing all other proteins to pass through. This is carried out in the presence of Ca^{2+}, since the binding of the Flag sequence to the monoclonal antibody is Ca^{2+} dependent. Once all unbound protein has been eluted from the column, the Flag-linked protein is released by passing EDTA through the column, which chelates the Ca^{2+}. Finally the Flag sequence is removed by the enzyme enterokinase, which recognises the following amino acid sequence and cleaves the C-terminal to the lysine residue:

N-Asp-Asp-Asp-Lys-C.

Using this approach, granulocyte-macrophage colony-stimulating factor (GMCSF) was cloned in and secreted from yeast, and purified in a single step. GMCSF was produced in the cell as signal peptide-Flag-gene. The signal sequence used was the signal sequence for the outer membrane protein OmpA. The Flag-gene protein was thus secreted into the periplasm, the fusion protein purified, and finally the Flag sequence removed, as described above.

Glutathione affinity agarose

In this method the protein of interest is expressed as a fusion protein with the enzyme glutathione *S*-transferase. The cell extract is passed through a column of glutathione-linked agarose beads, where the enzyme binds to the glutathione. Once all unbound protein has been washed through the column, the fusion protein is eluted by passing reduced glutathione through the column. Finally, cleavage of the fusion protein is achieved using human thrombin, which recognises a specific amino acid sequence in the linker region.

Protein A

Protein A binds to the Fc region of the immunoglobulin G (IgG) molecule. The protein of interest is cloned fused to the protein A gene, and the fusion protein purified by affinity chromatography on a column of IgG-Sepharose. The bound fusion protein is then eluted using either high salt or low pH, to disrupt the binding between the IgG molecule and the protein A–protein fusion product. Protein A is then finally removed by treatment with 70% (v/v) formic acid for 2 days, which cleaves an acid-labile Asp-Pro bond in the linker region.

Poly(arginine)

This method requires the addition of a series of arginine residues to the C terminus of the protein to be purified. This makes the protein highly basic (positively charged at neutral pH). The cell extract can therefore be fractionated using cation-exchange chromatography. Bound proteins are sequentially released from the column by applying a salt gradient, with the poly(Arg)-containing protein, because of its high overall positive charge, being the last to be eluted. The poly(Arg) tail is then removed by incubation with the enzyme carboxypeptidase B. Carboxypeptidase B is an exoprotease that sequentially removes arginine or lysine residues from the C terminus of proteins. The arginine residues are therefore sequentially removed from the C terminus, the removal of amino acid residues stopping when the 'normal' (i.e. non-arginine) C-terminal amino acid residue of the protein is reached.

8.4 PROTEIN STRUCTURE DETERMINATION

8.4.1 Relative molecular mass

There are three methods available for determining protein relative molecular mass, M_r, frequently referred to as molecular weight. The first two described here are quick and easy methods that will give a value to $\pm$ 5–10%. For many purposes one simply needs a rough

estimate of size and these methods are sufficient. The third method, mass spectrometry, requires expensive specialist instruments and can give accuracy to ± 0.001%. This kind of accuracy is invaluable in detecting postsynthetic modification of proteins.

SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

This form of electrophoresis, described in Section 10.3.1, separates proteins on the basis of their shape (size), which in turn relates to their relative molecular masses. A series of proteins of known molecular mass (molecular weight markers) are run on a gel on a track adjacent to the protein of unknown molecular mass. The distance each marker protein moves through the gel is measured and a calibration curve of log M_r versus distance moved is plotted. The distance migrated by the protein of unknown M_r is also measured, and from the graph its log M_r and hence M_r is calculated. The method is suitable for proteins covering a large M_r range (10 000–300 000). The method is easy to perform and requires very little material. If silver staining (Section 10.3.7) is used, as little as 1 ng of protein is required. In practice SDS–PAGE is the most commonly used method for determining protein M_r values.

Molecular exclusion (gel filtration) chromatography

The elution volume of a protein from a molecular exclusion chromatography column having an appropriate fractionation range is determined largely by the size of the protein such that there is a logarithmic relationship between protein relative molecular mass and elution volume (Section 11.7.1). By calibrating the column with a range of proteins of known M_r, the M_r of a test protein can be calculated. The method is carried out on HPLC columns ($\sim 1 \times 30$ cm) packed with porous silica beads. Flow rates are about 1 $cm^3\ min^{-1}$, giving a run time of about 12 min, producing sharp, well-resolved peaks. A linear calibration line is obtained by plotting a graph of log M_r versus K_d for the calibrating proteins. K_d is calculated from the following equation:

$$K_d = \frac{(V_e - V_o)}{(V_t - V_o)}$$

where V_o is the volume in which molecules that are wholly excluded from the column material emerge (the excluded volume), V_t is the volume in which small molecules that can enter all the pores emerge (the included volume) and V_e is the volume in which the marker protein elutes. This method gives values that are accurate to ± 10%.

Mass spectrometry

Using either electrospray ionisation (ESI) (Section 9.2.4) or matrix-assisted laser desorption ionisation (MALDI) (Section 9.3.8) intact molecular ions can be produced for proteins and hence their masses accurately measured by mass spectrometry. ESI produces molecular ions from molecules with molecular masses up to and in excess of 100 kDa, whereas MALDI produces ions from intact proteins up to and in excess of 200 kDa. In either case, only low picomole quantities of protein are needed. For example, $\alpha\beta_2$ crystallin gave a molecular mass value (20 200 ± 0.9), in excellent agreement with the deduced mass of 20 201. However, in addition about 10% of the analysed material produced an ion of mass 20 072.2. This showed that some of the purified protein molecules had lost their N-terminal amino acid (lysine). The deduced mass with

the loss of N-terminal lysine was 20 072.8. Clearly mass spectrometry has the ability to provide highly accurate molecular mass measurements for proteins and peptides, which in turn can be used to deduce small changes made to the basic protein structure.

8.4.2 Amino acid analysis

The determination of which of the 20 possible amino acids are present in a particular protein, and in what relative amounts, is achieved by hydrolysing the protein to yield its component amino acids and identifying and quantifying them chromatographically. Hydrolysis is achieved by heating the protein with 6 M hydrochloride acid for 14 h at 110 °C *in vacuo*. Unfortunately, the hydrolysis procedure destroys or chemically modifies the asparagine, glutamine and tryptophan residues. Asparagine and glutamine are converted to their corresponding acids (Asp and Glu) and are quantified with them. Tryptophan is completely destroyed and is best determined spectrophotometrically on the unhydrolysed protein.

The amino acids in the protein hydrolysate are then separated chromatographically. Nowadays this is normally done using the method of precolumn derivatisation, followed by separation by reverse-phase HPLC. In this approach the amino acid hydrolysate is first treated with a molecule that (i) reacts with amino groups in amino acids, (ii) is hydrophobic, thus allowing separation of derivatised amino acids by reversed-phase HPLC and (iii) is easily detected by its ultraviolet absorbance or fluorescence. Reagents routinely used for precolumn derivatisation include *o*-phthalaldehyde and 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate (AQC), which both produce fluorescent derivatives, and phenylisothiocyanate, which produces a phenylthiocarbamyl derivative that is detected by its absorbance at 254 nm. Analysis times can be as little as 20 min, and sensitivity is down to 1 pmole or less of amino acid.

8.4.3 Primary structure determination

For many years the amino acid sequence of a protein was determined from studies made on the purified protein alone. This in turn meant that sequence data available were limited to those proteins that could be purified in sufficiently large amounts. Knowledge of the complete primary structure of the protein was (and still is) a prerequisite for the determination of the three-dimensional structure of the protein, and hence an understanding of how that protein functions. However, nowadays the protein biochemist is normally satisfied with data from just a relatively short length of sequence either from the N terminus of the protein or from an internal sequence, obtained by sequencing peptides produced by cleavage of the native protein. The sequence data will then most likely be used for one of three purposes:

- To search sequence databases to see whether the protein of interest has already been isolated, and hence can therefore be identified. For this type of search extremely short lengths of sequence (three to five residues), known as sequence tags, need to be used. Examples of this type of data search are given in Sections 8.5.1 and 9.5.2.

- To search for sequence homology using computerised databases in order to identify the function of the protein. For example, the search may show significant sequence identity with the amino acid sequence of some known protein tyrosine kinases, strongly suggesting that the protein is also a tyrosine kinase.
- The sequence will be used to design an oligonucleotide probe for selecting appropriate clones from complementary DNA libraries. In this way the DNA coding for the protein can be isolated and the DNA sequence, and hence the protein sequence, determined. Obtaining a protein sequence in this way is far less laborious and time-consuming than having to determine the total protein sequence by analysis of the protein.

A further use of protein sequence data is in quality control in the biopharmaceutical industry. Many pharmaceutical companies produce products that are proteins, for example peptide hormones, antibodies, therapeutic enzymes, etc., and synthetic peptides also require analysis to confirm their identities. Sequence analysis, especially to determine sites and nature of postsynthetic modifications such as glycosylation, is necessary to confirm the structural integrity of these products.

Edman degradation

In 1950, Per Edman published a chemical method for the stepwise removal of amino acid residues from the N terminus of a peptide or protein. This series of reactions came to be known as the Edman degradation, and the method still remains the most effective chemical means for removing amino acid residues in a stepwise fashion from a polypeptide chain and thus determining the order of amino acids at the N-terminus of a protein or peptide.

However, the method is only infrequently used nowadays and will not be described in any detail here. Developments in the use of mass spectrometry over the past 20 years has led to mass spectrometry being the method of choice nowadays for determining protein sequences, and is discussed in more detail below and in Chapter 9.

Protein cleavage and peptide production

When studying proteins there are many occasions when one might wish to cleave a protein into peptide fragments (see, for example, peptide mass fingerprinting, Section 8.5.1). Peptides can be produced by either chemical or enzymatic cleavage of the native protein (see Table 8.5). Chemical methods tend to produce large fragments, as they cleave at the less common amino acids (often giving as few as two or three large peptides). Enzymatic methods tend to cleave adjacent to the more common amino acids (e.g. trypsin cleaves at every arginine and lysine residue in a protein), thus often producing as many as 50 or more peptides from a protein. Throughout this and other chapters, you will come across examples of where it is necessary to study peptide fragments of a protein.

Mass spectrometry

Because of the absolute requirement to produce ions in the gas phase for the analysis of any sample by mass spectrometry (MS), for many years MS analysis was applicable only to small, non-polar molecules ($< 500\, M_r$). However, in the early 1980s the

Table 8.5 **Specific cleavage of polypeptide**

Reagent	Specificity
Enzymic cleavage	
Chymotrypsin	C-terminal side of hydrophobic amino acid residues, e.g. Phe, Try, Tyr, Leu
Endoproteinase Arg-C	C-terminal side of arginine
Endoproteinase Asp-N	Peptide bonds N-terminal to aspartate or cysteine residues
Endoproteinase Glu-C	C-terminal side of glutamate residues and some aspartate residues
Endoproteinase Lys-C	C-terminal side of lysine
Thermolysin	N-terminal side of hydrophobic amino acid residues excluding Trp
Trypsin	C-terminal side of arginine and lysine residues but Arg-Pro and Lys-Pro poorly cleaved
Chemical cleavage	
BNPS skatole	C-terminal side of tryptophan residues
N-Bromosuccinimide	
o-Iodosobenzoate	
Cyanogen bromide	C-terminal side of methionine residues
Hydroxylamine	Asparagine–glycine bonds
2-Nitro-5-thiocyanobenzoate	N-terminal side of cysteine residues

introduction of fast atom bombardment (FAB) MS allowed the analysis of large, charged molecules such as proteins and peptides to be achieved for the first time. The further development of more sophisticated methods such as electrospray ionisation (ESI) and matrix-assisted laser desorption ionisation (MALDI) (see Chapter 9) has led to the analysis of protein by mass spectrometry becoming routine. Although the Edman degradation still has occasional applications in protein structure analysis, mass spectrometry is now the method of choice for determining amino acid sequence data. When peptides are fragmented by MS it is fortunate that cleavage occurs predominantly at the peptide bond (although it must be noted that other fragmentations, such as internal cleavages, secondary fragmentations, etc. do occur, thus complicating the mass spectrum). This means that the peptide fragments produced each differ sequentially by the mass of one amino acid residue. The amino acid sequence can thus be readily deduced. In particular, if side-chain modifications occur, these can also be observed due to the corresponding increase in mass difference. The use of mass spectrometry to obtain sequence data from proteins and peptides is described more fully in Section 9.5. Tandem mass spectrometry (MS/MS or MS^2) is also increasingly being used to obtain sequence data. A digest of the protein (e.g. with

trypsin) is separated by MS. The ion corresponding to one peptide is selected in the first analyser and collided with argon gas in a collision cell to generate fragment ions. The fragment ions thus generated are then separated, according to mass, in a second analyser, identified, and the sequence determined as described in Section 9.5.2.

A further method, ladder sequencing, has been developed, and combines the Edman chemistry with MS. Edman sequencing is carried out using a mixture of PITC and phenylisocyanate (PIC) (at about 5% of the concentration of PITC). N-terminal amino groups that react with PIC are effectively blocked as they are not cleaved at the acid cleavage step. Consequently, at each cycle, approximately 5% of the protein molecules are blocked. Thus, after 20 to 30 cycles of Edman degradation, a nested set of peptides is produced, each differing by the loss of one amino acid. Analysis of the mass of each of these polypeptides using ESI or MALDI allows the determination of the molecular mass of each polypeptide and the difference in mass between each molecule identifies the lost amino acid residue.

Detection of disulphide linkages

For proteins that contain more than one cysteine residue it is important to determine whether, and if so how many, cysteine residues are joined by disulphide bridges. The most commonly used method involves the use of MS (Section 9.5.5). The native protein (i.e. with disulphide bridges intact) is cleaved with a proteolytic enzyme (e.g. trypsin) to produce a number of small peptides. The same experiment is also carried out on proteins treated with dithiothreitol (DTT) which reduces (cleaves) the disulphide bridges. MALDI spectra of the tryptic digest before and after reduction with DTT allows identification of disulphide-linked peptides. Linked peptides from the native protein will disappear from the spectrum of the reduced protein and reappear as *two* peptides of lower mass. Knowledge of the exact mass of each of the two peptides, and knowledge of the cleavage site of the enzyme used, will allow easy identification of the two peptides from the known protein sequence. Thus, if the mass of two disulphide-linked peptides is M, and this is reduced to two separate chains of masses A and B, respectively, then $A + B = M + 2$. The extra two mass units derive from the fact that reduction of the disulphide bond results in an increase of mass of $+1$ for both cysteine residues.

$$\text{-S} - \text{S-} \xrightarrow{2\text{H}} - \text{SH} + \text{HS-}$$

Hydrophobicity profile

Having determined the amino acid sequence of a protein, analysis of the distribution of hydrophobic groups along the linear sequence can be used in a predictive manner. This requires the products of a hydrophobicity profile for the protein, which graphs the average hydrophobicity per residue against the sequence number. Averaging is achieved by evaluating, using a predictive algorithm, the mean hydrophobicity within a moving window that is stepped along the sequence from each residue to the next. In this way, a graph comprising a series of curves is produced and reveals areas of minima and maxima in hydrophobicity along the linear polypeptide chain. For membrane proteins, such profiles allow the identification of potential membrane-spanning segments. For example, an analysis of a thylakoid membrane protein revealed seven general regions of the protein

sequence that contained spans of 20–28 amino acid residues, each of which contained predominantly hydrophobic residues flanked on either side by hydrophilic residues. These regions represent the seven membrane-spanning helical regions of the protein.

For membrane proteins defining aqueous channels, hydrophilic residues are also present in the transmembrane section. Pores comprise amphipathic α-helices, the polar sides of which line the channel, whereas the hydrophobic sides interact with the membrane lipids. More advanced algorithms are used to detect these sequences, since such helices would not necessarily be revealed by simple hydrophobicity analysis.

8.4.4 Glycoproteins

Glycoproteins result from the covalent attachment of carbohydrate chains (glycans), both linear and branched in structure, to various sites on the polypeptide backbone of a protein. These post-translational modifications are carried out by cytoplasmic enzymes within the endoplasmic reticulum and Golgi apparatus. The amount of polysaccharide attached to a given glycoprotein can vary enormously, from as little as a few per cent to more than 60% by weight. Glycoproteins tend to be found in the serum and in cell membranes. The precise role played by the carbohydrate moiety of glycoproteins includes stabilisation of the protein structure, protection of the protein from degradation by proteases, control of protein half-life in blood, the physical maintenance of tissue structure and integrity, a role in cellular adhesion and cell–cell interaction, and as an important determinant in receptor–ligand binding.

The major types of protein glycoconjugates are:

- N-linked;
- O-linked;
- glycosylphosphatidylinositol (GPI)-linked.

N-linked glycans are always linked to an asparagine residue side-chain (Fig. 8.3) at a consensus sequence Asn-X-Ser/Thr where X is any amino acid except proline. O-linked glycosylation occurs where carbohydrate is attached to the hydroxyl group of a serine or threonine residue (Fig. 8.3). However, there is no consensus sequence similar to that found for N-linked oligosaccharides. GPI membrane anchors are a more recently discovered modification of proteins. They are complex glycophospholipids that are covalently attached to a variety of externally expressed plasma membrane proteins. The role of this anchor is to provide a stable association of protein with the membrane lipid bilayer, and will not be discussed further here.

There is considerable interest in the determination of the structure of O- and N-linked oligosaccharides, since glycosylation can affect both the half-life and function of a protein. This is particularly important of course when producing therapeutic glycoproteins by recombinant methods as it is necessary to ensure that the correct carbohydrate structure is produced. It should be noted that prokaryotic cells do not produce glycoproteins, so cloned genes for glycoproteins need to be expressed in eukaryotic cells. The glycosylation of proteins is a complex subject. From one glycoprotein to another there are variations in the sites of glycosylation (e.g. only about

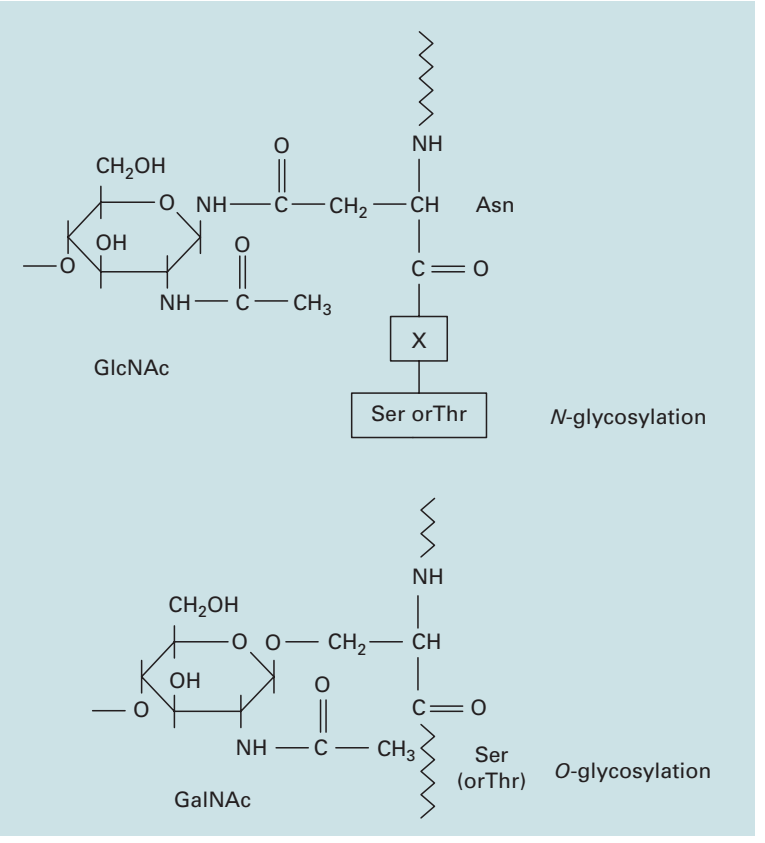


Fig. 8.3 The two types of oligosaccharide linkages found in glycoproteins.

30% of consensus sequences for N-linked attachments are occupied by polysaccharide; the nature of the secondary structure at this position also seems to play a role in deciding whether glycosylation takes place), variations in the type of amino acid–carbohydrate bond, variations in the composition of the sugar chains, and variations in the particular carbohydrate sequences and linkages in each chain. There are eight monosaccharide units commonly found in mammalian glycoproteins, although other less common units are also known to occur. These eight are *N*-acetyl neuraminic acid (NeuNAc), *N*-glycolyl neuraminic acid (NeuGc), D-galactose (Gal), *N*-acetyl-D-glucosamine (GlcNac), *N*-acetyl-D-galactosamine (GalNAc), D-mannose (Man), L-fucose (Fuc) and D-xylose (Xyl). To further complicate the issue, within any population of molecules in a purified glycoprotein there can be considerable heterogeneity in the carbohydrate structure (glycoforms). This can include some molecules showing increased branching of sugar side-chains, reduced chain length and further addition of single carbohydrate units to the same polypeptide chain. The complete determination of the glycosylation status of a molecule clearly requires considerable effort. However, the steps involved are fairly straight forward and the following therefore provides a generalised (and idealised) description of the overall procedures used.

The first question to be asked about a purified protein is 'Is it a glycoprotein?' Glycoprotein bands in gels (e.g. on SDS-polyacrylamide gels) can be stained with cationic dyes such as Alcian Blue, which bind to negatively charged glycosaminoglycan side-chains, or by the periodic acid-Schiff reagent (PAS), where carbohydrate is initially oxidised by periodic acid then subsequently stained with Schiff's reagent. However, although they are both carbohydrate specific (i.e. non-glycosylated proteins are not stained) both methods suffer from low sensitivity. A more sensitive, and informative, approach is to use the specific carbohydrate-binding proteins known as lectins. Blots from SDS-PAGE, dot blots of the glycoprotein sample, or the glycoprotein sample adsorbed onto the walls of a microtitre plate can be challenged with enzyme-linked lectins. Lectins that bind to the glycoprotein can be identified by the associated enzymic activity. By repeating the experiment with a range of different lectins, one can not only confirm the presence of a glycoprotein but also identify which sugar residues are, or are not, present. Having confirmed the presence of glycoprotein the following procedures would normally be carried out.

- *Identification of the type and amount of each monosaccharide*: Release of monosaccharides is achieved by hydrolysis in methanolic HCl at 80 °C for 18 h. The released monosaccharide can be separated and quantified by gas chromatography.
- *Protease digestion to release glycopeptide*: A protease is chosen that cleaves the glycoprotein into peptides and glycopeptides of ideally 5–15 amino acid residues. Glycopeptides are then fractionated by HPLC and purified glycopeptides subjected to N-terminal sequence analysis to allow identification of the site of glycosylation.
- *Oligosaccharide profiling*: Oligosaccharide chains are released from the polypeptide backbone either chemically, for example by hydrazinolysis to release N-linked oligosaccharide, or enzymatically using peptide-*N*-glucosidase F (PNGase F), which cleaves sugars at the asparagine link, or using endo-α-*N*-acetylgalactosaminidase (*O*-glycanase), which cleaves O-linked glycans. These released oligosaccharides can then be separated either by HPLC or by high performance anion exchange chromatography (HPAEC).
- *Structure analysis of each purified oligosaccharide*: This requires the determination of the composition, sequence and nature of the linkages in each purified oligosaccharide. A detailed description is beyond the scope of this book, but would involve a mixture of complementary approaches including analysis by FAB-MS, gas chromatography-MS, lectin analysis following partial release of sugars and nuclear magnetic resonance (NMR) analysis.

8.4.5 Tertiary structure

The most commonly used method for determining protein three-dimensional structure is X-ray crystallography. A detailed description of the theory and methodology is beyond the scope of this book, requiring a detailed mathematical understanding of the process and computer analysis of the extensive data that are generated. The following is therefore a brief and idealised description of the overall process, and ignores the multitude of pitfalls and problems inherent in determining three-dimensional structures.

- Clearly the first step must be to produce a crystal of the protein (a crystal should be thought of as a three-dimensional lattice of molecules). Protein crystallisation is

attempted using as homogeneous a preparation as possible, such preparations having a greater chance of yielding crystals than material that contains impurities. Because of our inadequate understanding of the physical processes involved in crystallisation, methods for growing protein crystals are generally empirical, but basically all involve varying the physical parameters that affect solubility of the protein–for example pH, ionic strength, temperature, presence of precipitating agents–to produce a state of supersaturation. The process involves extensive trial and error to find a procedure that results in crystals for a particular protein. Initially this involves a systematic screen of methods to identify those conditions that indicate crystallinity, followed by subsequent experiments that involve fine-tuning of these conditions. Basically, nucleation sites of crystal growth are formed by chance collisions of molecules forming molecular aggregates, and the probability that these aggregates will occur will be greater in a saturated solution. Clearly, to produce saturated solutions, tens of milligrams of proteins are required. This used to represent a considerable challenge for other than the most abundant proteins, but nowadays genetic engineering methodology allows the overproduction of most proteins from cloned genes almost on demand. The following are some of the methods that have proved successful.

(a) *Dialysis*. A state of supersaturation is achieved by dialysis of the protein solution against a solution containing a precipitant, or by a gradual change in pH or ionic strength. Because of frequent limitations on the amount of protein available, this approach often uses small volumes (<50 mm^3) for which a number of microdialysis techniques exist.

(b) *Vapour diffusion*. This process relies on controlled equilibration through the vapour phase to produce supersaturation in the sample. For example, in the hanging-drop method, a microdroplet (2–20 mm^3) of protein is deposited on a glass coverslip; then the coverslip is inverted and placed over a sealed reservoir containing a precipitant solution, with the droplet initially having a precipitant concentration lower than that in the reservoir. Vapour diffusion will then gradually increase the concentration of the protein solution. Because of the small volumes involved this method readily lends itself to screening large numbers of different conditions.

When produced, crystals may not be of sufficient size for analysis. In this case larger crystals can be obtained by using a small crystal to seed a supersaturated protein solution, which will result in a larger crystal.

- Once prepared, the crystal (which is extremely fragile) is mounted inside a quartz or glass capillary tube, with a drop of either mother liquor (the solution from which it was crystallised) or a stabilising solution drawn into one end of the capillary tube to prevent the crystal from drying out. The tube is then sealed and the crystal exposed to a beam of X-rays. Since the wavelength of X-rays is comparable to the planar separation of atoms in a crystal lattice, the crystal can be considered to act as a three-dimensional grating. The X-rays are therefore diffracted, interfering both in phase and out of phase to produce a diffraction pattern as shown in Fig. 8.4.
Data collection technology necessary for recording the diffraction pattern is now highly sophisticated. Originally, conventional diffractometers and photographic film were used to detect diffracted X-rays. This involved wet developing of the film and subsequent digital scanning of the negative. Data collection by this method took many weeks. By contrast, modern area-detectors can collect data in under 24 h.

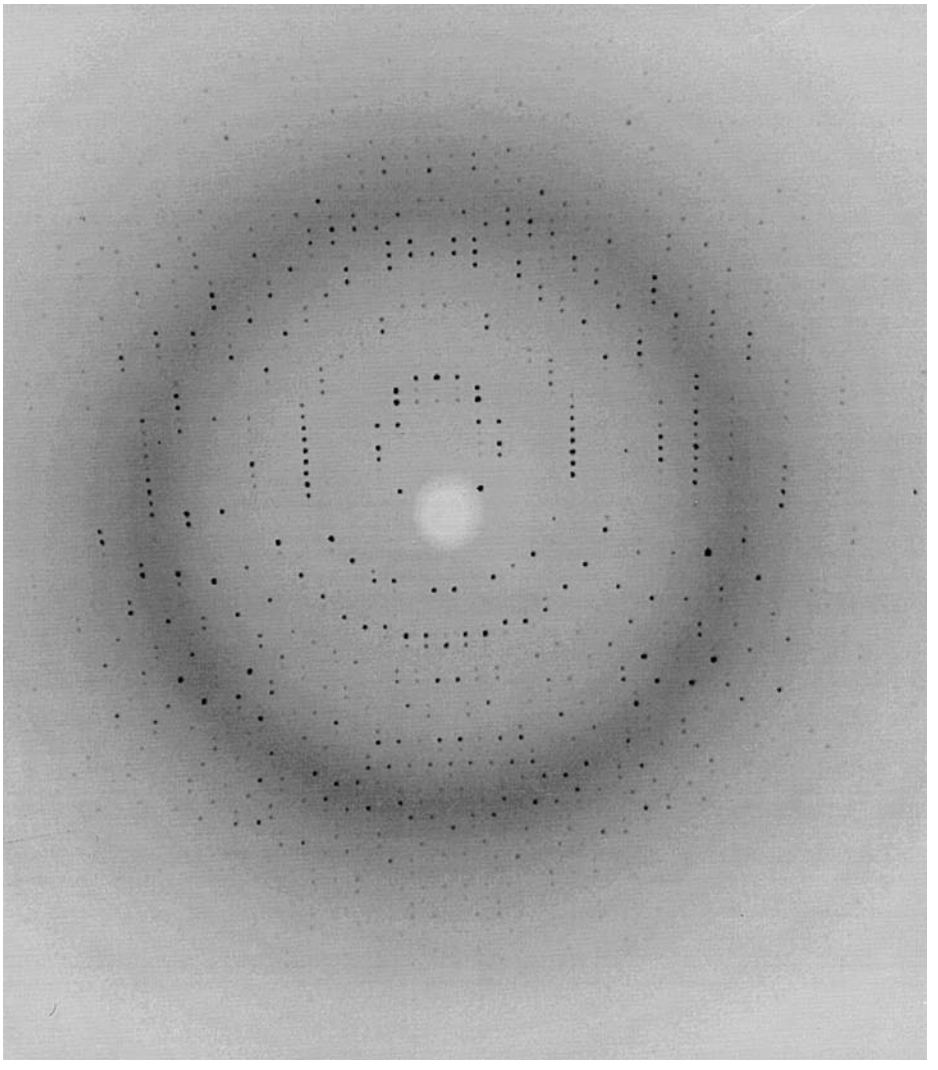


Fig. 8.4 X-ray diffraction frame of data from a crystal of herpes simplex virus type 1 thymidine kinase, complexed with substrate deoxythymidine, at 2 Å resolution. (Picture provided by John N. Champness, Matthew S. Bennett and Mark R. Sanderson of King's College London.)

- Unfortunately the diffraction pattern alone is insufficient to determine the crystal structure. Each diffraction maximum has both an amplitude and a phase associated with it, and both need to be determined. But the phases are not directly measurable in a diffraction experiment and must be estimated from further experiments. This is usually done by the method of isomorphous replacement (MIR). The MIR method requires at least two further crystals of the protein (derivatives), each being crystallised in the presence of a different heavy-metal ion (e.g. Hg^{2+}, Cu^{2+}, Mn^{2+}). Comparison of the diffraction patterns from the crystalline protein and the crystalline heavy-metal atom derivative allows phases to be estimated. A more recent approach to producing a heavy-metal derivative is to clone the protein of interest into a methionine

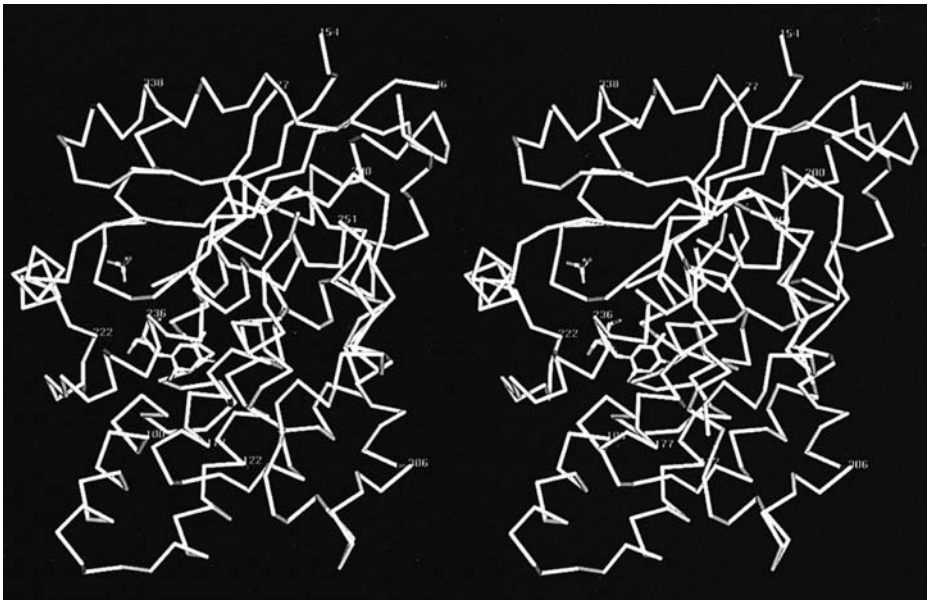


Fig. 8.5 (*Relaxed-eye stereo pair*): A C^{α}-trace of herpes simplex virus type 1 thymidine kinase from a crystallographic study of a complex of the enzyme with one of its substrates, deoxythymidine. The enzyme is an α–β protein, having a five-stranded parallel β-sheet surrounded by 14 α-helices. The active site, occupied by deoxythymidine, is a volume surrounded by four of the helices, the C-terminal edge of the β-sheet and a short 'flap' segment; a sulphate ion occupies the site of the β-phosphate of the absent co-substrate ATP. (Short missing regions of chain indicate where electron density calculated from the X-ray data could not be interpreted.) (Picture provided by John N. Champness, Matthew S. Bennett and Mark R. Sanderson of King's College London.)

auxotroph, and then grow this strain in the presence of selenomethionine (a selenium-containing analogue of methionine). Selenomethionine is therefore incorporated into the protein in the place of methionine, and the final purified and crystallised protein has the selenium heavy metal conveniently included in its structure.

- Diffraction data and phase information having been collected, these data are processed by computer to construct an electron density map. The known sequence of the protein is then fitted into the electron density map using computer graphics, to produce a three-dimensional model of the protein (Fig. 8.5.). In the past there had been concern that the three-dimensional structure determined from the rigid molecules found in a crystal may differ from the true, more flexible, structure found in free solution. These concerns have been effectively resolved by, for example, diffusing substrate into an enzyme crystal and showing that the substrate is converted into product by the crystalline enzyme (there is sufficient mother liquor within the crystal to maintain the substrate in solution). In a more recent development, it is now becoming possible to determine the solution structure of protein using NMR. At present the method is capable of determining the structure of a protein up to about 20 000 kDa but will no doubt be developed to study larger proteins. Although the time-consuming step of producing a crystal is obviated, the methodology and data analysis involved are at present no less time-consuming and complex than that for X-ray crystallography.

8.5 PROTEOMICS AND PROTEIN FUNCTION

In order to completely understand how a cell works, it is necessary to understand the function (role) of every single protein in that cell. The analysis of any specific disease (e.g. cancer) will also require us to understand what changes have taken place in the protein component of the cell, so that we can use this information to understand the molecular basis of the disease, and thus design appropriate drug therapies and develop diagnostic methods. (Just about every therapeutic drug that is currently in use has a protein as its target.) The completion of the Human Genome Project might suggest that it is not now necessary to study proteins directly, since the amino acid sequence of each protein can be deduced from the DNA sequence. This is not true for the following reasons:

- First, although the DNA in each cell type in the body is the same, different sets of genes are expressed in different tissues, and hence the protein component of a cell varies from cell type to cell type. For example, some proteins are found in nearly all cells (the so-called house-keeping genes) such as those involved in glycolysis, whereas specific cell types such as kidney, liver, brain, etc. contain specific proteins unique to that tissue and necessary for the functioning of that particular tissue/organ. It is therefore only by studying the protein component of a cell directly that we can identify which proteins are actually present.
- Secondly, it is now appreciated that a single DNA sequence (gene) can encode multiple proteins. This can occur in a number of ways:
  (i) Alternative splicing of the mRNA transcript.
  (ii) Variation in the translation 'stop' or 'start' sites.
  (iii) Frameshifting, where a different set of triplet codons is translated, to give a totally different amino acid sequence.
  (iv) Post-translational modifications. The genome sequence defines the amino acid sequence of a protein, but tells us nothing of any post-translational modifications (Sections 8.2.1 and 9.5.5) that can occur once the polypeptide chain is synthesised at the ribosome. Up to 10 different forms (variants) of a single polypeptide chain can be produced by phosphorylation, glycosylation, etc.

The consequence of the above is that the total protein content of the human body is an order of magnitude more complex than the genome. The human genome sequence suggests there may be 30 000–40 000 genes (and hence proteins) whereas estimates of the actual number of proteins in human cells suggests possibly as many as 200 000 or even more. The dogma that one gene codes for one protein has been truly demolished!

From the above, I hope it is easy to appreciate the need to directly analyse the protein component of the cell, and the need for an understanding of the function of each individual protein in the cell. In recent years, development of new techniques (discussed below) has enhanced our ability to study the protein component of the cell and has led to the introduction of the terms proteome and proteomics. The total DNA composition

of a cell is referred to as the genome, and the study of the structure and function of this DNA is called genomics. By analogy, the proteome is defined as the total protein component of a cell, and the study of the structure and function of these proteins is called proteomics. The ultimate aim of proteomics is to catalogue the identity and amount of every protein in a cell, and determine the function of each protein.

Earlier sections of this chapter and Chapter 11 describe the traditional, but still very valid approach to studying proteins, where individual proteins are extracted from tissue and purified so that studies can be made of the structure and function of the purified proteins. The subject of proteomics has developed from a different approach, where modern techniques allow us to view and analyse much of the total protein content of the cell in a single step. The development of these newer techniques has gone hand-in-hand with the development of techniques for the analysis of proteins by mass spectrometry, which has revolutionised the subject of protein chemistry. The cornerstone of proteomics has been two-dimensional (2-D) PAGE (described in Section 10.3) and the applications of this technique in proteomics are described below. However, although 2-D PAGE remains central to proteomics, the study of proteomics has stimulated the development of further methods for studying proteins and these will also be described below.

8.5.1 2-D PAGE

2-D PAGE has found extensive use in detecting changes in gene expressions between two different biological states, for example comparing normal and diseased tissue. In this case, a 2-D gel pattern would be produced of an extract from a diseased tissue such as a liver tumour and compared with the 2-D gel patterns of an extract from normal liver tissue. The two gel patterns are then compared to see whether there are any differences in the two patterns. If it is found that a protein is present (or is absent) only in the liver tumour sample, then by identifying this protein we are directed to the gene for this protein and can thus try to understand why this gene is expressed (or not) in the diseased state. In this way it is possible to obtain an understanding of the molecular basis of diseases. This approach can be taken to study *any* disease process where normal and diseased tissue can be compared, for example arthritis, kidney disease, or heart valve disease.

Under favourable circumstances up to 5000 protein spots can be identified on a large format 2-D gel. Thus with 2-D PAGE we now have the ability to follow changes in the expression of a significant proportion of the proteins in a cell or tissue type, rather than just one or two, which has been the situation in the past. The potential applications of proteome analysis are vast. Initially one must produce a 2-D map of the proteins expressed by an organism, tissue or cell under 'normal' conditions. This 2-D reference map and database can then be used to compare similar information from 'abnormal' or treated organisms, tissues or cells. For example, as well as comparing normal tissue with diseased tissue (as described above), we can:

- analyse the effects of drug treatment or toxins on cells;
- observe the changing protein component of the cell at different stages of tissue development;

- observe the response to extracellular stimuli such as hormones or cytokines;
- compare pathogenic and non-pathogenic bacterial strains;
- compare serum protein profiles from healthy individuals and Alzheimer or cancer patients to detect proteins, produced in the serum of patients, which can then be developed as diagnostic markers for diseases (e.g. by setting up an enzyme-linked immunosorbent assay (ELISA) to measure the specific protein).

As a typical example, a research group studying the toxic effect of drugs on the liver can compare the 2-D gel patterns from their 'damaged' livers with the normal liver 2-D reference map, thus identifying protein changes that occur as a result of drug treatment.

The sheer complexity and amount of data available from 2-D gel patterns is daunting, but fortunately there is a range of commercial 2-D gel analysis software, compatible with personal computer workstations, which can provide both qualitative and quantitative information from gel patterns, and can also compare patterns between two different 2-D gels (see below). This has allowed the construction of a range of databases of quantitative protein expression in a range of tissue and cell types. For example, an extensive series of 2-DE databases, known as SWISS-2D PAGE, is maintained at Geneva University Hospital and is accessible via the World Wide Web at <http://au.expasy.org/ch2d/>. This facility therefore allows an individual laboratory to compare their own 2-D protein database with that in another laboratory.

The comparison of two gel patterns is made by using any one of a number of software packages designed for this purpose. One of the more interesting approaches to comparing gel patterns is the use of the Flicker program, which is available on the Web at <http://open2dprot.sourceforge.net/Flicker>. This program superimposes the two 2-D patterns to be compared and then alternately, and rapidly, displays one pattern and then the other. Spots that appear on both gel patterns (the majority) will be seen as fixed spots, but a spot that appears on one gel and not the other will seen to be flashing (hence 'flicker'). When one has compared two 2-DE patterns and identified any proteins spot(s) of interest, it is then necessary to identify each specific protein. In the majority of cases this is done by peptide mass-fingerprinting. The spot of interest is cut out of the gel and incubated in a solution of the proteolytic enzyme trypsin, which cleaves the protein C-terminal to each arginine and lysine residue. In this way the protein is reduced to a set of peptides. This collection of peptides is then analysed by MALDI-MS (see Section 9.3.8) to give an accurate mass measurement for each of the peptides in the sample. This set of masses, derived from the tryptic digestion of the protein, is highly diagnostic for this protein, as no other protein would give the same set of peptide masses (fingerprint). Using Web-based programs such as Mascot or Protein Prospector this experimentally derived peptide mass-fingerprint is compared with databases of tryptic peptide mass-fingerprints generated from sequences of known proteins (or predicted sequences deduced from nucleotide sequences). If a match is found with a fingerprint from the database then the protein will be identified.

However, sometimes results from peptide mass-fingerprinting can be ambiguous. In this case it is necessary to obtain some partial amino acid sequence data from one of the peptides. This is done by tandem mass spectrometry (MS/MS; Section 9.5),

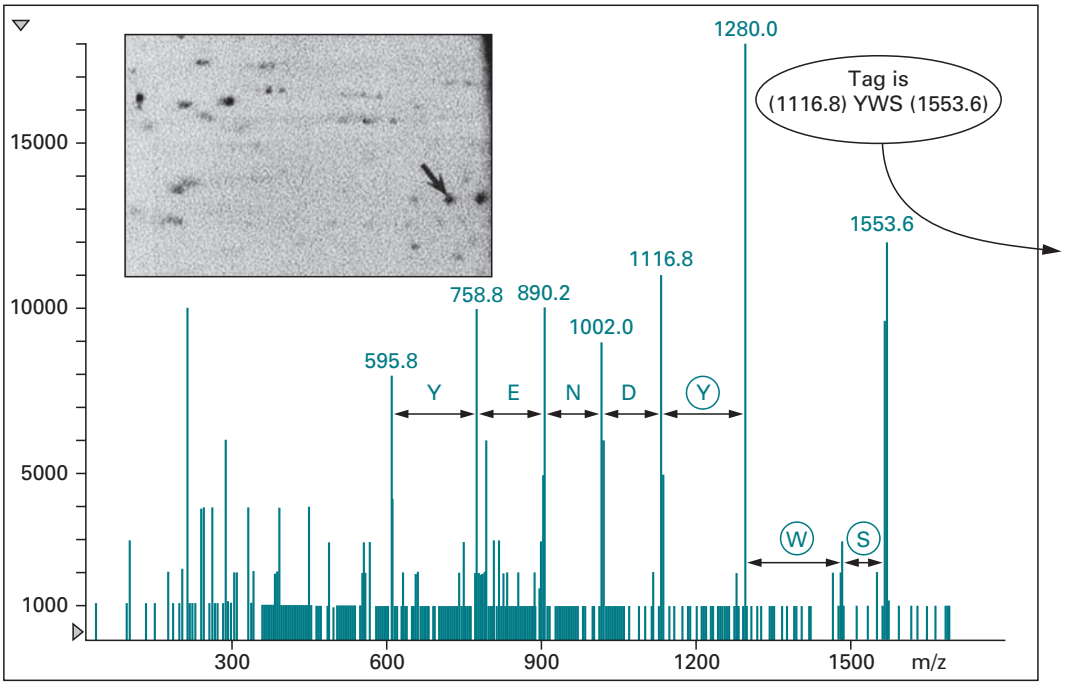


Fig. 8.6 Nano-ESI MS^2 spectrum of *m/z* 890 from RBL spot 2 showing construction of a sequence tag. The *y*-axis shows relative intensity. (Courtesy of Glaxo SmithKline, Stevenage, UK.)

where one of the peptides separated for mass-fingerprinting is further fragmented in a second analyser, and from the fragmentation pattern sequence data can be deduced (mass spectrometry conveniently fragments peptides at the peptide bond, such that the difference in the mass of fragments produced can be related to the loss of specific amino acids; Section 9.5.2). This partial sequence data is then used to search the protein sequence databases for sequence identity. Universal databases are available that store information on all types of protein from all biological species. These databases can be divided into two categories: (i) databases that are a simple repository of sequence data, mostly deduced directly from DNA sequences, for example the Tr EMBL database; and (ii) annotated databases where information in addition to the sequence is extracted by the biologist (the annotator) from the literature, review article, etc., for example the SWISS-PROT database.

An example of how sequence data can be produced is shown in Fig. 8.6. A lysate of 2×10^6 rat basophil leukaemic (RBL) cells were separated by 2-D electrophoresis and spot 2 chosen for analysis. This spot was digested *in situ* using trypsin and the resultant peptides extracted. This sample was then analysed by tandem MS using a triple quadrupole instrument (ESI-MS^2). MS of the peptide mixture showed a number of molecular ions relating to peptides. One of these (*m/z* 890) was selected for further analysis, being further fragmented in a quadrupole mass spectrometer to give fragment ions ranging from *m/z* 595.8 to 1553.6 (Fig. 8.6). The ions at *m/z* 1002.0, 1116.8, 1280.0, 1466.2 and 1553.6 are likely to be part of a Y ion series (see Fig. 8.6) as they appear at higher *m/z* than the precursor at *m/z* 890. The gap between adjacent Y ions is

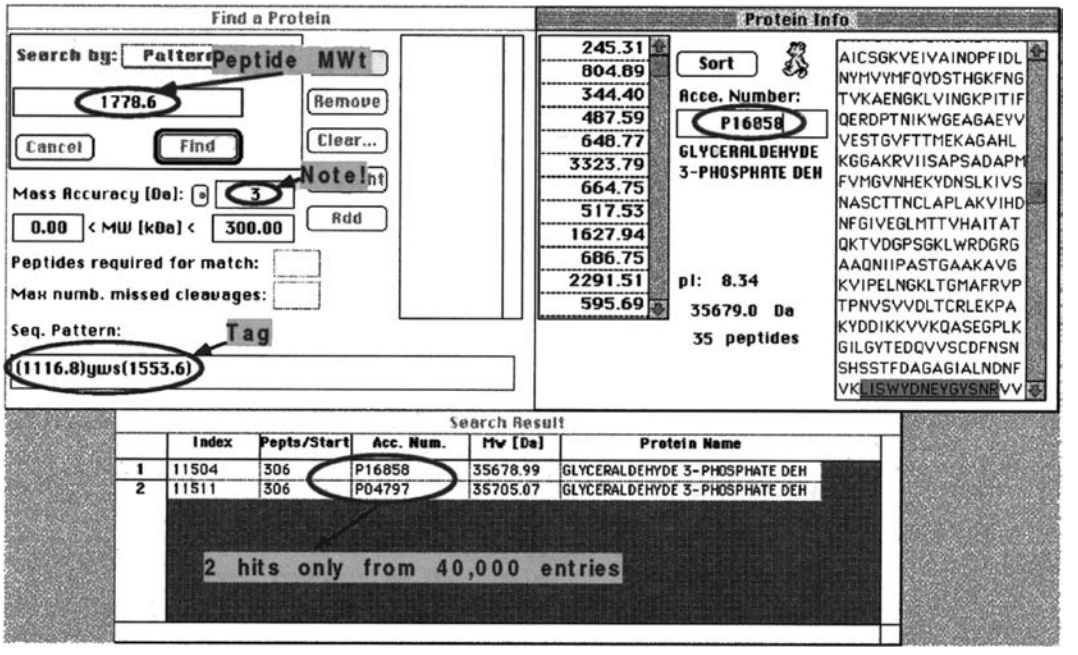


Fig. 8.7 The PeptideSearch™ input form and search result based on data obtained from nano-ESI MS2 of m/s 890 from RBL Spot 2. (Courtesy of Glaxo SmithKline, Stevenage, UK.)

related directly to an amino acid residue because the two flanking Y ions result from cleavage of two adjacent amide bonds. Therefore, with a knowledge of the relative molecular masses of each of the 20 naturally occurring amino acids, it is possible to determine the presence of a particular residue at any point within the peptide. The position of the assigned amino acid is deduced by virtue of the *m/z* ratio of the two ions. By reading several amino acids it was possible to assemble a sequence of amino acids, in this case (using the one-letter code) YWS. Database searching was then possible using the peptide 1778 Da, the position of the lower *m/z* Y ion (1116.8), the proposed amino acid sequence (YWS) and the higher Y ion at *m/z* 1553.6. This provides a sequence tag, which is written as (1116.8) YWS (1553.6).

A search of the SWISS-PROT database (Fig. 8.7), showed just two 'hits' from 40 000 entries, suggesting the protein is glyceraldehyde-3-phosphate dehydrogenase. The full sequence of this peptide is LISWYDNEYGYSNR and the MS/MS fragmentation data give a perfect match. Other peptides in the sample can also be analysed in the same manner, confirming the identity of the protein.

A further development of 2-D PAGE has been the introduction of difference gel electrophoresis (DIGE). This again allows the comparison of protein components of similar mixtures, but has the advantage that only one 2-D gel has to be run rather than two. In this method the two samples to be compared are each treated with one of two different, yet structurally very similar, fluorescent dyes (cy3 and cy5). Each dye reacts with amino groups, so that each protein is fluorescently labelled by the dye binding to lysine residues and the N-terminal amino groups. The two protein solutions to be compared are then mixed and run on a *single* 2-D gel. Thus every protein in one

sample superimposes with its differentially labelled identical counterpart in the other sample. Scanning of the gel at two different wavelengths that excite the two dye molecules reveals whether any individual spot is associated with only one dye molecule rather than two. Most spots will, of course, fluoresce at both wavelengths, but if a spot is associated with only one dye molecule then this tells us that that protein can have been present in only one of the extracts, and the wavelength at which it fluoresces tells you which extract it was originally in.

8.5.2 Isotope-coded affinity tags (ICAT)

Isotope-coded affinity tags (ICAT) uses mass spectrometry (rather than 2-D gels) to identify differences in the protein content of two complex mixtures. For example, the method can be used to identify protein differences between tumour and normal tissue, in the same way that 2-D PAGE can be used to address the same question (Section 8.5.1). This method uses two protein 'tags' that, whilst being in every other respect identical, differ slightly in molecular mass; hence one is 'heavy' and one is 'light'. Both contain (a) a chemical group that reacts with the amino acid cysteine, and (b) a biotin group. In both molecules these groups are joined by a linker region, but in one case the linker contains eight hydrogen atoms, in the other, eight deuterium atoms; one molecule (tag) is thus heavier than the other by 8 Da (see Fig. 9.26). One cell extract (e.g. from cancer cells) is thus treated with one tag (which binds to cysteine residues in all the proteins in the extract) and the second tag is used to treat the second extract (e.g. from normal cells). Both extracts are then treated with trypsin to produce mixtures of peptides, those peptides that contain cysteine having been 'tagged'. The two extracts are then combined and an avidin column used to affinity-purify the labelled peptides by binding to the biotin moiety. When released from the column this mixture of labelled peptides will contain pairs of identical peptides (derived from identical proteins) from the two cell extracts, each pair differing by a mass of 8 Da.

Analysis of this peptide mixture by liquid chromatography–MS will then reveal a series of peptide mass signals, each one existing as a 'pair' of signals separated by eight mass units. These data will reveal the relative abundance of each peptide in the pair. Since most proteins present in the two samples originally being compared will be present at much the same levels, most peptide pairs will have equal signal strengths. However, for proteins that exist in greater or lesser amounts in one of the extracts, different signal strengths will be observed for each of the peptides in the pair, reflecting the relative abundance of this protein in the two samples. Further analysis of either of these pairs via tandem mass spectrometry will provide some sequence data that should allow the protein to be identified. ICAT is discussed in more detail in Section 9.6.2.

8.5.3 Determining the function of a protein

Successfully applied, the methods described in the preceding section will have provided the amino acid sequence (or partial sequence) of a protein of interest. The next step is to identify the function and role of this protein. The first step is invariably to search the databases of existing protein sequences to find a protein or proteins that have sequence homology with the protein of interest (the homology method). This is

done using programs such as BLAST and PSI-BLAST. If sequence homology is found with a protein of known function, either from the same or different species, then this invariably identifies the function of the protein. However, this approach does not always work. For example, when the genome of the yeast *Saccharomyces cerevisiae* was completely sequenced in 1996, 6000 genes were identified. Of these, approximately 2000 coded for proteins that were already known to exist in yeast (i.e. had been purified and studied in previous years), 2000 had homology with known sequences and hence their function could be deduced by the homology method but 2000 could not be matched to any known genes, i.e. they were 'new', previously undiscovered genes. In these cases, there are a number of other computational methods that can be used to help to identify the protein's function. These include:

- *Phylogenic profile method:* This method aims to identify any other protein(s) that has the same phylogenic profile (i.e. the same pattern of presence or absence) as the unknown protein, in all known genomes. If such proteins are found it is inferred that the unknown protein is involved in the same cellular process as these other protein(s) (i.e. they are said to have a functional link) and will give a strong clue as to the function of the unknown protein. This method is based on the premise that two proteins would not always both be inherited into a new species (or neither inherited) unless the two proteins have a functional link. At the time of writing there are over 100 published genome sequences that can be surveyed with this method. Fig. 8.8 shows a simple, hypothetical example, where just five genomes are analysed.
- *Method of correlated gene neighbours:* If two genes are found to be neighbours in several different genomes, a functional linkage may be inferred between the two proteins. The central assumption of this approach is based on the observation that functionally related genes in prokaryotes tend to be linked to form operons (e.g. the *lac* operon). Although operons are rare in eukaryotic species, it does appear that proteins involved in the same biological process/pathway within the cell have their genes situated in close proximity (e.g. within 500 bp) in the genome. Thus, if two genes are found to be in close proximity across a number of genomes, it can be inferred that the protein products of these genes have a functional linkage. This method is most robust for microbial genomics but works to some extent in human cells where operon-like clusters are also observed. As an example, this method correctly identified a functional link between eight enzymes in the biosynthetic pathway for the amino acid arginine in *Mycobacterium tuberculosis*.
- *Analysis of fusion:* This method is based on the observation that two genes may exist separately in one organism, whereas the genes are fused into a single multifunctional gene in another organism. The existence of the protein product of the fused gene, in which the two functions of the protein clearly interact (being part of the same protein molecule), suggests that in the first organism the two separate proteins also interact. It has been suggested that gene fusion events occur to reduce the regulational load of multiple interacting gene products.
- *Protein–protein interactions:* A further clue to identifying protein function can come from identifying protein–protein interactions, and methods to identify these are described in the next section.

	A	B	C	D	E
P1	1	1	1	0	0
P2	0	0	1	1	1
P3	1	0	1	1	0
P4	0	1	1	0	1
P5	1	1	0	0	1
P6	0	1	1	0	1
P7	1	0	0	1	0
P8	1	0	1	1	0

Fig. 8.8 Phylogenic profile method. Five genomes, A–E, are shown (e.g. *E. coli*, *S. cerevisiae*, etc.). The presence (1) or absence (0) of eight proteins (P1–P8) in each of these genomes is shown. It can be seen that proteins P3 and P8 have the same phylogenic profile and therefore may have a functional linkage. P4 and P6 are similarly linked.

8.5.4 Protein–protein interactions

Given the complex network of pathways that exist in the cell (signalling pathways, biosynthetic pathways, etc.), it is clear that all proteins must interact with other molecules to fulfil their role. Indeed, it is now apparent that proteins do not exist in isolation in the cell; proteins involved in a common pathway appear to exist in a loose interaction, sometimes referred to as a biomodule. Therefore, if one can identify an interaction between our unknown protein and a well-characterised protein, it can be inferred that the former has a function somehow related to the latter. For example, if the unknown protein is shown to interact with one or more proteins involved in the biosynthetic pathways for arginine, then this strongly suggests that the unknown protein is also involved in this pathway. Using this approach networks of

interacting proteins are being identified in individual organisms. This has led to the development of the Database of Interacting Proteins (DIP), which can be found at <http://dip.doe-mbi.ucla.edu>. Given the current fad for inventing new words ending in 'ome', some refer to these maps of protein interactions as the interactome.

One of the most widely used, and successful, methods for investigating protein–protein interaction is the yeast two-hybrid (Y2H) system, which exploits the modular architecture of transcription factors. A transcription factor gene (GAL4) is split into the coding regions for two domains, a DNA-binding domain and a *trans*-activation domain. Both these domains are expressed, each linked to a different protein (one being the unknown protein, the other a protein with which it may interact), in separate yeast cells, which are then mated to produce diploid cells (the two proteins being studied are often referred to as the bait and prey). If, in this diploid cell, the bait and prey proteins bind to each other, they will bring together the two domains of the transcription factor, which will then be active and will bind to the promoter of a reporter gene (e.g. the *his* gene), inducing its expression. Identification of cells expressing the reporter gene product is evidence that the bait and prey proteins interact. In practice, following mating, diploids are selected on deficient medium (in this case, medium deficient in histidine), thus only yeast cells expressing interacting proteins survive (as they are capable of synthesising histidine). Once such a positive interaction is identified, the two interacting open reading frames (ORFs) are simply identified by sequencing a small part of the protein gene.

Using this approach, all 6000 ORFs from *S. cerevisiae* were individually cloned as both bait and prey. When the pool of 6000 prey clones was screened against each of the 6000 bait clones, 691 interactions were identified, only 88 of which were previously known. This therefore gave an indication of the function of over 600 proteins whose function was previously unknown. On a much larger scale, the same approach was used to identify protein–protein interactions in the fruit fly, *Drosophila melanogaster*. All 14 000 predicted *D. melanogaster* ORFs were amplified using the polymerase chain reaction (PCR) and each cloned into two-hybrid bait and prey vectors. A total of 45 417 two-hybrid positive colonies were obtained, from which 10 021 protein interactions involving 4500 proteins were obtained. The yeast two-hybrid system is described in greater detail in Section 6.8.3.

8.5.5 Protein arrays

A newly developing area for studying protein–protein interactions is the use of protein arrays (chips). Although the basic principle for screening and identifying interacting molecules is much the same as for DNA arrays (Section 6.8.8), the production of protein arrays is more technically demanding owing mainly to the difficulty of binding proteins to a surface and ensuring that the protein is not denatured at any stage of the assay procedure.

In a protein array, proteins are immobilised as small spots (150–200 μm) onto a solid support (typically glass or a nitrocellulose membrane), using high precision contact printing (not unlike a dot-matrix printer) at a spot density of the order of 1500 spots cm^{-2}. A solution of the protein of unknown function is then incubated on

the array surface for a period of time, then washed off, and the position(s) where the protein has bound, identified (see below). Since it is known which protein was immobilised in each position of the chip, each pair of interacting proteins can be identified.

Saccharomyces cerevisiae again provides a good example of the successful use of this technology where a protein array was used to identify yeast proteins that bind to the protein calmodulin (an important protein involved in calcium regulation). Five thousand eight hundred yeast ORFs were cloned into a yeast high copy expression vector, and each of the expressed proteins purified. Each protein was then spotted at high density onto nickel-coated glass microscope slides. Since each protein also contained a $(His)_6$-Tag (which binds to nickel) introduced at the C terminus, proteins were attached to the surface in an orientated manner, the C terminus being linked to the nickel-coated glass through the $(His)_6$ sequence, while the rest of the molecule was therefore suitably orientated away from the surface of the array to be available for interaction with another protein. The array was then incubated in a solution of calmodulin that had been labelled with biotin. The calmodulin was then washed off and the positions where calmodulin had bound to the array were identified by incubating the array with a solution of fluorescently labelled avidin (the protein avidin binds strongly to the small-molecular-mass vitamin biotin: see Section 10.3.8). The use of ultraviolet light thus identified fluorescence where the screening molecules had bound. In total, 33 new proteins that bind calmodulin were discovered in this way.

Figure 8.9 (see also colour section) shows an interaction map of the yeast proteome. The authors constructed the map from published data on protein–protein interactions in yeast. The map contains 1584 proteins and 2358 interactions. Proteins are coloured according to their functional role, e.g. proteins involved in membrane fusion (blue), lipid metabolism (yellow), cell structure (green), etc. If one views the electronic version of this publication it is possible for the reader to zoom in and search for protein names and to read interactions more clearly.

Figure 8.10 (see also colour section) is a summary of Fig. 8.9 showing the number of interactions of proteins from each functional group with proteins of their own and other groups. The word function means the cellular role of the protein. Numbers in parentheses indicate, first, the number of interactions within a group and, secondly, the number of proteins within a group. Numbers on connecting lines indicate the numbers of interactions between proteins of the two connected groups. For example, in the upper left-hand corner, there are 77 interactions between the 21 proteins involved in membrane fusion and 141 proteins involved in vesicular transport. Looking at the bottom right of the diagram it can be seen that some proteins involved in RNA processing/modification not surprisingly also interact with proteins involved in RNA turnover, RNA splicing, RNA transcription and protein synthesis.

8.5.6 Systems biology

It can be seen from the section on proteomics that the study of proteins is moving away from methods that involve the purification and study of individual proteins. Nowadays proteins are more likely to be studied as a stained spot on a complex 2-D gel pattern, often present in as little as nanogram amounts, more often than not using

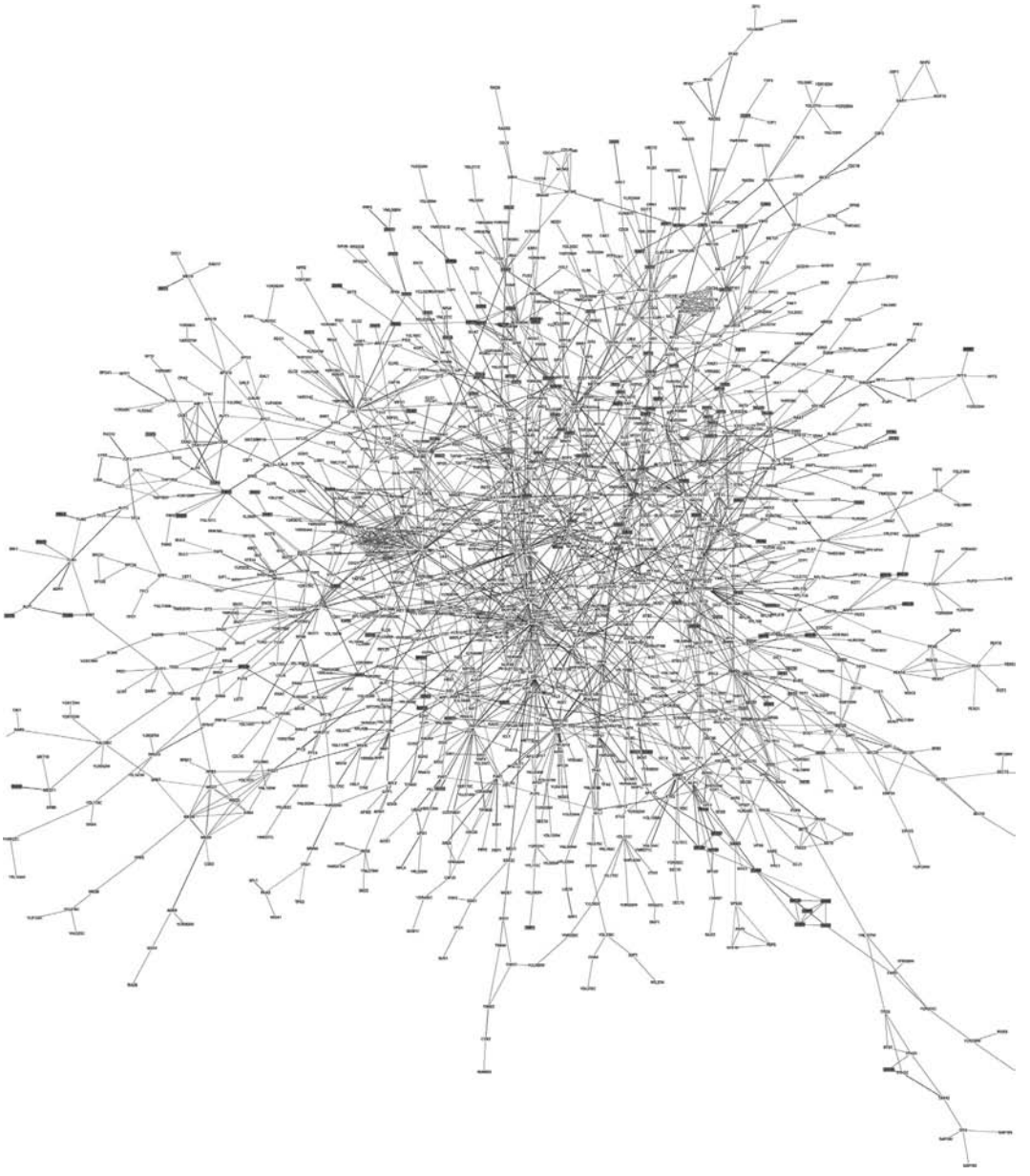


Fig. 8.9 An interaction map of the yeast proteome, assembled from published interactions (see text for details). (Courtesy of Benno Schwikowski, Peter Uetz and Stanley Fields. Reprinted with the permission of Nature Publishing Group.) (See also colour plate.)

analytical techniques such as mass spectrometry (see Chapter 9) and invariably requiring the interrogation of protein and genome sequence data on the Web (bioinformatics, Section 5.8). It is then necessary to determine which other proteins interact with the protein being studied. Proteomics is thus moving us away from studying proteins in isolation and encouraging us to consider the proteins in the cell as part

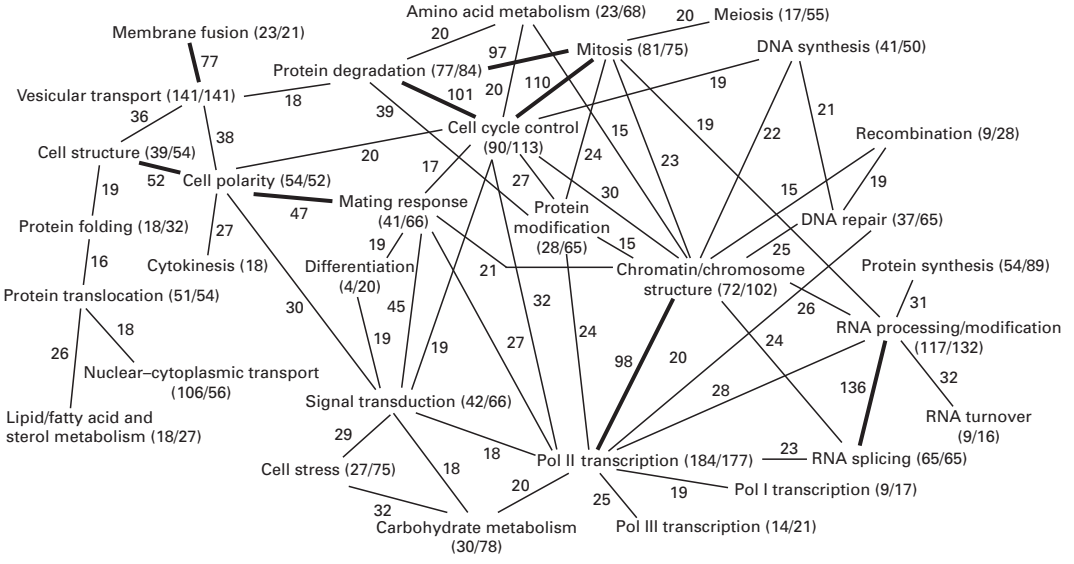


Fig. 8.10 A simplification of Fig. 8.9 identifying interactions between functional groups of proteins (see text for details). (Courtesy of Benno Schwikowski, Peter Uetz and Stanley Fields. Reprinted with the permission of Nature Publishing Group.) (See also colour plate.)

of a dynamic interacting system. This has led to the development of the concept of systems biology, which can be defined as the study of living organisms in terms of their underlying network structure rather than just their individual molecular components. Since systems biology requires a study of all interacting components in the cell the new high throughput and quantitative.

8.6 SUGGESTIONS FOR FURTHER READING

Cutler, P. (2004). *Protein Purification Protocols*. Totowa, NJ: Humana Press. (Detailed theory and practical procedures for a range of protein purification techniques.)

Walker, J. M. (2005). *Proteomics Protocols Handbook*. Totowa, NJ: Humana Press. (Theory and techniques of a spectrum of methods applied to proteomics.)

Nedelkov, D. (2006). *New and Emerging Proteomics Techniques*. New York: Humana Press. (In-depth details of a range of proteomics techniques.)

Thompson, J. D. (2008). *Functional Proteomics*. New York: Humana Press. (Comprehensive coverage of functional proteomics including protein analysis and mass spectrometry.)

Simpson, R. J., Adams, P. D. and Golemis, E. A. (2008). *Basic Methods in Protein Purification and Analysis: A Laboratory Manual*. New York: CSH Press. (A comprehensive collection of protein purification methods.)