9

Protein expression, purification and characterization

To study the structure of *any* protein successfully it is necessary to purify the molecule of interest. This is often a formidable task especially when some proteins are present within cells at low concentration, perhaps as few as 10 molecules per cell. Frequently, this involves purifying a single protein from a cell paste containing over 10 000 different proteins. The ideal objective is to obtain a single protein retaining most, if not all, of its native (*in vivo*) properties. This chapter focuses on the methods currently employed to isolate and purify proteins.

Although functional studies often require small amounts of protein (often less than 1 ng or 1 pmol), structural techniques require the purification of proteins on a larger scale so that 10 mg of pure protein is sometimes needed. Taken together, the requirement for purity and yield often places conflicting demands on the experimentalist. However, the successful generation of high-resolution structures is testament to the increased efficacy of methods of isolation and purification used today in protein biochemistry. Many of these methods evolved from simple, comparatively crude, protocols into sophisticated computer-controlled and enhanced procedures where concurrent advances in bioinformatics and material science have supported rapid progress in this area.

Proteins: Structure and Function by David Whitford
© 2005 John Wiley & Sons, Ltd

The isolation and characterization of proteins

Two broad alternative approaches are available today for isolating proteins. We can either isolate the protein conventionally by obtaining the source cell or tissue directly from the host organism or we can use molecular biology to express the protein of interest, often in a host such as *Escherichia coli*. Today molecular biology represents the most common route where DNA encoding the protein of interest is inserted into vectors facilitating the high level expression of protein in *E. coli.*

Recombinant DNA technology and protein expression

Prior to the development of recombinant protein expression systems the isolation of a protein from mammalian systems required either the death of the organism or the removal of a small selected piece of tissue in which it was suspected that the protein was found in reasonable concentrations. For proteins such as haemoglobin removal of small amounts of blood does not present ethical or practical problems. Given that haemoglobin is found in a soluble state and at very high natural concentrations (~145 g l^{-1} in humans) it is not surprising that this protein was amongst the first to be studied at a molecular level. However, many proteins are found at much lower concentrations and in tissues or cells that are not easily obtained except post mortem.

Recombinant systems of protein expression have alleviated most problems of this type and allow the study of proteins that were previously inaccessible. Once DNA encoding the protein of interest has been obtained it is relatively routine to obtain proteins via recombinant hosts in quantities one could previously only dream about (i.e. mg–g amounts). This does not mean that it is always possible to express proteins in *E. coli*. For example, membrane proteins still present numerous technical problems in heterologous expression, as do proteins rich in disulfide bonds or those bearing post-translational modifications such as myristoylation and glycosylation. Eukaryotic cells carry out many of these modifications routinely and this has led to the development of alternative expression systems besides *E. coli*. Alternative expression hosts include simple eukaryotes such as *Saccharomyces cerevisiae* and *Pichia pastoris*, as well as more complex cell types. The latter group includes cultured insect cells such as *Spodoptera frugiperda* infected with baculovirus vectors, *Drosophila* cell-based expression systems and mammalian cell lines often derived from immortalized carcinomas.

However, the most common approach and the route generally taken in any initial investigation is to express 'foreign' proteins in *E. coli*. These methods have revolutionized studies of protein structure and function by allowing the expression and subsequent isolation of proteins that were formerly difficult to study perhaps as a result of low intrinsic concentration within cells or tissues. Of importance to the development of recombinant DNA technology was the discovery of restriction enzymes (see Chapter 7) and their use with DNA ligases to create new (recombinant) DNA molecules that could be inserted into a carrier molecule or vector and then introduced into host cells such as *E. coli* (Figure 9.1). The process of constructing and propagating new DNA molecules by inserting into host cells is referred to as 'cloning'. A clone is an exact replica of the parent molecule, and in the case of *E. coli* this results, after several cycles of replication, in all cells containing new and identical DNA molecules.

The generation of new recombinant DNA molecules involves at least five discernable steps (Figure 9.2):

1. *Preparation of DNA*. Today DNA is usually generated using the polymerase chain reaction (PCR). The technique, described in Chapter 6, involves the action of a thermostable DNA polymerase with forward and

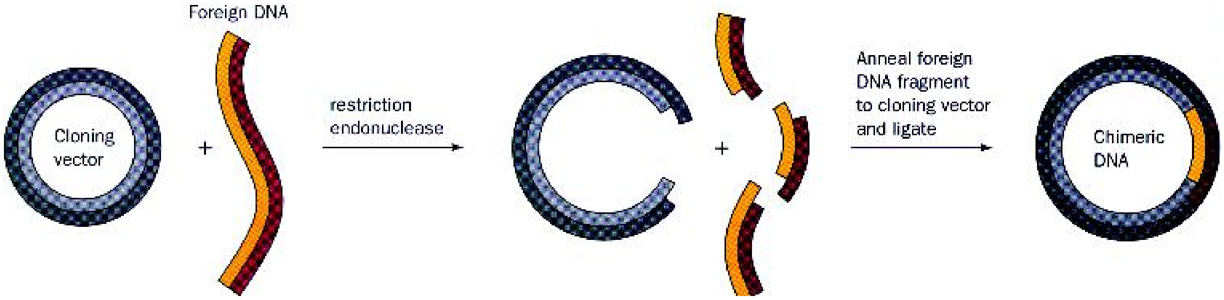

Figure 9.1 The use of restriction enzymes to generate DNA fragments and the use of DNA ligase to join together vector and insert together to form a new recombinant (chimeric) circular DNA molecule. The diagram shows the joining together of a cut vector and a cut fragment described as having sticky ends. This type of end occurs with most restriction enzymes and is contrasted with the smaller group of type II restriction endonucleases that produce blunt-ended fragments (reproduced with permission from Voet, D. Voet, J.G and Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons Ltd., Chichester, 1999)

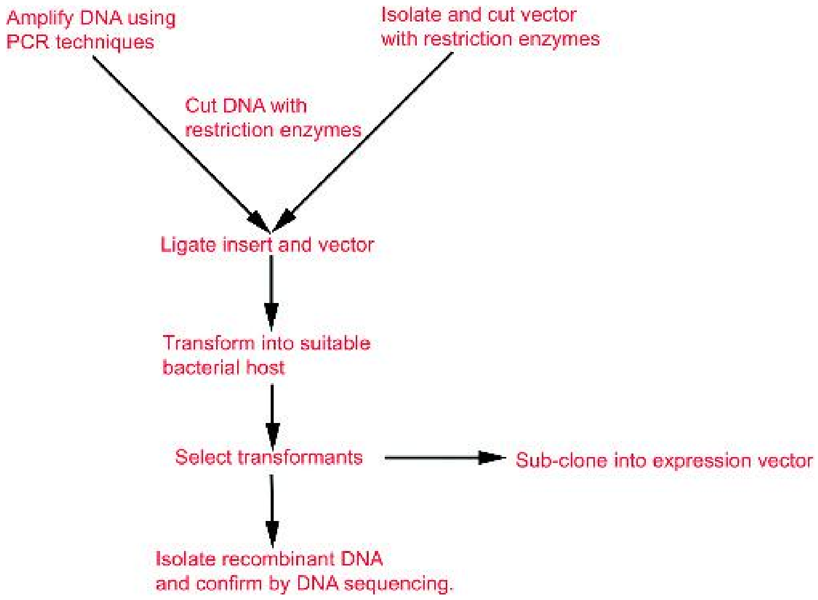
Amplify DNA using PCR techniques
Isolate and cut vector with restriction enzymes
Cut DNA with restriction enzymes
Ligate insert and vector
Transform into suitable bacterial host
Select transformants
Sub-clone into expression vector
Isolate recombinant DNA and confirm by DNA sequencing.

Figure 9.2 Generalized scheme for the creation of recombinant DNA molecules

reverse primers in a series of cycles involving denaturation, annealing and extension. The result is large quantities of amplified and copied DNA. Template DNA is preferably derived from cDNA via the action of reverse transcriptase on processed mRNA. This avoids the presence of introns in amplified sequences. In other instances DNA is derived from the products of restriction enzyme action.

2. *Restriction enzyme cleavage*. The generation of DNA fragments of specific length and with 'tailored' ends was vital to the success of recombinant DNA technology. This allowed the 'insert' (the name given to the DNA fragment) to be joined into a vector containing compatible sites (see Figure 9.1). Today the generation of PCR DNA fragments normally includes sites for restriction enzymes that facilitate subsequent cloning reactions and today a wide variety of restriction enzymes are commercially available.

3. *Ligation of DNA fragments*. As part of the overall cloning process insert and vector DNA are joined in a process called ligation. Without covalent linkage of pieces of DNA and the formation of closed circular DNA subsequent transformation reactions (see step 4) occur with very low efficiency.

4. *Transformation*. The propagation of recombinant DNA molecules requires the introduction of circular DNA (ligated vector + insert) into a suitable host cell. Replication of the cells generates many copies of the new recombinant DNA molecules. The whole process is called transformation and represents the uptake of circular DNA by host cells with the acquisition of new altered properties. The exact process by which cells take up DNA is unclear but two common methods used for *E. coli* involve the application of a high electric field (electroporation) or a mild heat shock. The result is that membranes are made 'leaky' and take up small circular DNA molecules. For *E. coli* the process of transformation requires special treatment and cells are often described as 'competent' meaning they are in a state to be transformed by plasmid DNA.

5. *Identification of recombinants*. After transformation cells such as *E. coli* can be grown on nutrient rich plates and it is here that marker genes found in many

vectors are most useful. Normal *E. coli* cells used in the laboratory lack resistance to antibiotics such as ampicillin, tetracycline or kanamycin. However, transformation with a vector containing one or more antibiotic resistance genes leads to resistant cells. As a result when cells are grown on agar plates containing an antibiotic only those cells transformed with a plasmid conferring resistance will grow. This selection procedure allows the screening of cells so that only transformed cells are studied further. By transforming with known amounts of plasmid DNA it is possible to calculate transformation efficiencies and it is not unusual for this value to reach 10^9-10^{10} transformants/μg DNA. Further identification of the recombinant DNA might involve isolation of the plasmid DNA and its digestion with restriction enzymes to verify an expected pattern of fragments. Definitive demonstration of the correct DNA fragment would involve nucleotide sequencing.

One of the first vectors to be developed was called pBR322 and it is the 'ancestor' of many vectors still used today for cloning. The vector pBR322 (Figure 9.3) contains distinctive sequences of DNA that includes an origin of replication (*ori*) gene as well as a gene (*rop*) that control DNA replication and the number of copies of plasmid found within a cell. In addition pBR322 contains two 'marker' genes that code for resistance to the antibiotics tetracycline and ampicillin. These useful markers are denoted by tet^R and amp^R.

The next step requires the transfer of DNA from the cloning vector to a vector optimized specifically for protein expression. A cloning vector such as pBR322 will allow several copies of the plasmid to be present in cells and may facilitate sequence analysis or analysis of restriction sites but they are rarely used for protein expression. Other common cloning vectors used widely in molecular biology and frequently described in scientific literature are pUC, pGEM®, pBluescript®. Each vector has many variants containing different restriction enzyme sites that allow the orientation of the insert in either direction as well as common features such as one or more antibiotic resistance genes, multiple cloning sites, an origin of replication, and a medium to high copy number.

In contrast, expression vectors are designed with additional features beneficial to protein production. The vectors retain many of the features of 'cloning vehicles'

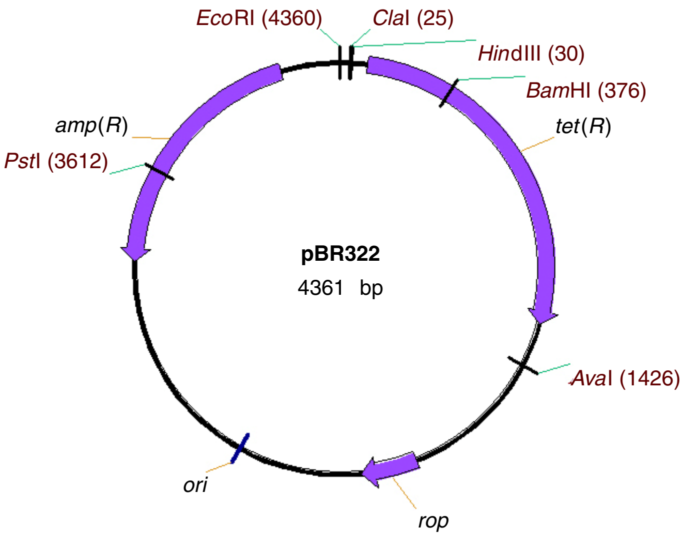

Figure 9.3 Plasmid vector pBR322. The small circular DNA molecule contains 4361 bp, the *rop*, *ori*, amp^R and tet^R genes as well as numerous sites for restriction enzymes such as *Eco*RI, *Bam*HI, *Ava*I, *Pst*I, *Cla*I and *Hind*III

such as multicloning sites, antibiotic resistance genes and the *rop* and *ori* genes but have in most cases:

1. *An inducible promoter*. Expressed proteins may prove toxic to the host organism. To limit this effect the activity of the promoter can be 'turned on' by an inducer. Although the mechanism of action of inducers varies from one system to another the most common system in use involves the promoter derived from the *lac* operon of *E. coli*. In many systems the sequence of this promoter has been modified to contain a 'consensus' sequence of bases for optimal activity (see Chapters 3 and 8). The *lac* operon (Figure 9.4) is a group of three genes – β-galactosidase, lactose permease and thiogalactoside transacetylase – whose activities are induced when lactose is present in the medium. The *lac* repressor is a homotetrameric protein with an N-terminal domain containing the HTH motif found in many transcription factors. It specifically binds DNA sequences in the operator region inhibiting transcription. A non-metabolized structural analogue of lactose, the natural inducer, is used to induce protein expression. Isopropylthiogalactose (IPTG) binds the lac repressor removing the inhibition of transcription (Figure 9.5). This system is the basis for most inducible expression vectors used in genetic engineering studies today.

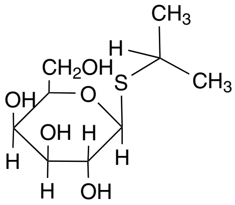

Figure 9.5 The structure of isopropylthiogalactose (IPTG)

2. *Restriction enzyme sites*. A large number of restriction sites are often clustered together in so called multi-cloning sites. These sites allow the insert to be placed 'in frame' with the start codon and in a correct relationship with the Shine–Dalgarno sequence. Two restriction enzymes, *Nco*I and *Nde*I, cut at the sequence C↓CATGG and CA↓TATG respectively. Careful inspection of this sequence reveals that ATG is part of the start codon and fragments cut with *Nco*I/*Nde*I will, if ligated correctly, be in frame with the start codon. This will yield the correct protein sequence upon translation.

3. *A prokaryotic ribosome binding site containing the Shine–Dalgarno sequence*. The Shine–Dalgarno sequence has a consensus sequence of AGGAGA located 10–15 bp upstream of the ATG codon.

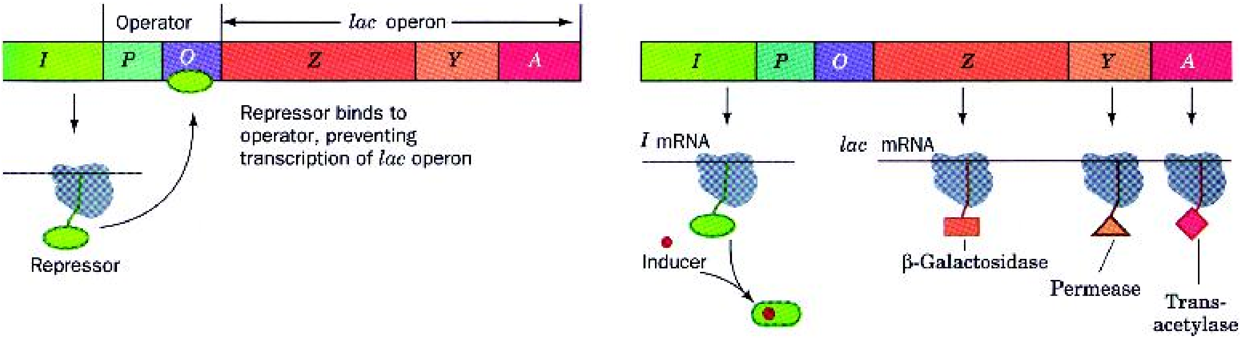

Figure 9.4 The expression of the lac operon. In the absence of inducer (IPTG) the repressor binds to the operator preventing transcription of the three genes of the lac operon. IPTG binds to the repressor allowing transcription to proceed. Most expression vectors contain the *lacI* gene together with the lac promoter as their inducible transcription system. (Reproduced with permission Voet, D. Voet, J.G and Pratt, C.W. *Fundamentals of Biochemistry* 1999. John Wiley & Sons Ltd., Chichester)

Successful expression requires the use of the host cells synthetic machinery and in particular formation of an initiation complex for translation. The 16S rRNA binds to purine rich sequences (AGGAGA) as a result of sequence complementarity enhancing formation of the initiation complex. The composition and length of the intervening sequences between the ribosomal binding site and the start codon are important with structures such as hairpin loops decreasing expression.

4. *A transcriptional terminator*. In an ideal situation mRNA is transcribed only for the desired protein and is then halted. To achieve this aim many expression vectors contain transcriptional terminators. Terminators are DNA sequences that cause the disassembly of RNA polymerase complexes, limit the overall mRNA length and are G-C rich and palindromic. The formation of RNA hairpin structures destabilizes DNA–RNA transcripts causing termination. In other cases termination results from the action of proteins like the *Rho*-dependent terminators in *E. coli*.
5. *Stop codons*. In frame stop codons prevent translation of mRNA by ribosomes and their presence as part of the 'insert' or as part of the expression vector is desirable to ensure the absence of extra residues in the expressed protein.
6. *Selectable markers*. One or more selectable markers such as the genes for ampicillin or tetracycline resistance allows the expression vector to be selected and maintained.

Extensive discussion of all of these features is beyond the scope of this book but one set of widely used expression systems are the pET-based vectors. These vectors are based on the bacteriophage T7 RNA polymerase promoter, a tightly regulated promoter, that allows high levels of protein expression. A large number of variants exist on a basic theme. The pET vector is approximately 4600 bp in size, contains either the Amp^R and Kan^R antibiotic resistance genes, as well as the normal *ori* and *rop* genes. The pET vectors (Figure 9.6) have regions containing sites for many Type II restriction enzymes and genes are cloned downstream under control of T7 RNA polymerase promoter. It might be thought that host RNA polymerases would initiate transcription from this promoter but it turns out that this promoter is highly selective for T7 RNA polymerase. One consequence is that background protein expression in the absence of T7 RNA polymerase is extremely low and this is desirable when expressed proteins are toxic to host cells. However, expression of foreign cDNA requires T7 RNA polymerase – an enzyme not normally found in *E. coli* – and to overcome this problem all *host* cells contain a chromosomal copy of the gene for T7 RNA polymerase. The addition of IPTG to cultures of *E. coli* (*DE3*) leads to the production of T7 RNA polymerase that in turn allows transcription of the target DNA.

One of the major advantages of bacterial cells is their ease of culture allowing growth in a sterile manner, in large volumes and with short doubling times (~40 minutes under optimal conditions). *E. coli* cells are easy to transform with vector DNA, especially when compared with eukaryotic cells, and growth media are relatively cheap. Large amounts of expressed protein can be obtained from *E. coli* and it is not unusual for the expressed protein to exceed 10 percent of the total cell mass. The success of genetic engineering has been demonstrated by production of large amounts of therapeutically important eukaryotic proteins using *E. coli* expression systems. Human insulin, growth hormones, blood coagulation factors IX and X as well as tissue plasminogen activator have all been successfully expressed and used to treat diseases such as diabetes, haemophilia, strokes and heart attacks.

Purification of proteins

Advances in recombinant DNA technology allow large-scale protein expression in host cells. Demonstration of protein expression is usually confirmed by SDS–polyacrylamide gel electrophoresis (see below) and is then followed by the question: how is the protein to be purified? An effective purification strategy requires protein recovery in high yield *and* with high purity. Achieving these two objectives can be difficult and is usually resolved in an empirical manner.

The first step in any purification strategy involves the fractionation of cells by mechanical disruption

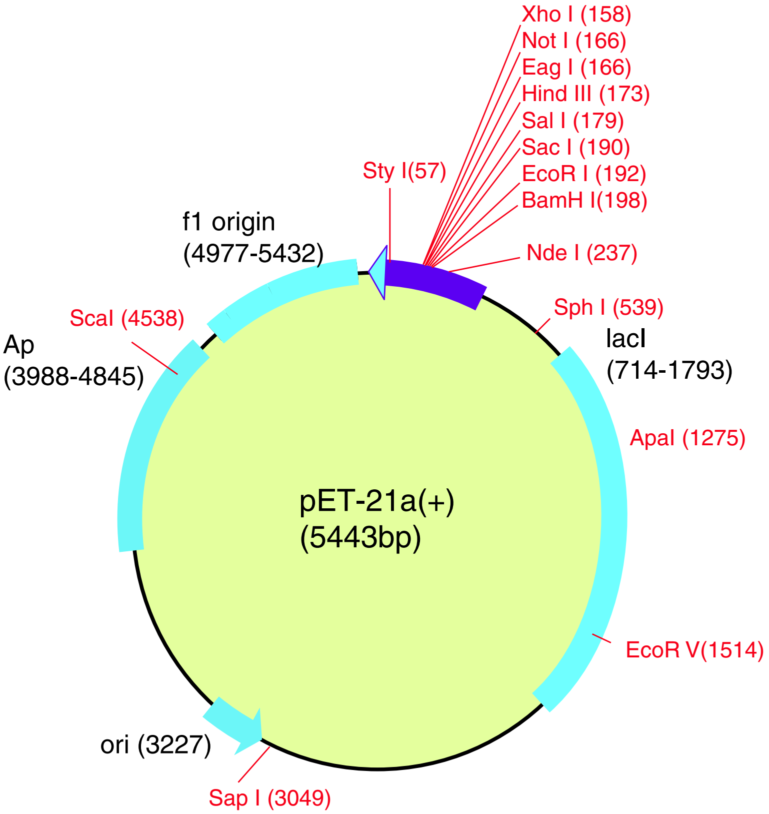

Figure 9.6 The pET series of vectors. Identifiable elements within the vector include the T7 promoter (311–327), T7 transcription start (310), T7 Tag coding sequence (207–239), Multiple cloning sites (*Bam*HI–*Xho*I) (158–203), His-Tag coding sequence (140–157), T7 terminator region (26–72), *lacI* coding sequence (714–1793), pBR322 origin of replication (3227), and the antibiotic resistance coding sequence (3988–4845)

or the use of chemical agents such as enzymes. Mechanical disruption methods can involve sonication where high frequency sound waves (ultrasound frequencies > 20 kHz) disrupt membrane structure. Other methods include mechanical shearing by rupturing membranes by passing cells through narrow orifices under extremely high pressure or grinding cells with abrasive material such as sand or glass beads to achieve similar results.

A different method of disruption involves the use of enzymes such as lysozyme to disrupt cell walls in Gram-negative bacteria like *E. coli* or chitinase to degrade the complex polysaccharide coats associated with yeast. Lysozyme acts by splitting the complex polysaccharide coat between *N*-acetylmuramic acid and *N*-acetylglucosamine units in the peptidoglycan layer (Figures 9.7 and 9.8). Multiple glycosidic bond cleavage weakens the cell envelope and the *E. coli* membrane 'sac' is easily broken by osmotic shock or mild sonication to release cytoplasmic contents that are separated from membranes by differential centrifugation.

Purification strategies normally exploit the physical properties of the protein as a means to isolation. If primary sequence information is available then predictions about overall charge, disulfide content, solubility, mass or hydrophobicity can be obtained. Bioinformatics may allow recognition of similar proteins

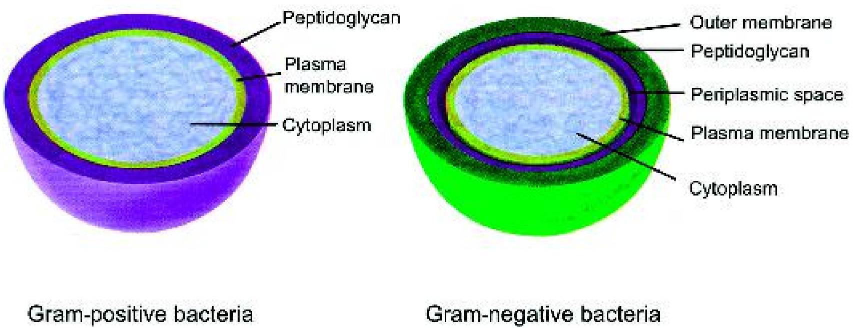

Figure 9.7 A diagram of the organization of the peptidoglycan layer in Gram-positive and Gram-negative bacteria. The peptidoglycan layer is a complex structure consisting of covalently linking carbohydrate and protein components and is more extensive in gram positive bacteria. Bacteria are defined as Gram-positive or -negative according to whether heat fixed cells retain a dye (crystal violet) and iodine after destaining with alcohol. Gram-positive bacteria allow the dye into the cells and heat fixing causes changes in the peptidoglycan layer that prevent the escape of the dye during destaining. In contrast the smaller peptidoglycan layer of Gram-negative bacteria allows the escape of dye during destaining (reproduced with permission Voet, D. Voet, J.G and Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons Ltd., Chichester, 1999)

and therefore the implementation of comparable purification methods. However, in some circumstances the protein is 'new' and sequence information is absent. To obtain sequence information it is necessary to purify the protein yet to purify the protein it is often advantageous to have sequence information! Exploiting physical properties and judicious use of selective methods allows the purification of a protein from cell extracts containing literally thousands of unwanted proteins to a single protein (homogeneity) in a few steps.

Centrifugation

The principle of centrifugation is to separate particles of different mass. Strictly speaking this is not entirely accurate because centrifugation depends on other factors such as molecular shape, temperature and solution density. Heavy particles sediment faster than light particles – a fact demonstrated by mixing sand and water in a container and waiting for a few minutes after which time the sand has settled to the bottom of the container. This reflects our everyday experience that heavier particles are pulled downwards by the earth's gravitational force. However, a close inspection of the solution would reveal that small (microscopically small) grains remain suspended in solution. This also happens to macromolecules such as proteins where, as a result of Brownian diffusion, their buoyant density counteracts the effect of gravity.

Fractionation procedures (Figure 9.9) are important in cell biology and are used extensively to purify cells, subcellular organelles or smaller components for further biochemical analysis. Disruption of *E. coli* cells produces membrane and cytosolic fractions, each containing particles of varying mass. These particles are separated by centrifugation, a process enhancing rates of sedimentation by increasing the force acting on particles within solutions. Centrifugation involves the rotation of solutions in tubes within specially designed rotors at frequencies ranging from 100 revolutions per minute up to ~80 000 r.p.m. More commonly centrifuges are used at much lower g values and a force of 20 000 g is sufficient to pellet most membranes found in cells as well as larger organelles such as mitochondria, chloroplasts, and the ER systems of animal and plant cells.

In the preceding section centrifugation has been described as a preparative tool but it also finds an

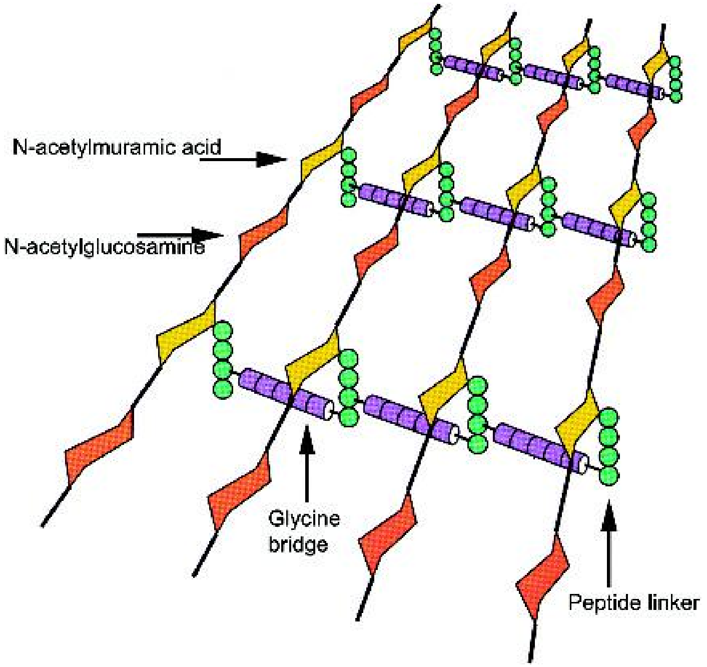

Figure 9.8 The arrangement of β(1–4) linked *N*-acetylglucosamine (NAG) and *N*-acetylmuramic acid (NAM) together with the linkage to the tetrapeptide bridge of L-Ala, D-isoglutamate, L-Lys and D-Ala peptidoglycan. The tetrapeptide bridges are joined together by five Gly residues that extend from the carboxyl group of one tetrapeptide to the ε-amino group of lysine in another. The isoglutamate residue is so called because the side chain or γ carboxyl group forms a peptide bond with the next amino group. The arrangement of units within the peptidoglycan layers has been studied particularly extensively for *Staphylococcus aureus* but considerable variation in structure can occur. The presence of D-amino acids renders the peptidoglycan layer resistant to most proteases (Reproduced with permission from Voet, D. Voet, J.G. Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons Ltd., Chichester, 1999)

application as a precise analytical tool. Centrifuged solutions are subjected to strong forces that vary along the length of the tube or more properly the radius (r) about which rotation occurs. The centrifugal force acting on any particle of mass m is the product of a particle's mass multiplied by the centrifugal acceleration (from Newton's second law of motion). This centrifugal acceleration is itself the product of the radius of rotation and the angular velocity ($r\omega^2$). For any particle the *net* force acting on it will be a balance between centrifugal forces causing particles to pellet and buoyant forces acting in the opposite direction. Particles that are less buoyant than the solvent will sink whilst those that are lighter than the solvent will float. This results in the following expression describing the forces acting on a particle during centrifugation:

$$\text{Net force} = \text{centrifugal force} - \text{buoyant force} \quad (9.1)$$

$$= \omega^2 rm - \omega^2 rm_s \quad (9.2)$$

$$= \omega^2 rm - \omega^2 r\nu\rho \quad (9.3)$$

where m_s is the mass of solvent displaced by the particle, ν is the volume of the particle and ρ is the density of the solvent. If, for the moment, the second term on the right hand side of the equation is neglected the centrifugal force is simply the angular velocity multiplied by the radius of rotation. Usually, the value of the force applied to particles during centrifugation is compared to the earth's gravitational force ($\boldsymbol{g}$, ~9.8 m^2 s^{-1}) and the dimensionless quantity

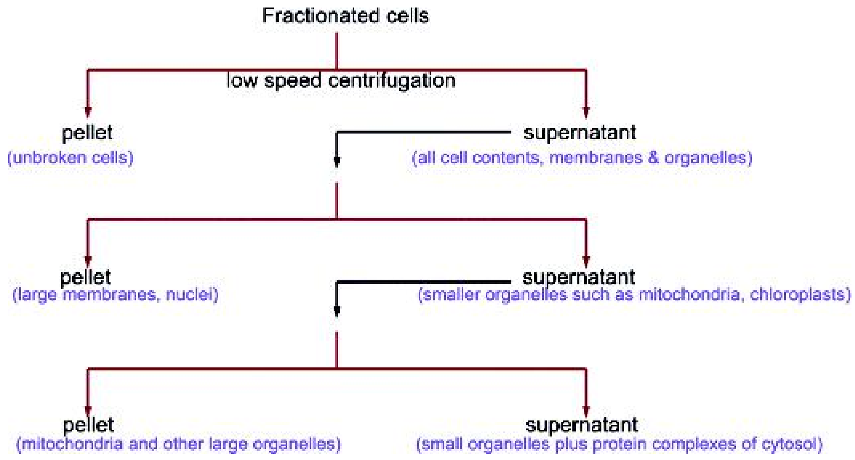
Fractionated cells
low speed centrifugation
pellet
(unbroken cells)
supernatant
(all cell contents, membranes & organelles)
pellet
(large membranes, nuclei)
supernatant
(smaller organelles such as mitochondria, chloroplasts)
pellet
(mitochondria and other large organelles)
supernatant
(small organelles plus protein complexes of cytosol)

Figure 9.9 The fractionation of cells. After initial tissue maceration the solution is filtered to remove large debris and centrifuged at low speeds to remove unbroken cells ($<1000\ \boldsymbol{g}$). Progressive increases in centrifugal forces pellet smaller cellular organelles and eventually at the highest speeds large protein complexes may be isolated

called the relative centrifugal force (RCF) is used.

$$\text{RCF} = \text{centrifugal force/force due to gravity} \quad (9.4)$$

$$= \omega^2 rm/m\ \boldsymbol{g} \quad (9.5)$$

$$= \omega^2 r/\boldsymbol{g} \quad (9.6)$$

To determine the maximum relative centrifugal force it is therefore only necessary to know the radius of the rotor and the speed of rotation. The radius is fixed for a given rotor and the speed of rotation is a user-defined parameter. Although RCF is a dimensionless quantity it is traditionally quoted as a numerical value; for example 5000 $\boldsymbol{g}$ meaning 5000 × the earth's gravitational force. Since ω the angular velocity, in radians per second, is the number of revolutions per second multiplied by 2π the RCF of a 10 cm radius rotor rotating at 8000 r.p.m. is

$$\text{RCF} = (2\pi \times 8000/60)^2 \times 0.1/9.8 \sim 7160\ \boldsymbol{g} \quad (9.7)$$

This is actually the $\boldsymbol{g}$ force at the end of the rotor tube. More careful consideration reveals that the $\boldsymbol{g}$ force will vary along the radius of the tube. In the above example the $\boldsymbol{g}$ force at a position 5 cm along the tube is approximately 3580 $\boldsymbol{g}$ and this has frequently led to the use of a term $\boldsymbol{g}_{\text{av}}$ to reflect average centrifugal force.

In analytical methods of centrifugation the second term of Equation 9.2, the buoyant force, must be considered. In most applications this term is 'invisible' since rotor speeds are such that membrane or large organelles are forced to sediment whilst the lighter particles remain in solution. An extension of this relationship allows the size or molecular mass of proteins to be determined.

During centrifugation at high speeds the initial acceleration of a particle due to the net force is relatively short in duration (~1 ns), after which time the particle moves at a constant velocity. This constant velocity arises because the solution exerts a frictional force (f) on the particle that is proportional to the sedimentation velocity ($\mathrm{d}r/\mathrm{d}t$). Once the steady state is reached we can rewrite Equation 9.3 as

$$f\ \mathrm{d}r/\mathrm{d}t = \omega^2 rm - \omega^2 r\nu\rho \quad (9.8)$$

In Equation 9.8 the volume of a particle is a difficult quantity to measure and is replaced by a term called the partial specific volume ($\overline{\nu}$). This is defined as the increase in volume when 1 g of solute is dissolved in a large volume of solvent. The quantity $m\nu$ is defined as the increase in volume caused by the addition of one molecule of mass m and it is equal to the volume of the particle.[1] Introduction of these terms allows Equation 9.8 to be simplified

$$f\,\mathrm{d}r/\mathrm{d}t = \omega^2 rm - \omega^2 rm\ \bar{v}\rho \tag{9.9}$$

$$f\,\mathrm{d}r/\mathrm{d}t = \omega^2 rm\ (1 - \bar{v}\rho) \tag{9.10}$$

and rearrangement leads to

$$S = \frac{\mathrm{d}r/\mathrm{d}t}{\omega^2 r} = \frac{m\ (1 - \bar{v}\rho)}{f} \tag{9.11}$$

where the term $\mathrm{d}r/\mathrm{d}t/\omega^2 r$ is defined as the sedimentation coefficient (S called the Svedburg unit after the pioneer of ultra-centrifugation Theodor Svedburg). The Svedburg has units of 10^{-13} s, and as an example haemoglobin has a sedimentation coefficient of 4×10^{-13} s or 4 S. Since the mass m can be described as the molecular weight of the solute divided by Avogadro's constant (M/N_o) and by assuming the particle to be spherical we can utilize Stokes law to describe the frictional coefficient, f, as

$$f = 6\pi\ \eta\ r_s \tag{9.12}$$

where r_s is the radius of the particle and η is the viscosity of the solvent (normally water).[2] This leads to the following equation

$$S = \frac{M(1 - \bar{v}\rho)}{N_o\ 6\pi\ \eta\ r_s} \tag{9.13}$$

that can be rearranged to give

$$M = SN_o\ (6\pi\ \eta\ r_s)/(1 - \bar{v}\rho) \tag{9.14}$$

If we do not make any assumptions about shape we can use the relationship

$$D = k_B T/f \tag{9.15}$$

where D is the diffusion coefficient, f is the frictional coefficient, k_B is Boltzmann's constant and T the absolute temperature. This leads to

$$M = SRT/D\ (1 - \bar{v}\rho) \tag{9.16}$$

and allows the molecular weight to be calculated if S, the sedimentation velocity, is measured since all of the remaining terms are constants or are derived from other measurements. From Equation 9.11 S is defined as $\frac{\mathrm{d}r}{\mathrm{d}t}\frac{1}{\omega^2 r}$ or

$$S\,\mathrm{d}t = 1/\omega^2 \mathrm{d}r/r \tag{9.17}$$

with integration of the above equation leading to the relationship

$$\ln r/r_0 = S\ t\ \omega^2 \tag{9.18}$$

By measuring the time taken for the particle to travel between two points, r_o (at time $t = 0$) and r (at time t) and knowing ω the angular velocity (in radians/s) it is relatively straightforward to calculate S. Measurement of protein sedimentation rates is performed by recording changes in refractive index as the protein migrates through an optical cell. Table 9.1 lists the sedimentation coefficients for various proteins.

A more direct way of measuring molecular weight using ultracentrifugation is to measure not the rate of sedimentation but the equilibrium established after many hours of centrifugation at relatively low speeds. At equilibrium a steady concentration gradient will occur with the flow of proteins towards sedimentation balanced by reverse flow as a result of diffusion. It can be demonstrated that for a homogeneous protein the gradient is described by the equation

$$c(r)/c(r_0) = \exp M(1 - \bar{v}\rho)\omega^2\ (r^2 - r_0{}^2)/2RT \tag{9.19}$$

where $c(r)$ is the concentration at a distance r from the axis of rotation and $c(r_0)$ is the concentration of protein at the meniscus or interface (r_0). From Equation 9.19 a plot of ln $c(r)$ against $(r^2 - r_0{}^2)$ will yield a straight line whose slope is $M(1 - \bar{v}\rho)\omega^2/2RT$, from which M is subsequently estimated (Figure 9.10).

Solubility and 'salting out' and 'salting in'

One of the most common and oldest methods of protein purification is to exploit the differential solubility of

[1] For most proteins $\bar{v}$ has a value of 7.4×10^{-4} m^3/kg or 0.74 ml/g.

[2] For non-spherical particles correction of this relationship must consider anisotropy where one axis may be much longer than the other two. An example would be a cylindrically shaped molecule. Hydrodynamic analysis can in some instances allow the shape of the molecule to be deduced or at least distinguished from the simple spherical situation.

Table 9.1 Sedimentation coefficients associated different proteins

Protein	Molecular mass (kDa)	Sedimentation coefficient ($S_{20,\mathrm{w}}$)
Lipase	6.7	1.14
Ribonuclease A	12.6	2.00
Myoglobin	16.9	2.04
Concanavalin B	42.5	3.50
Lactate Dehydrogenase	150	7.31
Catalase	222	11.20
Fibrinogen	340	7.63
Glutamate Dehydrogenase	1015	26.60
Turnip yellow mosaic virus	3013	48.80
Large ribosomal subunit	1600	50

Notice the 'abnormal' value for fibrinogen and the absence of a simple correlation between mass and S value (after Smith in Sober H.A. (ed.) *Handbook of Biochemistry and Molecular Biology*, 2nd edn. CRC Press).

proteins at various ionic strengths. In concentrated solutions the solubility of many proteins decreases differentially leading to precipitation whilst others remain soluble. This allows protein separation since as solubility decreases precipitation occurs with the precipitant removed from solution by centrifugation. The use of differential solubility to fractionate proteins is often used as one part in a strategy towards overall purification.

The solubility of proteins in aqueous solutions differs dramatically. Although membrane proteins are clearly insoluble in aqueous solution many structural proteins such as collagen are also essentially insoluble under normal physiological conditions. For small globular proteins with masses below 10 kDa intrinsic solubility is often high and concentrations of 10 mM are easily achieved. This corresponds to approximately 100 g l^{-1} or about 10 percent w/v. In solution globular proteins are surrounded by a tightly bound layer of water that differs in properties from the 'bulk' water component. This water, known as the hydration layer, is ordered over the surface of proteins and appears to interact preferentially with charged side chains and polar residues. In contrast, this water is rarely found close to hydrophobic patches on the surface of proteins. Altering the properties of this water layer effects the solubility of proteins. Similarly protein solubility is frequently related to the isoelectric point or pI. As the pH of the aqueous solution approaches the pI precipitation of a protein can occur. As the pH shifts away from the pI the solubility of a protein generally increases. This is most simply correlated with the overall charge of the protein and its interaction with water. Since the pI of proteins can vary from pH 2–10 it is clear that pH ranges for protein solubility will also vary widely.

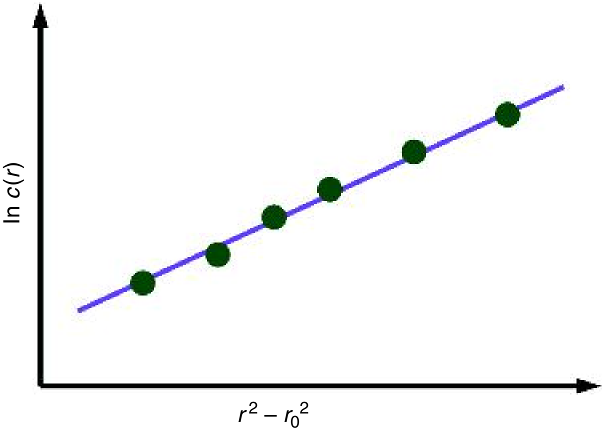

Figure 9.10 A plot of ln $c(r)$ against $(r^2 - r_0{}^2)$ yields a straight line that allows direct estimation of M, the protein molecular weight. The slope of the line is given by $M(1 - \overline{\nu}\rho)\omega^2/2RT$

The ionic strength (I) is defined as

$$I = 1/2\ \Sigma\ m_i z_i{}^2 = 1/2\ (m_+ z_+{}^2 + m_- z_-{}^2) \quad (9.20)$$

where m and z are the concentration and charge of all ionic species in solution. Thus a solution of 0.5 M NaCl has an ionic strength of 0.5 whilst a solution of 0.5 M $MgCl_2$ has an ionic strength of

1.5. The concept of ionic strength was introduced by G.N. Lewis to account for the non-ideal behaviour of electrolyte solutions that arose from the total number of ions present along with their charges rather than the chemical nature of each ionic species. A physical interpretation of properties of electrolytes from the studies of Peter Debye and Erich Hückel, although mathematical and beyond the scope of this book, assumed that electrolytes are completely dissociated into ions in solution, that concentrations of ions were dilute below 0.01 M and that on average each ion was surrounded by ions of opposite charge. This last concept was known as an ionic atmosphere and is relevant to changes in protein solubility as a function of ionic strength. Trends in protein solubility vary from one protein to another and show a strong dependence on the salts used to alter the ionic strength (Figure 9.11).

At low ionic strength a protein is surrounded by excess counter-ions of the opposite charge known as the ionic atmosphere. These counter-ions screen charges on the surface of proteins. Addition of extra ions increases the shielding of surface charges with the result that there is a decrease in intermolecular attraction followed by an increase in protein solubility as more dissolves into solution. This is the 'salting-in' phenomenon (Figure 9.12). However, the addition of more ions, reverses this trend. When the ionic concentration is very high (>1 M) each ion must be hydrated with the result that bulk solvent (water) is sequestered from proteins leading to decreases in solubility. If continued decreased solubility leads to protein aggregation and precipitation. Ionic effects on protein solubility were first recognized by Franz Hofmeister around 1888, and by analysing the effectiveness of anions and cations in precipitating serum proteins he established an order or priority that reflected each ions 'stabilizing' properties.

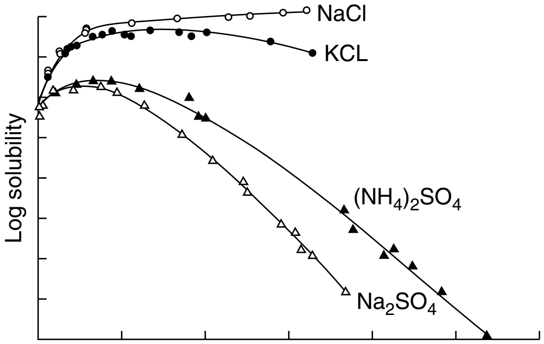

Figure 9.11 The solubility of haemoglobin in different electrolytes illustrates the 'salting in' and 'salting out' effects as a function of ionic strength. Derived from original data by Green, A.A. *J. Biol. Chem.* 1932, **95**, 47

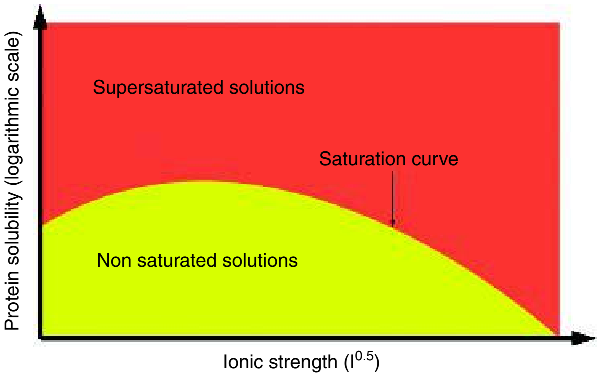
Supersaturated solutions
Saturation curve
Non saturated solutions
Protein solubility (logarithmic scale)
Ionic strength ($I^{0.5}$)

Figure 9.12 Model profile for solubility as a function of ionic strength. Saturation occurs when solid and solution phases are in equilibrium. 'Salting-out' is seen on the right-hand side of the diagram where there is a reduction in protein solubility as the concentration of salt increases whilst 'salting in' is apparent on the left-hand side of the diagram where there is an increase in protein solubility with ionic strength

Cations: $N(CH_3)_3^+ > NH_4^+ > K^+ > Na^+ > Li^+ > Mg^{2+} > Ca^{2+} > Al^{3+}$ > guanidinium

Anions: $SO_4^{2-} > HPO_4^{2-} > CH_3COO^-$ > citrate > tartrate > $F^- > Cl^- > Br^- > I^- > NO_3^- > ClO_4^- > SCN^-$

The priority is known as the Hofmeister series (Table 9.2); the first ions in each series are the most stabilizing with substituted ammonium ions, ammonium itself and potassium being more stabilizing

Table 9.2 Properties of the Hofmeister series of cations and anions

Stabilizing ions	↔ Destabilizing ions
Strongly hydrated anions	↔ Strongly hydrated cations
Weakly hydrated cations	↔ Weakly hydrated anions
Kosmotropic	↔ Chaotropic
Increase surface tension	↔ No effect on surface tension
Decrease solubility of non-polar side-chains	↔ Increase solubility of non-polar side-chains
Salting out	↔ Salting in

than calcium or guanidinium, for example. Similarly for anions of the Hofmeister series sulfate, phosphate and acetate are more stabilizing than perchlorate or thiocyanate.

At the heart of this series is the effect of ions on the ordered structure of water. Ordered structure in water is perturbed by ions since they disrupt natural hydrogen-bond networks. The result is that the addition of ions has a similar effect to increasing temperature or pressure. Ions with the greatest disruptive effect are known as structure-breakers or chaotropes. In contrast, ions exhibiting strong interactions with water molecules are called kosmotropes. It is immediately apparent that molecules such as guanidinium thiocyanate are extremely potent chaotropes whilst ammonium sulfate is a stabilizing molecule or kosmotrope. Na^+ and Cl^- are frequently viewed as the 'border zone' having neutral effects on proteins.

Ammonium sulfate is commonly used to selectively precipitate proteins since it is very soluble in water thereby allowing high concentrations (>4 M). Under these conditions harmful effects on proteins such as irreversible denaturation are absent and NH_4^+ and SO_4^{2-} are both at the favourable, non-denaturing, end of the Hofmeister series. The mechanism of 'salting out' proteins resides in the disruption of water structure by added ions that lead to decreases in the solubility of non-polar molecules. By using ammonium sulfate it is possible to quantitatively precipitate one protein from a mixture. The remaining proteins are left in solution and such methods are widely used to purify soluble proteins from crude cell extracts.

Chromatography

Chromatographic methods form the core of most purification protocols and have a number of common features. First, the protein of interest is dissolved in a mobile phase normally an aqueous buffered solution. This solution is often derived from a cell extract and may be used directly in the chromatographic separation. The mobile phase is passed over a resin (called the immobile phase) that selectively absorbs components in a process that depends on the type of functional group and the physical properties associated with a protein. These properties include charge, molecular mass, hydrophobicity and ligand affinity. The choice of the immobile phase depends strongly on the properties of the protein exploited during purification.

The general principle of chromatography involves a column containing the resin into which buffer is pumped to equilibrate the resin prior to sample application. The sample is applied to the column and separation relies on interactions between protein and the functional groups of the resin. Interaction slows the progress of one or more proteins through the column whilst other proteins flow unhindered through the column. As a result of their faster progress through the column these proteins are separated from the remaining protein (Figure 9.13). In some instances protein mixtures vary in their interaction with functional groups; a strong interaction results in slow progress through the column, a weak interaction leads to faster rates of progress whilst no interaction results in unimpeded flow through the resin.

Pioneering research into the principles of chromatography performed by A.J.P. Martin and R.L.M. Synge in the 1940s developed the concept of theoretical plates in chromatography. In part this theory explains migration rates and shapes of eluted zones and views the chromatographic column as a series of continuous, discrete, narrow layers known as theoretical plates. Within each plate equilibration of the solute (protein) between the mobile and stationary (immobile) phases occurs with the solute and solvent moving through the

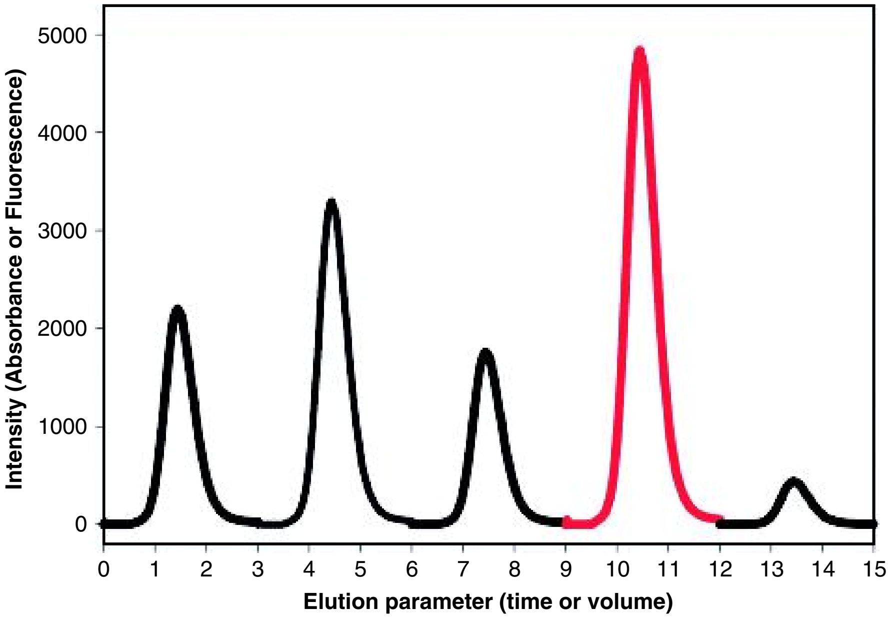

Figure 9.13 An 'ideal' separation of a mixture of proteins. The peaks representing different components are separated with baseline resolution. A less than ideal separation will result in peak overlap

column in a stepwise manner from 'plate' to 'plate'. The resolution of a column will increase as the theoretical plate number increases. Practically, there are a number of ways to increase the theoretical plate size, and these include making the column longer or decreasing the size of the beads packing the column. The value of increasing resolution is that two similarly migrating solutes may be separated with reasonable efficiency.

Today, sophisticated equipment is available to separate proteins. This equipment, controlled via a computer, includes pumps that allow flow rates to be automatically adjusted over a wide range from perhaps as low as 10 μl min^{-1} to 100 ml min^{-1}. In addition, complex gradients of solutions can be constructed by controlling the output from more than one pump, and under some circumstances these gradients vastly improve chromatographic separations. The addition of the sample to the top of the column is controlled via a system of valves whilst material flowing out of the column (eluate) is detected usually via fluorescence- or absorbance-based detectors (Figure 9.14).

A major advance has been the development of resins that withstand high pressures (sometimes in excess of 50 MPa) generated using automated or high pressure liquid chromatography (HPLC) systems (Figure 9.15). Higher pressures are advantageous in allowing higher flow rates and decreased separation times. Previously, manual chromatographic methods relied on the flow of solution through the column under the influence of gravity and resulted in separations taking several days. Today it is possible to routinely perform chromatographic separations with high levels of resolution in less than 10 min.

Modern chromatographic systems contain many of the components shown schematically in Figure 9.14. Two or more pumps drive buffer onto columns at controlled rates usually in a range from 0.01–10 ml min^{-1}. The pumps are linked to the column via chemically resistant tubing that conveys buffers to the column. Between the pump and the column is a 'mixer' that adjusts the precise volumes dispensed from the pumps to allow the desired buffer composition. This is particularly important in establishing buffer gradients of ionic strength or pH. Buffer of defined composition is pumped through a series of motorized valves that

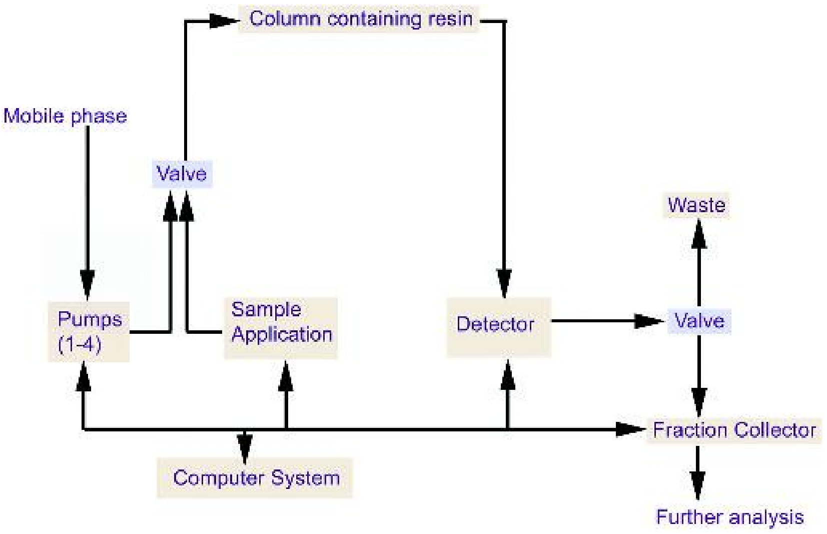

Figure 9.14 Schematic representation of chromatographic systems indicating flow of mobile phase and sample together with the flow of information between all devices and a computer workstation

Figure 9.15 A modern automated chromatography system embodying the flow scheme of Figure 9.14. (Reproduced courtesy of Amersham Biosciences)

allow addition of sample to the column or its diversion to a waste outlet. The column itself is made of precision-bored glass, stainless steel or titanium and contains chromatographic resin.

Elution of material from the column leads to a detection system that usually measures absorbance based around the near UV region at 280 nm for proteins. Frequently, detection systems will also measure pH and ionic strength of the material eluting from the column. The detection system is interfaced with a fraction collector that allows samples of precise volume to be collected and peak selection on the basis of rates of change of absorbance with time. In any system all of the components are controlled via a central computer and the user inputs desired parameters such as flow rate, fraction volume size, wavelength for detection, run duration, etc.

Ion exchange chromatography

Ion exchange chromatography separates proteins on the basis of overall charge and depends on the relative numbers of charged side chains. Simple calculations of the number of charge groups allow an estimation of the isoelectric point and an assessment of the charge at pH 7.0. Anionic proteins have a pI < 7.0 whilst cationic proteins have isoelectric points >7.0. In ion exchange chromatography anionic proteins bind to cationic groups of the resin in the process of *anion exchange* chromatography. Similarly, positively charged proteins bind to anionic groups within a supporting matrix in *cation exchange* chromatography.

The first ion exchangers developed for use with biological material were based on a supporting matrix of cellulose. Although hydrophilic and non-denaturing cellulose was hindered by its low binding capacity for proteins due to the small number of functional groups attached to the matrix, biodegradation and poor flow properties. More robust resins were required and the next generation were based around cross-linked dextrans, cross-linked agarose, or cross-linked acrylamide in the form of small, bead-like, particles. These resins offered significant improvements in terms of pH stability, flow rates and binding capacity. The latest resins are based around extremely small homogeneous beads of average diameter $<10\ \mu m$ containing hydrophilic polymers. One implementation of this technology involves polystyrene cross-linked with divinylbenzene producing resins resistant to biodegradation, usable over wide pH ranges (from 1–14) and tolerating high pressures (>10 MPa).

For cation exchange chromatography the common functional groups (Figure 9.16) involve weak acidic groups such as carboxymethyl or a strong acidic group such as sulfopropyl or methyl sulfonate. Similarly for anion exchange, the functional groups are diethylaminoethyl and the stronger exchange group of diethyl-(2-hydroxypropyl)amino ethyl. The binding of proteins to either anion or cation exchange resins will depend critically on pH and ionic strength. The pH of the solution remains important since it influences the overall charge on a protein. Fairly obviously, it is not sensible to perform ion exchange chromatography at a pH close to the isoelectric point of a protein. At this pH the protein is uncharged and will not stick to any ion exchange resin.

In ion exchange chromatography samples are applied to columns at low ionic strengths (normally $I < 0.05$ M) to maximize interactions between protein and matrix. The column is then washed with further solutions of constant pH and low ionic strength to remove proteins lacking any affinity for the resin. However, during any purification protocol many proteins are present in a sample and exhibit a range of interactions with ion exchange resins. This includes proteins that bind very tightly as well as those with varying degrees of affinity for the resin. These proteins are separated by slowly increasing the ionic strength of the solution. Today's sophisticated chromatography systems establish gradients where the ionic strength is increased from low to high levels with the ions competing with the protein for binding sites on the resin. As a result of competition proteins are displaced from the top of the column and forced to migrate downwards through the column. Repetition of this process occurs continuously along the column with the result that weakly bound proteins are eluted. Eventually at high ionic strengths even tightly bound proteins are displaced from the resin and the cumulative effect is the separation of proteins according to charge. Collecting proteins at regular intervals with a fraction collector and recording their absorbance or fluorescence establishes an elution profile and assists in locating the protein of interest.

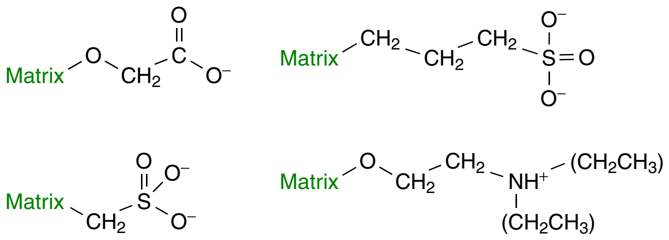

Figure 9.16 The common functional groups found in ion exchange resins. The terms strong and weak exchangers refer to the extent of ionization with pH. Strong ion exchangers such as QAE and SP are completely dissociated over a very wide pH range whereas with the weak exchangers (DEAE and CM) the degree of ionization varies and this is reflected in their binding properties

Size exclusion or gel filtration chromatography

Size exclusion separates proteins according to their molecular size. The principle of the method relies on separation by the flow of solute and solvent over beads containing a series of pores. These pores are of constant size and are formed by covalent cross-linking of polymers. High levels of cross-linking lead to a small pore size. In a mixture of proteins of different molecular mass some will be small enough to enter the pores and as a result their rate of diffusion through the column is slowed. Larger proteins will rarely enter these regions with the result that progress through the resin is effectively unimpeded. Very large proteins will elute in the 'void volume' and all proteins that fail to show an interaction with the pores will elute in this void volume irrespective of their molecular size. Migration rates through the column reflect the extent of interaction with the pores; small proteins spend a considerable amount of time in the solvent volume retained within pores and elute more slowly than those proteins showing marginal interactions with the beads. It is clear that this process represents an effective way of discriminating between proteins of different sizes (Figure 9.17).

In an ideal situation there is a linear relationship between the elution volume and the logarithm of the molecular mass (Figure 9.18). This means that if a size exclusion column is calibrated using proteins of known mass then the mass of an unknown protein can be deduced from its elution position. A reasonably accurate estimation of mass can be obtained if the unknown protein is of similar shape to the standard set. This occurs because the process of size exclusion or gel filtration is sensitive to the average volume occupied by the protein in solution. This volume is often represented by a term called the Stokes radius of the protein and this parameter is very sensitive to overall shape. So, for example, elongated molecules such as fibrous proteins show anomalous rates of migration. A rigorous description of size exclusion chromatography would view the separation as based on differences in hydrodynamic radius. Today, size exclusion chromatography is rarely used as an

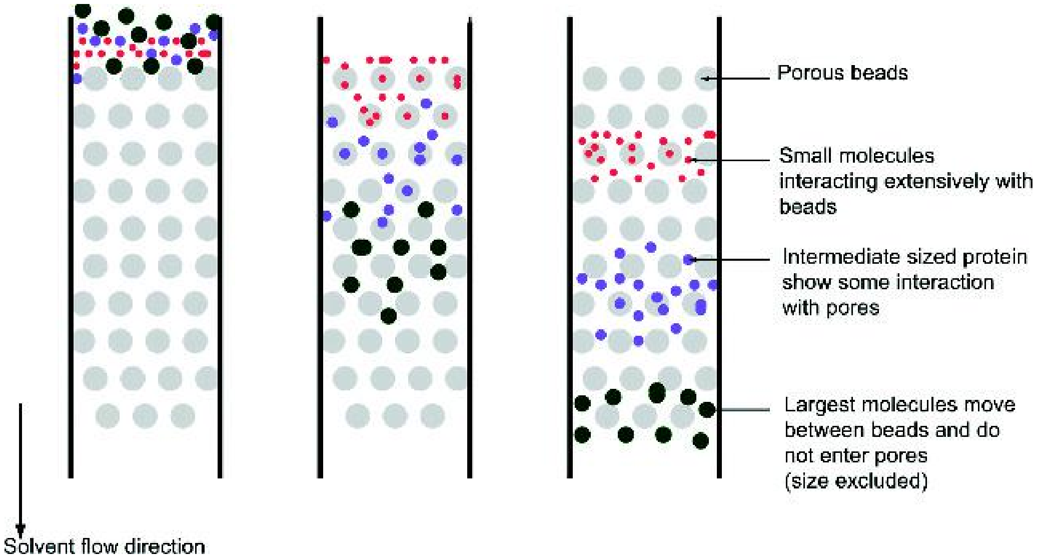
Porous beads
Small molecules interacting extensively with beads
Intermediate sized protein show some interaction with pores
Largest molecules move between beads and do not enter pores (size excluded)
Solvent flow direction

Figure 9.17 The principle of gel filtration showing separation of three proteins of different mass at the beginning, middle and end of the process. Resins are produced by cross linking agarose, acrylamide and dextran polymers to form different molecular size ranges from 100–10 000 (peptide–small proteins), 3000–70 000 (small proteins), 10^4–10^5 (larger proteins, probably multi-subunit proteins), and 10^5–10^7 (macromolecular assemblies such as large protein complexes, viruses, etc

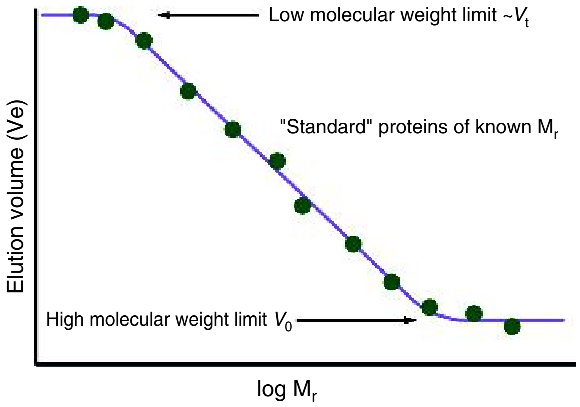

Figure 9.18 Estimation of molecular mass using size exclusion chromatography. The volume (or time) necessary to achieve elution from a column is plotted against the logarithm of molecular weight to yield a straight line that deviates at extreme (low and high) molecular weight ranges. These extremes correspond to the total volume of the column (V_t) and the void volume respectively (V_o)

analytical technique to estimate subunit molecular mass but it is frequently used as a preparative method to eliminate impurities from proteins. One common use is to remove salt from a sample of protein or to exchange the protein from one buffer to another.

Hydrophobic interaction chromatography

Hydrophobic interaction chromatography (HIC) is based on interactions between non-polar side chains and hydrophobic resins. At first glance it is not obvious how HIC is useful for separating and purifying proteins. In globular proteins the folded states are associated with buried hydrophobic residues well away from the surrounding aqueous solvent. However, close inspection of some protein structures reveals that a fraction of hydrophobic side chains remain accessible to solvent at the surface of folded molecules. In some cases these residues produce a distinct hydrophobic patch whose composition and size varies from one protein to another and leads to large differences in surface hydrophobicity. HIC separates proteins according to these differences. As a technique it exploits interactions between hydrophobic patches on proteins and hydrophobic groups covalently attached to an inert matrix. The most popular resins are cross-linked agarose polymers containing butyl, octyl or phenyl (hydrophobic) functional groups.

The hydrophobic interaction depends on solution ionic strength as well as chaotropic anions/cations. As a result HIC is usually performed under conditions of high ionic strength in solutions containing 1 M ammonium sulfate or 2 M NaCl. High ionic strength increases interactions between hydrophobic ligand and protein with the origin of this increased interaction arising from decreased protein solvation as a result of the large numbers of ions each requiring a hydration shell. The net result is that hydrophobic surfaces interact strongly as ions ‘unmask’ these regions on proteins.

Samples are applied to columns containing resin equilibrated in a solution of high ionic strength and proteins containing hydrophobic patches stick to the resin whilst those lacking hydrophobic interactions are not retained by the column. Elution is achieved by decreasing favourable interactions by lowering the salt concentration via a gradient or a sudden ‘step’. Proteins with different hydrophobicity are separated during elution with the most hydrophobic protein retained longest on the column. However, the prediction of the magnitude of surface hydrophobicity on proteins remains difficult and the technique tends to be performed in an empirical manner where a variety of resins are tried for a particular purification protocol. One advantage of HIC is the relatively mild conditions that preserve folded structure and activity.

Reverse phase chromatography

A second chromatographic method that exploits hydrophobicity is reverse phase chromatography (RPC) although the harsher conditions employed in this method can lead to denaturation. The basis of RPC is the hydrophobic interaction between proteins/peptides and a stationary phase containing alkyl or aromatic hydrocarbon ligands immobilized on inert supports. The original support polymer for reverse phase methods was silica to which hydrophobic ligands were linked. Although still widely used as a base matrix it has a disadvantage of instability at alkaline pH and this restricts its suitability in biological studies. As a result, synthetic organic polymers based on polystyrene beads have proved popular and show excellent chemical stability over a wide range of conditions.

The separation mechanism in RPC depends on hydrophobic binding interactions between the solute molecule (protein or peptide) in the mobile phase and the immobilized hydrophobic ligand (stationary phase). The initial mobile phase conditions used to promote binding in RPC are aqueous solutions and lead to a high degree of organized water structure surrounding protein and immobilized ligand. The protein binds to the hydrophobic ligand leading to a reduction in exposure of hydrophobic surface area to solvent. The loss of organized water structure is accompanied by a favourable increase in system entropy that drives the overall process. Proteins partition between the mobile and stationary phases in a process similar to that observed by an organic chemist separating iodine between water and methanol. The distribution of the protein between each phase reflects the binding properties of the medium, the hydrophobicity of the protein and the overall composition of the mobile phase. Initial experimental conditions are designed to favour adsorption of the protein from the mobile phase to the stationary phase. Subsequently, the mobile phase composition is modified by decreasing its polarity to favour the reverse process of desorption leading to the protein partitioning preferentially back into the mobile phase. Elution of the protein or peptide occurs through the use of gradients that increase the hydrophobicity of the mobile phase normally through the addition of an organic modifier. This is achieved through the addition of increasing amounts of non-polar solvents such as acetonitrile, methanol, or isopropanol. At the beginning a protein sample is applied to a column in 5 percent acetonitrile, 95 percent water, and this is slowly changed throughout the elution stage to yield at the end of the process a solution containing, for example, 80 percent acetonitrile/20 percent water.

For both silica and polystyrene beads the hydrophobic ligands are similar and involve linear hydrocarbon chains (*n*-alkyl groups) with chain lengths of C2, C4, C8 and C18 (Figure 9.19). C18 is the most hydrophobic ligand and is used for small peptide purification whilst the C2/C4 ligands are more suitable for proteins. Although proteins have been successfully purified using RPC the running conditions often lead to denaturation and reverse phase methods are more valuable for purifying short peptides arising from chemical synthesis or enzymatic digestion of larger proteins (see Chapter 6).

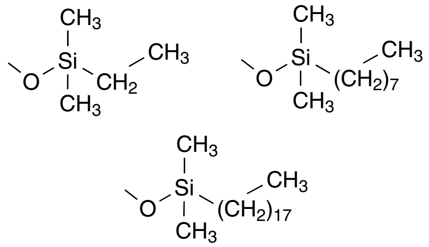

Figure 9.19 The hydrophobic C2, C8 and C18 ligands widely used in RPC

Affinity chromatography

An important characteristic of some proteins is their ability to bind ligands tightly but non-covalently. This property is exploited as a method of purification with the general principle involving covalently attachment of the ligand to a matrix. When mobile phase is passed through a column containing this group proteins showing high affinity for the ligand are bound and their rate of progress through the column is decreased. Most proteins will not show affinity for the ligand and flow straight through the column. This protocol therefore offers a particularly selective route of purification and affinity chromatography has the great advantage of exploiting biochemical properties of proteins, as opposed to a physical property.

Affinity chromatography relies on the interaction between protein and immobilized ligand. Binding must be sufficiently specific to discriminate between proteins but if binding is extremely tight (high affinity) it can leads to problems in recovering the protein from the column. Most methods of recovery involve competing for the binding site on the protein. Consequently, a common method of elution is to add exogenous or free ligand to compete with covalently linked ligand for the binding site of the protein. A second method is to change solution conditions sufficiently to discourage protein–ligand binding. Under these conditions the protein no longer shows high affinity for the ligand and is eluted. The last method must of course not use conditions that promote denaturation of the protein, and frequently it involves only moderate changes in solution pH or ionic strength. Enzymes are particularly suited to this form of chromatography since co-factors such as NAD, NADP, and ADP are clear candidate ligands for use in affinity chromatography.

One limitation in the development of affinity chromatography is that methods must exist to link the ligand covalently to the matrix. This modification must not destroy its affinity for the protein and must be reasonably stable i.e not broken under moderate temperature, pH and solution conditions used during chromatography. Recently, a great expansion in the number and type of affinity based chromatographic separations has occurred due to the development of expression systems producing proteins with histidine tags or fused to other protein such as glutathione-*S*-transferase or maltose binding protein (Table 9.3). Histidine tags can be located at either the N- or C-terminal regions of a protein and bind to metal chelate columns containing the ligand imidoacetic acid. Similarly glutathione-*S*-transferase is a protein that binds the ligand glutathione which itself can be covalently linked to agarose based resins.

Dialysis and ultrafiltration

The basis of dialysis and ultrafiltration is the presence of a semipermeable membrane that allows the flux of small (low molecular weight) compounds at the same time preventing the diffusion of larger molecules such as proteins. Both methods are commonly used in the purification of proteins and in the case of ultrafiltration particularly in the concentration of proteins.

Dialysis is commonly used to remove low molecular weight contaminants or to exchange buffer in a solution containing protein. A solution of protein is placed in dialysis tubing that is sealed at both ends and added to a much larger volume of buffer (Figure 9.20). Low molecular weight contaminants diffuse across the membrane whilst larger proteins remain trapped. Similarly, the buffers inside and outside the dialysis tubing will exchange by diffusion until equilibrium is reached between the compartments. This equilibrium is governed by the Donnan effect, which maintains electrical neutrality on either side of the membrane. For polyvalent ions such as proteins this means that ions of the opposite charge will remain in contact with protein and the solution ionic strength will not fall to exactly that in the bulk solution. For dilute solutions of protein coupled with higher exterior ion concentrations the Donnan effect becomes negligible but for high concentrations of protein or low ionic strength buffers it remains a significant effect on the colligative properties of ions. Despite this complication, buffer inside the dialysis tubing is gradually replaced by the solution found in the larger exterior volume. By replacing the exterior volume of buffer at regular intervals the process rapidly achieves the removal of low molecular weight contaminants or the exchange of solvent.

Ultrafiltration is used to concentrate solutions of proteins by the application of pressure to a sealed vessel whose only outlet is via a semipermeable membrane (Figure 9.20). The molecular weight limit for the semipermeable membrane can be closely controlled. Solvent and low molecular weight molecules pass through the membrane whilst larger molecules remain inside the vessel. As solvent is forced across the membrane proteins inside the ultrafiltration cell are progressively concentrated. Frequently, ultrafiltration is used to concentrate a protein during the final stages of purification when it may be required at high concentrations; in crystallization trials for example.

Polyacrylamide gel electrophoresis

The most common method of analysing the purity of an isolated protein is to use polyacrylamide gel electrophoresis (PAGE) in the presence of the detergent sodium dodecyl sulfate (SDS) and a reductant of disulfide bridges such as β-mercaptoethanol. The technique has the additional advantage of allowing the monomeric (subunit) molecular mass to be determined

Table 9.3 Ligands used for affinity chromatography together with the natural co-factor/substrate and the proteins/enzymes purified by these methods

Aim of affinity reaction	Ligand/natural substrate	Example
To bind nucleotide binding domains	Natural ligands are NADP or NAD. Most affinity-based methods covalently bind either ADP or AMP to resins and this is sufficient to bind the nucleotide binding domain	Dehydrogenases bind NAD or NADH and are frequently purified via this route
Separation of proteins based on the ability of some side chains to chelate divalent metal ions	Iminodiacetic acid covalently linked to gel binds metal ions such as Ni^{2+} or Zn^{2+} leaving a partially unoccupied coordination sphere. Histidine residues are the usual target but tryptophan and cysteine residues can also bind	Any recombinant protein containing a His-tag. Binding occurs via imidazole nitrogen and protonation destroys binding and hence is a route towards elution
Immobilized lectins such as concanavalin A	Binds α-D-mannose, α-D-glucose and related sugars via interaction with free OH groups	Separation of glycoproteins particularly viral glycoproteins and cell surface antigens
Glutathione binding proteins	Glutathione is the natural substrate and can be covalently linked to column	Fusion proteins containing glutathione-*S*-transferase
Affinity of monoclonal and polyclonal IgG for protein A, protein G or a fusion protein, protein A/G	Protein A is a protein of molecular weight 42 000 derived from *S. aureus*. It consists of six major regions five of which bind to IgG. Other affinity columns exploit the use of protein G – a protein derived from streptococci. Both proteins exhibit affinity for the Fc region of immunoglobulins	Separation of IgG subclasses from serum or cell culture supernatants
Reversible coupling of proteins containing a free thiol group	Thiol group immobilized on column matrix. Forms mixed disulfide with protein containing free–SH group. Reversed by addition of excess thiol or other reductant	Cysteine proteinases a family of enzymes with reactive thiols at their catalytic centres are purified by these methods
Affinity for oligonucleotides	Polyadenylic acid linked via N6 amino group or polyuridylic acid linked to matrix	Purification of viral reverse transcriptases and mRNA binding proteins

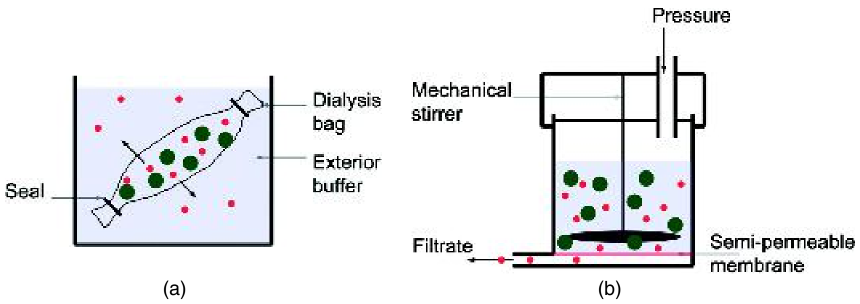

Figure 9.20 Dialysis (a) and ultrafiltration (b) assemblies used in protein purification

with reasonable accuracy. SDS–PAGE is widely used to assess firstly if an isolated protein is devoid of contaminating proteins and secondly whether the purified protein has the expected molecular mass. Both of these parameters are extremely useful during protein purification.

The principle of SDS–PAGE is the separation of proteins (or their subunits) according to molecular mass by their movement through a polyacrylamide gel of closely defined composition under the influence of an electric field. Looking at the individual components of this system allows the principles of electrophoresis to be illustrated and to emphasize the potential of this technique to provide accurate estimations of mass, composition and purity. The gel component is a porous matrix of cross-linked polyacrylamide. The most common method of formation involves the reaction between acrylamide and N, N'-methylenebisacrylamide (called 'bis'; Figure 9.21) catalysed by two additional compounds ammonium persulfate (APS) and N,N,N',N'-tetramethylethylenediamine (TEMED). Polyacrylamide gels form when a dissolved mixture of acrylamide and the bifunctional 'bis' cross-linker polymerizes into long, covalently linked, chains. The polymerization of acrylamide is a free-radical catalysed reaction in which APS acts as an initiator. Initiation involves generating a persulfate free-radical that activates the quaternary amine TEMED, which in turns acts as a catalyst for the polymerization of acrylamide monomers. The concentration of acrylamide can be varied to alter 'resolving power' and this contributes to the definition of a 'pore' size through which proteins migrate under the influence of an electric field. A smaller pore size is generated by high concentrations of acrylamide with the result that only small proteins migrate effectively. This action of the gel is therefore similar to size exclusion chromatography where there is an effective filtration or molecular sieving process by pores.

The mobility of a protein through polyacrylamide gels is determined by a combination of overall charge, molecular shape and molecular weight. Native PAGE (performed in the absence of SDS) yields the mass of native proteins and is subject to deviations caused by non-spherical shape and residual charge as well as interactions between subunits. The technique is less widely used than SDS–PAGE. In the presence of SDS the parameters of shape and charge become unimportant and separation is achieved solely on the basis of protein molecular weight. The underlying reason for this observation is that SDS binds to almost all proteins destroying native conformation (Figure 9.22). SDS causes proteins to unfold, forming rod-like protein micelles that migrate through gels. Since almost all proteins form this rod-like unfolded structure with an excess of negative charge due to the bound SDS the effect of shape and charge is eliminated. SDS binds to proteins at a relatively constant ratio of 1.4 g detergent/g protein leading in a protein of molecular weight 10 000 (i.e ~100 residues) to approximately one SDS molecule for every two amino acid residues.

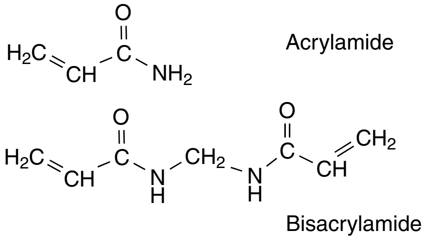

Figure 9.21 The structure of acrylamide and bisacrylamide

A consequence of detergent binding is to coat all proteins with negative charge and to eliminate the charge found on the native protein. As a result of SDS binding all proteins become highly negatively charged, adopt an extended rod-like conformation and migrate towards the anode under the influence of an electric field.

A protein mixture is applied to the top of a gel and migrates through the matrix as a result of the electric field with 'lighter' components migrating faster than 'heavier' components. Over time the component

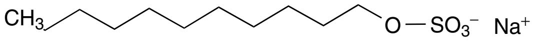

Figure 9.22 Sodium dodecyl sulfate, also called sodium lauryl sulfate, has the structure of a long acyl chain containing a charged sulfate group; it is an extremely potent denaturant of proteins

proteins are separated and the resolving power of the techniques is sufficiently high that heterogeneous mixtures of proteins can be separated and distinguished from each other. In practice most gels contain two components – a short ‘focusing’ gel containing a low percentage of acrylamide (<5 percent) that assists in the ordering and entry of polypeptides into a longer ‘resolving’ gel (Figure 9.23). The above discussion has assumed a constant acrylamide concentration throughout the gel but it is now possible to ‘tailor’ gel performance by varying the concentration of acrylamide. This creates a gradient gel where acrylamide concentrations vary linearly, for example, from 10 to 15 percent. The effect of this gradient is to enhance the resolving power of the gel over a narrower molecular weight range. Although an initial investigation is usually performed with gels containing a constant acrylamide concentration (say 15 percent) increased resolution can be obtained by using gradient gels. The range of gradient usually depends on the mass of the polypeptides under investigation.

The migration of proteins through a gel will depend on the voltage/current conditions used as well as the temperature and it is most common to compare the mobility of an unknown protein or mixture of proteins with pure components of known molecular mass. For example, in Figure 9.24 the migration of proteins in an extract derived from recombinant *E. coli* is compared to the mobility of standard proteins whose molecular masses are known and range from 6 kDa to 200 kDa. The fractionation of the protein of interest can be followed clearly at each stage.

A large number of proteins ranging in mass from less than 10 kDa to greater than 100 kDa are seen in the starting material. Progressive purification removes many of these proteins and allows bands of similar monomeric mass to be identified.

After separation the gel is colourless but profiles such as those of Figure 9.24 result from gel staining that locates the position of each protein band. A number of different stains exist but the two most common dyes are Coomassie Brilliant Blue R250 (see Chapter 3) and silver nitrate. For routine use Coomassie blue staining is preferable and has a detection limit of approximately 100 ng protein. Coomassie Brilliant Blue reacts non-covalently with all proteins and the success of the staining procedure relies on

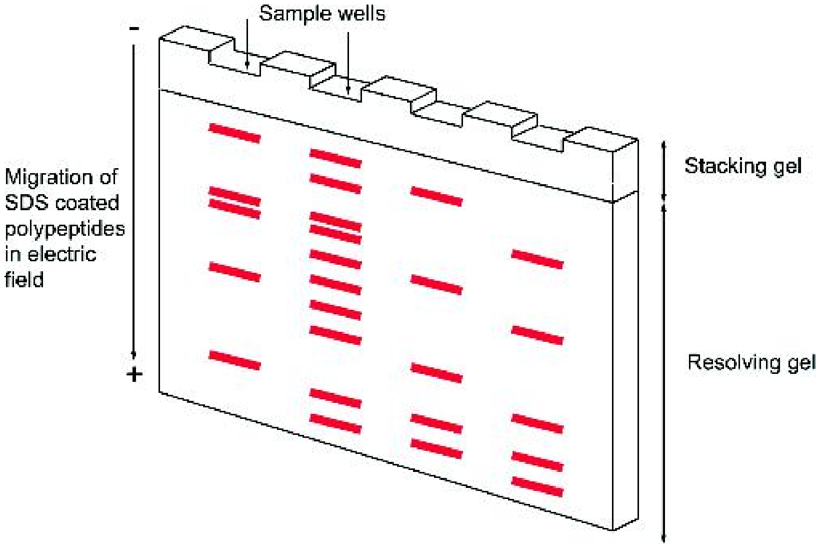

Figure 9.23 A conventional polyacrylamide gel. A top ‘stacking’ gel containing a lower concentration of acrylamide facilitates entry of SDS-coated polypeptides into the ‘resolving’ gel where the bands separate according to molecular mass. Separation is shown by the different migration of bands (shown in red) within the gel

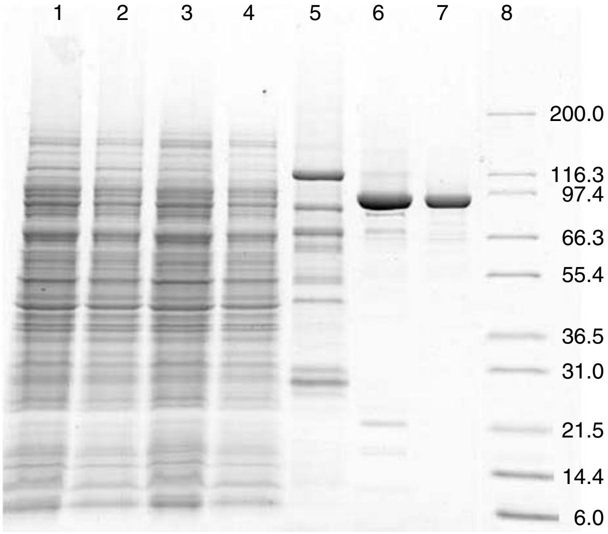

Figure 9.24 A 'real' gel showing the fractionation of a foreign protein expressed originally as a fusion protein from *E. coli* cell lysates. The gel was stained using Coomassie blue. From left to right the protein undergoes progressive purification to eventually yield a pure protein. The cell lysates are shown in lanes 1 and 2, lanes 3 and 4 show the non bound material from application to a glutathione affinity column whilst lane 5 represents the bound material eluted from this column. Lane 6 represents the protein released from the GST fusion by limited proteolysis whilst lane 7 shows the result of further purification by anion exchange chromatography. Lane 8 shows molecular weight markers over a range of 200–6 kDa (reproduced with permission from Chaddock, J.A. *et al. Protein Expression and Purification*. 2002, **25**, 219–228. Academic Press)

'fixing' by precipitation the migrated proteins with acetic acid followed by removal of excess dye using methanol–water mixtures. Silver nitrate has the advantage of increased sensitivity with a detection limit of ~1 ng of protein. The molecular basis for silver staining remains less than clear but involves adding silver nitrate solutions to the gel previously washed under mildly acidic conditions. Today, purpose-built machines have automated the running and staining of gels (Figure 9.25).

Figure 9.25 An automated electrophoresis system. The left-hand compartment performs electrophoretic separation of proteins on thin, uniformly made, gels whilst the second compartment performs the staining reaction to visualize protein components. The gels are run horizontally. (Reproduced courtesy of Amersham Biosciences)

Further use of polyacrylamide gels in studying proteins occurs in the techniques of Western blotting and two-dimensional (2D) electrophoresis. In the first area SDS–PAGE is combined with 'blotting' techniques to allow bands to be identified via further assays. These procedures involve enzyme-linked immunosorbent assays (ELISA), in which antibody binding is linked to a second reaction involving an enzyme-catalysed colour change. Since the reaction will only occur if a specific antigen is detected this process is highly specific. This combination of techniques is called Western blotting (Figure 9.26). Western blotting was given its name as a pun on the technique Southern blotting, pioneered by E.M. Southern, where DNA is transferred from an agarose gel and immobilized on nitrocellulose matrices. In Western blotting the protein is eluted/transferred from SDS–polyacrylamide gels either by capillary action or by electroelution. Electroelution results in the movement of all of the separated proteins out of the gel and onto a second supporting media such as polyvinylidene difluoride (PVDF) or nitrocellulose. Immobilized on this second supporting matrix the proteins undergo further reactions that assist with identification (Figure 9.27). A common technique reacts the eluted protein with previously raised antibodies. Since antibodies will only bind to a specific antigen this reaction immediately

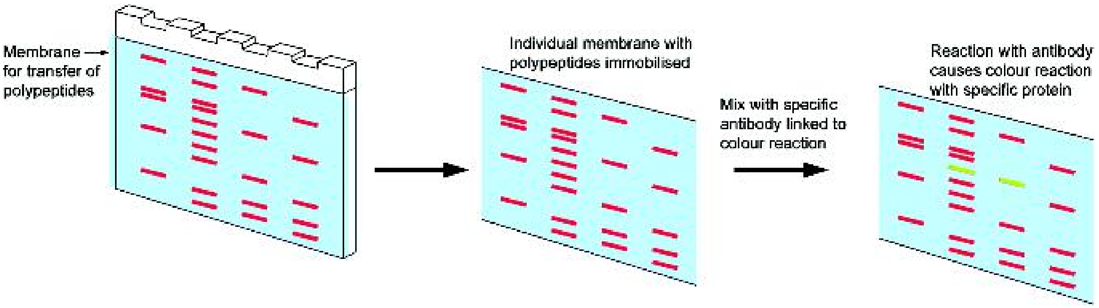
Membrane for transfer of polypeptides
Individual membrane with polypeptides immobilised
Mix with specific antibody linked to colour reaction
Reaction with antibody causes colour reaction with specific protein

Figure 9.26 Western blotting and identification of a protein via cross reaction with a specific antibody. In the example above a single protein in sample wells 2 and 3 is identified

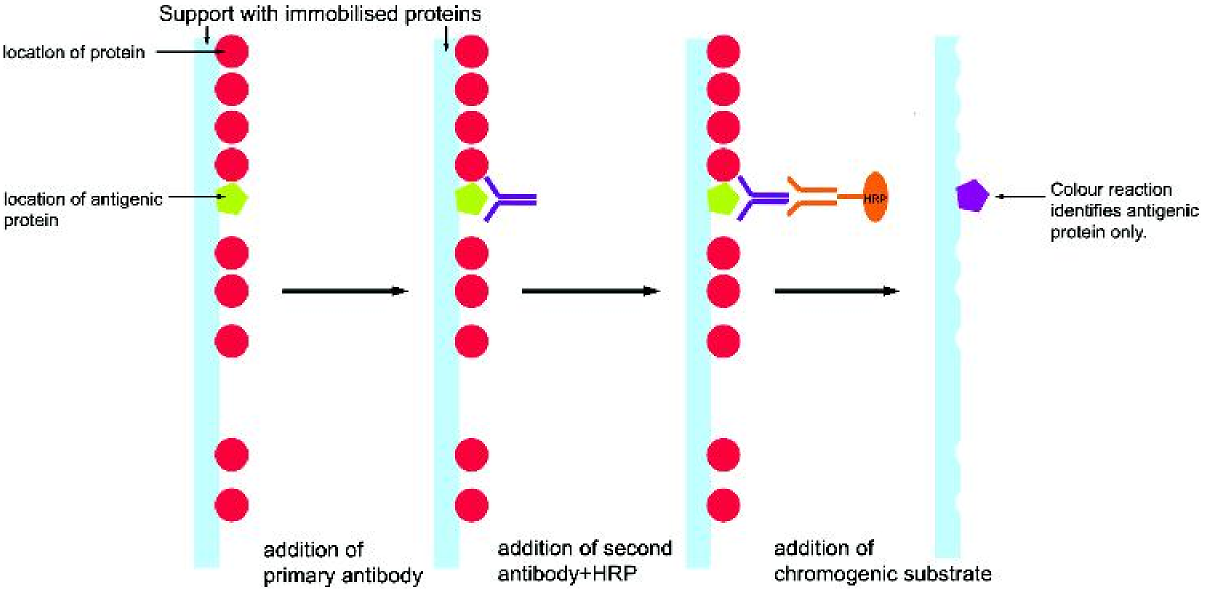
Support with immobilised proteins
location of protein
location of antigenic protein
HRP
Colour reaction identifies antigenic protein only.
addition of primary antibody
addition of second antibody+HRP
addition of chromogenic substrate

Figure 9.27 An antibody raised against the protein of interest is used to locate the polypeptide on the PVDF membrane. This membrane is essentially a replica of the polyacrylamide gel. A second antibody containing covalently bound horseradish peroxidase reacts with the first or primary antibody and in the presence of a chromogenic substrate (3,3′,5,5′-tetramethylbenzidine and hydrogen peroxide) causes a reaction that identifies the location of the polypeptide via the production of a coloured band. Care has to be taken to avoid non-specific binding, but to a first approximation the intensity of colour indicates the quantity of antigenic polypeptide as well as its location on the support. Since this matrix is a replica of the original SDS–polyacrylamide gel staining of the gel will enable identification of this band within a complex mixture of polypeptides as a result of the specific antigen–antibody reaction

identifies the protein of interest. However, by itself the reaction does not lead to any easily detectable signal on the supporting matrix. To overcome this deficiency the antibody reaction is usually carried out with a second antibody cross-reacting to the initial antibody and coupled to an enzyme. Two common enzymes linked to antibodies in this fashion are horseradish peroxidase (HRP) and alkaline phosphatase. These enzymes were chosen because the addition of substrates (benzidine derivatives and hydrogen peroxide for HRP together with 5-bromo-4-chloro-3-indoyl phosphate or 4-chloro-1-naphthol substrates with alkaline phosphatase) leads to an insoluble product and a colour reaction that is detected visually. The appearance of a colour on the nitrocellulose support allows the position of the antibody–antigen complex to be identified and to be correlated with the result of SDS–PAGE.

An alternative procedure to the reaction with antibodies is to identify proteins absorbed irreversibly to a nitrocellulose sheet using radioisotope-labelled antibodies. Iodine-125 is commonly used to label antibodies and the protein is identified from the exposure of photosensitive film. Together these two techniques allow a protein to be detected on the basis of molecular mass (SDS–PAGE) and also native conformation (Western blotting) at very low levels. These techniques are now the basis for many clinical diagnostic tests.

A second route of analysis involves spreading protein separation into a second dimension. This technique is called two-dimensional gel electrophoresis and separates proteins according to their isoelectric points in the first dimension followed by a second separation based on subunit molecular mass using SDS–PAGE.

The bioinformatics revolution currently underway in protein biochemistry is providing ever-increasing amounts of sequence data that requires analysis at the molecular level. The area of genomics has led naturally to proteomics – the wish to characterize proteins expressed by the genome. This involves analysis of the structure and function of all proteins within a single organism, including not only individual properties but also their collective properties through interactions with physiological partners. Proteomics requires a rapid way of identifying many proteins within a cell, and in this context 2D electrophoresis is advancing as a powerful method for analysis of complex protein mixtures (Figure 9.28). This technique sorts proteins according to two independent properties. The first-dimension relies on isoelectric focusing (IEF) and centres around the creation of a pH gradient under the influence of an electric field. A protein will move through this pH gradient until it reaches a point where its net charge is zero. This is the pI of that protein and the protein will not migrate further towards either electrode. The separation of protein mixtures by IEF is usually performed with gel 'strips' which are layered onto a second SDS gel to perform the 2D step. The result is a 2D array where each 'spot' should represent a single protein species in the sample. Using this approach thousands of different proteins within a cell or tissue can be separated and relevant information such as isoelectric point (pI), subunit or monomeric mass and the amount of protein present in cells can be derived.

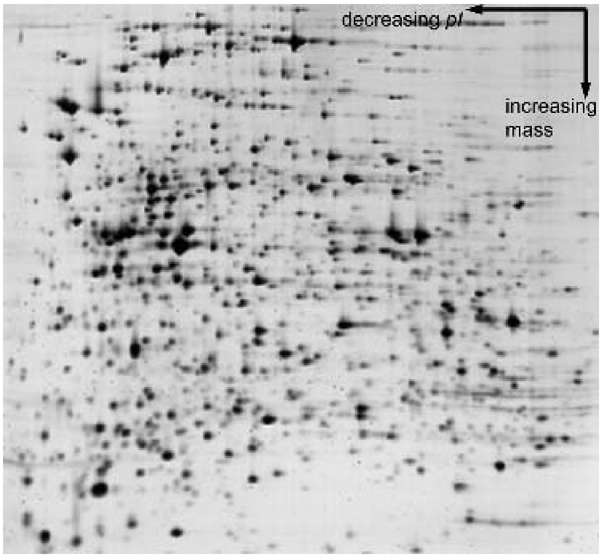

Figure 9.28 Separation of proteins using 2D electrophoresis. The proteins separate according to isoelectric point in dimension 1 and according to molecular mass in dimension 2. Often the proteins are labelled via the incorporation of radioisotopes and exposure to a sensitive photographic film allows identification of each 'spot'

Although the value of 2D electrophoresis as an analytical technique was recognized as long ago as

1975 it is only in the last 10 years that technical improvements allied to advances in computers have seen this approach reach the forefront of research. The observation of a single spot on 2D gels is sufficient to allow more detailed analysis by eluting or transferring the ‘spot’ and subjecting it to microsequencing methods based on the Edman degradation, amino acid analysis or even mass spectrometry. Computers and software are now available that allow digital evaluations of the complex 2D profile and electrophoresis results can be readily compared between different organisms. One obvious advantage of these techniques is their ability to detect post-translational modifications of proteins. Such modifications are not readily apparent from protein sequence analysis and are not easily predicted from genome analysis. 2D electrophoresis is also permitting differential cell expression of proteins to be investigated, and in some cases the identification of disease markers from abnormal profiles.

Mass spectrometry

Mass spectrometry has become a valuable tool in studying covalent structure of proteins. This has arisen because mass determination methods have become more accurate with the introduction of new techniques such as matrix assisted laser desorption time of flight (MALDI-TOF) and electrospray methods. This allows accurate mass determination for primary sequences as well as the detection of post-translational modifications, the analysis of protein purity and where appropriate the detection of single residue mutations in genetically engineered proteins.

In mass spectrometry the fundamental observed parameter is the mass to charge ratio (*m*/*z*) of gas phase ions. Although this technique has its origins in studies originally performed by J.J. Thomson at the beginning of the 20th century, over the last 20 years the technique has advanced rapidly and specifically in the area of biological analysis. The technique is important to protein characterization, as demonstrated by the award of the Nobel prize for Chemistry in 2002 to two pioneers of this technique, John Fenn and Koichi Tanaka, for the development of methods permitting analysis of structure and identification of biological macromolecules and for the development of ‘soft’ desorption–ionization methods in mass spectrometry.

All mass spectrometers consist of three basic components: an ion source, a mass analyser and a detector. Ions are produced from samples generating charged states that the mass analyser separates according to their mass/charge ratio whilst a detector produces quantifiable signals. Early mass spectrometry studies utilized electron impact or chemical ionization methods to generate ions and with the extreme conditions employed this frequently led to the fragmentation of large molecules into smaller ions (Table 9.4). Improvements in biochemical mass spectrometry arise from the development of ‘soft’ ionization methods that generate molecular ions without fragmentation (Table 9.4). In particular, two methods of soft ionization are widely used in studies of proteins and these are matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI).

An initial step in all mass spectrometry measurements is the introduction of the sample into a vacuum

Table 9.4 The various mass spectrometry methods have different applications

Ionization method	Analyte	Upper limit for mass range (kDa)
Electron impact	Volatile	∼1
Chemical ionization	Volatile	∼1
Electrospray ionization	Peptides and proteins	∼200
Fast atom bombardment	Peptides, small proteins	∼10
Matrix assisted laser desorption ionization (MALDI)	Peptides or proteins	∼500

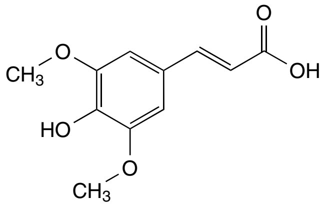

Figure 9.29 Structure of sinapinic acid used as matrix in MALDI methods of ionization. Sinapinic acid (3,5-dimethoxy-4-hydroxycinnamic acid) is the matrix of choice for protein of $M_r > 10\,000$, whilst for smaller proteins α-cyano 4-hydroxycinnamic acid is frequently preferred as a matrix

and the formation of a gas phase. Since proteins are not usually volatile it was thought that mass spectrometry would not prove useful in this context. Although early studies succeeded in making volatile derivatives by modifying polar groups the biggest advance has arisen through the use of alternative strategies. One of the most important of these is the MALDI-TOF method where the introduction of a matrix assists in the formation of gas phase protein ions (Figure 9.29). A common matrix is sinapinic acid (M_r 224) where the protein (analyte) is dissolved in the matrix at a typical molar ratio of ~1:1000. The protein is frequently present at a concentration of 1–10 μM. This mixture is dried on a probe and introduced into the mass spectrometer. The next stage involves the use of a laser beam to initiate desorption and ionization and in this area the matrix prevents degradation of the protein by absorbing energy. The matrix also assists the irradiated sample to vaporize by forming a rapidly expanding matrix plume that aids transfer of the protein ions into the mass analyser. Most commercially available mass spectrometers coupled to MALDI ion sources are 'time-of-flight' instruments (TOF-MS) in which the processes of ion formation, separation and detection occur under a very stringent high vacuum. The ionization of proteins does not cause fragmentation of the protein but generates a series of positively charged ions the most common of which occurs through the addition of a proton and is often designated as the $(M + H)^+$ ion. A second ionized species with half the previous m/z ratio denoted as $(M + 2H)^{2+}$ is also commonly formed. If Na^+ or K^+ ions are present in samples then additional molecular ions can form such as $(M + Na)^+$ and $(M + K)^+$ and will be observed with characteristic m/z ratios. The gaseous protein ions formed by laser irradiation are rapidly accelerated to the same (high) kinetic energy by the application of an electrostatic field and then expelled into a field-free region (called the flight tube) where they physically separate from each other according to their m/z ratios. A detector at the end of the flight tube records the arrival time and intensity of signals as groups of mass-resolved ions exit the flight tube. Small molecular ions will exit the flight tube first followed by proteins with progressively greater m/z ratios. Commonly, the matrix ions will be the first ions to reach the detector followed by low molecular weight impurities. To estimate the size of each molecular ion samples of known size are usually measured as part of an internal calibration. By estimating the TOF for these standard proteins the TOF and hence the mass for an unknown protein can be estimated with very high accuracy.

The mass of the intact protein is one of the most useful attributes in protein identification procedures. Although SDS–PAGE can yield a reasonably accurate mass for a protein such measurements are prone to error. Some proteins run with anomalous mobility on these gels leading to either over or under estimation of mass that can range in error from 5 to 25 percent. As a result, mass spectrometry has been used to refine estimations of molecular weight by excision of a protein from 1 or 2D gels. MALDI-TOF is the mass spectrometry technique of choice for rapid determination of molecular weights for gel 'spots' and can be performed on samples directly from the gel, from electroblotting membranes or from the protein extracted into solution. In fast-atom bombardment (FAB) a high-energy beam of neutral atoms, typically Xe or Ar, strikes a solid sample causing desorption and ionization. It is used for large biological molecules that are difficult to get into the gas phase. FAB causes little fragmentation of the protein and usually gives a large molecular ion peak, making it ideal for molecular weight determination. A third method of ion generation is electrospray ionization (ESI) mass spectrometry and involves a solution of protein molecules suspended in a volatile solvent that are sprayed at atmospheric pressure through a narrow channel (a capillary of diameter ~0.1 mm) whose outlet is maintained within an electric field operating at voltages of ~3–4 kV. This field causes the liquid to disperse into fine droplets that can pass down a potential and pressure gradient towards the mass analyser. Although the exact mechanism of droplet and ion formation remains unclear further desolvation occurs to release molecular ions into the gas-phase. Like the MALDI-TOF procedure the electrospray method is a 'soft' ionization technique capable of ionizing and transferring large proteins into the gas-phase for subsequent mass spectrometer detection. In the previous discussion the mass analyser has not been described in detail but three principal designs exist: the magnetic and/or electrostatic sector mass analyser, the time of flight (TOF) analyser and the quadrupole analyser. Each analyser has advantages and disadvantages (Table 9.5).

The principle of using mass spectrometry in protein sequencing is shown for a 185 kDa protein derived from SDS–PAGE and digested with a specific protease (Figure 9.30). One peptide species is 'selected' for collision (m/z 438) and the derived fragments of this peptide are measured to give the product mass spectrum. Some of these fragments differ in their mass by values equivalent to individual amino acids. The sequence is deduced with the exception that leucine and isoleucine having identical masses are not distinguished. This sequence information is used together with the known specificity of the protease to create a 'peptide sequence tag'. Such tag sequences can then be compared with databases of protein sequences to identify either the protein itself or related proteins.

Mass spectrometry has very rapidly become an essential tool for all well-funded protein biochemistry laboratories and the range of applications is increasing steadily as the proteomics revolution continues.

Table 9.5 Different methods of detection employed in mass spectrometry

Analyser	Features of system
Quadrupole	Unit mass resolution, fast scan, low cost
Sector (magnetic and/or electrostatic)	High resolution, exact mass
Time-of-flight (TOF)	Theoretically, no limitation for m/z maximum, high flux systems

How to purify a protein?

Armed with the knowledge described in preceding sections it should be possible to attempt the isolation of any protein with the expectation that a highly enriched pure sample will be obtained. In practice, numerous difficulties present themselves and often render it difficult to achieve this objective. Assuming a protein is soluble and found either in the cytoplasm or a subcellular compartment then perhaps the major problem likely to be encountered in isolation is the potentially low concentration of some proteins in tissues. The use of molecular biology avoids this problem and allows soluble proteins to be over-expressed in high yields in hosts such as *E. coli* (Table 9.6). If cDNA encoding the protein of interest has been cloned it is logical to start with an expression system as the source of biological material.

The first step will involve disrupting the host cells (*E. coli*) using either lysozyme or mild sonication followed by centrifugation to remove heavier components such as membranes from the soluble cytoplasmic fraction which is presumed to contain the protein of interest. At this stage with the removal of substantial amounts of protein it may be possible to assay for the presence of the protein especially if it has an enzymatic activity that can be readily measured.

Alternatively the protein may have a chromophore with distinctive absorbance or fluorescence spectra that can be quantified. However, it is worth noting that at this stage the protein is just one of many hundreds of soluble proteins found in *E. coli*. These measurements remain important in any purification because they define the initial concentration of protein. All subsequent steps are expected to increase the concentration of protein at the expense of unwanted proteins. Concurrently, most investigators would also run SDS–PAGE gels of the initial starting material to obtain a clearer picture of the number and size range of proteins present in the cell lysate (Figure 9.31). Again, successive steps in the purification would be expected to remove many, if not all, of these proteins leaving the protein of interest as a single homogeneous band on a gel.

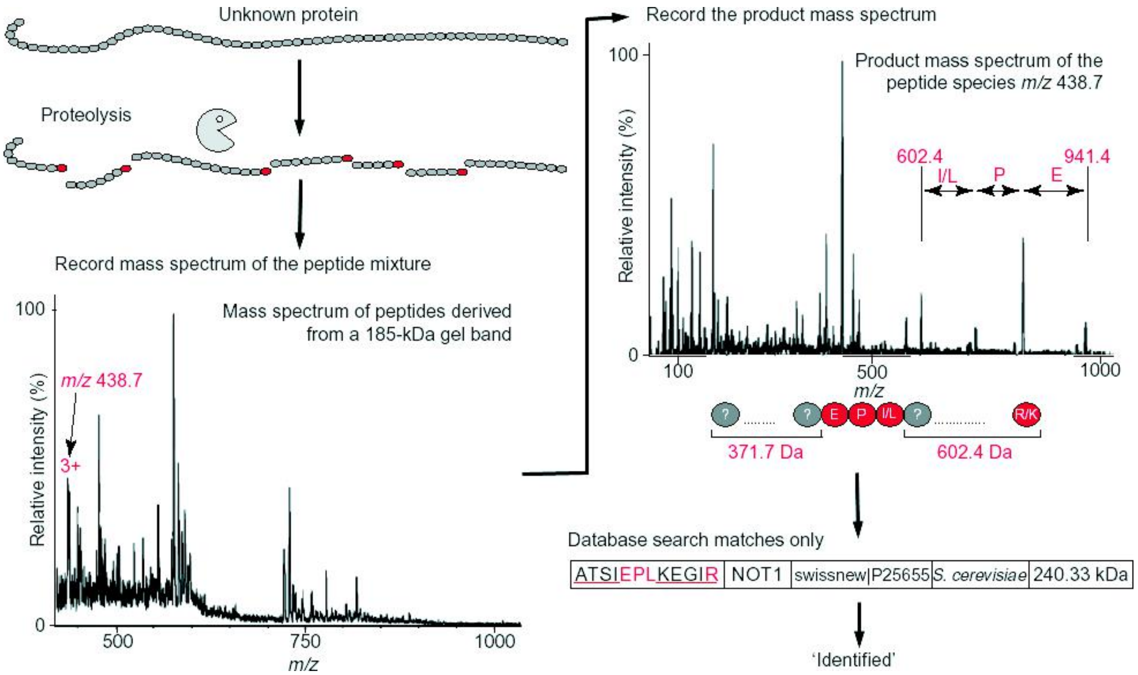
Unknown protein
Proteolysis
Record mass spectrum of the peptide mixture
Mass spectrum of peptides derived from a 185-kDa gel band
100
0
Relative intensity (%)
m/z 438.7
3+
500
750
1000
m/z
Record the product mass spectrum
Product mass spectrum of the peptide species m/z 438.7
602.4
941.4
I/L
P
E
100
500
1000
m/z
?
E
P
I/L
R/K
371.7 Da
602.4 Da
Database search matches only
ATSIEPLKEGIR
NOT1
swissnew|P25655
S. cerevisiae
240.33 kDa
'Identified'

Figure 9.30 Mass spectrometric identification of an unknown protein can be obtained from determining the sequence of proteolytic fragments and comparing the sequence with available databases (reproduced and adapted with permission from Rappsilber, J. and Mann, M. *Trends Biochem. Sci.* 2002, **27**, 74–78. Elsevier)

Table 9.6 Example of the purification of an over-expressed enzyme from *E. coli*

Purification of the enzyme NADH-cytochrome b_5 reductase					
Purification step	Activity (units)	Protein (mg)	Specific activity (units/mg)	Recovery (% total)	Enrichment (over initial level)
Cell supernatant	26 247	320	82	100	1
Ammonium sulfate precipitation (50 %)	19 340	93	208	74	3
Ammonium sulfate precipitation (90 %)	18 627	72	259	71	3
Affinity	12 762	10	1276	49	16
Size exclusion	11 211	8	1400	43	17

The enzyme's activity is measured at all stages along with the total protein concentration. The total protein falls during purification but the specific activity increases significantly. Activity involved measurement of the rate of oxidation of NADH (Data adapted from Barber, M. and Quinn, G.B. *Prot. Expr. Purif.* 1996, 8, 41–47). The enzyme binds NAD and can therefore be purified by affinity chromatography using agarose containing covalently bound ADP

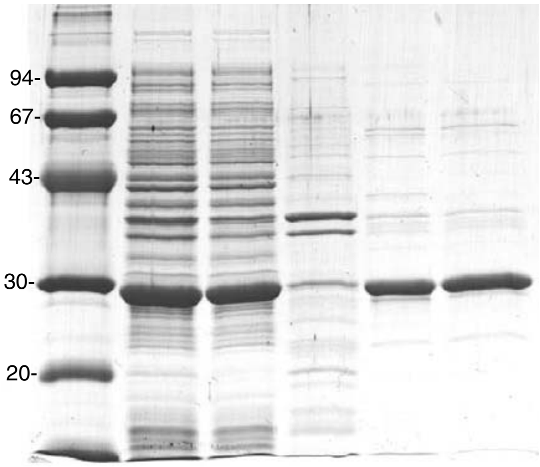

Figure 9.31 Progressive purification of a protein monitored by SDS–PAGE. Analysis was carried out on 12 % acrylamide gels with molecular mass markers shown in the first lane. The lanes contain from left to right total extract; polypeptide composition of supernatant after centrifugation at 14 000 g; polypeptide composition of pellet fraction after centrifugation at 14 000 g; material obtained after elution from an anion-exchange column and finally material obtained after gel filtration. The final protein is nearly pure, although a few impurities can still be seen (reproduced with permission from Karwaski, M.F. *et al*. *Protein Expression and Purification*. 2002, **25**, 237–240. Academic Press)

It is rare to achieve purification in one step using a single technique. More frequently the strategy involves 'capture' of the material from a crude lysate or mixture of proteins followed by one or more steps designed to enrich the protein of interest followed by a final 'polishing' stage that removes the last traces of contaminants.

It is not unreasonable to ask which method should be used first? In the absence of any preliminary information there are no set rules and the purification must be attempted in an empirical fashion by assaying for protein presence via enzyme activity, biological function or other biophysical properties at each step of the purification procedure. However, it is rare that purifications are attempted without some ancillary information obtained from other sources. Such information should generally be used to guide the procedure in the direction of a possible purification strategy. So, for example, if a sequentially homologous protein has been previously purified using anion exchange chromatography and size exclusion to homogeneity it is almost certain to prove successful again.

Membrane proteins present major difficulties in isolation. Removing the protein from the lipid bilayer will invariably lead to a loss of structure *and* function whilst attempting many of the isolation procedures described above with hydrophobic proteins is more difficult as a result of the conflicting solvent requirements of such proteins. However, it is by no means impossible to overcome these difficulties and many hydrophobic proteins have been purified to homogeneity by combining procedures described in the above sections with the judicious use of detergents. In many instances detergents allow hydrophobic proteins to remain in solution without aggregating and to be amenable to chromatographic procedures that are applicable to soluble proteins.

Summary

Purification requires the isolation of a protein from a complex mixture frequently derived from cell disruption. The aim of a purification strategy is the isolation of a single protein retaining most, if not all, biological activity and the absence of contaminating proteins.

Purification methods have been helped enormously by the advances in cloning and recombinant DNA technology that allow protein over-expression in foreign host cells. This allows proteins that were difficult to isolate to be studied where previously this had been impossible.

Methods of purification rely on the biophysical properties of proteins with the properties of mass, charge, hydrophobicity, and hydrodynamic radius being frequently used as the basis of separation techniques. Chromatographic methods form the most common group of preparative techniques used in protein purification.

In all cases chromatography involves the use of a mobile, usually aqueous, phase that interact with an inert support (resin) containing functional groups that enhance interactions with some proteins. In ion exchange chromatography the supporting matrix contains negatively or positively charged groups. Similar methods allow protein separation on the basis of hydrodynamic radius (size exclusion), ligand binding (affinity), and non-polar interactions (HIC and RPC).

Alongside preparative techniques are analytical methods that establish the purity and mass of the product. SDS–PAGE involves the separation of polypeptides under the influence of an electric field solely on the basis of mass. This technique has proved of widespread value in ascertaining subunit molecular mass as well as overall protein purity.

An extension of the basic SDS–PAGE technique is Western blotting. This method allows the identification of an antigenic polypeptide within a mixture of size-separated components by its reaction with a specific antibody.

2D electrophoresis allows the separation of proteins according to mass and overall charge. Consequently, through the use of these methods for individual cell types or organisms, it is proving possible to identify large numbers of different proteins within proteomes of single-celled organisms or individual cells.

Of all analytical methods mass spectrometry has expanded in importance as a result of technical advances permitting accurate identification of the mass/charge ratio of molecular ions. The most popular methods are MALDI-TOF and electrospray spectrometry. Using modern instrumentation the mass of proteins can be determined in favourable instances to within 1 a.m.u.

In combination with 1 and 2D gel methods mass spectrometry is proving immensely valuable in characterization of proteomes. The expansion of proteomics in the post-genomic revolution has placed greater importance on preparative and analytical techniques. When the methods described here are combined with the techniques such as NMR spectroscopy, X-ray crystallography or cryo-EM it is possible to go from gene identification to protein structure within a comparatively short space of time.

Problems

1. Calculate the relative centrifugal force for a rotor spun at 5000 r.p.m at distances of 9, 6 and 3 cm from the centre of rotation. Calculate the forces if the rotor is now rotating at 20 000 r.p.m.

2. Subunit A has a mass of 10 000, a p*I* of 4.7; subunit B has a mass of 12 000, a p*I* of 10.2, subunit C has a mass of 30 000, a p*I* of 7.5 and binds NAD; subunit D has a mass of 30 000, a p*I* of 4.5 and subunit E has a mass of 100 000, with a p*I* of 5.0. Subunits A, B and E are monomeric, subunits C and D are heptameric. A preliminary fractionation of a cell lysate reveals that all proteins are present in approximately equivalent quantities. Outline an efficient possible route towards purifying each protein. How would you assess effective homogeneity?

3. Why is it helpful to know amino acid composition prior to Edman sequencing?

4. A colleague has failed to sequence using Edman degradation an oligopeptide believed to be part of a larger polypeptide chain. Provide reasonable explanations why this might have happened. Suggest alternative strategies that could work. After further investigation a partial amino acid sequence Cys-Trp-Ala-Trp-Ala-Cys-$CONH_2$ is obtained. Does this highlight additional problems. Describe the methods you might employ to (i) identify this protein and (ii) isolate the complete protein?

5. Discuss the underlying reasons for the use of ammonium sulfate in protein purification.

6. It has been reported that some proteins retain residual structure in the presence of SDS. What are the implications of this observation for SDS–PAGE. What other properties of proteins might cause problems when using this technique?

7. In gel filtration the addition of polyethylene glycol to the running buffer has been observed to make proteins elute at a later stage as if they were of smaller size. Explain this observation. Suggest other co-solvents that would have a similar effect.

8. How would you confirm that a protein is normally found as a multimer?

9. Four mutated forms of protein X have been expressed and purified. Mutant 1 is believed to contain the substitution Gly>Trp, Mutant 2 Glu>Asp, Mutant 3 Ala>Lys, and Mutant 4 Leu>Ile. How would you confirm the presence of these substitutions in the product protein?

10. Describe the equipment a well-equipped laboratory would require to isolate and purify to homogeneity a protein from bacteria.