2

Protein Structure

Introduction

Proteus, a Greek sea god and keeper of all knowledge, would not give up information easily. Even while being held down, he would struggle mightily, assuming different forms before revealing any of his secrets. Although proteins were not named after Proteus, the description could not be more appropriate.

Proteins are complex macromolecules made up of successive amino acids that are covalently bonded together in a head-to-tail arrangement through substituted amide linkages called peptide bonds. Each protein molecule is composed of an exact sequence of amino acids arranged in a linear, unbranched fashion. Protein molecules have the property of acquiring a distinctive 3-dimensional configuration.

The amino acid backbone may have many posttranslational modifications contributing to the size, charge and function of the mature protein. A functionally active protein, being the product of posttranslational structural modifications, cannot be determined by reference to any single gene. Rather, the active form is often the result of complex biochemical reactions performed on the target protein that are not all controlled by any one gene product.

A. The Amino Acids

A representative amino acid is shown in Figure 2.1. The α-amino acids contain an α-carbon to which an amino group ($—NH_3^+$) and a carboxylate group ($—COO^-$) are attached. The 20 amino acids that make up the building blocks of proteins differ in the structure of their R groups, which may be hydrophilic or hydrophobic, acidic, basic, or neutral. Of the 20 amino acids normally used to build proteins, 19 have the general structure shown in Figure 2.1. Proline, the 20th natural amino acid, is similar but has the side chain bonded to the nitrogen atom to give the amino acid. The chemical composition of the unique R groups of the amino acids is responsible for the most important

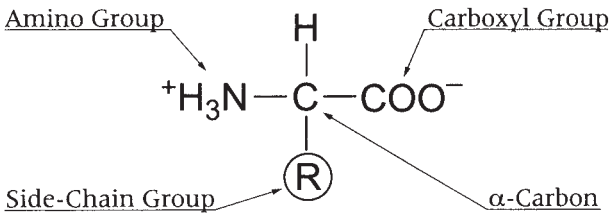


FIGURE 2.1 A schematic version of the general formula of an amino acid. At pH 7.0, both the amino and carboxyl groups are ionized.

characteristics of the amino acids: chemical reactivity, ionic charge, and relative hydrophobicity.

The charged amino acids may be either acidic or basic. At low pH, proteins are positively charged due to the basic groups on lysine and arginine, whereas, at high pH, proteins are negatively charged due to the acidic groups on aspartic and glutamic acids.

The peptide backbone of proteins is composed of amino acids having polar, non-polar, aromatic and charged residues. In their native states, proteins are compact structures with only a small number of amino acids exposed to the surface. The amino acids that are exposed are well suited for their microenvironment. For proteins in an aqueous milieu charged and polar residues are commonly exposed. For proteins that are embedded in membranes, nonpolar, lipophilic amino acids are found in the interface.

The amino acids are presented in Table 2.1 according to their polarity and charge. There are eight amino acids with nonpolar hydrophobic R groups. Five of them have aliphatic hydrocarbon R groups (alanine, leucine, isoleucine, valine and proline), two have aromatic rings (phenylalanine and tryptophan) and one of the nonpolar amino acids contains sulfur (methionine). As a group, these amino acids are less soluble in water than the polar amino acids.

The polar amino acids are more soluble in water because their R groups can form hydrogen bonds with water. The polarity of serine, threonine, and tyrosine is contributed by their hydroxyl groups (—OH); that of asparagine and glutamine by their carboxy-amide groups and cysteine by its sulfhydryl group (—SH). Glycine falls in this group by default. Its R group, a hydrogen atom, is too small to influence the high degree of polarity contributed by the α-amino and carboxyl groups.

Amino acids that carry a net negative charge at pH 6.0–7.0 contain a second carboxyl group. These are the acidic amino acids, aspartic acid, and glutamic acid. The basic amino acids, in which the R groups have a net positive charge at pH 7.0, are lysine, arginine, and histidine which is weakly basic. At pH 6.0 more than 50% of the histidine molecules are positively charged (protonated), but at pH 7.0 less than 10% have a positive charge.

Historically, the amino acids were designated by three-letter abbreviations. Subsequently, to make them more computer-friendly, a set of one-letter symbols has been adopted to facilitate comparative display

TABLE 2.1 Amino Acids

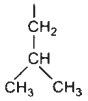
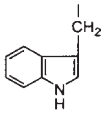
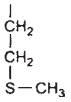
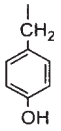
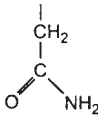
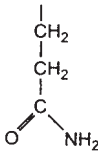
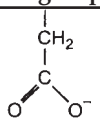
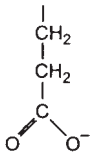
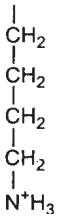
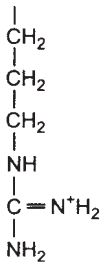
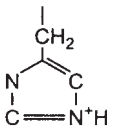
Amino acid	Triple letter code	Single letter code	R group
Amino acids with nonpolar R groups			
Alanine	Ala	A	
Valine	Val	V	
Leucine	Leu	L	
Isoleucine	Ile	I	
Proline	Pro	P	
Phenylalanine	Phe	F	
Tryptophan	Trp	W	
Methionine	Met	M	
Amino acids with uncharged polar R groups			
Glycine	Gly	G	
Serine	Ser	S	
Threonine	Thr	T	
Cysteine	Cys	C	
Tyrosine	Tyr	Y	
Asparagine	Asn	N	
Glutamine	Gln	Q	

TABLE 2.1 *Continued*

Amino acid	Triple letter code	Single letter code	R group
Acidic amino acids (negatively charged)			
Aspartic Acid	**Asp**	**D**	
Glutamic Acid	**Glu**	**E**	
Basic amino acids (positively charged)*			
Lysine	**Lys**	**K**	
Arginine	**Arg**	**R**	
Histidine	**His**	**H**	

*at physiological pH.

of amino acid sequences of homologous proteins. General information about the amino acids is presented in Tables 2.1 and 2.2.

B. The Four Levels of Protein Structure

The structure of a protein can be resolved into several different levels, each of which is built upon the one below it in a hierarchical fashion. Proteins are so precisely built that the change of a few atoms in one amino acid can disrupt the structure and have catastrophic consequences.

Primary Structure

Linderstrøm–Lang and his coworkers (1959) were the first to recognize structural levels of organization within a protein. They introduced the

TABLE 2.2 Hydropathy Values, Solubility in Water, and pK Values of the Amino Acids

Amino acid	Hydropathy index*	Solubility (g/100 ml water)	Side Chain pK	α-NH_2 pK	α-COOH pK
Isoleucine	4.5	3.36		9.76	2.32
Valine	4.2	5.6		9.72	2.29
Leucine	3.8	2.37		9.6	2.36
Phenylalanine	2.8	2.7		9.24	2.58
Cysteine	2.5	unlimited	8.33	10.78	1.71
Methionine	1.9	5.14		9.21	2.28
Alanine	1.8	15.8		9.87	2.35
Glycine	–0.4	22.5		9.13	2.17
Threonine	–0.7	unlimited		9.12	2.15
Tryptophan	–0.9	1.06		9.39	2.38
Serine	–0.8	36.2		9.15	2.21
Tyrosine	–1.3	0.038	10.1	9.11	2.2
Proline	–1.6	154		10.6	1.99
Histidine	–3.2	4.19	6.0	8.97	1.78
Glutamic acid	–3.5	0.72	4.25	9.67	2.19
Glutamine	–3.5	2.6		9.13	2.17
Aspartic acid	–3.5	0.42	3.65	9.6	1.88
Asparagine	–3.5	2.4		8.8	2.02
Lysine	–3.9	66.6	10.28	8.9	2.2
Arginine	–4.5	71.8	13.2	9.09	2.18

*The higher the magnitude of the hydropathy index number the more hydrophobic the amino acid. Hydropathy values from Kyte J and Doolittle RF (1982).

terms primary, secondary, and tertiary structure. Primary structure refers to the sequence of amino acids that make up a specific protein. The number, chemical nature, and sequential order of amino acids in a protein chain determine the distinctive structure and confer the characteristics that define its chemical behavior. Although most protein sequences have been established by direct amino acid sequence analysis, the vast majority of primary sequences are being deduced directly from the DNA sequence. Often, one method serves to confirm the other.

The covalent bond that links amino acids together is called a peptide bond. Depicted in Figure 2.2, the peptide bond is formed by a reaction between the α-NH_3^+ group of one amino acid and the α-COO^- group

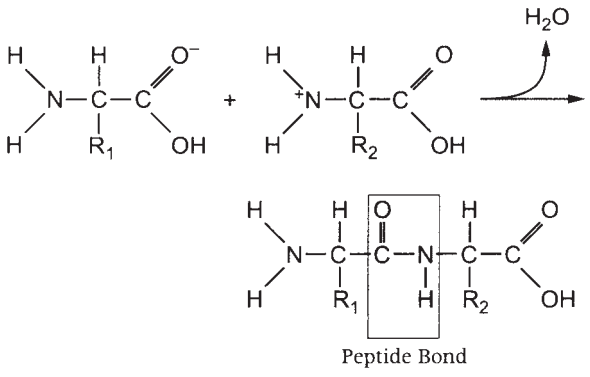


FIGURE 2.2 The formation of the peptide bond; a condensation reaction between two amino acids resulting in a loss of water. The peptide bond is highlighted.

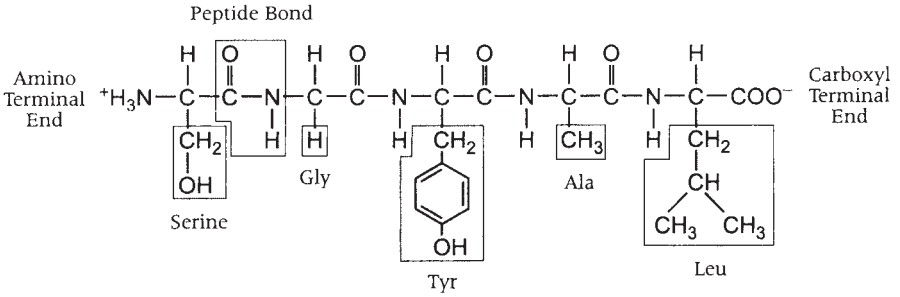


FIGURE 2.3 A schematic linear representation of a pentapeptide.

of another amino acid. One water molecule is removed as each peptide bond forms. The polypeptide backbone is simply a linear ordered array of amino acid units incorporated into a polypeptide chain. All proteins and polypeptides have this fundamental linear order and aside from the modifications to the amino acids, differ only in the number of amino acids linked together in the chain and in the sequence in which the various amino acids occur in the polypeptide chain. Additional covalent bonds, found primarily in secreted and cell-surface proteins, are disulfide bonds (also called S–S bonds) between cysteine residues. These bonds are stable at physiological pH in the absence of reductants and oxidants.

The terminal residue of a polypeptide chain which contains a free α-NH_3^+ group is referred to as the amino terminal or N-terminal residue. The carboxyl-terminal or C-terminal residue has a free COO^- group. In biological systems, proteins are synthesized from NH_3^+ terminal to COO^- terminal, and the accepted convention is to write the amino acid sequence of a polypeptide from left to right starting with the N-terminal residue. A short peptide composed of five amino acids is shown in Figure 2.3.

The term peptide generally refers to a structure with only a small number (typically 2–20) of amino acids linked together. The term polypeptide generally refers to longer chains. The term protein applies to those chains with a specific sequence, length, and folded conformation in their native states. Denatured long chain polypeptides that have lost their quaternary structure are also referred to as proteins.

Secondary Structure

The polypeptide chain of a protein is folded into a specific 3-dimensional structure producing a protein's unique conformation. Secondary structure refers to the regular local structural arrays found in proteins that can be referred to as independent folding units. The secondary structure is determined by the chemical interactions (mainly hydrogen bonding) of amino acid residues with other amino acids in close proximity. The secondary structure is identifiable as substructures, usually α-helices and β-structures. The α-helix, shown in Figure 2.4, is a common secondary structure in proteins. Helices are characterized by their pitch (rise per residue), period (number of residues per turn), handedness (right or left), and diameter. The typical α-helix is right-

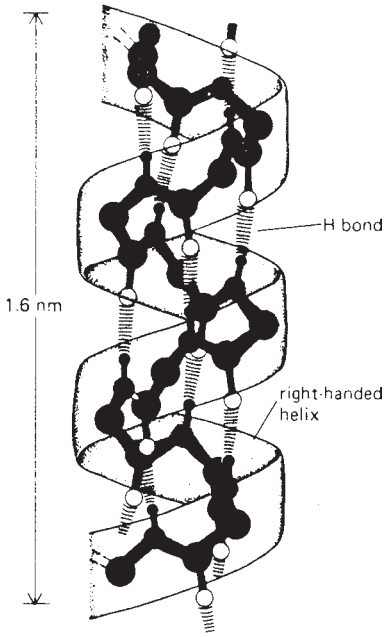


FIGURE 2.4 An α helix. Note the extensive hydrogen bonding. (Reprinted from Alberts et al, 1983 with permission).

handed with a pitch of 1.5 Å and a period of 3.6 residues per turn. All side chains point away from the helix. The length per turn is 5.4 angstroms or 1.5 angstroms per residue. Neglecting the side chains, the α-helix has a diameter of about 6 angstroms. The diameter of the α-helix is sufficiently small that the structure is more like a filled cylinder than an open spring (Cohen, 1993). The lengths of α-helices vary in proteins, but on average, there are ~10 residues per helical segment. Theoretically, helices can be right- or left-handed. However, α-helices comprised of L-amino acids are never left-handed. If a right-handed helix were a spiral staircase and you were climbing up, the banister would be on your right-hand side. Some amino acids do not accommodate themselves to helices whereas others lead to helix formation.

There are variations on the α-helix in which the chain is either more tightly or more loosely coiled. Hydrogen bonds between corresponding groups that are closer or further apart in the primary structure by one residue are designated the 3_{10} helix or the π helix, respectively. One or two turns of the 3_{10} helix is a common occurrence at the ends of helical portions of proteins. In the 3_{10} helix, every n and n + 3 amino acid is linked by hydrogen bonds. The name arises because there are 10 atoms in the ring closed by the hydrogen bond, and there are only three residues per turn. The 3_{10} helix is only about 5% as abundant as the α-helix.

Two types of β pleated sheet or β structure (Figure 2.5) commonly occur in proteins. Antiparallel pleated sheets are formed by extended, adjacent segments of polypeptide chain, whose sequences with respect to the direction from $—NH_3^+$ to $—COO^-$ run in the opposite direction. Antiparallel sheets may be formed by a single chain folding back on itself or by two or more segments of chain from remote parts of the

molecule. Chains running in the same direction form the parallel pleated sheet.

Along a single β strand, the amino acid side chains are positioned alternately up and down. When β strands are assembled into a sheet, the side chains are aligned in rows with all the side chains in a row pointing up or down. This is true for both parallel and antiparallel β sheets.

The polypeptide chains in many proteins often fold sharply back upon themselves, giving rise to secondary structures called β-bends, which are stabilized by hydrogen bonds. The amino acids occurring most frequently at the bend are glycine, proline, aspartic acid, asparagine, and tryptophan. An example of a polypeptide chain spontaneously folding to a more favorable conformation is shown in Figure 2.6.

Proteins are generally built up from combinations of secondary structure elements, α-helices and β–sheets, connected by loop regions of various lengths. Secondary structural elements combine in ways that result in the formation of a stable hydrophobic core. They are often arranged in motifs, simple, commonly occurring super-secondary elements with characteristic geometric arrangements. Secondary structure elements and motifs combine to form domains. A domain is part of the polypeptide chain that can fold independently into a stable

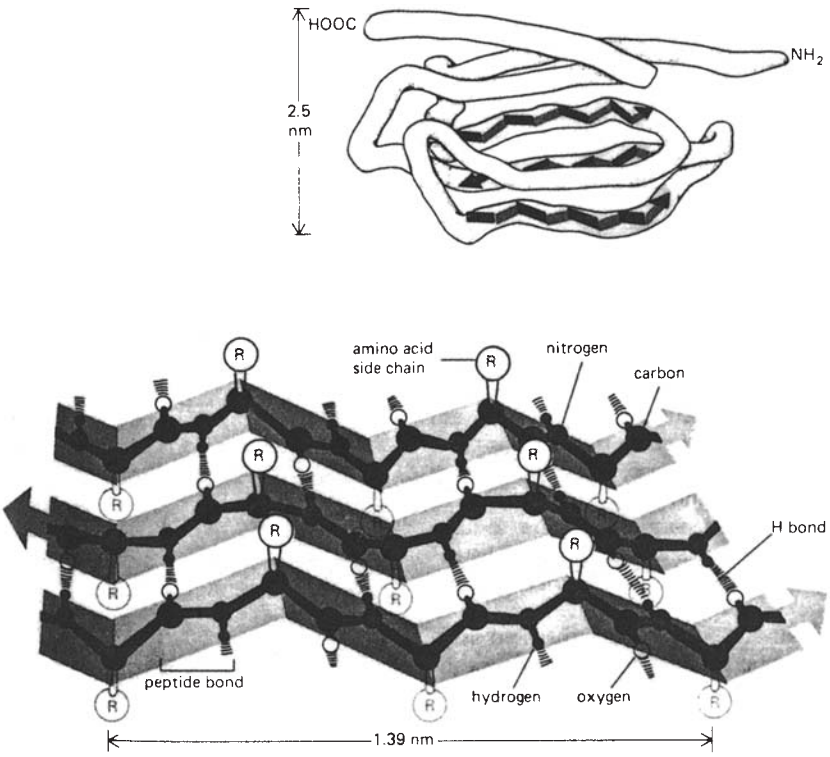


FIGURE 2.5 An idealized antiparallel β sheet is shown in detail. Note that every peptide bond is hydrogen-bonded to a neighboring peptide bond. (Reprinted from Alberts et al, 1983 with permission).

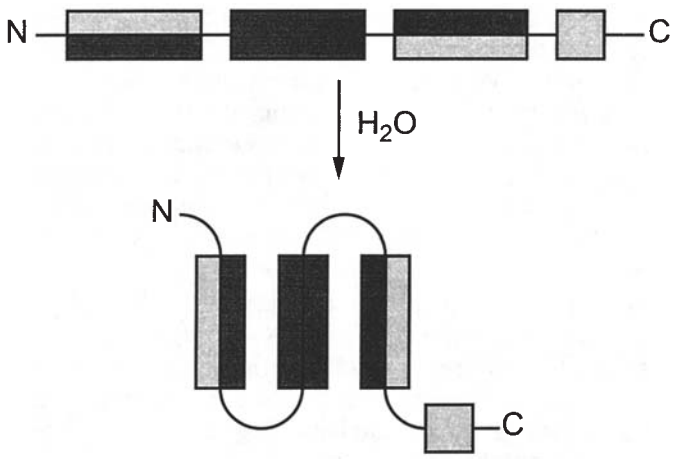


FIGURE 2.6 Folding of a protein chain in water. The hydrophobic segments (shaded areas) rearrange to form a compact core surrounded by the hydrophilic portions. (Reprinted from Fletterick et al, 1985 with permission).

tertiary structure. A protein can consist of a single domain or multiple domains.

In the evolution of proteins, domains are important units that have been shuffled, duplicated and fused to create larger, more complex proteins. Although the possibilities of distinct amino acid sequences are unlimited, the number of different folding patterns for the domains is not. Extrapolation based on existing databases of protein sequence and structure indicates that most of the natural domain sequences assume one of a few thousand folds (Govindarajan et al, 1999), of which ~1,000 are already known (Orengo et al, 2002).

Tertiary Structure

Unraveling the pathway by which unstructured proteins spontaneously fold to their native functional form has been a central goal of protein chemists since before Anfinsen's landmark paper (1973). The sequences of existing proteins have been selected through evolution not only to adopt a functional 3-dimensional structure after folding but also to optimize the protein folding process both temporally and spatially, given the constraints of the cellular context. An attempt to understand the complicated structure of a polypeptide is greatly simplified by realizing that much of the complex 3-dimensional (tertiary) structure can be described as an assembly of regular secondary structural elements. Tertiary structure is the intramolecular arrangement of the secondary structure independent folding units with respect to each other. The 3-dimensional organization of several hundred polypeptide chains has been revealed by crystallography and nuclear magnetic resonance spectroscopy. This level of organization is determined by the noncovalent interactions between helices and β-structures together with the side-chain and backbone interactions unique to a given protein. The tertiary structure can be inferred from an analysis of the packing of these secondary structural elements. Essential for the tertiary structures is a delicate balance between many noncovalent inter-

actions: hydrophobic regions formed by nonpolar R groups of the amino acids; ionic; van der Waals interactions; and hydrogen bonds. Disulfide bonds are a major force as they most likely stabilize the conformation after folding occurs. These bonds form spontaneously when the appropriate thiol groups are brought into juxtaposition as the result of cooperative interactions of the R groups that lead to correct folding.

Quaternary Structure

Many proteins in solution exist as aggregates of two or more polypeptide chains, either identical or different. Polypeptide chains are designated by letters, an example being hemoglobin, referred to as $\alpha_2\beta_2$. Many proteins are dimers, trimers, tetramers, or even higher-order aggregates of identical polypeptide chains. Quaternary structure refers to the stoichiometry and spatial arrangement of the subunits and is determined by the amino acid sequences of the subunits. The subunit stoichiometry of a protein is the number of each type of polypeptide that has combined to produce the specific structure. Each subunit interacts with the other subunits through hydrophobic and polar interactions suggesting that the interacting surfaces must be highly complementary. Dissociated subunits can recombine to give a functionally competent native protein. Not only do the specific amino acids confer secondary and tertiary structure but the quaternary structure is also dictated by the amino acid sequence of the polypeptide chains.

Two fundamental types of interactions can take place between identical subunits, isologous and heterologous, as demonstrated in Figure 2.7A,B. In an isologous interaction, the interacting surfaces are identical, giving rise to a closed, dimeric structure (Morgan et al, 1979). In the heterologous association, the interfaces that interact are not identical. The surfaces must be complementary but need not be symmetric (Degani and Degani, 1980). This association is potentially open-ended unless the geometry of the interaction is such as to produce a closed cyclical structure as shown in Figure 2.7C.

A multimeric protein is composed of more than one folded polypeptide. Each of the polypeptides composing such a protein folds to form a tertiary structure unique to the amino acid sequence of that polypeptide. These folded polypeptides will then associate with each other to form the multimeric protein. The arrangement in space of these folded polypeptides in the protein is its quaternary structure.

A protein quaternary structure database exists that contains ~10,000 structurally defined proteins of presumed biological importance (http://pqs.ebi.ac.uk). Each assembly consists of at least two protein chains (Sali et al, 2003).

C. Chemical Characteristics of Proteins

Proteins have ionic and hydrophobic sites both internally (within the folds of the tertiary structure) and on the surface where the primary structure comes in contact with the environment. The ionic sites of a

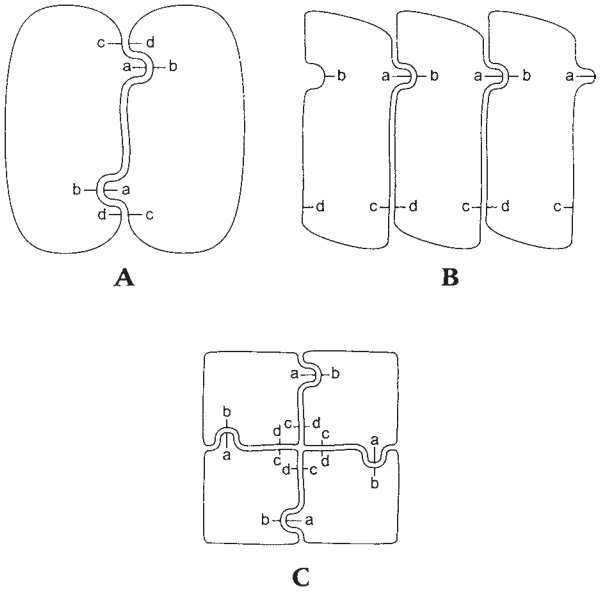


FIGURE 2.7 Schematic illustration of isologous and heterologous association between protein subunits. (A) Isologous association to form a dimer. (B) Heterologous association leading to an infinitely long polymer. (C) Heterologous association to form a closed, finite structure, in this case a tetramer. (Adapted from Monod et al, 1965 with permission).

protein are provided by the charged amino acids at physiological pH and by covalently attached modifying groups (e.g., carbohydrates and phosphate).

Proteins are polyelectrolytes containing positively and negatively charged groups. The net charge on a protein is contributed by the free α-amino group of the N-terminal residue, the free α-carboxyl group of the C-terminal residue, those R groups capable of ionization, and by the unique array of modifications attached to the protein. In proteins, the ionizing R groups greatly out-number the single ionizing groups at the two terminal residues. At physiological pH the α-COO^- and α-NH_3^+ groups are ionized, with the deprotonated carboxyl group bearing a negative charge and the protonated amino group a positive charge. Therefore, an amino acid in its dipolar state is called a zwitterion.

At a certain pH, referred to as the isoelectric point (pI), the numbers of positive and negative charges on a protein are equal and the protein is electrically neutral. A protein has a net positive charge at pH values below its pI and a net negative charge above the pI. If a protein has a high pI (e.g., 10), this implies that it is basic and that the excess positive charges contributed by arginine and lysine residues are rendered neutral at a high pH. Conversely, a protein with a low pI is rendered neutral when the charges are neutralized at low pH, which occurs when the protein is in an environment with many free H^+ ions which will protonate the negative charges on aspartic and glutamic acid residues.

Hydrophobicity

The hydrophobic character of a protein is due to nonpolar amino acids (e.g., valine and phenylalanine) and covalently attached hydrophobic groups (e.g., lipids and fatty acids). The hydrophobicity of a protein can be changed by partially denaturing the protein and exposing the protein's interior primary structure which contains most of the hydrophobic amino acids.

The final folded structure of a protein is a thermodynamic compromise between allowing hydrophilic side-chains access to the aqueous solvent and minimizing contact between hydrophobic side-chains and water. It follows that the interiors of water-soluble proteins are predominantly composed of hydrophobic amino acids, while the hydrophilic side-chains are on the exterior where they can interact with water.

A hydropathy (strong feeling about water) index, presented in Table 2.2, has been devised where each amino acid is assigned a value reflecting its relative hydrophilicity or hydrophobicity. An appropriate evaluation of a given amino acid sequence should be able to predict whether a given peptide segment is sufficiently hydrophobic to interact with or reside within the interior of the membrane which itself is hydrophobic. A computer program has been devised that uses a moving-segment approach that continuously determines the average hydropathy within a segment of predetermined length as it advances through the sequence (Kyte and Doolittle, 1982). The consecutive scores are plotted from the amino to the carboxy terminus. The procedure gives a graphic visualization of the hydropathic character of the chain from one end to the other. A midpoint line is printed that corresponds to the grand average of the hydropathy of the amino acid compositions found in most sequenced proteins, a universal midline. For most proteins there is an excellent agreement between the interior portions of the protein and the region appearing on the hydrophobic side of the midpoint line, as well as the exterior regions appearing on the hydrophilic side of the line. Potential membrane spanning segments can be identified by this procedure.

Figure 2.8 shows the hydropathy profile of erythrocyte glycophorin, illustrating the concept of hydropathic indices. Glycophorin has an easily recognizable membrane-spanning segment in the region of residues 75–94. As can be seen from Figure 2.8, a positive hydropathic stretch is indicative of hydrophobic amino acids.

Consensus Sequences

It is clearly not sufficient to determine the primary structure of a protein or deduce it from the DNA sequence and expect that this will explain all the properties of a protein. Elucidation of the complete covalent structure of a protein includes the knowledge of the primary structure, as described above, and the chemical nature and positions of all the modifications to the protein that take place in the cell and are necessary for its correct function, regulation, and antigenicity.

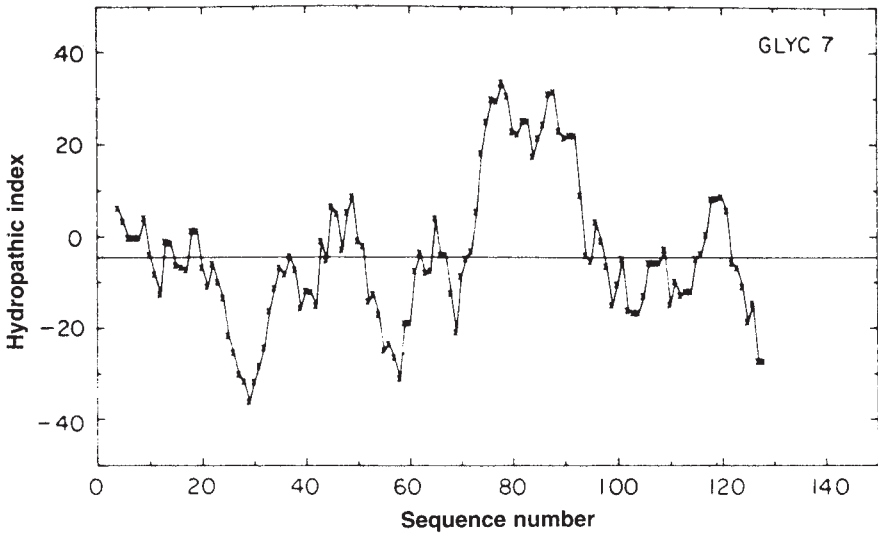


FIGURE 2.8 Hydropathy profile of erythrocyte glycophorin which has an easily recognizable membrane-spanning segment in the region of residues 75 to 94. (Reprinted from Kyte and Doolittle, 1982 with permission).

When an investigator has succeeded in defining the amino acid sequence of the target protein or translates a cDNA sequence, a careful analysis of the primary amino acid structure may yield a great deal of structural and functional information about the protein. Consensus sequences have been defined for many posttranslational modifications and motifs which may suggest a function for a previously undefined protein.

The term "consensus sequence" refers to those sequence elements containing and surrounding the site to be modified. These sequence elements are considered essential for its recognition. It generally takes the form of a short linear sequence of amino acids indicating the identity of the minimum set of amino acids comprising such a site (Kennelly and Krebs, 1991).

The sequences of many proteins contain short, conserved motifs which are involved in recognition and targeting. These sequences, which appear in the primary, linear structure of the protein, are often separate from other functional properties of the molecule in which they are found. They do not include elements from different polypeptide chains or from widely scattered portions of a single polypeptide chain. Therefore, they are not the result of distant segments being brought together as the protein assumes its native conformation. The conservation of these motifs varies; some are highly conserved while others permit substitutions that retain only a certain pattern of charge across the motif.

The usefulness of consensus sequence analysis is based on its simplicity. Summarizing the complexities of the substrate recognition sequence as sets of short sequences has facilitated the evaluation of a large body of observations. A word of caution: the existence of an

apparent consensus sequence does not assure that a protein is modified. Consensus sequence information functions best as a guide whose implications must be confirmed or refuted experimentally. A partial list of consensus sequences for a variety of posttranslational modifications and cell functions is presented in Appendix E.

Proteomics

A View of the Proteome

The annotation of the human genome indicates that the number of genes is between 30,000 and 40,000. However, the estimated number of proteins encoded by these genes is two to three orders of magnitude higher. Therefore, organism complexity is generated more by a complex proteome than by a complex genome.

Proteomics is many things to many investigators. The proteome is defined as the time and cell specific protein complement of the genome. This encompasses all proteins that are expressed in a cell at one time, including isoforms and protein modifications. Therefore, expression proteomics aims to measure up-and-down-regulation of protein levels. Whereas the genome is constant for one cell, being largely identical for all cells of an organism, the proteome is very dynamic with time and in response to external factors, and differs markedly between cell types (Rappsilber and Mann, 2002).

Every cell type or tissue has its own proteome, which represents only part of the genome. There are many ways that the same polypeptide backbone can be posttranslationally altered making a famous biological dogma, one gene one enzyme, put forth by Beadle and Tatum, no longer tenable.

Functional proteomics attempts to characterize cellular components, multiprotein complexes, and signaling pathways. Proteomics attempts to provide a snapshot of protein synthesis rate, degradation rate, posttranslational modification, subcellular distribution and physical interactions with other cell components.

Several diverse mechanisms can result in the expression of many protein variants from the same gene locus: single nucleotide polymorphisms (SNPS), gene splicing, alternative splicing of pre-mRNA, proteolytic cleavage, and co-and posttranslational modification of amino acids. Even though the difference between two proteins can be very small, a single amino acid change or the modification of a single amino acid, this difference could be crucial for the function of the protein, classically demonstrated by hemoglobin where a single amino acid substitution results in sickle cell anemia.

Protein abundance cannot be predicted from mRNA abundance, and posttranslational modifications cannot be predicted from deduced amino acid sequence. Simply predicting a polypeptide sequence can even be problematic because of alternative splicing of mRNAs or frameshifts (Gygi et al, 1999). The identity of a protein must therefore be defined by its structural formula (the connectivity of all atoms), not just the amino acid sequence, but any and all of the modifications.

Proteomics would not be possible without the previous achievements of genomics that provided the "blueprint" for all possible gene products (Tyers and Mann, 2003). However, unlike DNA analysis, which spawned technological advances like automated sequencing and the polymerase chain reaction, proteomics must deal with a set of problems for which technological shortcuts do not as yet exist. The biological protein analyst must deal with the problems of limited and variable sample material, degradation and posttranslational modifications due to developmental and temporal conditions.

One fundamental challenge to proteomics is the accurate detection of proteins that are present in vanishingly small amounts. The abundance of individual proteins differs by as much as four orders of magnitude. Low abundance proteins such as transcription factors and kinases that are present at 1–2000 copies per cell represent molecules that perform important regulatory functions (Hucho and Buchner, 1997). The protein diversity that can result from a single gene demands a more precise analysis.

An organized effort has begun to develop an infrastructure in proteomics that is aimed at unraveling the complexity of the proteome in health and disease. To this end the Human proteome Organization (HUPO, http://www.hupo.org) was founded. HUPO's mission is threefold: to consolidate proteome organizations into a worldwide organization, to encourage the spread of proteomics technologies; and to coordinate proteome initiatives aimed at characterizing specific cell and tissue proteomes (Hanash, 2003).

Developing Proteomic Technologies

Some of the major technology platforms that have developed along with proteomics include 2-D gel electrophoresis and mass spectrometry. Historically, 2-D electrophoresis provided the technology that illustrated the potential for cataloging expressed proteins in databases (Celis et al, 1996). After obtaining the protein fraction, the method of choice for proteomic studies is one- or 2-D gel electrophoresis. The advantage of one dimensional SDS-PAGE is that virtually all proteins are soluble in SDS. Also, the range of relative molecular mass from 10 kD–300 kD is covered, and extremely acidic and basic proteins are easily visualized.

Because even the best 2-D gels can routinely resolve no more than 1,000 proteins, it is clear that only the most abundant proteins can be visualized if a crude protein mixture is used. Proteins present in low abundance, like regulatory proteins, will probably go undetected if a total cell lysate is used. Applying larger amounts of sample is usually not an option since the resolution will breakdown if the gel is overloaded. One solution is to enrich for the target protein by using a purification step prior to electrophoresis. For example, a 66 kD protein that is present at about 1,000 copies per cell, a protein of medium abundance, less than 2 picomoles (100 ng) in one liter of cell culture would not be visualized from whole cell extracts. But if it could be affinity enriched it could probably be detected by silver stain. For 2-D gel electrophoresis-based proteomics the protein mixture is fractionated

by 2-D electrophoresis. 2-D image analysis software is commercially available. Typically, the software packages can digitize images from 2-D gels, detect, quantify, and identify spots of differential intensity between gels.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD, eds. (1983): *Molecular Biology of the Cell.* New York: Garland Publishing, Inc.

Anfinsen CB (1973): Principles that govern the folding of protein chains. *Science* 181:223–230

Celis JE, Gromov P, Ostergaard M, Madsen P, Honore B, Vandekerckhove J, Rasmussen HH (1996): Human 2-D PAGE databases for proteome analysis in health and disease. *FEBS Lett* 398:120–134

Cohen FE (1993): The parallel β helix of pectate lyase C: Something to sneeze at. *Science* 260:1444–1445

Degani Y, Degani C (1980): Enzymes with asymmetrically arranged subunits. *Trends Biochem Sci* 5:337–341

Fletterick RJ, Schroer T, Matela RJ, Staples J, eds. (1985): *Molecular Structure: Macro molecules in Three Dimensions.* Boston: Blackwell Scientific Publications

Govindarajan S, Recabarren R, Goldstein RA (1999): Estimating the total number of protein folds. *Proteins* 35:408–414

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999): Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 19:1720–1730

Hanash S (2003): Disease proteomics. *Nature* 422:226–232

Hucho F, Buchner K (1997): Signal transduction and protein kinases: the long way from the plasma membrane into the nucleus. *Naturwissenschaften* 84:281–290

Kennelly PJ, Krebs EG (1991): Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. *J Biol Chem* 266:15555–15558

Kyte J, Doolittle RIF (1982): A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157:105–132

Linderstrøm–Lang KU, Schellman JA (1959): Protein structure and enzyme activity. In: *The Enzymes, Vol. 1.* Boyer PD, Lardy H, Myrbäck K, eds. New York: Academic Press

Monod J, Wyman J, Changeux J-P (1965): On the nature of allosteric transitions: A plausible model. *J Mol Biol* 12:88–118

Morgan RS, Miller SL, McAdon JM (1979): The symmetry of self-complementary surfaces. *J Mol Biol* 127:31–39

Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D, Pearl FM, Sillitoe I, Todd AE, Thornton JM (2002): The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* 2:11–21

Rappsilber J, Mann M (2002): What does it mean to identify a protein in proteomics? *Trends Biochem Sci* 27:74–78

Sali A, Glaeser R, Earnest T, Baumeister W (2003): From words to literature in structural proteomics. *Nature* 422:2126–225

Tyers M, Mann M (2003): From genomics to proteomics. *Nature* 422:193–197

General References

Creighton TE (1984): *Protein Structures and Molecular Properties*. New York: W H Freeman Publishing Co.

Rossmann MG, Argos P (1981): Protein folding. *Ann Rev Biochem* 50:497–532

http://www.springer.com/978-0-8176-4340-9

Protein Analysis and Purification
Benchtop Techniques
Rosenberg, I.M.
2005, XXVI, 520 p., Hardcover
ISBN: 978-0-8176-4340-9
A product of Birkhäuser Basel