B1 AMINO ACIDS

Key Notes

Amino acids

All proteins are made up from the same set of 20 standard amino acids. A typical amino acid has a primary amino group, a carboxyl group, a hydrogen atom and a side-chain (R group) attached to a central α-carbon atom (C_α). Proline is the exception to the rule in that it has a secondary amino group.

Enantiomers

All of the 20 standard amino acids, except for glycine, have four different groups arranged tetrahedrally around the C_α atom and thus can exist in either the D or L configuration. These two enantiomers are nonsuperimposable mirror images that can be distinguished on the basis of their different rotation of plane-polarized light. Only the L isomer is found in proteins.

The 20 standard amino acids

The standard set of 20 amino acids have different side-chains or R groups and display different physicochemical properties (polarity, acidity, basicity, aromaticity, bulkiness, conformational inflexibility, ability to form hydrogen bonds, ability to cross-link and chemical reactivity). Glycine (Gly, G) has a hydrogen atom as its R group. Alanine (Ala, A), valine (Val, V), leucine (Leu, L), isoleucine (Ile, I) and methionine (Met, M) have aliphatic side-chains of differing structures that are hydrophobic and chemically inert. The aromatic side-chains of phenylalanine (Phe, F), tyrosine (Tyr, Y) and tryptophan (Trp, W) are also hydrophobic in nature. The conformationally rigid proline (Pro, P) has its aliphatic side-chain bonded back on to the amino group and thus is really an imino acid. The hydrophobic, sulfur-containing side-chain of cysteine (Cys, C) is highly reactive and can form a disulfide bond with another cysteine residue. The basic amino acids arginine (Arg, R) and lysine (Lys, K) have positively charged side-chains, whilst the side-chain of histidine (His, H) can be either positively charged or uncharged at neutral pH. The side-chains of the acidic amino acids aspartic acid (Asp, D) and glutamic acid (Glu, E) are negatively charged at neutral pH. The amide side-chains of asparagine (Asn, N) and glutamine (Gln, Q), and the hydroxyl side-chains of serine (Ser, S) and threonine (Thr, T) are uncharged and polar, and can form hydrogen bonds.

Related topics Acids and bases (B2) Protein structure (B3)

Amino acids

Amino acids are the building blocks of **proteins** (see Topic B3). Proteins of all species, from bacteria to humans, are made up from the same set of **20 standard amino acids**. Nineteen of these are α-amino acids with a **primary amino group** (–NH_3^+) and a **carboxylic acid** (carboxyl; –COOH) group attached to a central carbon atom, which is called the **α-carbon atom** (C_α) because it is adjacent to the carboxyl group (*Fig. 1a*). Also attached to the C_α atom is a hydrogen atom and a

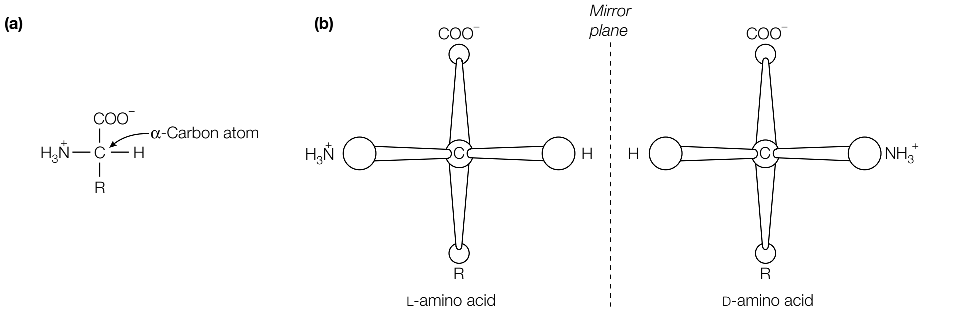

Fig. 1. (a) Basic structure of an amino acid showing the four different groups around the central α-carbon atom, (b) the two enantiomers of an amino acid.

variable side-chain or 'R' group. The one exception to this general structure is proline, which has a secondary amino group and is really an **α-imino acid**. The names of the amino acids are often abbreviated, either to three letters or to a single letter. Thus, for example, proline is abbreviated to Pro or P (see *Fig. 2*).

Enantiomers

All of the amino acids, except for glycine (Gly or G; see *Fig. 2*), have four different groups arranged **tetrahedrally** around the central C_α atom which is thus known as an **asymmetric center** or **chiral center** and has the property of **chirality** (Greek; *cheir*, hand) (*Fig. 1b*). The two **nonsuperimposable, mirror images** are termed **enantiomers**. Enantiomers are physically and chemically indistinguishable by most techniques, but can be distinguished on the basis of their different optical rotation of plane-polarized light. Molecules are classified as dextrorotatory (D; Greek '*dextro*' = right) or levorotatory (L; Greek '*levo*' = left) depending on whether they rotate the plane of plane-polarized light clockwise or anticlockwise. D- and L-amino acids can also be distinguished by enzymes which usually only recognize one or other enantiomer. Only the **L-amino acids** are found in proteins. **D-Amino acids** rarely occur in nature, but are found in bacterial cell walls (see Topic A1) and certain antibiotics.

The 20 standard amino acids

The standard 20 amino acids differ only in the structure of the **side-chain** or 'R' group (*Figs 2* and *3*). They can be subdivided into smaller groupings on the basis of similarities in the properties of their side-chains. They display different **physicochemical properties** depending on the nature of their side-chain. Some are acidic, others are basic. Some have small side-chains, others large, bulky side-chains. Some have aromatic side-chains, others are polar. Some confer conformational inflexibility, others can participate either in hydrogen bonding or covalent bonding. Some are chemically reactive.

Hydrophobic, aliphatic amino acids

Glycine (Gly or G) (*Fig. 2a*), the smallest amino acid with the simplest structure, has a hydrogen atom in the side-chain position, and thus does not exist as a pair of stereoisomers since there are two identical groups (hydrogen atoms) attached to the C_α atom. The aliphatic side-chains of **alanine** (Ala or A), **valine** (Val or V), **leucine** (Leu or L), **isoleucine** (Ile or I) and **methionine** (Met or M) (*Fig. 2a*) are chemically unreactive, but hydrophobic in nature. **Proline** (Pro or P) (*Fig. 2a*) is

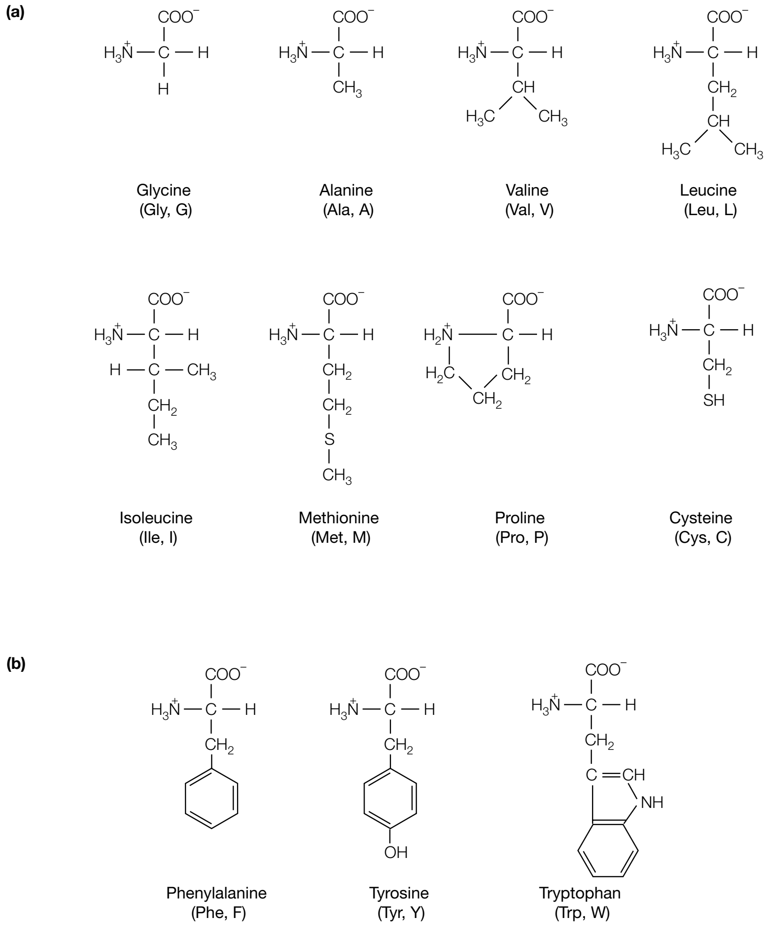

Fig. 2. The standard amino acids. (a) Hydrophobic, aliphatic R groups, (b) hydrophobic, aromatic R groups. The molecular weights of the amino acids are given in Topic B2, Table 1.

also hydrophobic but, with its aliphatic side-chain bonded back on to the amino group, it is conformationally rigid. The sulfur-containing side-chain of **cysteine** (Cys or C) (*Fig. 2a*) is also hydrophobic and is highly reactive, capable of reacting with another cysteine to form a disulfide bond (see Topic B3).

Hydrophobic, aromatic amino acids

Phenylalanine (Phe or F), **tyrosine** (Tyr or Y) and **tryptophan** (Trp or W) (*Fig. 2b*) are hydrophobic by virtue of their aromatic rings.

Polar, charged amino acids

The remaining amino acids all have polar, hydrophilic side-chains, some of which are charged at neutral pH. The amino groups on the side-chains of the basic amino acids **arginine** (Arg or R) and **lysine** (Lys or K) (*Fig. 3a*) are protonated and thus positively charged at neutral pH. The side-chain of **histidine** (His or H) (*Fig. 3a*) can be either positively charged or uncharged at neutral pH. In contrast, at neutral pH the carboxyl groups on the side-chains of the acidic amino acids **aspartic acid** (aspartate; Asp or D) and **glutamic acid** (glutamate; Glu or E) (*Fig. 3a*) are de-protonated and possess a negative charge.

Polar, uncharged amino acids

The side-chains of **asparagine** (Asn or N) and **glutamine** (Gln or Q) (*Fig. 3b*), the amide derivatives of Asp and Glu, respectively, are uncharged but can participate in hydrogen bonding. **Serine** (Ser or S) and **threonine** (Thr or T) (*Fig. 3b*) are polar amino acids due to the reactive hydroxyl group in the side-chain, and can also participate in hydrogen bonding (as can the hydroxyl group of the aromatic amino acid Tyr).

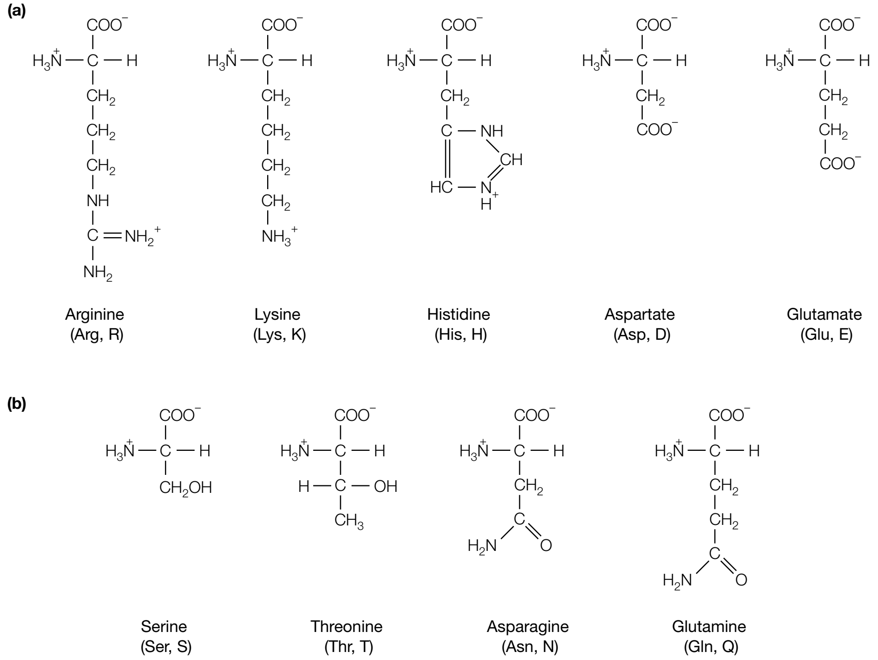

Fig. 3. The standard amino acids. (a) Polar, charged R groups, (b) polar, uncharged R groups. The molecular weights of the amino acids are given in Topic B2, Table 1.

B2 ACIDS AND BASES

Key Notes

Acids, bases and pH

pH is a measure of the concentration of H^+ in a solution. An acid is a proton donor, a base is a proton acceptor. Ionization of an acid yields its conjugate base, and the two are termed a conjugate acid–base pair, for example acetic acid (CH_3COOH) and acetate (CH_3COO^-). The p*K* of an acid is the pH at which it is half dissociated. The Henderson–Hasselbalch equation expresses the relationship between pH, p*K* and the ratio of acid to base, and can be used to calculate these values.

Buffers

An acid-base conjugate pair can act as a buffer, resisting changes in pH. From a titration curve of an acid the inflexion point indicates the p*K* value. The buffering capacity of the acid–base pair is the p*K* ± 1 pH unit. In biological fluids the phosphate and carbonate ions act as buffers. Amino acids, proteins, nucleic acids and lipids also have some buffering capacity. In the laboratory other compounds, such as TRIS, are used to buffer solutions at the appropriate pH.

Ionization of amino acids

The α-amino and α-carboxyl groups on amino acids act as acid–base groups, donating or accepting a proton as the pH is altered. At low pH, both groups are fully protonated, but as the pH is increased first the carboxyl group and then the amino group loses a hydrogen ion. For the standard 20 amino acids, the p*K* is in the range 1.8–2.9 for the α-carboxyl group and 8.8–10.8 for the α-amino group. Those amino acids with an ionizable side-chain have an additional acid–base group with a distinctive p*K*.

Related topics Amino acids (B1)

Acids, bases and pH

The pH of a solution is a measure of its concentration of protons (H^+), and pH is defined as:

$$\mathrm{pH} = \log_{10}\frac{1}{\mathrm{H}^+} = -\log_{10}[\mathrm{H}^+]$$

in which the square brackets denote a **molar concentration**.

An **acid** can be defined as a proton donor and a **base** as a proton acceptor:

$$\text{Acid} \rightleftharpoons \mathrm{H}^+ + \text{base}$$

For example;

$$\underset{\text{Acetic acid}}{\mathrm{CH_3COOH}} \rightleftharpoons \mathrm{H}^+ + \underset{\text{Acetate}}{\mathrm{CH_3COO^-}}$$

$$\underset{\text{Ammonium ion}}{\mathrm{NH_4^+}} \rightleftharpoons \mathrm{H}^+ + \underset{\text{Ammonia}}{\mathrm{NH_3}}$$

The species formed by the **ionization** of an acid is its conjugate base. Conversely, protonation of a base yields its conjugate acid. So, for example, acetic acid and acetate are a **conjugate acid–base pair**.

The ionization of a weak acid is given by:

$$\mathrm{HA} \rightleftharpoons \mathrm{H^+} + \mathrm{A^-}$$

The apparent **equilibrium constant** (*K*) for this ionization is defined as:

$$K = \frac{[\mathrm{H^+}][\mathrm{A^-}]}{[\mathrm{HA}]} \qquad \text{(Equation 1)}$$

The **p*K*** of an acid is defined as:

$$\mathrm{p}K = -\log K = \log \frac{1}{K}$$

The p*K* of an acid is the pH at which it is half dissociated, i.e. when [A⁻] = [HA].

The **Henderson–Hasselbalch equation** expresses the relationship between pH and the ratio of acid to base. It is derived as follows. Rearrangement of Equation 1 gives:

$$\frac{1}{[\mathrm{H^+}]} = \frac{1}{K} \times \frac{[\mathrm{A^-}]}{[\mathrm{HA}]}$$

Taking the logarithm of both sides of this equation gives:

$$\log \frac{1}{[\mathrm{H^+}]} = \log \frac{1}{K} + \log \frac{[\mathrm{A^-}]}{[\mathrm{HA}]}$$

Substituting pH for log $1/[\mathrm{H^+}]$ and p*K* for log $1/K$ gives:

$$\mathrm{pH} = \mathrm{p}K + \log \frac{[\mathrm{A^-}]}{[\mathrm{HA}]}$$

which is the Henderson–Hasselbalch equation. This equation indicates that the p*K* of an acid is numerically equal to the pH of the solution when the molar concentration of the acid is equal to that of its conjugate base. The pH of a solution can be calculated from the Henderson–Hasselbalch equation if the molar concentrations of $\mathrm{A^-}$ and HA, and the p*K* of HA are known. Similarly, the p*K* of an acid can be calculated if the molar concentrations of $\mathrm{A^-}$ and HA, and the pH of the solution are known.

Buffers

An acid–base conjugate pair, such as acetic acid and acetate, is able to resist changes in the pH of a solution. That is, it can act as a **buffer**. On addition of hydroxide ($\mathrm{OH^-}$) to a solution of acetic acid the following happens:

$$\mathrm{CH_3COOH} + \mathrm{OH^-} \rightleftharpoons \mathrm{CH_3COO^-} + \mathrm{H_2O}$$

A plot of the dependence of the pH of this solution on the amount of $\mathrm{OH^-}$ added is called a titration curve (*Fig. 1*). There is an inflection point in the curve at pH 4.8 which is the p*K* of acetic acid. In the vicinity of this pH, a relatively large amount of $\mathrm{OH^-}$ (or $\mathrm{H^+}$) produces little change in pH as the added $\mathrm{OH^-}$ (or $\mathrm{H^+}$) reacts with $\mathrm{CH_3COOH}$ (or $\mathrm{CH_3COO^-}$), respectively. Weak acids are most effective in buffering against changes in pH within 1 pH unit of the p*K* (see *Fig. 1*), often referred to as p*K* ± 1, the **buffering capacity**.

Biological fluids, including the cytosol and extracellular fluids such as blood, are buffered. For example, in healthy individuals the pH of the blood is carefully controlled at pH 7.4. The major buffering components in most biological fluids are the phosphate ion ($\mathrm{H_2PO_4^-}$, p*K* 6.82) and the carbonate ion ($\mathrm{HCO_3^-}$, p*K* 6.35) because they have p*K* values in this range. However, many biological molecules,

Outer membrane

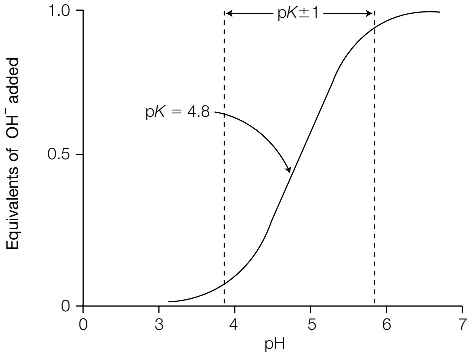

Fig. 1. Titration curve of acetic acid.

including amino acids, proteins, nucleic acids and lipids, have multiple acid–base groups that are effective at buffering in the physiological pH range (pH 6–8).

When working with enzymes, proteins and other biological molecules it is often crucial to buffer the pH of the solution in order to avoid **denaturation** (loss of activity) of the component of interest (see Topic C3). Numerous buffers are used in laboratories for this purpose. One of the commonest is tris(hydroxymethyl)aminomethane or **TRIS** which has a p*K* of 8.08.

Ionization of amino acids

The 20 standard amino acids have two **acid–base groups**: the α-amino and α-carboxyl groups attached to the C_α atom. Those amino acids with an **ionizable side-chain** (Asp, Glu, Arg, Lys, His, Cys, Tyr) have an additional acid–base group. The **titration curve** of Gly is shown in *Fig. 2a*. At low pH (i.e. high hydrogen ion concentration) both the amino group and the carboxyl group are fully protonated so that the amino acid is in the cationic form $H_3N^+CH_2COOH$ (*Fig. 2b*). As the amino acid in solution is titrated with increasing amounts of a strong base (e.g. NaOH), it loses two protons, first from the carboxyl group which has the lower **p*K*** value (p*K* = 2.3) and then from the amino group which has the higher p*K* value (pK = 9.6). The pH at which Gly has no net charge is termed its **isoelectric point, pI**. The α-carboxyl groups of all the 20 standard amino acids have p*K* values in the range 1.8–2.9, whilst their α-amino groups have p*K* values in the range 8.8–10.8 (*Table 1*). The side-chains of the acidic amino acids Asp and Glu have p*K* values of 3.9 and 4.1, respectively, whereas those of the basic amino acids Arg and Lys, have p*K* values of 12.5 and 10.8, respectively. Only the side-chain of His, with a p*K* value of 6.0, is ionized within the physiological pH range (pH 6–8). It should be borne in mind that when the amino acids are linked together in proteins, only the side-chain groups and the terminal α-amino and α-carboxyl groups are free to ionize.

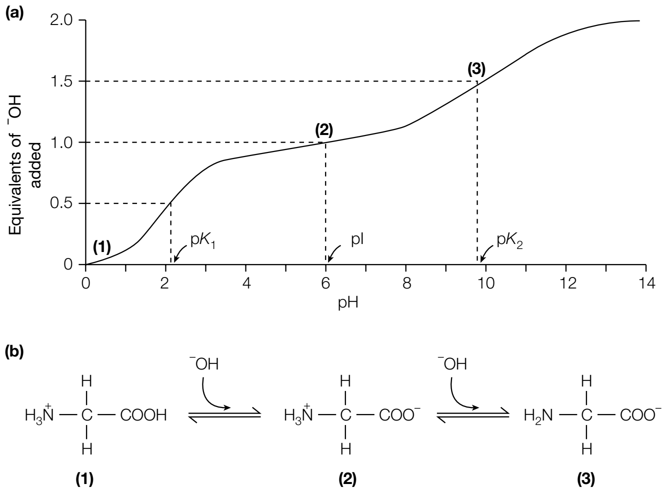

Fig. 2. Ionization of glycine. (a) Titration curve of glycine, (b) dissociation of glycine. Numbers in bold in parentheses in (a) correspond to the structures in (b).

Table 1. p*K* *values and molecular weights of the 20 standard amino acids*

Amino acid	Mol. Wt	p*K* α-COOH	p*K* α-NH_3^+	p*K* side-chain
Alanine	89.1	2.35	9.87	
Arginine	174.2	1.82	8.99	12.48
Asparagine	132.1	2.14	8.72	
Aspartic acid	133.1	1.99	9.90	3.90
Cysteine	121.2	1.92	10.70	8.37
Glutamic acid	147.1	2.10	9.47	4.07
Glutamine	146.2	2.17	9.13	
Glycine	75.1	2.35	9.78	
Histidine	155.2	1.80	9.33	6.04
Isoleucine	131.2	2.32	9.76	
Leucine	131.2	2.33	9.74	
Lysine	146.2	2.16	9.06	10.54
Methionine	149.2	2.13	9.28	
Phenylalanine	165.2	2.20	9.31	
Proline	115.1	1.95	10.64	
Serine	105.1	2.19	9.21	
Threonine	119.1	2.09	9.10	
Tryptophan	204.2	2.46	9.41	
Tyrosine	181.2	2.20	9.21	10.46
Valine	117.1	2.29	9.74	

B3 Protein structure

Key Notes

Peptide bond

A protein is a linear sequence of amino acids linked together by peptide bonds. The peptide bond is a covalent bond between the α-amino group of one amino acid and the α-carboxyl group of another. The peptide bond has partial double bond character and is nearly always in the *trans* configuration. The backbone conformation of a polypeptide is specified by the rotation angles about the C_α–N bond (*phi*, ϕ) and C_α–C bond (*psi*, ψ) of each of its amino acid residues. The sterically allowed values of ϕ and ψ are visualized in a Ramachandran plot. When two amino acids are joined by a peptide bond they form a dipeptide. Addition of further amino acids results in long chains called oligopeptides and polypeptides.

Primary structure

The linear sequence of amino acids joined together by peptide bonds is termed the primary structure of the protein. The position of covalent disulfide bonds between cysteine residues is also included in the primary structure.

Secondary structure

Secondary structure in a protein refers to the regular folding of regions of the polypeptide chain. The two most common types of secondary structure are the α-helix and the β-pleated sheet. The α-helix is a cylindrical, rod-like helical arrangement of the amino acids in the polypeptide chain which is maintained by hydrogen bonds parallel to the helix axis. In a β-pleated sheet, hydrogen bonds form between adjacent sections of polypeptides that are either running in the same direction (parallel β-pleated sheet) or in the opposite direction (antiparallel β-pleated sheet). β-Turns reverse the direction of the polypeptide chain and are often found connecting the ends of antiparallel β-pleated sheets.

Tertiary structure

Tertiary structure in a protein refers to the three-dimensional arrangement of all the amino acids in the polypeptide chain. This biologically active, native conformation is maintained by multiple noncovalent bonds.

Quaternary structure

If a protein is made up of more than one polypeptide chain it is said to have quaternary structure. This refers to the spatial arrangement of the polypeptide subunits and the nature of the interactions between them.

Protein stability

In addition to the peptide bonds between individual amino acid residues, the three-dimensional structure of a protein is maintained by a combination of noncovalent interactions (electrostatic forces, van der Waals forces, hydrogen bonds, hydrophobic forces) and covalent interactions (disulfide bonds).

Protein structure determination

The three-dimensional structure of a protein can be determined using complex physical techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryoelectron microscopy.

Protein folding

Proteins spontaneously fold into their native conformation, with the primary structure of the protein dictating its three-dimensional structure. Protein folding is driven primarily by hydrophobic forces and proceeds through an ordered set of pathways. Accessory proteins, including protein disulfide isomerases, peptidyl prolyl *cis–trans* isomerases, and molecular chaperones, assist proteins to fold correctly in the cell.

Related topics

Eukaryote cell structure (A2)
Amino acids (B1)
Myoglobin and hemoglobin (B4)
Collagen (B5)
The genetic code (H1)

Peptide bond

Proteins are linear sequences of amino acids linked together by peptide bonds. The peptide bond is a chemical, covalent bond formed between the α-amino group of one amino acid and the α-carboxyl group of another (*Fig. 1a*) (see Topic B1). Once two amino acids are joined together via a peptide bond to form a dipeptide, there is still a free amino group at one end and a free carboxyl group at the other, each of which can in turn be linked to further amino acids. Thus, long, unbranched chains of amino acids can be linked together by peptide bonds to form oligopeptides (up to 25 amino acid residues) and polypeptides (> 25 amino acid residues). Note that the polypeptide still has a free α-amino group and a free α-carboxyl group. Convention has it that peptide chains are written down with the free α-amino group on the left, the free α-carboxyl group on the right and a hyphen between the amino acids to indicate the peptide bonds. Thus, the tripeptide $^{+}H_3N$-serine–leucine–phenylalanine-COO^{-} would be written simply as Ser-Leu-Phe or S-L-F.

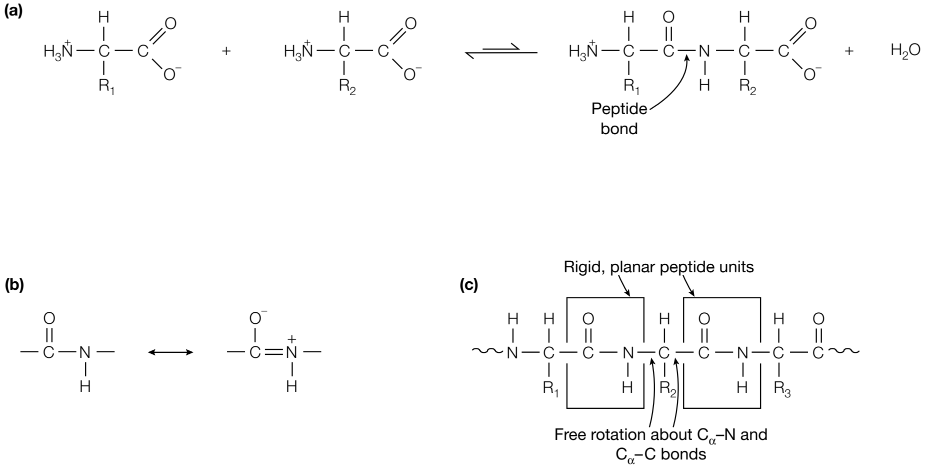
(a)
Peptide bond
H_2O
(b)
(c)
Rigid, planar peptide units
Free rotation about C_α–N and C_α–C bonds

Fig. 1. (a) Formation of a peptide bond, (b) resonance structures of the peptide bond, (c) peptide units within a polypeptide.

The peptide bond between the carbon and nitrogen exhibits **partial double-bond character** due to the closeness of the carbonyl carbon–oxygen double-bond allowing the **resonance structures** in *Fig. 1b* to exist. Because of this, the C–N bond length is also shorter than normal C–N single bonds. The **peptide unit** which is made up of the CO–NH atoms is thus relatively rigid and planar, although free rotation can take place about the C_α–N and C_α–C bonds (the bonds either side of the peptide bond), permitting adjacent peptide units to be at different angles (*Fig. 1c*). The hydrogen of the amino group is nearly always on the opposite side (*trans*) of the double bond to the oxygen of the carbonyl group, rather than on the same side (*cis*).

The backbone of a protein is a linked sequence of rigid planar peptide groups. The backbone conformation of a polypeptide is specified by the **rotation angles** or **torsion angles** about the C_α–N bond (*phi*, ϕ) and C_α–C bond (*psi*, ψ) of each of its amino acid residues. When the polypeptide chain is in its planar, fully extended (all-*trans*) conformation the ϕ and ψ angles are both defined as 180°, and increase for a clockwise rotation when viewed from C_α (*Fig. 2*). The **conformational range** of the torsion angles, ϕ and ψ, in a polypeptide backbone are restricted by steric hindrance. The sterically allowed values of ϕ and ψ can be determined by calculating the distances between the atoms of a tripeptide at all values of ϕ and ψ for the central peptide unit. These values are visualized in a steric contour diagram, otherwise known as a conformation map or **Ramachandran plot** (*Fig. 3*). From *Fig. 3* it can be seen that most areas of the Ramachandran plot (most combinations of ϕ and ψ) are conformationally inaccessible to a polypeptide chain. Only three small regions of the conformation map are physically accessible to a polypeptide chain, and within these regions are the ϕ–ψ values that produce the right-handed α-helix, the parallel and antiparallel β-pleated sheets and the collagen helix (see below and Topic B5).

The polypeptide chain folds up to form a specific shape (**conformation**) in the protein. This conformation is the **three-dimensional arrangement** of atoms in the structure and is determined by the amino acid sequence. There are four levels of structure in proteins: **primary**, **secondary**, **tertiary** and, sometimes but not always, **quaternary**.

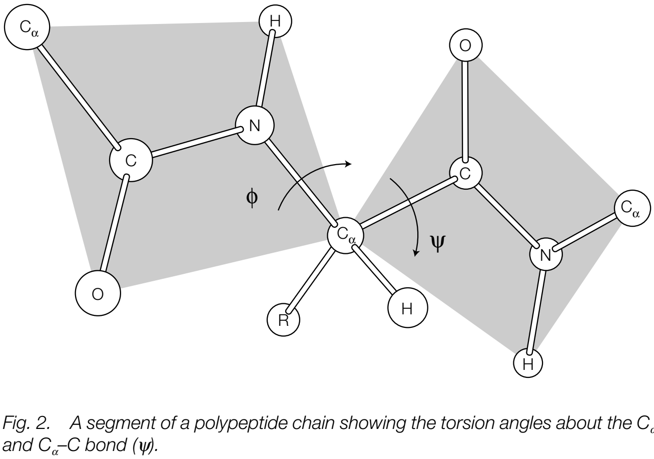

Fig. 2. A segment of a polypeptide chain showing the torsion angles about the C_α–N bond (ϕ) and C_α–C bond (ψ).

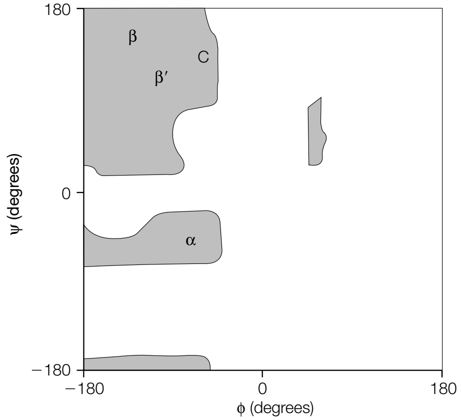

Fig. 3. Ramachandran plot showing the allowed angles for poly-L-alanine (grey regions). α, ϕ–ψ values that produce the right-handed α-helix; β, the antiparallel β-pleated sheet; β', the parallel β-pleated sheet; C, the collagen helix.

Primary structure

The primary level of structure in a protein is the **linear sequence of amino acids** as joined together by peptide bonds. This sequence is determined by the sequence of nucleotide bases in the gene encoding the protein (see Topic H1). Also included under primary structure is the location of any other **covalent bonds**. These are primarily **disulfide bonds** between cysteine residues that are adjacent in space but not in the linear amino acid sequence. These covalent cross-links between separate polypeptide chains or between different parts of the same chain are formed by the oxidation of the SH groups on cysteine residues that are juxtaposed in space (*Fig. 4*). The resulting disulfide is called a **cystine** residue. Disulfide bonds are often present in extracellular proteins, but are rarely found in intracellular proteins. Some proteins, such as collagen, have covalent cross-links formed between the side-chains of Lys residues (see Topic B5).

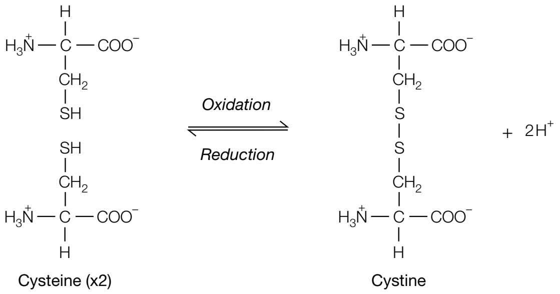

Fig. 4. Formation of a disulfide bond between two cysteine residues, generating a cystine residue.

Secondary structure

The secondary level of structure in a protein is the regular folding of regions of the polypeptide chain. The two most common types of protein fold are the α-**helix** and the **β-pleated sheet**. In the rod-like α-helix, the amino acids arrange themselves in a regular helical conformation (*Fig. 5a*). The carbonyl oxygen of each peptide bond is **hydrogen bonded** to the hydrogen on the amino group of the fourth amino acid away (*Fig. 5b*), with the hydrogen bonds running nearly parallel to the axis of the helix. In an α-helix there are 3.6 amino acids per turn of the helix covering a distance of 0.54 nm, and each amino acid residue represents an advance of 0.15 nm along the axis of the helix (*Fig. 5a*). The side-chains of the amino acids are all positioned along the outside of the cylindrical helix (*Fig. 5c*). Certain amino acids are more often found in α-helices than others. In particular, Pro is rarely found in α-helical regions as it cannot form the correct pattern of hydrogen bonds due to the lack of a hydrogen atom on its nitrogen atom. For this reason, Pro is often found at the end of an α-helix, where it alters the direction of the polypeptide chain and terminates the helix. Different proteins have a different amount of the polypeptide chain folded up into α-helices. For example, the single polypeptide chain of myoglobin has eight α-helices (see Topic B4).

In the **β-pleated sheet,** hydrogen bonds form between the peptide bonds either in different polypeptide chains or in different sections of the same polypeptide chain (*Fig. 6a*). The planarity of the peptide bond forces the polypeptide to be pleated with the side-chains of the amino acids protruding above and below the sheet (*Fig. 6b*). Adjacent polypeptide chains in β-pleated

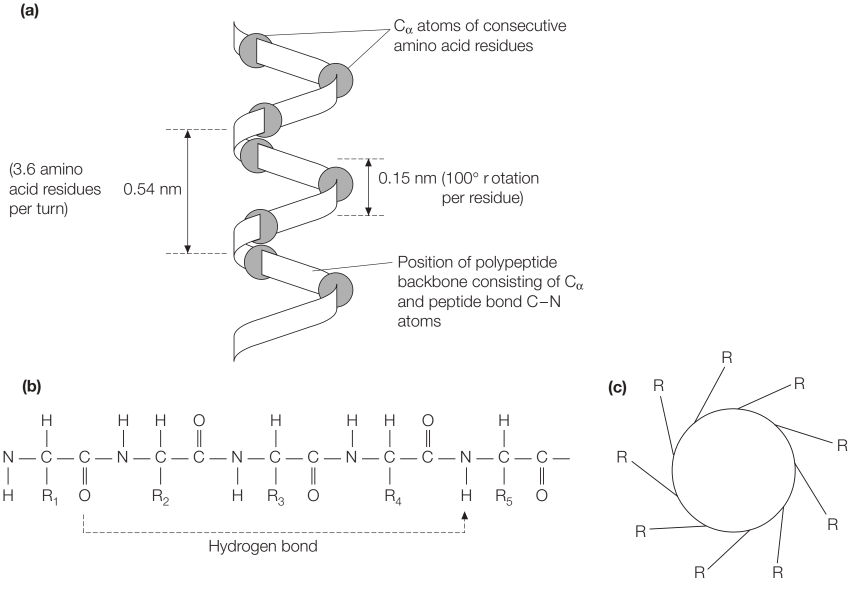
(a)
Cα atoms of consecutive amino acid residues
(3.6 amino acid residues per turn)
0.54 nm
0.15 nm (100° r otation per residue)
Position of polypeptide backbone consisting of Cα and peptide bond C–N atoms
(b)
—N—C—C—N—C—C—N—C—C—N—C—C—N—C—C—
H
R1
O
R2
R3
R4
R5
Hydrogen bond
(c)
R

Fig. 5. The folding of the polypeptide chain into an α-helix. (a) Model of an α-helix with only the C_α atoms along the backbone shown; (b) in the α-helix the CO group of residue n is hydrogen bonded to the NH group on residue (n + 4); (c) cross-sectional view of an α-helix showing the positions of the side-chains (R groups) of the amino acids on the outside of the helix.

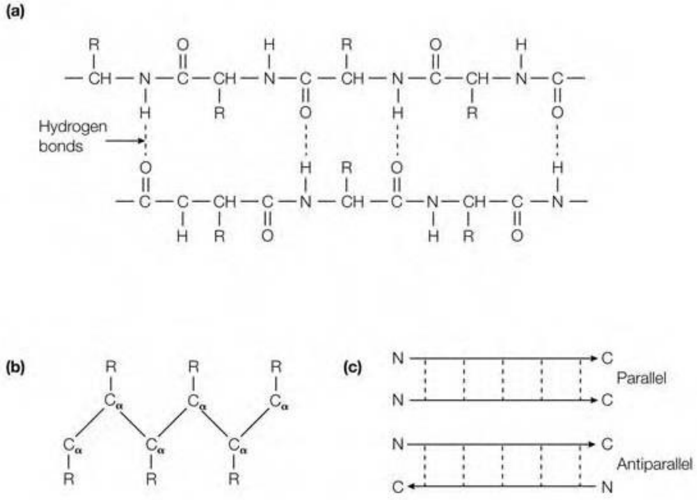

Fig. 6. The folding of the polypeptide chain in a β-pleated sheet. (a) Hydrogen bonding between two sections of a polypeptide chain forming a β-pleated sheet; (b) a side-view of one of the polypeptide chains in a β-pleated sheet showing the side-chains (R groups) attached to the C_α atoms protruding above and below the sheet; (c) because the polypeptide chain has polarity, either parallel or antiparallel β-pleated sheets can form.

sheets can be either **parallel** or **antiparallel** depending on whether they run in the same direction or in opposite directions, respectively (*Fig. 6c*). The polypeptide chain within a β-pleated sheet is fully extended, such that there is a distance of 0.35 nm from one C_α atom to the next. β-Pleated sheets are always slightly curved and, if several polypeptides are involved, the sheet can close up to form a **β-barrel**. Multiple β-pleated sheets provide strength and rigidity in many structural proteins, such as silk fibroin, which consists almost entirely of stacks of antiparallel β-pleated sheets.

In order to fold tightly into the compact shape of a globular protein, the polypeptide chain often reverses direction, making a hairpin or **β-turn**. In these β-turns the carbonyl oxygen of one amino acid is hydrogen bonded to the hydrogen on the amino group of the fourth amino acid along (*Fig. 7*). β-Turns

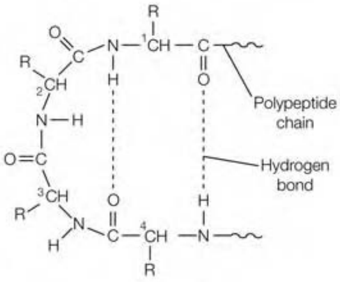

Fig. 7. The folding of the polypeptide chain in a β-turn.

are often found connecting the ends of antiparallel β-pleated sheets. Regions of the polypeptide chain that are not in a regular secondary structure are said to have a **coil** or **loop conformation**. About half the polypeptide chain of a typical globular protein will be in such a conformation.

Tertiary structure

The third level of structure found in proteins, tertiary structure, refers to the spatial arrangement of amino acids that are far apart in the linear sequence as well as those residues that are adjacent. Again, it is the sequence of amino acids that specifies this final **three-dimensional structure** (*Figs. 8 and 9*). In water-soluble globular proteins such as myoglobin (see Topic B4), the main driving force behind the folding of the polypeptide chain is the energetic requirement to bury the nonpolar amino acids in the hydrophobic interior away from the surrounding aqueous, hydrophilic medium. The polypeptide chain folds spontaneously so that the majority of its hydrophobic side-chains are buried in the interior, and the majority of its polar, charged side-chains are on the surface. Once folded, the **three-dimensional biologically-active (native) conformation** of the protein is maintained not only by hydrophobic interactions, but also by electrostatic forces, hydrogen bonding and, if present, the covalent disulfide bonds. The electrostatic forces include salt bridges between oppositely charged groups and the multiple weak van der Waals interactions between the tightly packed aliphatic side-chains in the interior of the protein.

Quaternary structure

Proteins containing more than one polypeptide chain, such as hemoglobin (see Topic B4), exhibit a fourth level of protein structure called **quaternary structure** (*Fig. 8*). This level of structure refers to the spatial arrangement of the polypeptide **subunits** and the nature of the interactions between them. These interactions may be covalent links (e.g. disulfide bonds) or noncovalent interactions (electrostatic forces, hydrogen bonding, hydrophobic interactions).

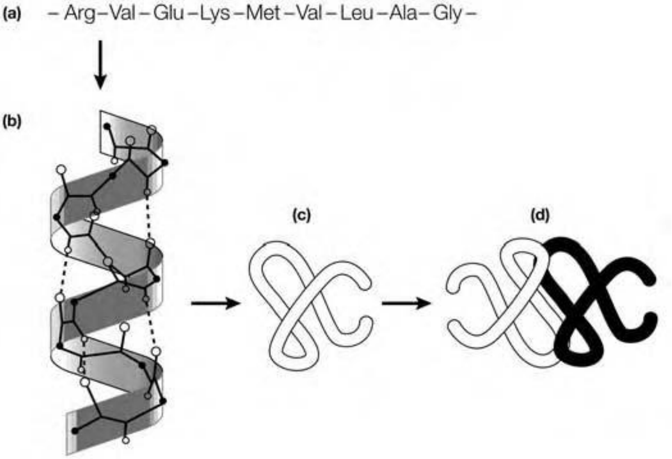

Fig. 8. The four levels of structure in proteins. (a) Primary structure (amino acid sequence), (b) secondary structure (α-helix), (c) tertiary structure, (d) quaternary structure.

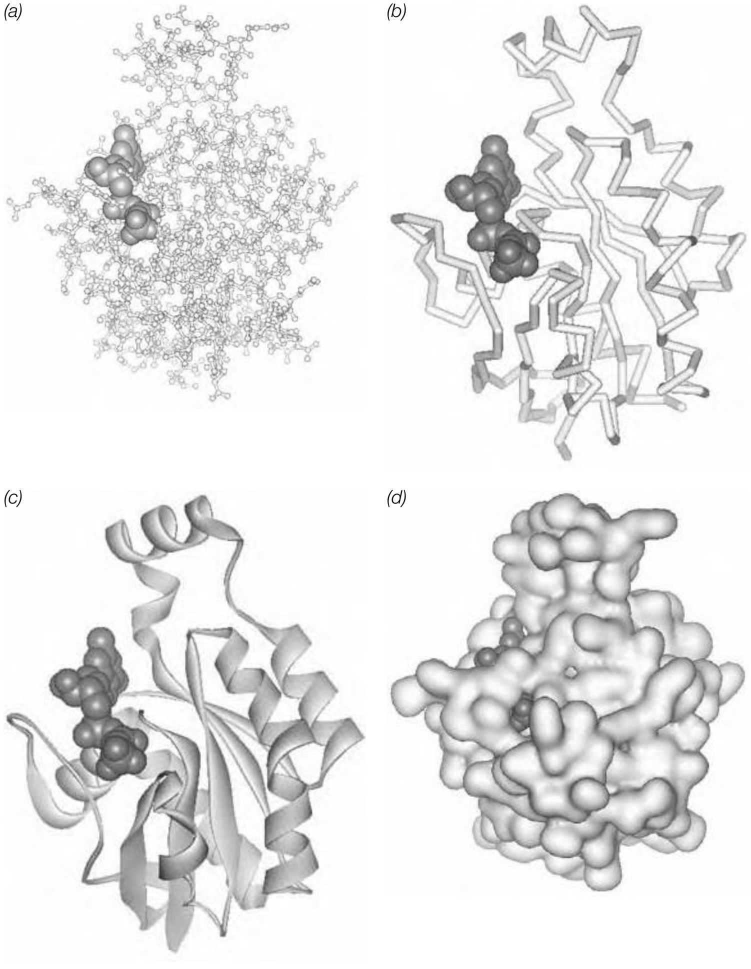

Fig. 9. Various graphic representations of the structure of RND3/RHOE a small GTP-binding protein complexed with GTP (guanosine triphosphate in spacefill representation). (a) The ball-and-stick representation reveals the location of all the atoms in the protein. (b) C_α backbone trace shows how the polypeptide chain is folded. (c) The ribbon representation emphasizes how α-helices and β-strands are organized in the protein. (d) A model of the water-accessible surface reveals the numerous bumps and crevices on the surface of the protein.

Protein stability

The native three-dimensional conformation of a protein is maintained by a range of noncovalent interactions (electrostatic forces, hydrogen bonds, hydrophobic forces) and covalent interactions (disulfide bonds) in addition to the peptide bonds between individual amino acids.

- **Electrostatic forces**: these include the interactions between two ionic groups of opposite charge, for example the ammonium group of Lys and the carboxyl group of Asp, often referred to as an **ion pair** or **salt bridge**. In addition, the noncovalent associations between electrically neutral molecules, collectively referred to as **van der Waals forces**, arise from electrostatic interactions between permanent and/or induced dipoles, such as the carbonyl group in peptide bonds.

- **Hydrogen bonds**: these are predominantly electrostatic interactions between a weakly acidic donor group and an acceptor atom that bears a lone pair of electrons, which thus has a partial negative charge that attracts the hydrogen atom. In biological systems the donor group is an oxygen or nitrogen atom that has a covalently attached hydrogen atom, and the acceptor is either oxygen or nitrogen (*Fig. 10*). Hydrogen bonds are normally in the range 0.27–0.31 nm and are highly directional, i.e. the donor, hydrogen and acceptor atoms are colinear. Hydrogen bonds are stronger than van der Waals forces but much weaker than covalent bonds. Hydrogen bonds not only play an important role in protein structure, but also in the structure of other biological macromolecules such as the DNA double helix (see Topic F1) and lipid bilayers (see Topic E1). In addition, hydrogen bonds are critical to both the properties of water and to its role as a biochemical solvent.
- **Hydrophobic forces**: The **hydrophobic effect** is the name given to those forces that cause nonpolar molecules to minimize their contact with water. This is clearly seen with amphipathic molecules such as lipids and detergents which form micelles in aqueous solution (see Topic E1). Proteins, too, find a conformation in which their nonpolar side-chains are largely out of contact with the aqueous solvent, and thus hydrophobic forces are an important determinant of protein structure, folding and stability. In proteins, the effects of hydrophobic forces are often termed **hydrophobic bonding**, to indicate the specific nature of protein folding under the influence of the hydrophobic effect.
- **Disulfide bonds**: These covalent bonds form between Cys residues that are close together in the final conformation of the protein (see *Fig. 4*) and function to stabilize its three-dimensional structure. Disulfide bonds are really only formed in the oxidizing environment of the endoplasmic reticulum (see Topic A2), and thus are found primarily in extracellular and secreted proteins.

Protein structure determination

Although the presence of α-helices and β-pleated sheets in proteins can often be predicted from the primary amino acid sequence, it is not possible to predict the precise three-dimensional structure of a protein from its amino acid sequence,

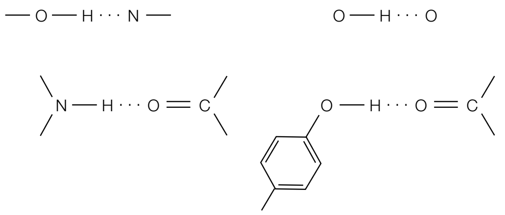

Fig. 10. Examples of hydrogen bonds (shown as dotted lines).

unless its sequence is very similar to that of a protein whose three-dimensional structure is already known. Sophisticated physical methods and complex analyses of the experimental data are required to determine the conformation of a protein. The three-dimensional structure of a protein can be determined to the atomic level by the techniques of X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryoelectron microscopy.

In **X-ray crystallography** the first requirement are **crystals** of highly purified protein. In the crystal millions of protein molecules are precisely aligned with one another in a rigid array characteristic of that particular protein. **Beams of X-rays** are then passed through the crystal (*Fig. 11*). The wavelengths of X-rays are 0.1—0.2 nm, short enough to resolve the atoms in the protein crystal. The atoms in the crystal scatter the X-rays, producing a **diffraction pattern** of discrete spots on photographic film. The intensities of the diffraction maxima (the darkness of the spots on the film) are then used mathematically to construct the three-dimensional image of the protein crystal.

Nuclear magnetic resonance (NMR) spectroscopy can be used to determine the three-dimensional structures of small (up to approximately 30 kDa) proteins in **aqueous solution**. It does not require the crystallization of the protein. In this technique, a concentrated protein solution is placed in a **magnetic field** and the effects of different radio frequencies on the spin of different atoms in the protein measured. The behavior of any particular atom is influenced by neighboring atoms in adjacent residues, with closer residues causing more perturbation than distant ones. From the magnitude of the effect, the distances between residues can be calculated and then used to generate the three-dimensional structure of the protein.

Cryoelectron microscopy is often used to determine the three-dimensional structures of proteins, particularly multisubunit proteins, that are difficult to crystallize. In this technique, the protein sample is **rapidly frozen** in liquid helium to preserve its structure. The frozen, hydrated protein is then examined

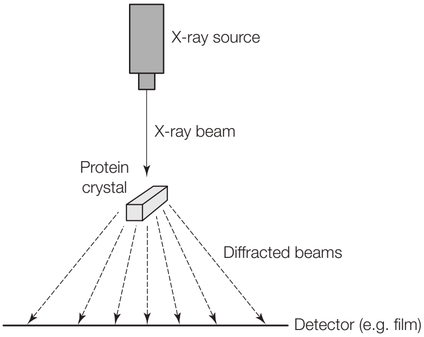

Fig. 11. X-ray crystallography. When a narrow beam of X-rays strikes a crystal, part of it passes straight through and the rest is scattered (diffracted) in various directions. The intensity of the diffracted waves is recorded on photographic film or with a solid-state electronic detector. From the diffraction data the three-dimensional structure of the protein can be determined.

in a cryoelectron microscope using a low dose of **electrons** to prevent radiation-induced damage to the structure. The resulting images are analyzed by complex computer programs and the three-dimensional structure of the protein reconstructed.

Protein folding

Under appropriate physiological conditions, proteins **spontaneously fold** into their native conformation. As there is no need for external templates, this implies that the primary structure of the protein dictates its three-dimensional structure. From experiments with the protein **RNase A** it has been observed that it is mainly the internal residues of a protein that direct its folding to the native conformation. Alteration of surface residues by mutation is less likely to affect the folding than changes to internal residues. It has also been observed that protein folding is driven primarily by **hydrophobic forces**. Proteins fold into their native conformation through an **ordered set of pathways** rather than by a random exploration of all the possible conformations until the correct one is stumbled upon.

Although proteins can fold *in vitro* (in the laboratory) without the presence of accessory proteins, this process can take minutes to days. *In vivo* (in the cell) this process requires only a few minutes because the cells contain **accessory proteins** which assist the polypeptides to fold to their native conformation. There are three main classes of protein folding accessory proteins:

- **protein disulfide isomerases** catalyze disulfide interchange reactions, thereby facilitating the shuffling of the disulfide bonds in a protein until they achieve their correct pairing.
- **peptidyl prolyl *cis–trans* isomerases** catalyze the otherwise slow interconversion of Xaa–Pro peptide bonds between their *cis* and *trans* conformations, thereby accelerating the folding of Pro-containing polypeptides. One of the classes of peptidyl prolyl *cis–trans* isomerases is inhibited by the **immunosuppressive drug** cyclosporin A.
- **molecular chaperones**, which include proteins such as the **heat shock proteins** 70 (Hsp 70), the **chaperonins**, and the **lectins calnexin and calreticulin**. These prevent the improper folding and aggregation of proteins that may otherwise occur as internal hydrophobic regions are exposed to one another.

B4 MYOGLOBIN AND HEMOGLOBIN

Key Notes

Oxygen-binding proteins

Hemoglobin and myoglobin are the two oxygen-binding proteins present in large multicellular organisms. Hemoglobin transports oxygen in the blood and is located in the erythrocytes; myoglobin stores the oxygen in the muscles.

Myoglobin

Myoglobin was the first protein to have its three-dimensional structure solved by X-ray crystallography. It is a globular protein made up of a single polypeptide chain of 153 amino acid residues that is folded into eight α-helices. The heme prosthetic group is located within a hydrophobic cleft of the folded polypeptide chain.

Hemoglobin

Hemoglobin has a quaternary structure as it is made up of four polypeptide chains; two α-chains and two β-chains ($\alpha_2\beta_2$), each with a heme prosthetic group. Despite little similarity in their primary sequences, the individual polypeptides of hemoglobin have a three-dimensional structure almost identical to the polypeptide chain of myoglobin.

Binding of oxygen to heme

The heme prosthetic group consists of a protoporphyrin IX ring and a central Fe^{2+} atom which forms four bonds with the porphyrin ring. In addition, on one side of the porphyrin ring the Fe^{2+} forms a bond with the proximal histidine (His F8); a residue eight amino acids along the F-helix of hemoglobin. The sixth bond from the Fe^{2+} is to a molecule of O_2. Close to where the O_2 binds is another histidine residue, the distal histidine (His E7), which prevents carbon monoxide binding most efficiently.

Allostery

Hemoglobin is an allosteric protein. The binding of O_2 is cooperative; the binding of O_2 to one subunit increases the ease of binding of further O_2 molecules to the other subunits. The oxygen dissociation curve for hemoglobin is sigmoidal whereas that for myoglobin is hyperbolic. Myoglobin has a greater affinity for O_2 than does hemoglobin.

Mechanism of the allosteric change

Oxyhemoglobin has a different quaternary structure from deoxyhemoglobin. As O_2 binds to the Fe^{2+} it distorts the heme group and moves the proximal histidine. This in turn moves helix F and alters the interactions between the four subunits.

The Bohr effect

H^+, CO_2 and 2,3-bisphosphoglycerate are allosteric effectors, promoting the release of O_2 from hemoglobin. H^+ and CO_2 bind to different parts of the polypeptide chains, while 2,3-bisphosphoglycerate binds in the central cavity between the four subunits.

Fetal hemoglobin

Hemoglobin F (HbF) which consists of two α-chains and two γ-chains ($\alpha_2\gamma_2$) is present in the fetus. HbF binds 2,3-bisphosphoglycerate less strongly than adult hemoglobin (HbA) and thus has a higher affinity for O_2 which promotes the transfer of O_2 from the maternal to the fetal circulation.

Hemoglobinopathies

Comparison of hemoglobin sequences from different species reveals that only nine amino acid residues are invariant. Some residues are subject to conservative substitution of one residue by another with similar properties, others to nonconservative substitution where one amino acid residue is replaced by another with different properties. Hemoglobinopathies are diseases caused by abnormal hemoglobins. The best characterized of these is the genetically transmitted, hemolytic disease sickle-cell anemia. This is caused by the nonconservative substitution of a glutamate by a valine, resulting in the appearance of a hydrophobic sticky patch on the surface of the protein. This allows long aggregated fibers of hemoglobin molecules to form which distort the shape of the red blood cells. Heterozygotes carrying only one copy of the sickle-cell gene are more resistant to malaria than those homozygous for the normal gene.

Related topics

Cytoskeleton and molecular motors (A3)
Bioimaging (A4)
Protein structure (B3)
Regulation of enzyme activity (C5)
The DNA revolution (I1)
Electron transport and oxidative phosphorylation (L2)
Hemes and chlorophylls (M4)

Oxygen-binding proteins

Hemoglobin is one of two **oxygen-binding proteins** found in vertebrates. The function of hemoglobin is to carry O_2 in the blood from the lungs to the other tissues in the body, in order to supply the cells with the O_2 required by them for the oxidative phosphorylation of foodstuffs (see Topic L2). The hemoglobin is found in the blood within the **erythrocytes** (red blood cells). These cells essentially act, amongst other things, as a sack for carrying hemoglobin, since mature erythrocytes lack any internal organelles (nucleus, mitochondria, etc.). The other O_2-binding protein is **myoglobin**, which stores the oxygen in the tissues of the body ready for when the cells require it. The highest concentrations of myoglobin are found in skeletal and cardiac **muscle** which require large amounts of O_2 because of their need for large amounts of energy during contraction (see Topic A3).

Myoglobin

Myoglobin is a relatively small protein of mass 17.8 kDa made up of 153 amino acids in a single polypeptide chain. It was the first protein to have its **three-dimensional structure** determined by **X-ray crystallography** (see Topic B3) by John Kendrew in 1957. Myoglobin is a typical **globular protein** in that it is a highly folded compact structure with most of the hydrophobic amino acid residues buried in the interior and many of the polar residues on the surface. X-ray crystallography revealed that the single polypeptide chain of myoglobin consists entirely of **α-helical secondary structure** (see Topic B3). In fact there are eight α-helices (labeled A–H) in myoglobin (*Fig. 1a*). Within a hydrophobic

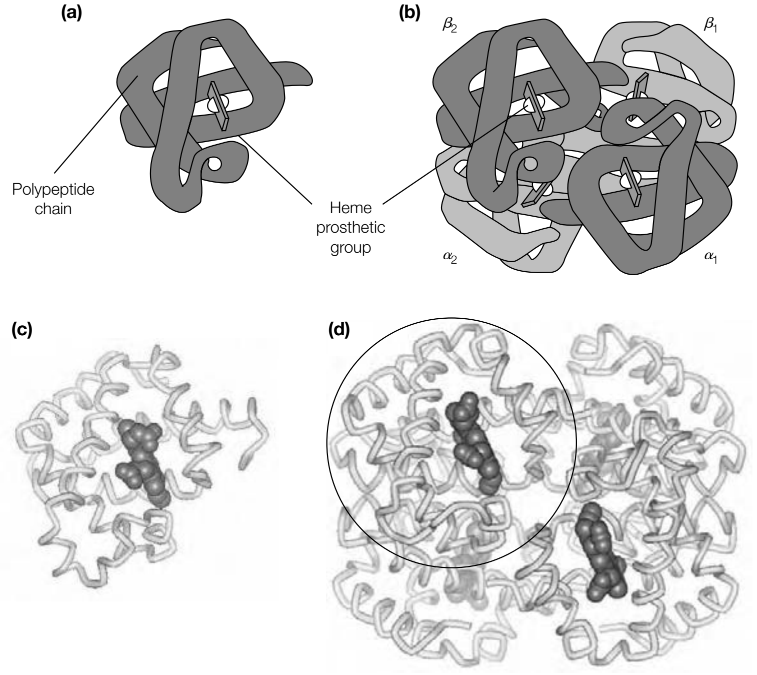

Fig. 1. Structure of (a) myoglobin and (b) hemoglobin, showing the α and β polypeptide chains. C_α-backbone traces of (c) human myoglobin and (d) human hemoglobin, showing the α-helices, the heme prosthetic group in space-filling representation and how the monomer of myoglobin maps onto the structure of hemoglobin (circle).

crevice formed by the folding of the polypeptide chain is the **heme prosthetic group** (*Fig. 1a*). This nonpolypeptide unit is noncovalently bound to myoglobin and is essential for the biological activity of the protein (i.e. the binding of O_2).

Hemoglobin

The three-dimensional structure of hemoglobin was solved using **X-ray crystallography** (see Topic B3) in 1959 by Max Perutz. This revealed that hemoglobin is made up of four polypeptide chains, each of which has a very similar three-dimensional structure to the single polypeptide chain in myoglobin (*Fig. 1b*) despite the fact that their amino acid sequences differ at 83% of the residues. This highlights a relatively common theme in protein structure: that very different primary sequences can specify very similar three-dimensional structures. The major type of hemoglobin found in adults (HbA) is made up of two different polypeptide chains: the **α-chain** that consists of 141 amino acid residues, and the **β-chain** of 146 residues ($\alpha_2\beta_2$; *Fig. 1b*). Each chain, like that in myoglobin, consists of eight α-helices and each contains a heme prosthetic group (*Fig. 1b*). Therefore, hemoglobin can bind four molecules of O_2. The four polypeptide chains, two α and two β, are packed tightly together in a tetrahedral array to form an overall spherically shaped molecule that is held together by multiple noncovalent interactions (see Topic B3).

Binding of oxygen to heme

The **heme prosthetic group** in myoglobin and hemoglobin is made up of a **protoporphyrin IX** ring structure with an **iron atom** in the ferrous (Fe^{2+}) oxidation state (see Topic M4; *Fig. 2*). This Fe^{2+} bonds with four nitrogen atoms in the center of the protoporphyrin ring and forms two additional bonds on either side of the plane of the protoporphyrin ring. One of these is to a histidine residue which lies eight residues along helix F of hemoglobin, the **proximal histidine** (His F8) (*Fig. 2*). The sixth bond is to one of the oxygen atoms in a molecule of $\mathbf{O_2}$ (*Fig. 2*). Near to where the O_2 binds to the heme group is another histidine residue, the **distal histidine** (His E7) (*Fig. 2*). This serves two very important functions. First, it prevents heme groups on neighboring hemoglobin molecules coming into contact with one another and oxidizing to the Fe^{3+} state in which they can no longer bind O_2. Second, it prevents **carbon monoxide** (CO) binding with the most favorable configuration to the Fe^{2+}, thereby lowering the affinity of heme for CO. This is important because once CO has bound irreversibly to the heme, the protein can no longer bind O_2. Thus, although the oxygen binding site in hemoglobin and myoglobin is only a small part of the whole protein, the polypeptide chain modulates the function of the heme prosthetic group.

Allostery

Hemoglobin is an **allosteric protein** (see Topic C5 for a fuller discussion of allostery). This means that the binding of O_2 to one of the subunits is affected by its interactions with the other subunits. In fact the binding of O_2 to one hemoglobin subunit induces conformational changes (see below and *Fig. 2*) that are relayed to the other subunits, making them more able also to bind O_2 by raising their affinity for this molecule. Thus binding of O_2 to hemoglobin is said to be **cooperative**. In contrast, the binding of O_2 to the single polypeptide unit of myoglobin is **noncooperative**. This is clearly apparent from the **oxygen dissociation curves** for the two proteins: that for hemoglobin is **sigmoidal**, reflecting this cooperative binding, whereas that for myoglobin is **hyperbolic** (*Fig. 3*).

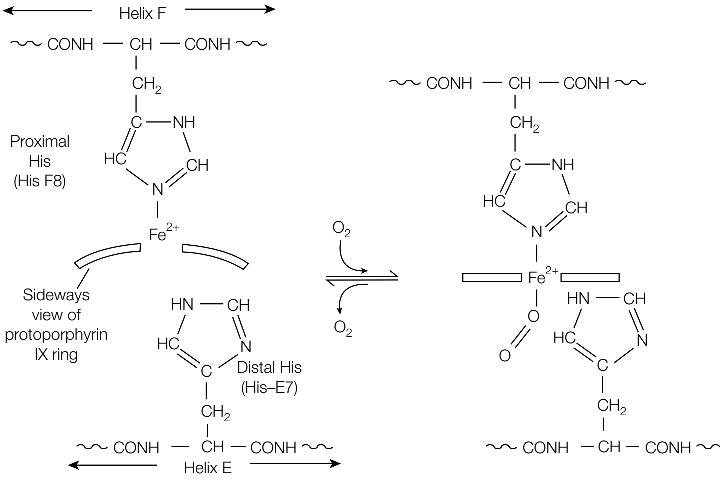

Fig. 2. Binding of O_2 to heme. The Fe^{2+} of the protoporphyrin ring is bonded to His F8 but not to His E7 which is located nearby. As the heme Fe^{2+} binds O_2, helix F moves closer to helix E (see the text for details).

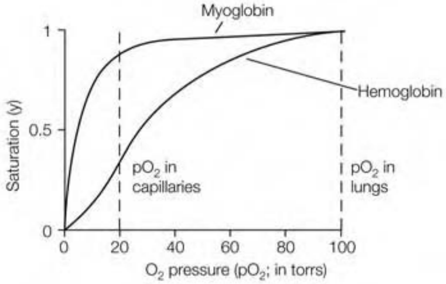

Fig. 3. Oxygen dissociation curves for hemoglobin and myoglobin.

From the O_2 dissociation curve it can also be seen that for any particular oxygen pressure the degree of saturation of myoglobin is higher than that for hemoglobin. In other words, myoglobin has a higher affinity for O_2 than does hemoglobin. This means that in the blood capillaries in the muscle, for example, hemoglobin will release its O_2 to myoglobin for storage there.

Mechanism of the allosteric change

X-ray crystallography revealed that **oxyhemoglobin**, the form that has four O_2 molecules bound, differs markedly in its **quaternary structure** from **deoxyhemoglobin**, the form with no O_2 bound. In the absence of bound O_2, the Fe^{2+} lies slightly to one side of the porphyrin ring, which itself is slightly curved (*Fig. 2*). As a molecule of O_2 binds to the heme prosthetic group it pulls the Fe^{2+} into the plane of the porphyrin ring (*Fig. 2*), flattening out the ring in the process. Movement of the Fe^{2+} causes the **proximal histidine** to move also. This, in turn, shifts the position of helix F and regions of the polypeptide chain at either end of the helix. Thus, movement in the center of the subunit is transmitted to the surfaces, where it causes the ionic interactions holding the four subunits together to be broken and to reform in a different position, thereby altering the quaternary structure, leading to the cooperative binding of O_2 to Hb.

The Bohr effect

The binding of O_2 to hemoglobin is affected by the concentration of **H^+ ions** and **CO_2** in the surrounding tissue; the Bohr effect. In actively metabolizing tissue, such as muscle, the concentrations of these two substances are relatively high. This effectively causes a shift of the O_2 dissociation curve for hemoglobin to the right, promoting the release of O_2. This comes about because there are H^+ binding sites, primarily His146 in the β-chain, which have a higher affinity for binding H^+ in deoxyhemoglobin than in oxyhemoglobin. An increase in CO_2 also causes an increase in H^+ due to the action of the enzyme **carbonic anhydrase** which catalyzes the reaction:

$$CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+$$

In addition, CO_2 can react with the primary amino groups in the polypeptide chain to form a negatively charged carbamate. Again, this change from a positive to a negative charge favors the conformation of deoxyhemoglobin. On returning in the blood to the lungs, the concentrations of H^+ and CO_2 are relatively lower and that of O_2 higher, so that the process is reversed and O_2 binds to hemoglobin. Thus, it can be seen that not only does hemoglobin carry O_2 but it also carries CO_2 back to the lungs where it is expelled.

2,3-Bisphosphoglycerate is a highly anionic organic phosphate molecule (*Fig.*

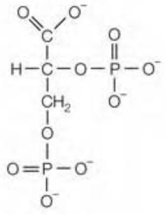

Fig. 4. 2,3-Bisphosphoglycerate.

4) that is present in erythrocytes along with the hemoglobin. This molecule promotes the release of O_2 from hemoglobin by lowering the affinity of the protein for O_2. 2,3-Bisphosphoglycerate binds in the small cavity in the center of the four subunits. In oxyhemoglobin this cavity is too small for it, whereas in deoxyhemoglobin it is large enough to accommodate a single molecule of 2,3-bisphosphoglycerate. On binding in the central cavity of deoxyhemoglobin it forms ionic bonds with the positively charged amino acid side-chains in the β-subunits, stabilizing the quaternary structure. H^+, CO_2 and 2,3-bisphosphoglycerate are all **allosteric effectors** (see Topic C5) as they favor the conformation of deoxyhemoglobin and therefore promote the release of O_2. Because these three molecules act at different sites, their effects are additive.

Fetal hemoglobin

In the fetus there is a different kind of hemoglobin, **hemoglobin F** (HbF) which consists of two α-chains and two γ-chains ($\alpha_2\gamma_2$), in contrast to adult hemoglobin (HbA, $\alpha_2\beta_2$). HbF has a **higher affinity** for O_2 under physiological conditions than HbA, which optimizes the transfer of oxygen from the maternal to the fetal circulation across the placenta. The molecular basis for this difference in O_2 affinity is that HbF binds 2,3-bisphosphoglycerate less strongly than does HbA. Near birth the synthesis of the γ-chain is switched off, and that of the β-chain (which is present in HbA) is switched on (*Fig. 5*).

Hemoglobinopathies

Comparison of the primary sequences of hemoglobin chains from more than 60 different species reveals that only nine residues in the polypeptide chain are

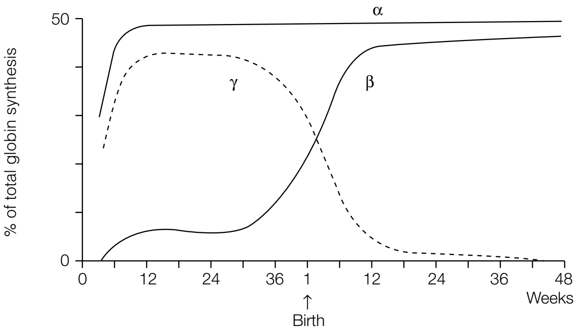

Fig. 5. The switch in human globin chain synthesis at birth.

invariant (i.e. the same) between all of the species. These nine residues include the **proximal and distal histidines** which are essential for the correct functioning of the protein. Many of the other residues are replaced from one species to another by residues with similar properties (e.g. the hydrophobic valine is replaced with the hydrophobic isoleucine, or the polar serine is replaced with the polar asparagine), so-called **conservative substitutions**. In contrast, only a few residues have changed between species to a completely different residue (e.g. a hydrophobic leucine to a positively charged lysine or a negatively charged glutamate to a positively charged arginine), so-called **nonconservative substitutions**, since this type of change could have a major effect on the structure and function of the protein.

Several hundred **abnormal hemoglobins** have been characterized, giving rise to the so-called **hemoglobinopathies**. Probably the best characterized hemoglobinopathy is **sickle-cell anemia** (sickle-cell hemoglobin; HbS). This disease is characterized by the patient's erythrocytes having a characteristic sickle or crescent shape. The molecular basis for this disease is the change of a glutamic acid residue for a valine at position 6 of the β-chain, resulting in the substitution of a polar residue by a hydrophobic one. This **nonconservative substitution** of valine for glutamate gives HbS a **sticky hydrophobic patch** on the outside of each of its β-chains. In the corner between helices E and F of the β-chain of deoxy-HbS is a hydrophobic site that is complementary to the sticky patch (*Fig. 6*). Thus the complementary site on one deoxy-HbS molecule can bind to the sticky patch on another deoxy-HbS molecule, resulting in the formation of **long fibers** of hemoglobin molecules that distort the erythrocyte. Electron microscopy (see Topic A4) has revealed that the fibers have a diameter of 21.5 nm and consist of a 14-stranded helix. Multiple polar interactions, in addition to the critical interaction between the sticky patches, stabilize the fiber. In oxy-HbS the complementary site is masked, so the formation of the long fibers occurs only when there is a high concentration of the deoxygenated form of HbS.

Sickle-cell anemia is a **genetically transmitted**, hemolytic disease. The sickled cells are more fragile than normal erythrocytes, lysing more easily and having a shorter half-life, which leads to severe anemia. As sickle-cell anemia is genetically transmitted, **homozygotes** have two copies of the abnormal gene whereas **heterozygotes** have one abnormal and one normal copy. Homozygotes often

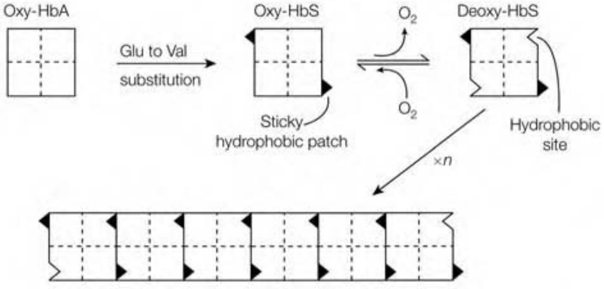

Fig. 6. Molecular basis for the aggregation of deoxyhemoglobin molecules in sickle-cell anemia.

have a reduced life-span as a result of infection, renal failure, cardiac failure or thrombosis, due to the sickled cells becoming trapped in small blood vessels leading to tissue damage. In contrast, heterozygotes are usually not symptomatic as only approximately 1% of their erythrocytes are sickled, compared with approximately 50% in a homozygote. The frequency of the sickle gene is relatively high in certain parts of Africa and correlates with the incidence of **malaria**. The reason for this is that heterozygotes are protected against the most lethal form of malaria, whereas normal homozygotes are more vulnerable to the disease. Inheritance of the abnormal hemoglobin gene can now be monitored by recombinant DNA techniques (see Topic I1).

B5 COLLAGEN

Key Notes

Function and diversity

Collagen is the name given to a family of structurally related proteins that form strong insoluble fibers. Collagens consist of three polypeptide chains, the identity and distribution of which vary between collagen types. The different types of collagen are found in different locations in the body.

Biosynthesis: overview

The collagen polypeptides are post-translationally modified by hydroxylation and glycosylation on transport through the rough endoplasmic reticulum and Golgi. The three polypeptides form the triple-helical procollagen which is secreted out of the cell. The extension peptides are removed to form tropocollagen which then aggregates into a microfibril and is covalently cross-linked to form the mature collagen fiber.

Composition and post-translational modifications

One-third of the amino acid residues in collagen are Gly, while another quarter are Pro. The hydroxylated amino acids 4-hydroxyproline (Hyp) and 5-hydroxylysine (Hyl) are formed post-translationally by the action of proline hydroxylase and lysine hydroxylase. These Fe^{2+}-containing enzymes require ascorbic acid (vitamin C) for activity. In the vitamin C deficiency disease scurvy, collagen does not form correctly due to the inability to hydroxylate Pro and Lys. Hyl residues are often post-translationally modified with carbohydrate.

Structure

Collagen contains a repeating tripeptide sequence of Gly–X–Y, where X is often Pro and Y is often Hyp. Each polypeptide in collagen folds into a helix with 3.3 residues per turn. Three polypeptide chains then come together to form a triple-helical cable that is held together by hydrogen bonds between the chains. Every third residue passes through the center of the triple helix, which is so crowded that only Gly is small enough to fit.

Secretion and aggregation

The extension peptides on both the N and C termini of the polypeptide chains direct the formation of the triple-helical cable and prevent the premature aggregation of the procollagen molecules within the cell. Following secretion out of the cell, the extension peptides are cleaved off by peptidases, and the resulting tropocollagen molecules aggregate together in a staggered array.

Cross-links

Covalent cross-links both between and within the tropocollagen molecules confer strength and rigidity on the collagen fiber. These cross-links are formed between Lys and its aldehyde derivative allysine. Allysine is derived from Lys by the action of the copper-containing lysyl oxidase which requires pyridoxal phosphate for activity.

Bone formation

Hydroxyapatite (calcium phosphate) is deposited in nucleation sites between the ends of tropocollagen molecules as the first step in bone formation.

Related topics		
	Protein structure (B3) Translation in eukaryotes (H3) Protein glycosylation (H5)	Citric acid cycle (L1) Amino acid metabolism (M2)

Function and diversity

Collagen, which is present in all multicellular organisms, is not one protein but a family of structurally related proteins. It is the most abundant protein in mammals and is present in most organs of the body, where it serves to hold cells together in discrete units. It is also the major **fibrous element** of skin, bones, tendons, cartilage, blood vessels and teeth. The different collagen proteins have very diverse functions. The extremely hard structures of bones and teeth contain collagen and a calcium phosphate polymer. In tendons, collagen forms **rope-like fibers of high tensile strength**, while in the skin collagen forms **loosely woven fibers that can expand in all directions**. The different types of collagen are characterized by **different polypeptide compositions** (*Table 1*). Each collagen is composed of three polypeptide chains, which may be all identical (as in types II and III) or may be of two different chains (types I, IV and V). A single molecule of type I collagen has a molecular mass of 285 kDa, a width of 1.5 nm and a length of 300 nm.

Table 1. Types of collagen

Type	Polypeptide composition	Distribution
I	$[\alpha1(\mathrm{I})]_2\ \alpha2(\mathrm{I})$	Skin, bone, tendon, cornea, blood vessels
II	$[\alpha1(\mathrm{II})]_3$	Cartilage, intervertebral disk
III	$[\alpha1(\mathrm{III})]_3$	Fetal skin, blood vessels
IV	$[\alpha1(\mathrm{IV})]_2\ \alpha2(\mathrm{IV})$	Basement membrane
V	$[\alpha1(\mathrm{V})]_2\ \alpha2(\mathrm{V})$	Placenta, skin

Biosynthesis: overview

Like other secreted proteins, collagen polypeptides are synthesized by ribosomes on the rough endoplasmic reticulum (RER; see Topic H3). The polypeptide chain then passes through the RER and Golgi apparatus before being secreted. Along the way it is **post-translationally modified**: Pro and Lys residues are hydroxylated and carbohydrate is added (*Fig. 1*). Before secretion, three polypeptide chains come together to form a triple-helical structure known as **procollagen**. The procollagen is then secreted into the extracellular spaces of the connective tissue where extensions of the polypeptide chains at both the N and C termini (**extension peptides**) are removed by peptidases to form **tropocollagen** (*Fig. 1*). The tropocollagen molecules aggregate and are extensively **cross-linked** to produce the mature **collagen fiber** (*Fig. 1*).

Composition and post-translational modifications

The amino acid composition of collagen is quite distinctive. Nearly *one-third* of its residues are **Gly**, while another *one-quarter* are **Pro**, significantly higher proportions than are found in other proteins. The hydroxylated amino acids **4-hydroxyproline** (**Hyp**) and **5-hydroxylysine** (**Hyl**) (*Fig. 2*) are found exclusively in collagen. These hydroxylated amino acids are formed from the parent amino acid by the action of **proline hydroxylase** and **lysine hydroxylase**, respectively (*Fig. 2*). These enzymes have an Fe^{2+} ion at their active site and require **ascorbic acid** (**vitamin C**) for activity. The ascorbic acid acts as an antioxidant, keeping

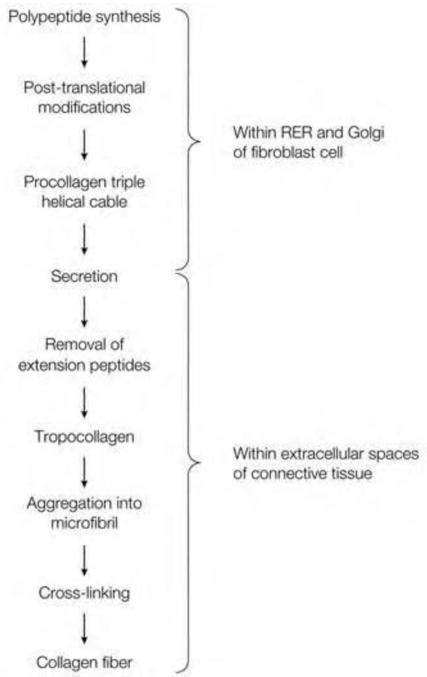
Polypeptide synthesis
Post-translational modifications
Procollagen triple helical cable
Secretion
Removal of extension peptides
Tropocollagen
Aggregation into microfibril
Cross-linking
Collagen fiber
Within RER and Golgi of fibroblast cell
Within extracellular spaces of connective tissue

Fig. 1. Overview of the biosynthesis of collagen.

the Fe^{2+} ion in its reduced state. Proline hydroxylase and lysine hydroxylase are dioxygenases, using a molecule of O_2. α-Ketoglutarate, the citric acid cycle intermediate (see Topic L1), is an obligatory substrate and is converted into succinate during the reaction (*Fig. 2*). Both enzymes will hydroxylate only Pro and Lys residues that are incorporated in a polypeptide chain, and then only when the residue is on the N-terminal side of Gly. Hyp is important in stabilizing the structure of collagen through hydrogen bond formation (see below). In **vitamin C deficiency**, Hyp (and Hyl) are not synthesized, resulting in the weakening of the collagen fibers. This leads to the skin lesions, fragile blood vessels and poor wound healing that are characteristic of the disease **scurvy**.

The other post-translational modification that occurs to collagen is **glycosylation**. In this case the sugar residues, usually only glucose, galactose and their disaccharides, are attached to the hydroxyl group in the newly formed Hyl residues, rather than to Asn or Ser/Thr residues as occurs in the more widespread N- and O-linked glycosylation (see Topic H5). The amount of attached carbohydrate in collagen varies from 0.4 to 12% by weight depending on the tissue in which it is synthesized.

Structure

The **primary structure** of each polypeptide in collagen is characterized by a repeating **tripeptide** sequence of **Gly–X–Y** where X is often, but not exclusively, Pro and Y is often Hyp. Each of the three polypeptide chains in collagen is some

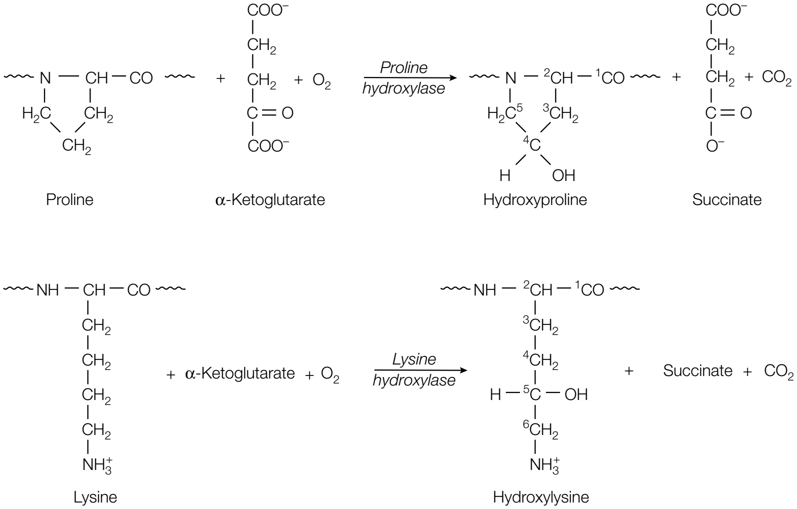

Fig. 2. Formation of hydroxyproline and hydroxylysine.

1000 residues long and they each fold up into a **helix** that has only 3.3 residues per turn, rather than the 3.6 residues per turn of an α-helix (see Topic B3). This **secondary structure** is unique to collagen and is often called the **collagen helix**. The three polypeptide chains lie parallel and wind round one another with a slight right-handed, rope-like twist to form a **triple-helical cable** (*Fig. 3*). Every third residue of each polypeptide passes through the center of the triple helix, which is so crowded that only the small side-chain of Gly can fit in. This explains the absolute requirement for Gly at every third residue. The residues in the X and Y positions are located on the outside of the triple-helical cable, where there is room for the bulky side-chains of Pro and other residues. The three polypeptide chains are also staggered so that the Gly residue in one chain is aligned with the X residue in the second and the Y residue in the third. The triple helix is held together by an extensive network of **hydrogen bonds**, in particular between the primary amino group of Gly in one helix and the primary carboxyl group of Pro in position X of one of the other helices. In addition, the hydroxyl groups of Hyp residues participate in stabilizing the structure. The relatively inflexible Pro and Hyp also confer rigidity on the collagen structure.

The importance of Gly at every third residue is seen when a **mutation** in the DNA encoding Type I collagen leads to the incorporation of a different amino acid at just one position in the 1000 residue polypeptide chain. For example, if a mutation leads to the incorporation of Cys instead of Gly, the triple helix is disrupted as the -CH_2-SH side-chain of Cys is too large to fit in the interior of the triple helix. This leads to a partly unfolded structure that is susceptible to excessive hydroxylation and glycosylation and is not efficiently secreted by the fibroblast cells. This, in turn, results in a defective collagen structure that can give rise to **brittle bones** and **skeletal deformities**. A whole spectrum of such mutations are known which cause the production of defective collagen and result in **osteogenesis imperfecta** (brittle bones).

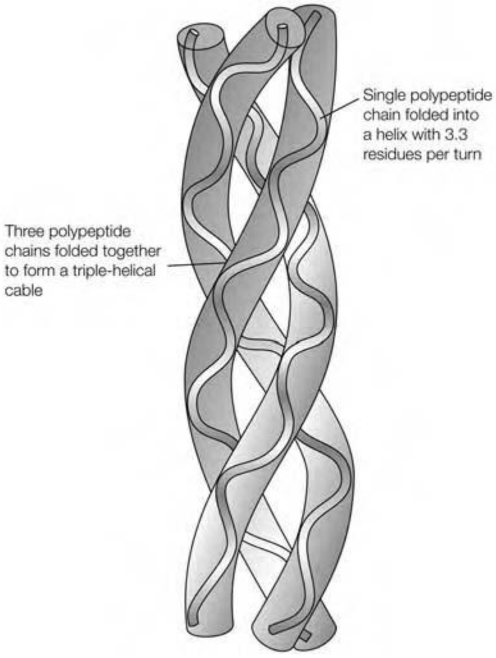
Single polypeptide chain folded into a helix with 3.3 residues per turn
Three polypeptide chains folded together to form a triple-helical cable

Fig. 3. Arrangement of the three polypeptide chains in collagen.

Secretion and aggregation

When the collagen polypeptides are synthesized they have additional amino acid residues (100–300) on both their N and C termini that are absent in the mature collagen fiber (*Fig. 4*). These **extension peptides** often contain Cys residues, which are usually absent from the remainder of the polypeptide chain. The extension peptides help to align correctly the three polypeptides as they come together in the triple helix, a process that may be aided by the formation of disulfide bonds between extension peptides on neighboring polypeptide chains. The extension peptides also prevent the premature aggregation of the procollagen triple helices within the cell. On **secretion** out of the fibroblast the extension peptides are removed by the action of extracellular **peptidases** (*Fig. 4*). The resulting tropocollagen molecules then **aggregate** together in a staggered head-to-tail arrangement in the collagen fiber (*Fig. 4*).

Cross-links

The strength and rigidity of a collagen fiber is imparted by **covalent cross-links** both between and within the tropocollagen molecules. As there are few, if any, Cys residues in the final mature collagen, these covalent cross-links are not disulfide bonds as commonly found in proteins, but rather are unique cross-links formed between **Lys** and its aldehyde derivative **allysine**. Allysine residues are formed from Lys by the action of the monooxygenase **lysyl oxidase** (*Fig. 5*). This **copper**-containing enzyme requires the coenzyme **pyridoxal phosphate**, derived from vitamin B_6 (see Topic M2), for activity. The aldehyde group on allysine then reacts spontaneously with either the side-chain amino group of Lys or with other allysine residues on other polypeptide chains to form covalent interchain bonds.

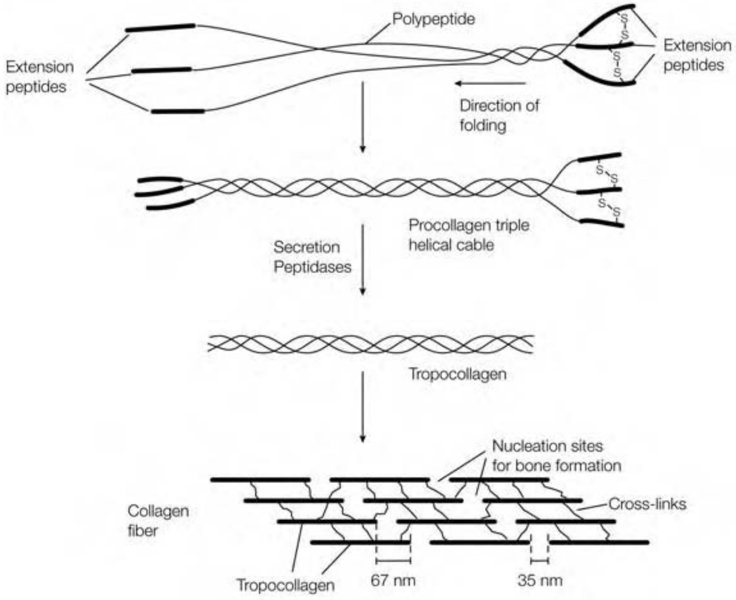

Fig. 4. Role of the extension peptides in the folding and secretion of procollagen. Once secreted out of the cell, the extension peptides are removed and the resulting tropocollagen molecules aggregate and are cross-linked to form a fiber.

The importance of cross-linking to the normal functioning of collagen is demonstrated by the disease **lathyrism**. This occurs in humans and other animals through the ingestion of sweet pea (*Lathyrus odoratus*) seeds which contain the chemical **β-aminopropionitrile**. This compound irreversibly inhibits lysyl oxidase, thereby preventing the cross-linking of the tropocollagen molecules, resulting in serious abnormalities of the bones, joints and large blood vessels due to the fragile collagen. One collagen deficiency disease, the **Ehlers–Danlos syndrome type V,** is due to a deficiency in lysyl oxidase and results in hypermobile joints and hyperextensibility of the skin.

Bone formation

The regular staggered array of spaces between the ends of the tropocollagen molecules in a collagen fiber (see *Fig. 4*) are the **nucleation sites** for the deposition of a form of **calcium phosphate, hydroxyapatite**, in bone formation. Further hydroxyapatite is added until the nucleation sites grow and join with one another to form the mature bone structure.

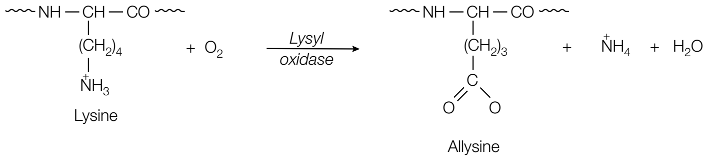

Fig. 5. Conversion of lysine to allysine by lysyl oxidase.

Section B – Amino acids and proteins

B6 Protein purification

Key Notes

Principles of protein purification

The aim of protein purification is to isolate one particular protein from all the others in the starting material. A combination of fractionation techniques is used that exploits the solubility, size, charge, hydrophobicity or/and specific binding affinity of the protein of interest.

Selection of a protein source

Because proteins have different distributions in biological materials, it is important to make the right choice of starting material from which to purify the protein. This will usually be a source that is relatively rich in the protein of interest and which is readily available. The use of recombinant DNA techniques means that large amounts of normally scarce proteins can be obtained by expression in bacterial or eukaryotic cells.

Homogenization and solubilization

The protein has to be obtained in solution prior to its purification. Thus tissues and cells must be disrupted by homogenization or osmotic lysis and then subjected to differential centrifugation to isolate the subcellular fraction in which the protein is located. For membrane-bound proteins, the membrane structure has to be solubilized with a detergent to liberate the protein.

Stabilization of proteins

Certain precautions have to be taken in order to prevent proteins being denatured or inactivated during purification by physical or biological factors. These include buffering the pH of the solutions, undertaking the procedures at a low temperature and including protease inhibitors to prevent unwanted proteolysis.

Assay of proteins

In order to monitor the progress of the purification of a protein, it is necessary to have an assay for it. Depending on the protein, the assay may involve measuring the enzyme activity or ligand-binding properties, or may quantify the protein present using antibodies directed against it.

Ammonium sulfate precipitation

The solubility of proteins decreases as the concentration of ammonium sulfate in the solution is increased. The concentration of ammonium sulfate at which a particular protein comes out of solution and precipitates may be sufficiently different from other proteins in the mixture to effect a separation.

Dialysis

Proteins can be separated from small molecules by dialysis through a semi-permeable membrane which has pores that allow small molecules to pass through but not proteins.

Gel filtration chromatography

Gel filtration chromatography separates proteins on the basis of their size and shape using porous beads packed in a column. Large or elongated proteins cannot enter the pores in the beads and elute from the bottom of the column first, whereas smaller proteins can enter the beads, have a larger volume of liquid accessible to them and move through the column more slowly, eluting later. Gel filtration chromatography can be used to de-salt a protein mixture and to estimate the molecular mass of a protein.

Ion exchange chromatography

In ion exchange chromatography, proteins are separated on the basis of their net charge. In anion exchange chromatography a column containing positively-charged beads is used to which proteins with a net negative charge will bind, whereas in cation exchange chromatography, negatively-charged beads are used to which proteins with a net positive charge will bind. The bound proteins are then eluted by adding a solution of sodium chloride or by altering the pH of the buffer.

Affinity chromatography

Affinity chromatography exploits the specific binding of a protein for another molecule, its ligand (e.g. an enzyme for its inhibitor, antigen for its antibody). The ligand is immobilized on an insoluble support which is then packed into a column. On adding a mixture of proteins, only the protein of interest binds to the ligand. All other proteins pass straight through the column. The bound protein is then eluted from the immobilized ligand in a highly purified form.

Related topics

Eukaryote cell structure (A2)
Cellular fractionation (A5)
Acids and bases (B2)
Electrophoresis of proteins (B7)
Introduction to enzymes (C1)
Enzyme inhibition (C4)
Antibodies as tools (D4)
Membrane proteins and carbohydrate (E2)
Signal transduction (E5)
The DNA revolution (I1)

Principles of protein purification

The basic aim in protein purification is to isolate one particular protein of interest from other contaminating proteins so that its structure and/or other properties can be studied. Once a suitable cellular **source** of the protein has been identified, the protein is liberated into solution and then separated from contaminating material by sequential use of a series of different **fractionation techniques** or **separations**. These separations exploit one or more of the following basic properties of the protein: its **solubility**, its **size**, its **charge**, its **hydrophobicity** or its **specific binding affinity**. These separations may be **chromatographic techniques** such as **ion exchange**, **gel filtration** or **affinity chromatography**, **hydrophobic interaction chromatography**, in which the protein binds to a hydrophobic material, or **electrophoretic techniques** such as **isoelectric focusing** (see Topic B7). Other electrophoretic procedures, mainly **sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE)** (see Topic B7), are used to monitor the extent of purification and to determine the molecular mass and subunit composition of the purified protein.

Selection of a protein source

Before attempting to purify a protein, the first thing to consider is the source of **starting material**. Proteins differ in their cellular and tissue distribution, and thus if a protein is known to be abundant in one particular tissue (e.g. kidney) it makes sense to start the purification from this source. Also, some sources are more readily available than others and this should be taken into account too. Nowadays, with the use of **recombinant DNA techniques** (see Topic I1), even scarce proteins can be expressed (synthesized) in bacteria or eukaryotic cells and relatively large amounts of the protein subsequently obtained.

Homogenization and solubilization

Once a suitable source has been identified, the next step is to obtain the protein in solution. For proteins in biological fluids, such as blood serum or cell culture medium, this is already the case, but for many proteins the tissues and cells need to be disrupted and broken open (lysed). **Homogenization** and subsequent **differential centrifugation** of biological samples is detailed in Topic A5. In addition to the procedures described there, another simple way of breaking open cells that do not have a rigid cell wall to release the cytosolic contents is **osmotic lysis**. When animal cells are placed in a hypotonic solution (such as water or a buffered solution without added sucrose), the water in the surrounding solution diffuses into the more concentrated cytosol, causing the cell to swell and burst. Differential centrifugation is then employed to remove contaminating subcellular organelles (see Topic A5). Those proteins that are bound to membranes require a further solubilization step. After isolation by differential centrifugation, the appropriate membrane is treated with a **detergent** such as Triton X-100 to disrupt the lipid bilayer and to release the integral membrane proteins into solution (see Topic E2 for more details).

Stabilization of proteins

Throughout the purification procedure, steps have to be taken to ensure that the protein of interest is not **inactivated** or **denatured** either by physical or biological factors. The pH of the solutions used needs to be carefully **buffered** (see Topic B2) at a pH in which the protein is stable, usually around pH 7. The temperature often needs to be maintained below 25°C (usually around 4°C) to avoid **thermal denaturation** and to minimize the activity of **proteases**. Upon homogenization, **proteases** within the source material that are normally in a different subcellular compartment will be liberated into solution and come into contact with the protein of interest and may degrade it. For example, the acid hydrolases in lysosomes (see Topic A2) could be liberated into solution and rapidly degrade the protein of interest. Thus, as well as carrying out the procedures at low temperature, **protease inhibitors** are often included in the buffers used in the early stages of the isolation procedure in order to minimize unwanted proteolysis (see Topic C4).

Assay of proteins

A suitable means of detecting (**assaying**) the protein must be available to monitor the success of each stage in the purification procedure. The most straightforward **assays** are those for enzymes that catalyze reactions with readily detectable products (for more details on enzyme assays see Topic C1). Proteins which are not enzymes may be assayed through the observation of their biological effects. For example, a receptor can be assayed by measuring its ability to bind its specific ligand. Immunological techniques are often used to assay for the protein of interest using antibodies that specifically recognize it [e.g. radioimmunoassay, enzyme-linked immunosorbent assay (ELISA), or Western blot analysis (see Topics B7 and D4)].

Ammonium sulfate precipitation

A commonly employed first separation step is **ammonium sulfate precipitation**. This technique exploits the fact that the **solubility** of most proteins is lowered at high salt concentrations. As the salt concentration is increased, a point is reached where the protein comes out of solution and precipitates. The concentration of salt required for this **salting-out effect** varies from protein to protein, and thus this procedure can be used to fractionate a mixture of proteins. For example, 0.8 M ammonium sulfate precipitates out the clotting protein fibrinogen from blood serum, whereas 2.4 M ammonium sulfate is required to precipitate albumin. However, many other proteins will also precipitate out at these concentrations of ammonium sulfate. Therefore this is a relatively crude separation technique, although it often provides a convenient concentration step. Salting out is also sometimes used at later stages in a purification procedure to **concentrate** a dilute solution of the protein since the protein precipitates and can then be redissolved in a smaller volume of buffer.

Dialysis

Proteins can be separated from small molecules by dialysis through a **semi-permeable membrane** such as cellophane (cellulose acetate). **Pores** in the membrane allow molecules up to approximately 10 kDa to pass through, whereas larger molecules are retained inside the dialysis bag (*Fig. 1*). As most proteins have molecular masses greater than 10 kDa, this technique is not suitable for fractionating proteins, but is often used to remove small molecules such as salts and ammonium sulfate from a protein solution. It should be noted that at equilibrium, the concentration of small molecules inside a dialysis bag will be equal to that outside (*Fig. 1b*), and so several changes of the surrounding solution are often required to lower the concentration of the small molecule in the protein solution sufficiently.

Gel filtration chromatography

In gel filtration chromatography (**size exclusion chromatography** or **molecular sieve chromatography**), molecules are separated on the basis of their **size and shape**. The protein sample in a small volume is applied to the top of a column of **porous beads** (diameter 0.1 mm) that are made of an insoluble but highly hydrated polymer such as polyacrylamide (Bio-Gel) or the carbohydrates dextran (Sephadex) or agarose (Sepharose) (*Fig. 2a*). Small molecules can enter the **pores** in the beads whereas larger or more elongated molecules cannot. The smaller molecules therefore have a larger volume of liquid accessible to them; both the liquid surrounding the porous beads and that inside the beads. In contrast, the larger molecules have only the liquid surrounding the beads accessible to them, and thus move through the column faster, emerging out of the bottom (**eluting**) first (*Fig. 2a* and *b*). The smaller molecules move more slowly through the column and elute later. Beads of differing pore sizes are available, allowing proteins of different sizes to be effectively separated. Gel filtration chromatography is often used to de-salt a protein sample (for example to remove the ammonium sulfate after ammonium sulfate precipitation), since the salt enters the porous beads and is eluted late, whereas the protein does not

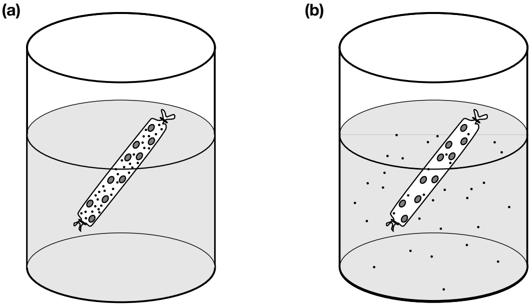

Fig. 1. Separation of molecules on the basis of size by dialysis. (a) Starting point, (b) at equilibrium.

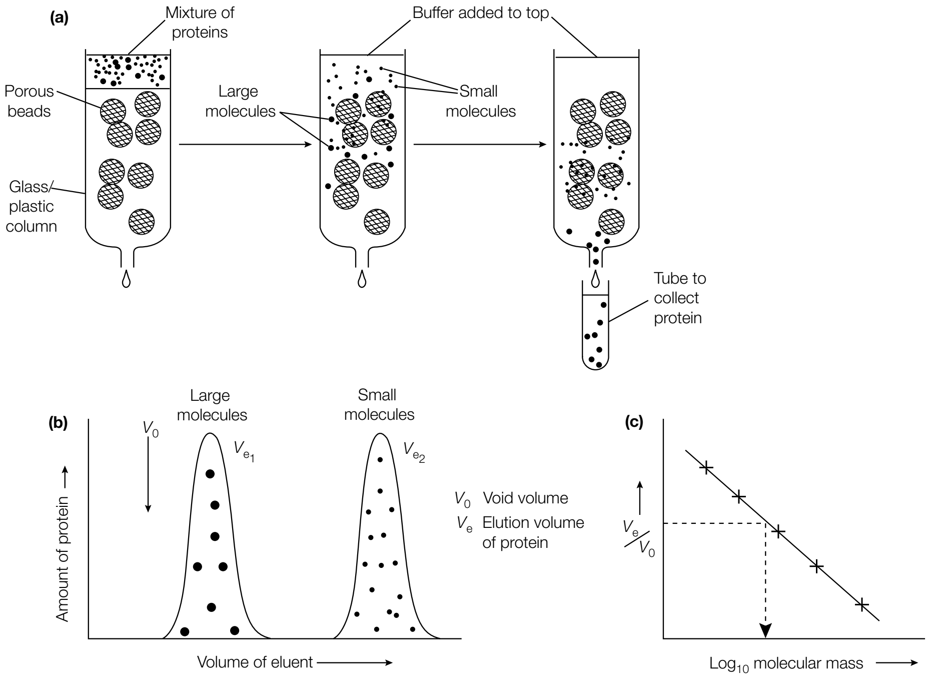

Fig. 2. Gel filtration chromatography. (a) Schematic illustration of gel filtration chromatography; (b) elution diagram indicating the separation; (c) a plot of relative elution volume versus the logarithm of molecular mass for known proteins, indicating how the molecular mass of an unknown can be read off when its relative elution volume is known.

enter the beads and is eluted early. Gel filtration chromatography can also be used to estimate the **molecular mass** of a protein. There is a linear relationship between the relative elution volume of a protein (V_e/V_o where V_e is the elution volume of a given protein and V_o is the void volume of the column, that is the volume of the solvent space surrounding the beads; *Fig. 2b*) and the logarithm of its molecular mass. Thus a 'standard' curve of V_e/V_o against $\log_{10}$ molecular mass can be determined for the column using proteins of known mass. The elution volume of any sample protein then allows its molecular mass to be estimated by reference to its position on the standard curve (*Fig. 2c*).

Ion exchange chromatography

In ion exchange chromatography, proteins are separated on the basis of their **overall (net) charge**. If a protein has a net negative charge at pH 7, it will bind to a column containing positively-charged beads, whereas a protein with no charge or a net positive charge will not bind (*Fig. 3a*). The negatively-charged proteins bound to such a column can then be eluted by washing the column with an increasing gradient (increasing concentration) of a solution of **sodium chloride** (Na^+ Cl^- ions) at the appropriate pH. The Cl^- ions compete with the protein for the positively-charged groups on the column. Proteins having a low density of negative charge elute first, followed by those with a higher density of negative charge (*Fig. 3b*). Columns containing positively-charged diethylaminoethyl (DEAE) groups (such as DEAE-cellulose or DEAE-Sephadex) are used for

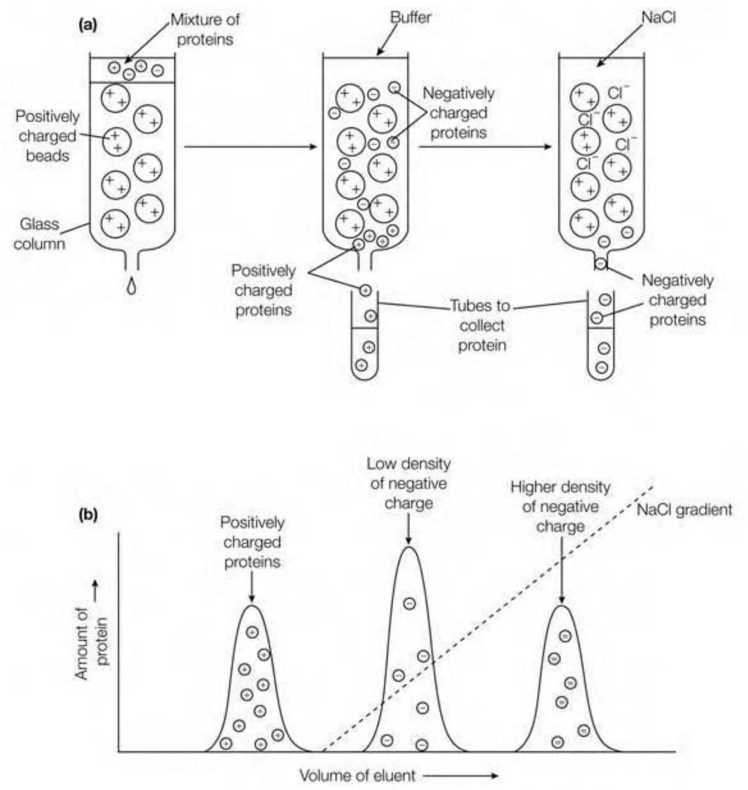
(a)
Mixture of proteins
Positively charged beads
Glass column
Buffer
Negatively charged proteins
NaCl
Cl⁻
Positively charged proteins
Tubes to collect protein
Negatively charged proteins
(b)
Low density of negative charge
Higher density of negative charge
NaCl gradient
Positively charged proteins
Amount of protein
Volume of eluent

Fig. 3. Ion exchange chromatography. (a) Schematic illustration of ion exchange chromatography; (b) elution diagram indicating the separation of a protein of net positive charge that does not bind to the positively-charged beads and passes straight through the column, and of two proteins with different net negative charges that bind to the positively-charged beads and are eluted on increasing the concentration of NaCl applied to the column. The protein with the lower density of negative charge elutes earlier than the protein with the higher density of negative charge.

separation of negatively-charged proteins (anionic proteins). This is called **anion exchange chromatography**. Columns containing negatively-charged carboxymethyl (CM) groups (such as CM-cellulose or CM-Sephadex) are used for the separation of positively-charged proteins (cationic proteins). This is called **cation exchange chromatography**. As an alternative to elution with a gradient of NaCl, proteins can be eluted from anion exchange columns by decreasing the pH of the buffer, and from cation exchange columns by increasing the pH of the buffer, thus altering the ionization state of the amino acid side-chains (see Topic B2) and hence the net charge on the protein.

Affinity chromatography

Affinity chromatography exploits the specific, high affinity, noncovalent binding of a protein to another molecule, the **ligand**. First, the ligand is covalently attached to an inert and porous matrix (such as Sepharose). The protein mixture is then passed down a column containing the **immobilized ligand**. The protein of interest will bind to the ligand, whereas all other proteins pass straight through the column (*Fig. 4*). After extensive washing of the column with buffer to remove nonspecifically bound proteins, the bound protein is released from the immobilized ligand either by adding soluble ligand which competes with the immobilized ligand for the protein, or by altering the properties of the buffer (changing the pH or salt concentration). If soluble ligand is used to elute the protein from the column, extensive **dialysis** often then has to be used to remove the small ligand from the larger protein. Because this technique exploits the specific, often unique, binding properties of the protein, it is often possible to separate the protein from a mixture of hundreds of other proteins in a single chromatographic step. Commonly employed combinations of immobilized ligand and protein to be purified used in affinity chromatographic systems include an **inhibitor** to purify an **enzyme** (see Topic C4), an **antibody** to purify its **antigen** (see Topic D4), a **hormone** (e.g. insulin) to purify its **receptor** (see Topic E5), and a **lectin** (e.g. concanavalin A) to purify a **glycoprotein** (see Topics E2 and H5). Advances in recombinant DNA technology (see Topic I1) mean that proteins can be engineered with specific sequences of amino acids at the C-terminal end, a so-called **tag**. The recombinant tagged protein can then be expressed in a suitable cell system and the affinity of the tag for an immobilized antibody or other molecule exploited to purify the protein.

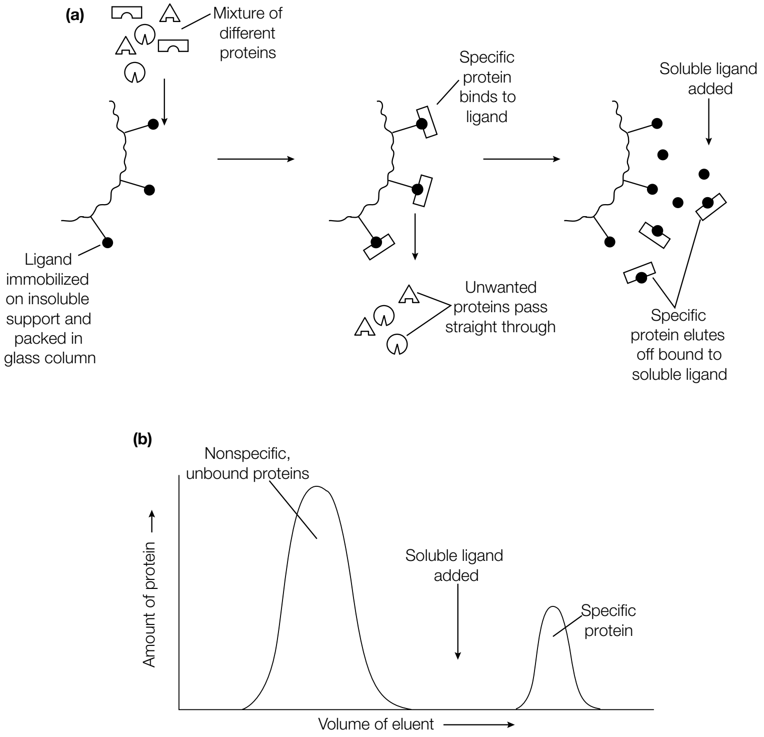
(a)
Mixture of different proteins
Ligand immobilized on insoluble support and packed in glass column
Specific protein binds to ligand
Unwanted proteins pass straight through
Soluble ligand added
Specific protein elutes off bound to soluble ligand
(b)
Nonspecific, unbound proteins
Soluble ligand added
Specific protein
Amount of protein →
Volume of eluent →

Fig. 4. Affinity chromatography. (a) Schematic diagram of affinity chromatography; (b) elution diagram indicating that nonspecific proteins that do not bind to the immobilized ligand pass straight through the column, while the specific protein binds to the immobilized ligand and is eluted from the column only on addition of soluble ligand.

Section B – Amino acids and proteins

B7 ELECTROPHORESIS OF PROTEINS

Key Notes

Electrophoresis

In polyacrylamide gel electrophoresis (PAGE) proteins are applied to a porous polyacrylamide gel and separated in an electric field on the basis of their net negative charge and their size. Small/more negatively-charged proteins migrate further through the gel than larger/less negatively-charged proteins.

SDS-PAGE

In SDS-PAGE, the protein sample is treated with a reducing agent to break disulfide bonds and then with the anionic detergent sodium dodecyl sulfate (SDS) which denatures the proteins and covers them with an overall negative charge. The sample is then fractionated by electrophoresis through a polyacrylamide gel. As all the proteins now have an identical charge to mass ratio, they are separated on the basis of their mass. The smallest proteins move farthest. SDS-PAGE can be used to determine the degree of purity of a protein sample, estimate the molecular mass of a protein and deduce the number of polypeptide subunits in a protein.

Isoelectric focusing

In isoelectric focusing, proteins are separated by electrophoresis in a gel containing polyampholytes which produce a pH gradient. They separate on the basis of their relative content of positively- and negatively-charged residues. Each protein migrates through the gel until it reaches the point where it has no net charge, its isoelectric point (pI).

Two-dimensional gel electrophoresis

In two-dimensional gel electrophoresis, proteins are subjected first to isoelectric focusing and then in the second direction to SDS-PAGE to produce a two-dimensional pattern of spots separated on the basis of charge and then mass. This technique can be used to compare the proteome of cells under different conditions.

Visualization of proteins in gels

Proteins can be visualized directly in gels by staining them with the dye Coomassie brilliant blue or with a silver stain. Radioactively-labeled proteins can be detected by overlaying the gel with X-ray film and observing the darkened areas on the developed autoradiograph that correspond to the radiolabeled proteins. A specific protein of interest can be detected by immunoblot (Western blot) following its transfer from the gel to nitrocellulose using an antibody that specifically recognizes it. This primary antibody is then detected with either a radiolabeled or enzyme-linked secondary antibody.

Related topics

Acids and bases (B2)
Protein structure (B3)
Protein purification (B6)
Protein sequencing and peptide synthesis (B8)
Antibodies as tools (D4)

Electrophoresis

When placed in an **electric field,** molecules with a net charge, such as proteins, will move towards one electrode or the other, a phenomenon known as **electrophoresis**. The greater the net charge the faster the molecule will move. In **polyacrylamide gel electrophoresis (PAGE)** the electrophoretic separation is carried out in a gel which serves as a molecular sieve. Small molecules move readily through the pores in the gel, whereas larger molecules are retarded. The gels are commonly made of **polyacrylamide** which is chemically inert and which is readily formed by the polymerization of acrylamide. The pore sizes in the gel can be controlled by choosing appropriate concentrations of acrylamide and the cross-linking reagent, methylene bisacrylamide. The higher the concentration of acrylamide used, the smaller the pore size in the final gel. The gel is usually cast between two glass plates separated by a distance of 0.5–1.0 mm (*Fig. 1*). The protein sample is added to wells in the top of the gel, which are formed by placing a plastic comb in the gel solution before it sets (*Fig. 1*). A blue dye (bromophenol blue) is mixed with the protein sample to aid its loading on to the gel. Because bromophenol blue is a small molecule, it also migrates quickly through the gel during electrophoresis and so indicates the progress of electrophoresis.

SDS-PAGE

In **sodium dodecyl sulfate (SDS)-PAGE**, the proteins are **denatured** and coated with an **overall negative charge** [due to bound sodium dodecyl sulfate (SDS) molecules] and thus the basis for their separation is only their **mass**. The protein mixture is first treated with a **reducing agent** such as 2-mercaptoethanol or dithiothreitol to break all the disulfide bonds (*Fig. 2*) (see Topic B3). The strong **anionic detergent SDS** is then added which disrupts nearly all the noncovalent interactions in the protein, unfolding the polypeptide chain. Approximately one molecule of SDS binds via its hydrophobic alkyl chain to the polypeptide

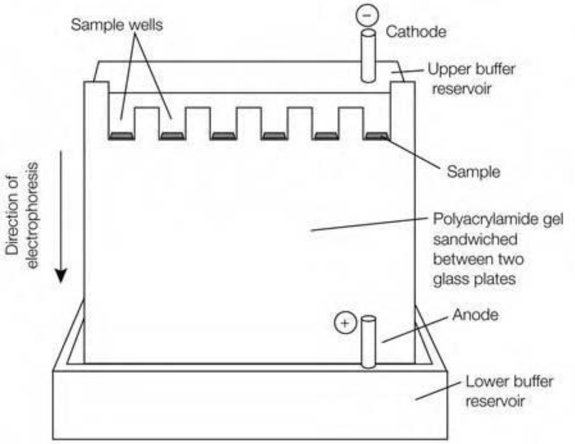
Sample wells
−
Cathode
Upper buffer reservoir
Sample
Direction of electrophoresis
Polyacrylamide gel sandwiched between two glass plates
+
Anode
Lower buffer reservoir

Fig. 1. Polyacrylamide gel electrophoresis. The protein samples are loaded into the sample wells formed in the top of the gel. An electric field is applied across the gel from top to bottom and the proteins migrate down through the gel. The smaller the protein the further it will migrate.

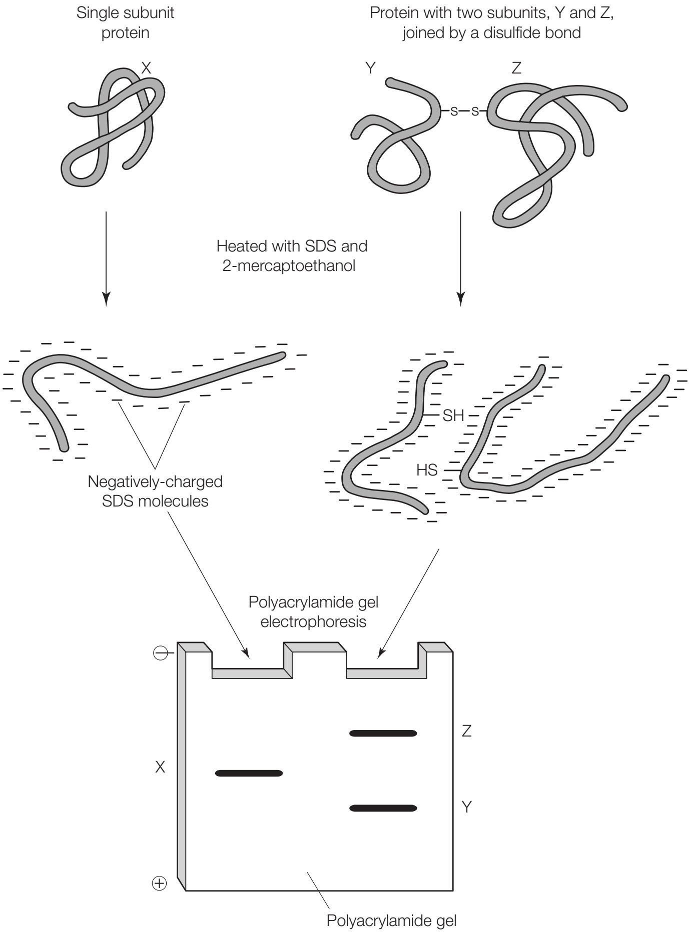
Single subunit protein
Protein with two subunits, Y and Z, joined by a disulfide bond
X
Y
Z
S–S
Heated with SDS and 2-mercaptoethanol
SH
HS
Negatively-charged SDS molecules
Polyacrylamide gel electrophoresis
⊖
Z
X
Y
⊕
Polyacrylamide gel

Fig. 2. SDS-PAGE. The protein mixture is heated in the presence of 2-mercaptoethanol, which breaks any disulfide bonds, and SDS. The unfolded polypeptide chains are coated with the negatively-charged molecules of SDS and will migrate towards the anode on polyacrylamide gel electrophoresis. Smaller polypeptides migrate further through the gel than larger ones.

backbone for every two amino acid residues, which gives the denatured protein a large net negative charge that is proportional to its mass. The SDS/protein mixture is then applied to sample wells in the top of a **polyacrylamide gel** (*Fig. 1*). The buffer, which is the same in both the upper and lower reservoirs and in the gel, has a pH of approximately 9, such that the proteins have **net negative charge** and will migrate towards the anode in the lower reservoir. An electric current (approximately 300 V) is applied across the gel from top to bottom for 30–90 min in order to move the proteins through the gel (*Fig. 1*). After carrying out electrophoresis, the gel is removed from the apparatus and the proteins visualized (*Fig. 3a*). Small proteins move furthest through the gel, whereas large ones move more slowly as they are held back by the cross-linking in the gel. Under these conditions, the mobility of most polypeptide chains is linearly proportional to the logarithm of their mass. Thus, if proteins of known molecular mass are electrophoresed alongside the samples, the mass of the unknown proteins can be determined as there is a linear relationship between $\log_{10}$ of molecular mass and distance migrated through the gel (*Fig. 3b*). Proteins that differ in mass by about 2% (e.g. 40 and 41 kDa; a difference of approximately 10 amino acid residues) can be distinguished under appropriate conditions. SDS-PAGE is a rapid, sensitive and widely-used technique which can be used to determine the degree of purity of a protein sample, to estimate the molecular mass of an unknown protein and to deduce the number of polypeptide subunits within a protein (see Topic B3).

Isoelectric focusing

Isoelectric focusing electrophoretically separates proteins on the basis of their relative content of positively and negatively charged groups. When a protein is at its **pI** (see Topic B2), its **net charge is zero** and hence it will not move in an electric field. In isoelectric focusing, a polyacrylamide gel is used which has large pores (so as not to impede protein migration) and contains a mixture of **polyampholytes** (small multicharged polymers that have many pI values). If an electric field is applied to the gel, the polyampholytes migrate and produce a **pH gradient**. To separate proteins by isoelectric focusing, they are electrophoresed through such a gel. Each protein will migrate through the gel until it reaches a position at which the pH is equal to its pI (*Fig. 4*). If a protein diffuses away from this position, its net charge will change as it moves into a region of different pH and the resulting electrophoretic forces will move it back to its isoelectric position. In this way each protein is focused into a narrow band (as thin as 0.01 pH unit) about its pI.

Two-dimensional gel electrophoresis

Isoelectric focusing can be combined with SDS-PAGE to obtain very high resolution separations in a procedure known as **two-dimensional gel electrophoresis**. The protein sample is first subjected to isoelectric focusing in a narrow strip of gel containing polyampholytes (*see Fig. 4*). This isoelectric focusing gel strip is then placed on top of an SDS-polyacrylamide gel and electrophoresed to

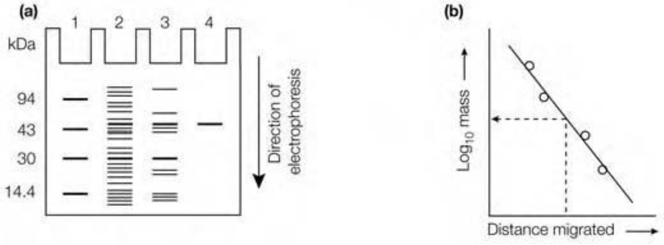

Fig. 3. SDS-PAGE. (a) Appearance of proteins after electrophoresis on an SDS polyacrylamide gel. Lane 1, proteins (markers) of known molecular mass; lane 2, unpurified mixture of proteins; lane 3, partially purified protein; lane 4, protein purified to apparent homogeneity; (b) determination of the molecular mass of an unknown protein by comparison of its electrophoretic mobility (distance migrated) with those of proteins (markers) of known molecular mass.

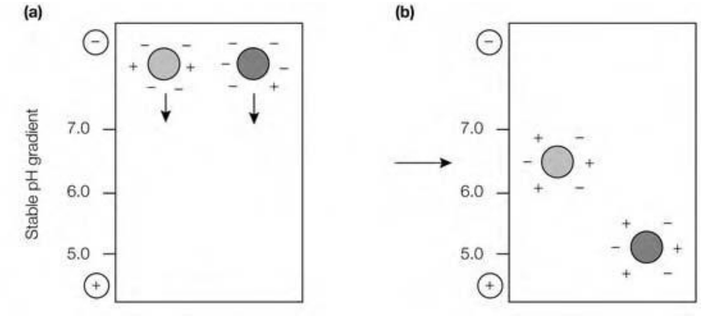

Fig. 4. Isoelectric focusing. (a) Before applying an electric current. (b) After applying an electric current the proteins migrate to a position at which their net charge is zero (isoelectric point, pI).

produce a two-dimensional pattern of spots in which the proteins have been **separated in the horizontal direction on the basis of their pI**, and **in the vertical direction on the basis of their mass** (*Fig. 5*). The overall result is that proteins are separated both on the basis of their size and their charge. Thus two proteins that have very similar or identical pIs, and produce a single band by isoelectric focusing, if they have different molecular masses will produce two spots by two-dimensional gel electrophoresis (see *Fig. 5*). Similarly, proteins with similar or identical molecular masses, which would produce a single band by SDS-PAGE, will also produce two spots if they have different pIs because of the initial separation by isoelectric focusing. The **high resolution separation** of proteins in a complex mixture that can be achieved by two-dimensional gel electrophoresis makes this technique extremely useful for comparing the **proteome** (the entire complement of proteins in a cell or organism) of cells or tissues under different conditions, e.g. differentiated versus undifferentiated cells, and is often used prior to the analysis of individual protein spots by **mass spectrometry** (see Topic B8).

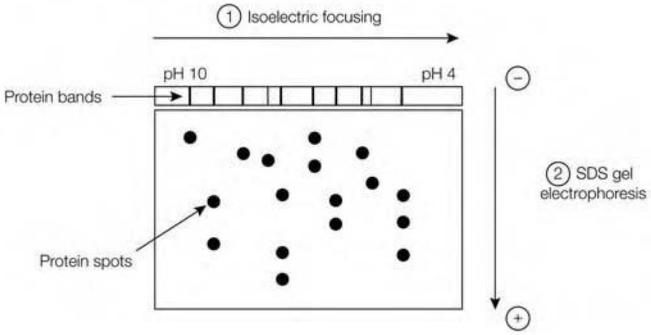

Fig. 5. Two-dimensional gel electrophoresis. The protein sample is first subjected to isoelectric focusing in one dimension and then to SDS-PAGE in the second dimension.

Visualization of proteins in gels

As most proteins are not directly visible on gels to the naked eye, a method has to be employed in order to visualize them following electrophoresis. The most commonly used protein stain is the dye **Coomassie brilliant blue**. After electrophoresis, the gel containing the separated proteins is immersed in an acidic alcoholic solution of the dye. This **denatures** the proteins, **fixes** them in the gel so that they do not wash out, and allows the dye to bind to them. After washing away excess dye, the proteins are visible as discrete blue bands (see *Fig. 3a*). As little as 0.1–1.0 μg of a protein in a gel can be visualized using Coomassie brilliant blue. A more sensitive general protein stain involves soaking the gel in a **silver** salt solution. However, this technique is rather more difficult to apply. If the protein sample is **radioactive** the proteins can be visualized indirectly by overlaying the gel with a sheet of **X-ray film**. With time (hours to weeks depending on the radioactivity of the sample proteins), the radiation emitted will cause a darkening of the film. Upon development of the film the resulting **autoradiograph** will have darkened areas corresponding to the positions of the radiolabeled proteins.

Another way of visualizing the protein of interest is to use an **antibody** against the protein in an **immunoblot** (**Western blot**) (see Topic D4 for more detail). For this technique, the proteins have to be transferred out of the gel on to a sheet of **nitrocellulose** or **nylon membrane**. This is accomplished by overlaying the gel with the nitrocellulose and **blotting** the protein on to it by applying an electric current. The nitrocellulose then has an exact image of the pattern of proteins that was in the gel. The excess binding sites on the nitrocellulose are then blocked with a nonspecific protein solution such as milk powder, before placing the nitrocellulose in a solution containing the antibody that recognizes the protein of interest (the **primary antibody**). After removing excess unbound antibody, the primary antibody that is now specifically bound to the protein of interest is detected with either a **radiolabeled**, **fluorescent** or **enzyme-coupled secondary antibody**. Finally, the secondary antibody is detected either by placing the nitrocellulose against a sheet of X-ray film (if a radiolabeled secondary antibody has been used), by using a fluorescence detector or by adding to the nitrocellulose a solution of a substrate that is converted into a colored insoluble product by the enzyme that is coupled to the secondary antibody.

B8 Protein sequencing and peptide synthesis

Key Notes

Amino acid composition analysis

The number of each type of amino acid in a protein can be determined by acid hydrolysis and separation of the individual amino acids by ion exchange chromatography. The amino acids are detected by colorimetric reaction with, for example, ninhydrin or fluorescamine.

Edman degradation

The N-terminal amino acid of a protein can be determined by reacting the protein with dansyl chloride or fluorodinitrobenzene prior to acid hydrolysis. The amino acid sequence of a protein can be determined by Edman degradation which sequentially removes one residue at a time from the N terminus. This uses phenyl isothiocyanate to label the N-terminal amino acid prior to its release from the protein as a cyclic phenylthiohydantoin amino acid.

Sequencing strategy

In order to sequence an entire protein, the polypeptide chain has to be broken down into smaller fragments using either chemicals (e.g. cyanogen bromide) or enzymes (e.g. chymotrypsin and trypsin). The resulting smaller fragments are then sequenced by Edman degradation. The complete sequence is assembled by analyzing overlapping fragments generated by cleaving the polypeptide with different reagents. The polypeptides in a multisubunit protein have to be dissociated and separated prior to sequencing using urea or guanidine hydrochloride which disrupt noncovalent interactions, and 2-mercaptoethanol or dithiothreitol that break disulfide bonds.

Protein fingerprint

Following digestion of a protein with trypsin, the resulting peptide map is diagnostic of the protein and is referred to as the protein’s fingerprint.

Mass spectrometry

Matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) spectrometry is used to determine the precise mass of peptides. The peptides are immobilized in an organic matrix and then blasted with a laser, causing them to be ejected in the form of an ionized gas. The ionized peptides in the gas are then accelerated in an electric field and separated. Tandem mass spectrometry (MS-MS) uses two mass spectrometers in tandem to fragment the peptides further.

Proteomics

Proteomics is the study of the entire complement of proteins, the proteome, in a cell or organism.

Recombinant DNA technology

The sequence of a protein can be determined using recombinant DNA technology to identify and sequence the piece of DNA encoding the protein. The amino acid sequence of the protein can then be deduced from its DNA sequence using the genetic code.

Information derived from protein sequences

The amino acid sequence of a protein not only reveals the primary structure of the protein but also information on possible protein families or groups and evolutionary relationships, potential gene duplication(s) and possible post-translational modifications. In addition, a knowledge of the amino acid sequence can be used to generate specific antibodies.

Peptide synthesis

In solid phase peptide synthesis, polypeptides are chemically synthesized by addition of free amino acids to a tethered peptide. To prevent unwanted reactions, the α-amino group and reactive side-chain groups of the free amino acids are chemically protected or blocked, and then deprotected or deblocked once the amino acid is attached to the growing polypeptide chain.

Related topics

Amino acids (B1)
Myoglobin and hemoglobin (B4)
Protein purification (B6)
Electrophoresis of proteins (B7)
Antibodies: an overview (D2)
Antibodies as tools (D4)
The genetic code (H1)
Protein glycosylation (H5)
The DNA revolution (I1)

Amino acid composition analysis

The number of each type of **amino acid** in a protein sample can be determined by amino acid composition analysis. The purified protein sample is hydrolyzed into its constituent amino acids by heating it in 6 M HCl at 110°C for 24 h in an evacuated and sealed tube. The resulting mixture (**hydrolysate**) of amino acids is subjected to **ion exchange chromatography** (see Topic B6) on a column of sulfonated polystyrene to separate out the **20 standard amino acids** (see Topic B1). The separated amino acids are then detected and quantified by reacting them with **ninhydrin**. The α-amino acids produce a blue color, whereas the imino acid proline produces a yellow color. The amount of each amino acid in an unknown sample can be determined by comparison of the optical absorbance with a known amount of each of the individual amino acids in a standard sample. With ninhydrin, as little as 10 nmol of an amino acid can be detected. A more sensitive detection system (detecting down to 10 pmol of an amino acid) uses **fluorescamine** to react with the α-amino group to form a fluorescent product. Amino acid composition analysis indicates the number of each amino acid residue in a peptide, but it does not provide information on the sequence of the amino acids. For example, the amino acid composition of the oligopeptide:

Val-Phe-Asp-Lys-Gly-Phe-Val-Glu-Arg

would be:

(Arg, Asp, Glu, Gly, Leu, Lys, Phe$_2$, Val$_2$)

where the parentheses and the commas between each amino acid denote that this is the amino acid composition, not the sequence.

Edman degradation

The **amino-terminal** (N-terminal) residue of a protein can be identified by reacting the protein with a compound that forms a stable covalent link with the free α-amino group, prior to hydrolysis with 6 M HCl. The labeled N-terminal amino acid can then be identified by comparison of its chromatographic properties with standard amino acid derivatives. Commonly used reagents for N-terminal analysis are **fluorodinitrobenzene** and **dansyl chloride**. If this technique was applied to the oligopeptide above, the N-terminal residue would be identified as Val, but the remainder of the sequence would still be unknown. Further reaction with dansyl chloride would not reveal the next residue in the sequence since the peptide is totally degraded in the acid hydrolysis step.

This problem was overcome by Pehr Edman who devised a method for labeling the N-terminal residue and then cleaving it from the rest of the peptide without breaking the peptide bonds between the other amino acids. In so-called **Edman degradation**, one residue at a time is sequentially removed from the N-terminal end of a peptide or protein and identified. The uncharged N-terminal amino group of the protein is reacted with **phenyl isothiocyanate** to form a phenylthiocarbamoyl derivative which is then released from the rest of the protein as a cyclic **phenylthiohydantoin (PTH) amino acid** under mildly acidic conditions (*Fig. 1*). This milder cleavage reaction leaves the remainder of the peptide intact, available for another round of labeling and release. The released PTH amino acid is identified by **high performance liquid chromatography** (HPLC). This sequencing technique has been **automated** and refined so that upwards of 50 residues from the N-terminus of a protein can be sequenced from picomole quantities of material.

Sequencing strategy

An 'average' sized protein of 50 kDa would contain approximately 500 amino acids. Thus, even with large amounts of highly purified material, only about the N-terminal one-tenth of the protein can be sequenced by Edman degradation. In

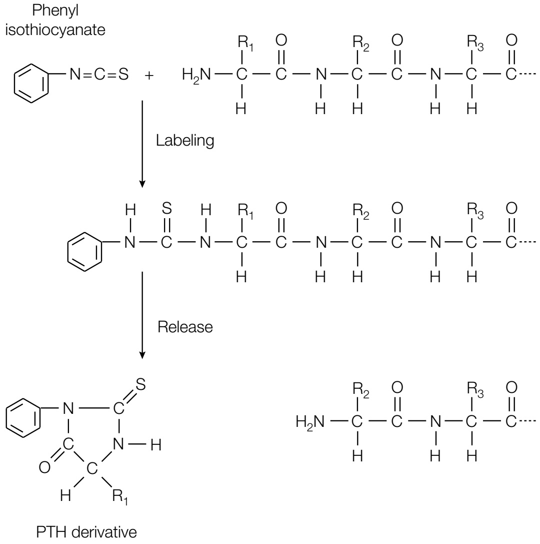

Fig. 1. Edman degradation. The N-terminal amino acid is labeled with phenyl isothiocyanate. Upon mild acid hydrolysis this residue is released as a PTH-derivative and the peptide is shortened by one residue, ready for another round of labeling and release.

order to sequence a larger protein, the first step is to **cleave** it into **smaller fragments** of 20–100 residues which are then separated and sequenced. Specific cleavage can be achieved by **chemical** or **enzymatic** methods. For example, the chemical **cyanogen bromide** (CNBr) cleaves polypeptide chains on the C-terminal side of Met residues, whereas the enzymes **trypsin** and **chymotrypsin** cleave on the C-terminal side of basic (Arg, Lys) and aromatic (Phe, Trp, Tyr) residues, respectively. On digestion with trypsin, a protein with six Lys and five Arg would yield 12 tryptic peptides, each of which would end with Arg or Lys, apart from the C-terminal peptide. The peptide fragments obtained by specific chemical or enzymatic cleavage are then separated by **chromatography** (e.g. ion exchange chromatography; see Topic B6) and the sequence of each in turn determined by Edman degradation.

Although the sequence of each peptide fragment would now be known, the order of these fragments in the polypeptide chain would not. The next stage is to generate **overlapping fragments** by cleaving another sample of the original polypeptide chain with a different chemical or enzyme (e.g. chymotrypsin), separating the fragments and then sequencing them. These **chymotryptic peptides** will overlap one or more of the **tryptic peptides**, enabling the order of the fragments to be established (*Fig. 2*). In this way, the entire length of the polypeptide chain can be sequenced.

To sequence the polypeptides in a **multisubunit protein**, the individual polypeptide chains must first be dissociated by disrupting the **noncovalent interactions** with **denaturing agents** such as **urea** or **guanidine hydrochloride**. The **disulfide bonds** in the protein also have to be broken by reduction with **2-mercaptoethanol** or **dithiothreitol**. To prevent the cysteine residues recombining, **iodoacetate** is added to form stable *S*-carboxymethyl derivatives. The individual polypeptide chains then have to be separated by, for example, ion exchange chromatography (see Topic B6) before sequencing each. Nowadays, as little as picomole amounts of proteins can be sequenced following their separation by SDS-PAGE either using the polyacrylamide gel containing the protein directly, or following their transfer to nitrocellulose (see Topic B7).

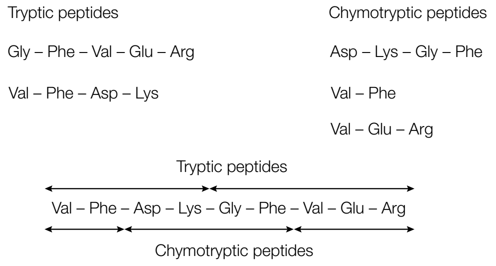

Fig. 2. The use of overlapping fragments to determine the sequence of a peptide. The protein is first digested with trypsin and the resulting peptides separated and sequenced. The protein is separately digested with chymotrypsin and the resulting peptides again separated and sequenced. The order of the peptide fragments in the protein can be determined by comparing the sequences obtained.

To summarize the key steps in the chemical sequencing of a protein:

1. Purify the protein.
2. Reduce any disulfide bonds and block their re-oxidation with iodoacetate.
3. Cleave the protein with cyanogen bromide or a protease.
4. Separate the fragments and sequence them.
5. Look for overlaps between the two sets of sequences in order to construct the full sequence.

Protein fingerprint

Following digestion of a protein with trypsin or another reagent, the resulting mixture of peptides can be separated by chromatographic or electrophoretic procedures (see Topics B6 and B7). The resulting pattern, or **peptide map**, is diagnostic of the protein from which the peptides were generated and is referred to as the protein's **fingerprint**. Comparison of protein fingerprints can be used to identify mutations in a protein as the altered amino acid may change the properties of one of the peptides.

Mass spectrometry

The precise mass of intact proteins and peptides derived from them can be determined by **mass spectrometry**. This is a very sensitive technique that requires only very small amounts of material. The most commonly used mass spectrometric method is called **matrix-assisted laser desorption ionization-time-of-flight spectrometry (MALDI-TOF)**. In this method, peptides are first mixed with an organic acid and then dried onto a ceramic or metal slide. The sample is then blasted with a **laser** which causes the peptides to be ejected from the slide in the form of an ionized gas in which each molecule carries one or more positive charges (*Fig. 3a*). The ionized peptides are then accelerated in an electric field and fly toward a detector. The time it takes for them to reach the detector is determined by their mass and their charge; large peptides move more slowly, and more highly charged peptides move more quickly. The precise mass is then determined by analysis of those peptides with a single charge. If a mixture of tryptic peptides is used, then the resulting masses measured in the MALDI-TOF can be used to search **protein sequence databases** for matches with theoretical massess calculated for all trypsin-released peptides for all proteins in a sequenced genome (*Fig. 3b*). Thus, the identity of the original protein and its sequence can readily be determined by a combination of mass spectrometry and protein sequence database searching.

A variation of this method can be used directly to determine the sequences of the individual peptides. Following trypsin digestion of the purified protein and determination of their masses by mass spectrometry as above, each peptide is further fragmented at the peptide bonds and the masses of these fragments measured in a coupled second mass spectrometer. This is so-called **tandem mass spectrometry (MS-MS)**. The mass differences between the fragments can be used to construct a partial amino acid sequence which, in turn, can be used to search protein sequence databases or provide the means for cloning the gene.

Sequencing of proteins by mass spectrometry has several advantages over traditional chemical Edman sequencing:

- Much smaller amounts of material are required.
- The sequence of the peptide can be obtained in only a few minutes compared with the hour required for just one cycle of Edman degradation.
- Mass spectrometry can be used to sequence several polypeptides in a mixture, alleviating the need to completely purify the sample prior to analysis.

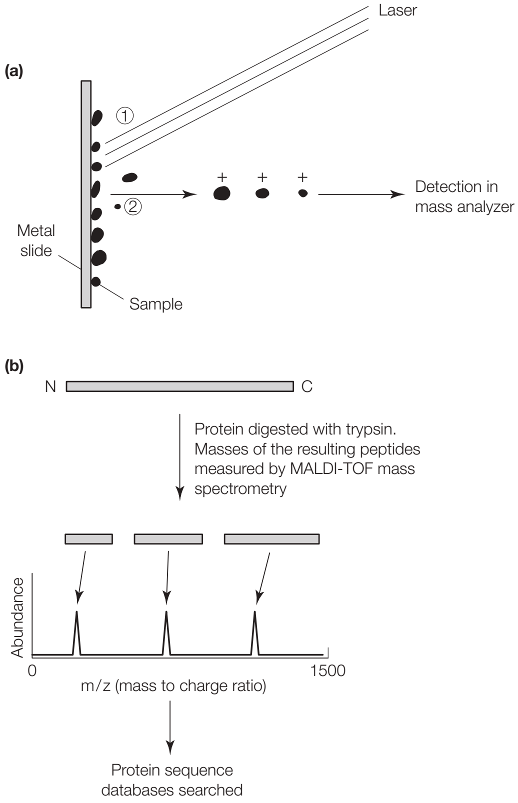
(a)
Laser
①
②
Detection in mass analyzer
Metal slide
Sample
(b)
N
C
Protein digested with trypsin. Masses of the resulting peptides measured by MALDI-TOF mass spectrometry
Abundance
0
1500
m/z (mass to charge ratio)
Protein sequence databases searched

Fig. 3. Mass spectrometry to determine protein sequence. (a) In a MALDI-TOF mass spectrometer, pulses of light from a laser ionize a peptide mixture that is absorbed on a metal slide (1). An electric field accelerates the molecules in the sample toward the detector (2). The time to the detector is inversely proportional to the mass of the protein. (b) Use of mass spectrometry to identify proteins. The protein of interest is digested with trypsin and the resulting peptide fragments are loaded into the mass spectrometer where their masses are measured. Sequence databases are then searched to identify the protein whose calculated tryptic digest profile matches the experimentally determined data.

- Mass spectrometry can be used to determine the sequence of peptides which have **blocked N-termini**, such as pyroglutamate, a derivative of glutamate in which the side-chain carboxyl group forms an amide bond with its primary amino group (a common eukaryotic post-translational modification that prevents Edman degradation) and to characterize other **post-translational modifications** such as glycosylation and phosphorylation.

Proteomics

The **proteome** is the entire complement of proteins in a cell or organism. **Proteomics** is the large-scale effort to identify and characterize all of the proteins encoded in an organism's genome, including their post-translational modifications. Increasingly, in the field of proteomics, proteins resolved by two-dimensional gel electrophoresis (see Topic B7) are subjected to trypsin digestion and the extremely accurate molecular masses of the peptides produced are used as a '**fingerprint**' to identify the protein from **databases** of real or predicted tryptic peptide sizes.

Recombinant DNA technology

Although numerous proteins have been sequenced by Edman degradation and mass spectrometry, the determination of the complete sequences of large proteins by these methods is a demanding and time-consuming process. **Recombinant DNA technology** (see Topic I1) has enabled the sequences of even very large proteins or of proteins that are difficult to purify to be determined by first sequencing the stretch of **DNA** encoding the protein and then using the **genetic code** to decipher the protein sequence (see Topic H1). Even so, some direct protein sequence data is often required to confirm that the protein sequence obtained is the correct one and to identify any post-translational modifications on the protein. Thus, protein sequencing and DNA sequencing techniques are often used together to determine the complete sequence of a protein.

Information derived from protein sequences

The amino acid sequence can provide information over and above the **primary structure** of the protein.

1. The sequence of interest can be compared with other known sequences to see whether there are similarities. For example, the sequences of hemoglobin and myoglobin indicate that they belong to the globin group or **family of proteins** (see Topic B4).
2. The comparison of the sequences of the same protein in different species can provide information about **evolutionary relationships**.
3. The presence of repeating stretches of sequence would indicate that the protein may have arisen by **gene duplication** (e.g. in antibody molecules; see Topic D2).
4. Within the amino acid sequence there may be specific sequences which act as signals for the **post-translational processing** of the protein (e.g. glycosylation or proteolytic processing; see Topic H5).
5. The amino acid sequence data can be used to prepare **antibodies** specific for the protein of interest which can be used to study its structure and function (see Topic D4).
6. The amino acid sequence can be used for designing **DNA probes** that are specific for the gene encoding the protein (see Topics I3 and I4).

Peptide synthesis

Polypeptides can be **chemically synthesized** by covalently linking amino acids to the end of a growing polypeptide chain. In **solid phase peptide synthesis** the growing polypeptide chain is covalently anchored at its C-terminus to an insoluble support such as polystyrene beads. The next amino acid in the sequence has to react with the free α-amino group on the tethered peptide, but it has a free α-amino group itself which will also react. To overcome this problem the free amino acid has its α-amino group **chemically protected** (blocked) so that it does not react with other molecules. Once the new amino acid is coupled, its now N-terminal α-amino group is **deprotected** (deblocked) so that the next peptide bond can be formed. Every cycle of amino acid addition therefore requires a **coupling step** and a **deblocking step**. In addition, reactive side-chain groups must also be blocked to prevent unwanted reactions occurring.