2 Biophysical and Biochemical Characteristics of Therapeutic Proteins

Wim Jiskoot and Daan J. A. Crommelin

INTRODUCTION

For a recombinant human protein to become a therapeutic product, its biophysical and biochemical characteristics must be well understood. These properties serve as a basis for understanding the behavior of the protein under various circumstances, e.g., for establishing the range of conditions to properly purify the protein and to stabilize it during production, storage and shipping.

PROTEIN STRUCTURE

Primary Structure

Most proteins which are developed for therapy perform specific functions by interacting with other small and large molecules, e.g., cell-surface receptors (mostly proteins), nucleic acids, carbohydrates, and lipids. The functional properties of proteins are derived from their folding into distinct three-dimensional structures. Each protein fold is based on its specific polypeptide sequence in which different natural amino acids are connected through peptide bonds in a specific way. This alignment of the 20 amino acid residues, called a primary sequence, has in general all the information necessary for folding into a distinct tertiary structure comprising different secondary structures such as α-helices and β-sheets (see below). Because the 20 amino acids possess different side chains, polypeptides with widely diverse properties are obtained.

All of the natural amino acids consist of a C_α carbon to which an amino group, a carboxyl group, a hydrogen, and a side chain are covalently attached. All natural amino acids, except glycine (having a proton as the side chain), are chiral and have an L-configuration (Fig. 2.1). In a polypeptide these amino acids are joined by condensation to yield peptide bonds consisting of the C_α-carboxyl group of an amino acid joined with the C_α-amino group of the next amino acid (Fig. 2.2).

The condensation gives an amide (NH) group at the N-terminal side of C_α and a carbonyl (C=O) group at the C-terminal side. These groups, as well as the side chains, play important roles in protein folding. Owing to their ability to form hydrogen bonds, they make major energetic contributions to the formation of two important secondary structures, α-helix and β-sheet. The peptide bonds between various amino acid residues are very much equivalent, however, so that they do not determine which part of a sequence should form an α-helix or β-sheet. Sequence-dependent secondary structure formation is determined by the side chains.

The 20 natural amino acids commonly found in proteins are shown in Fig. 2.3. They are described by their full names and three- and one-letter codes. Their side chains are structurally different in such a way that at physiological pH values, aspartic and glutamic acid are negatively charged and lysine and arginine are positively charged. At pH 7.4, a minor fraction of the histidine side chains is positively charged (pKa = 6). Tyrosine and cysteine are protonated and uncharged at physiological pHs, but become negatively charged above pH 10 and 8, respectively.

This text is a revised and abbreviated version of the chapter by Tsutomu Arakawa and John S. Pilo in the fourth, previous edition of this book. The discussion of techniques to physicochemically characterize protein structures was taken out and forms now a separate chapter, Chap. 3.

W. Jiskoot (✉)
Division of BioTherapeutics, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Leiden, The Netherlands
e-mail: w.jiskoot@lacdr.leidenuniv.nl

D. J. A. Crommelin
Department of Pharmaceutics, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
e-mail: d.j.a.crommelin@uu.nl

D. J. A. Crommelin et al. (eds.), *Pharmaceutical Biotechnology*, https://doi.org/10.1007/978-3-030-00710-2_2

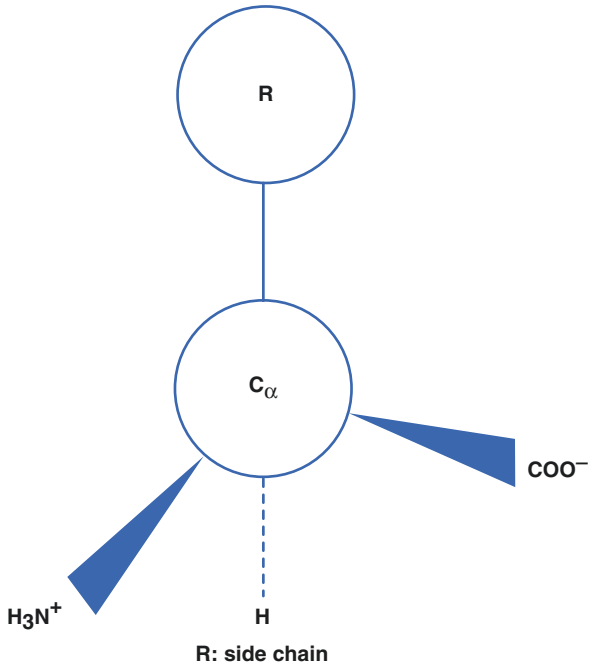

Figure 2.1 ■ Structure of L-amino acids

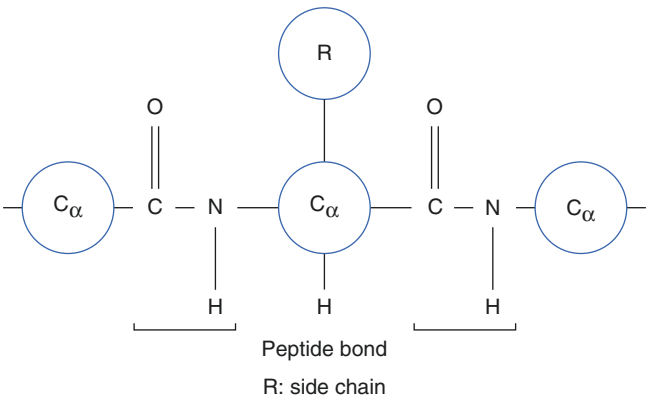

Figure 2.2 ■ Schematical structure of two sequential peptide bonds

Polar amino acids consist of serine, threonine, asparagine, and glutamine, as well as cysteine, while nonpolar amino acids consist of alanine, valine, phenylalanine, proline, methionine, leucine, and isoleucine. Glycine behaves neutrally while cystine, the oxidized form of two cysteines (i.e., a Cys-Cys, or disulfide bridge), is characterized as hydrophobic. Although tyrosine and tryptophan often enter into polar interactions, they are better characterized as nonpolar, or hydrophobic, as described later.

These 20 amino acids are incorporated into a unique sequence based on the genetic code, as the example in Fig. 2.4 shows. This is an amino acid sequence of human granulocyte-colony-stimulating factor (G-CSF), which selectively regulates proliferation and maturation of neutrophils. Although several properties of this protein depend on the location of each amino acid and hence the location of each side chain in the three-dimensional structure, some properties can be estimated simply from the amino acid composition, as shown in Table 2.1.

Using the pK_a values of these side chains, one amino terminus and one carboxyl terminus, one can calculate total charges (positive plus negative charges) and net charges (positive minus negative charges) of a protein as a function of pH, i.e., a titration curve. Since cysteine can be oxidized to form a disulfide bond or can be in a free form, accurate calculation above pH 8 requires knowledge of the status of cysteinyl residues in the protein. The titration curve thus obtained is only an approximation, since some charged residues may be buried and the effective pKa values depend on the local environment of each residue. Nevertheless, the calculated titration curve gives a first approximation of the overall charged state of a protein at a given pH and hence its solubility (Cf. Fig. 5.5, Chap. 5). Other molecular parameters, such as isoelectric point (pI, where the net charge of a protein becomes zero), molecular weight, extinction coefficient, partial specific volume, and hydrophobicity, can also be estimated from the amino acid composition, as shown in Table 2.1.

The primary structure of a protein, i.e., the sequence of the 20 amino acids, can lead to the three-dimensional structure because the various amino acids have diverse physicochemical properties. As an example, Fig. 2.5 shows a cartoon of the three-dimensional structure of filgrastim (recombinant human G-CSF). Each type of amino acid has the tendency to be more preferentially incorporated into certain secondary structures. The frequencies with which each amino acid is found in α-helix, β-sheet, and β-turn, secondary structures that are discussed later in this chapter, can be calculated as an average over a number of proteins whose three-dimensional structures have been solved. These frequencies are listed in Table 2.2. The β-turn has a distinct configuration consisting of four sequential amino acids and there is a strong preference for specific amino acids in these four positions. For example, asparagine has an overall high frequency of occurrence in a β-turn and is most frequently observed in the first and third position of a β-turn. This characteristic of asparagine is consistent with its side chain being a potential site of N-linked glycosylation (see below). Furthermore, effects of glycosylation on the biological and physicochemical properties of proteins are extremely important. However, their contribution to structure is not readily predictable.

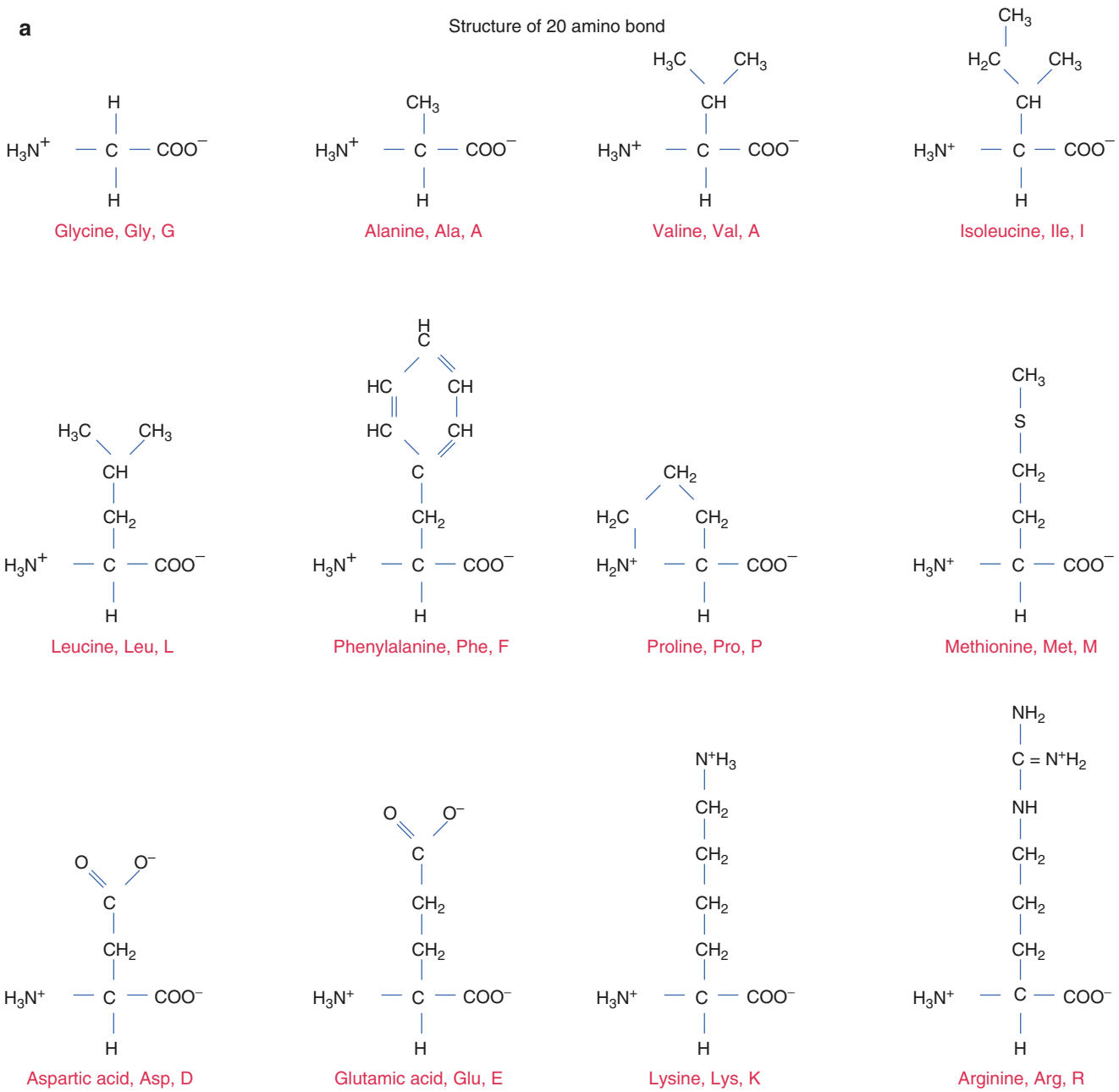

Figure 2.3 ■ **a**, **b** Chemical structure of the 20 natural amino acids, which are the building blocks commonly found in proteins

Another property of amino acids, which impacts protein folding, is the hydrophobicity of their side chains. Although nonpolar amino acids are basically hydrophobic, it is important to know how hydrophobic they are. This property has been determined by measuring the partition coefficient or solubility of amino acids in water and organic solvents and normalizing such parameters relative to glycine. Relative to the side chain of glycine, a single hydrogen, such normalization shows how strongly the side chains of nonpolar amino acids prefer the organic phase to the aqueous phase. A representation of such measurements is shown in Table 2.3. The values indicate that the free energy increases as the side chain of tryptophan and tyrosine are transferred from an organic solvent to water and that such transfer is thermodynamically unfavorable. Although it is unclear how comparable the hydrophobic property is between an organic solvent and the interior of protein molecules, the hydrophobic side chains favor clustering together, resulting

b Structure of 20 amino acide

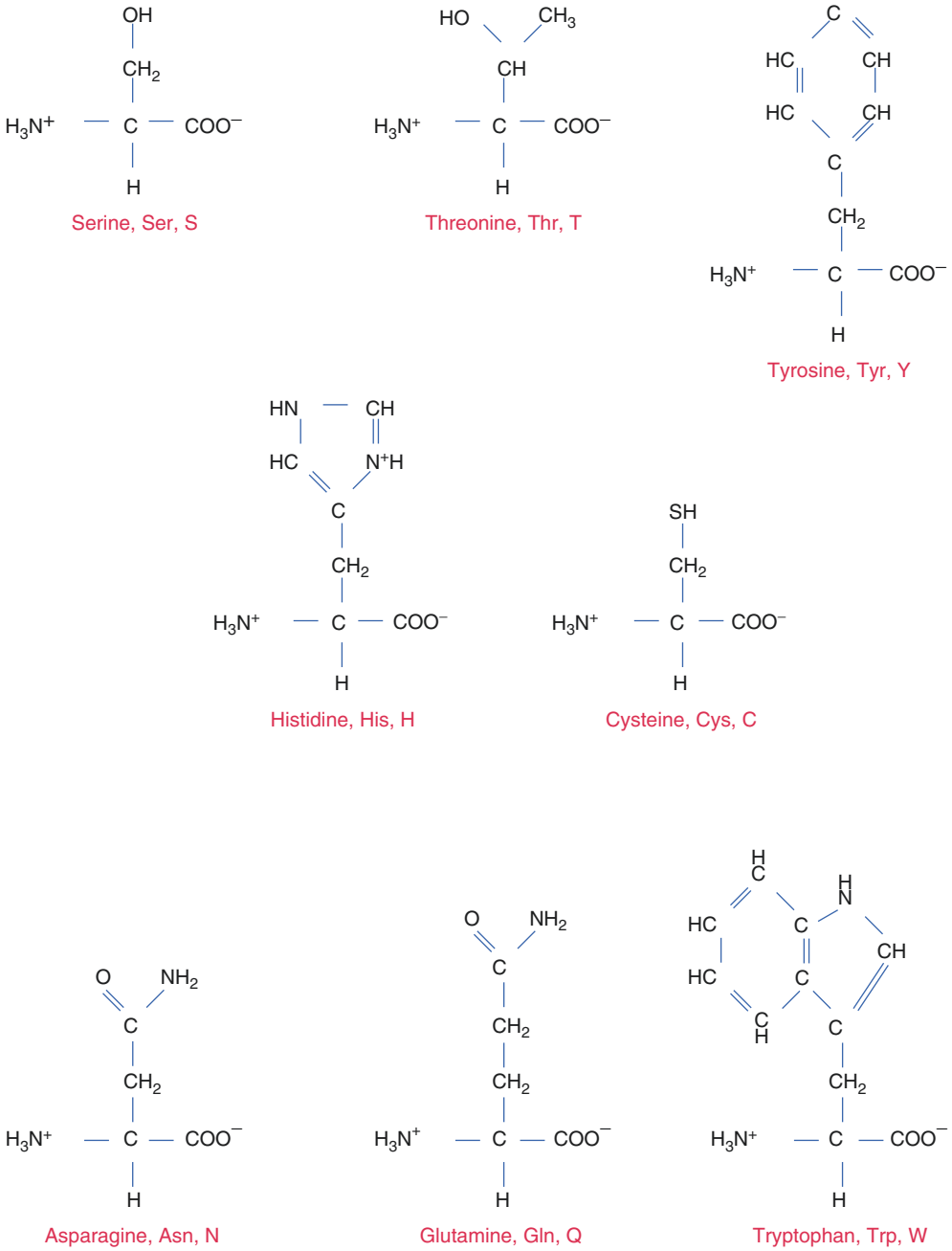

Figure 2.3 ■ (continued)

```
TPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYK   40
LCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLSQLHS   80
GLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIWQQ  120
MEELGMAPALQPTQGAMPAFASAFQRRAGGVLVASHLQSF  160
LEVSYRVLRHLAQP
```

Figure 2.4 ■ Amino acid sequence of human granulocyte-colony-stimulating factor

Parameter		Value	
Molecular weight		18,673	
Total number of amino acids		174	
1 μg		53.5 pmol	
Molar extinction coefficient		15,820	
1 A (280)		1.18 mg/mL	
Isoelectric point		5.86	
Charge at pH 7		−3.39	
Amino acid	**Number**	**% By weight**	**% By frequency**
A Ala	19	7.23	10.92
C Cys	5	2.76	2.87
D Asp	4	2.47	2.30
E Glu	9	6.22	5.17
F Phe	6	4.73	3.45
G Gly	14	4.28	8.05
H His	5	3.67	2.87
1 Me	4	2.42	2.30
K Lys	4	2.75	2.30
L Leu	33	20.00	18.97
M Met	3	2.11	1.72
N Asn	0	0.00	0.00
P Pro	13	6.76	7.47
Q Gln	17	11.66	9.77
R Arg	5	4.18	2.87
S Ser	14	6.53	8.05
T Thr	7	3.79	4.02
V Val	7	3.71	4.02
W Trp	2	1.99	1.15
Y Tyr	3	2.62	1.72

Table 2.1 ■ Amino acid composition and physicochemical parameters of granulocyte-colony-stimulating factor

in a core structure with properties similar to those of an organic solvent. These hydrophobic characteristics of nonpolar amino acids and hydrophilic characteristics of polar amino acids generate a partition of amino acyl residues into a hydrophobic core and a hydrophilic surface, which contributes to the conformational stability of folded proteins.

■ Secondary Structure

Immediately evident in the primary structure of a protein is that each amino acid is linked by a peptide bond (Fig. 2.2). The amide, NH, is a hydrogen donor and the carbonyl, C=O, is a hydrogen acceptor, and they can form a stable hydrogen bond when they are positioned in an appropriate configuration of the polypeptide chain. Such structures of the polypeptide chain are called secondary structure. Two main structures, α-helix and β-sheet, accommodate such stable hydrogen bonds. Besides α-helices and β-sheets, loops and turns are common secondary structures found in proteins.

α-Helix

The α-helix is a right-handed helix that makes one turn per 3.6 residues. The overall length of α-helices can vary widely. Figure 2.6 shows an example of a short

α-helix. In this case, the C=O group of residue 1 forms a hydrogen bond with the NH group of residue 5 and the C=O group of residue 2 forms a hydrogen bond with the NH group of residue 6. All the hydrogen bonds are aligned along the helical axis. Since peptide NH and C=O groups both have electric dipole moments pointing in the same direction, they will add to a substantial dipole moment throughout the entire α-helix, with the negative partial charge at the C-terminal side and the positive partial charge at the N-terminal side.

The side chains project outward from the α-helix. This projection means that all the side chains surround the outer surface of an α-helix and interact both with each other and with side chains of other regions within the protein that come in contact with these side chains. These interactions, so-called long-range interactions, can stabilize the α-helical structure and help it to act as a folding unit. Often an α-helix serves as a building block for the three-dimensional structure of globular proteins by bringing hydrophobic side chains to one side of a helix and hydrophilic side chains to the opposite side of the same helix. Distribution of side chains along the α-helical axis can be viewed using the helical wheel. Since one turn in an α-helix is 3.6 residues long, each residue can be plotted every 360°/3.6 = 100° around a circle (viewed from the top of the α-helix), as shown in Fig. 2.7. Such a plot shows the projection of the position of the residues onto a plane perpendicular to the helical axis. One of the helices in erythropoietin is shown in Fig. 2.7, using an open circle for hydrophobic side chains and an open rectangle for hydrophilic side chains. It becomes immediately obvious that one side of the α-helix is highly hydrophobic, suggesting that this side forms an internal core, while the other side is relatively hydrophilic and is hence most likely exposed to the surface. Since many biologically important proteins function by interacting with other macromolecules, the information obtained from the helical wheel is extremely useful. For example, mutations of amino acids in the solvent-exposed side may lead to identification of regions responsible for biological activity, while mutations in the internal core may lead to altered protein stability.

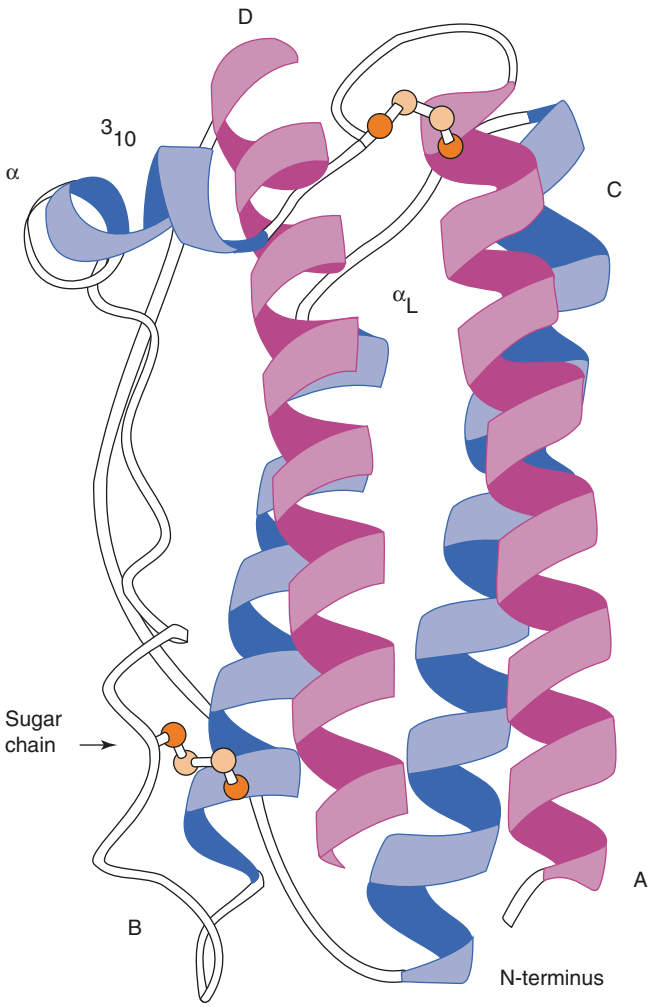

Figure 2.5 ■ Schematic illustration of the three-dimensional structure of filgrastim (recombinant human G-CSF). Filgrastim is a 175-amino acid protein. Its four antiparallel alpha helices (*A*, *B*, *C*, and *D*) and short 3-to-10 type helix (3_{10}) form a helical bundle. The two biologically active sites (α and α_L) are remote from modifications at the N-terminus of the α-helix and the sugar chain attached to loops *C–D*. Note: filgrastim is not glycosylated; the sugar chain is included to illustrate its location in endogenous G-CSF

β-*Sheet*

The second major secondary structural element found in proteins is the β-sheet. In contrast to the α-helix, which is built up from a continuous region with a peptide hydrogen bond linking every fourth amino acid, the β-sheet comprises peptide hydrogen bonds between different regions of the polypeptide that may be far apart in sequence. β-strands can interact with each other in one of the two ways shown in Fig. 2.8, i.e., either parallel or antiparallel. In a parallel β-sheet, each strand is oriented in the same direction with peptide hydrogen bonds formed between the strands, while in antiparallel β-sheets, the polypeptide sequences are oriented in the opposite direction. In both structures, the C=O and NH groups project into opposite sides of the polypeptide chain, and hence, a β-strand can interact from either side of that particular chain to form peptide hydrogen bonds with adjacent strands. Thus, more than two β-strands

α-Helix		β-Sheet		β-Turn		β-Turn position 1		β-Turn position 2		β-Turn position 3		β-Turn position 4	
Glu	1.51	Val	1.70	Asn	1.56	Asn	0.161	Pro	0.301	Asn	0.191	Trp	0.167
Met	1.45	Lie	1.60	Gly	1.56	Cys	0.149	Ser	0.139	Gly	0.190	Gly	0.152
Ala	1.42	Tyr	1.47	Pro	1.52	Asp	0.147	Lys	0.115	Asp	0.179	Cys	0.128
Leu	1.21	Phe	1.38	Asp	1.46	His	0.140	Asp	0.110	Ser	0.125	Tyr	0.125
Lys	1.16	Trp	1.37	Ser	1.43	Ser	0.120	Thr	0.108	Cys	0.117	Ser	0.106
Phe	1.13	Leu	1.30	Cys	1.19	Pro	0.102	Arg	0.106	Tyr	0.114	Gln	0.098
Gln	1.11	Cys	1.19	Tyr	1.14	Gly	0.102	Gln	0.098	Arg	0.099	Lys	0.095
Trp	1.08	Thr	1.19	Lys	1.01	Thr	0.086	Gly	0.085	His	0.093	Asn	0.091
Ile	1.08	Gln	1.10	Gln	0.98	Tyr	0.082	Asn	0.083	Glu	0.077	Arg	0.085
Val	1.06	Met	1.05	Thr	0.96	Trp	0.077	Met	0.082	Lys	0.072	Asp	0.081
Asp	1.01	Arg	0.93	Trp	0.96	Gln	0.074	Ala	0.076	Tyr	0.065	Thr	0.079
His	1.00	Asn	0.89	Arg	0.95	Arg	0.070	Tyr	0.065	Phe	0.065	Leu	0.070
Arg	0.98	His	0.87	His	0.95	Met	0.068	Glu	0.060	Trp	0.064	Pro	0.068
Thr	0.83	Ala	0.83	Glu	0.74	Val	0.062	Cys	0.053	Gln	0.037	Phe	0.065
Ser	0.77	Ser	0.75	Ala	0.66	Leu	0.061	Val	0.048	Leu	0.036	Glu	0.064
Cys	0.70	Gly	0.75	Met	0.60	Ala	0.060	His	0.047	Ala	0.035	Ala	0.058
Tyr	0.69	Lys	0.74	Phe	0.60	Phe	0.059	Phe	0.041	Pro	0.034	Ile	0.056
Asn	0.67	Pro	0.55	Leu	0.59	Glu	0.056	Ile	0.034	Val	0.028	Met	0.055
Pro	0.57	Asp	0.54	Val	0.50	Lys	0.055	Leu	0.025	Met	0.014	His	0.054
Gly	0.57	Glu	0.37	Ile	0.47	Ile	0.043	Trp	0.013	Ile	0.013	Val	0.053

Taken and edited from Chou PY, Fasman GD (1978) Empirical predictions of protein conformation. Ann Rev Biochem 47: 251–276 with permission from Annual Reviews, Inc.

Table 2.2 ■ Frequency of occurrence of 20 amino acids in α-helix, β-sheet, and β-turn

Amino acid side chain	Cal/mol
Tryptophan	3400
Norleucine	2600
Phenylalanine	2500
Tyrosine	2300
Dihydroxyphenylalanine	1800
Leucine	1800
Valine	1500
Methionine	1300
Histidine	500
Alanine	500
Threonine	400
Serine	–300

Taken from Nozaki Y, Tanford C (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J Biol Chem 246:2211–2217 with permission from American Society of Biological Chemists

Table 2.3 ■ Hydrophobicity scale: transfer free energies of amino acid side chains from organic solvent to water

can contact each other either in a parallel or in an antiparallel manner, or even in combination. Such clustering can result in all the β-strands lying in a plane as a sheet. The β-strands which are at the edges of the sheet may have unpaired alternating C = O and NH groups.

Side chains project perpendicularly to this plane in opposite directions and can interact with other side chains within the same β-sheet or with other regions of the molecule, or may be exposed to the solvent.

However, in almost all known protein structures, β-strands are right-handed twisted. This way, the β-strands adapt into widely different conformations. Depending on how they are twisted, all the side chains in the same strand or in different strands do not necessarily project in the same direction.

Loops and Turns

Loops and turns serve to connect other secondary structure elements, such as α-helices and β-strands. They are comprised of an amino acid sequence which is usually hydrophilic and exposed to the solvent.

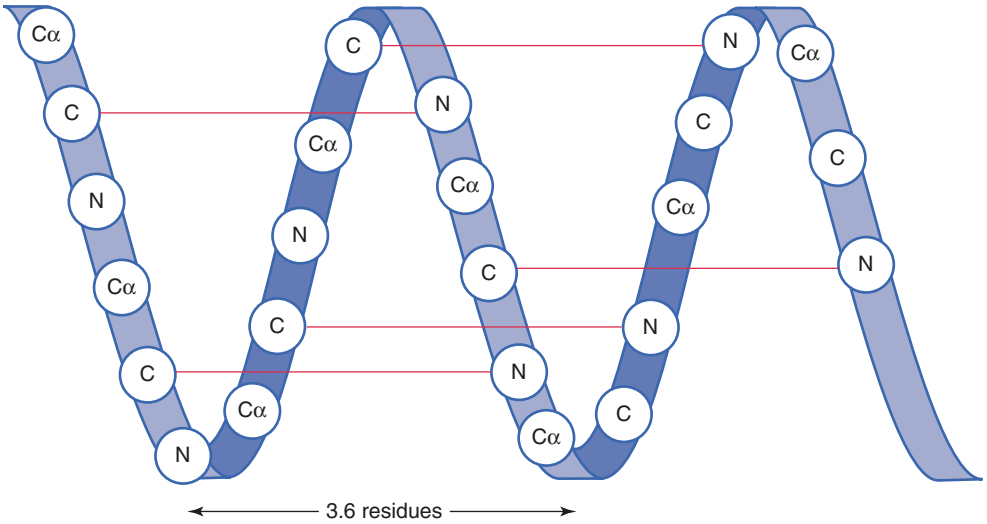

Figure 2.6 ■ Schematic illustration of the structure of an α-helix

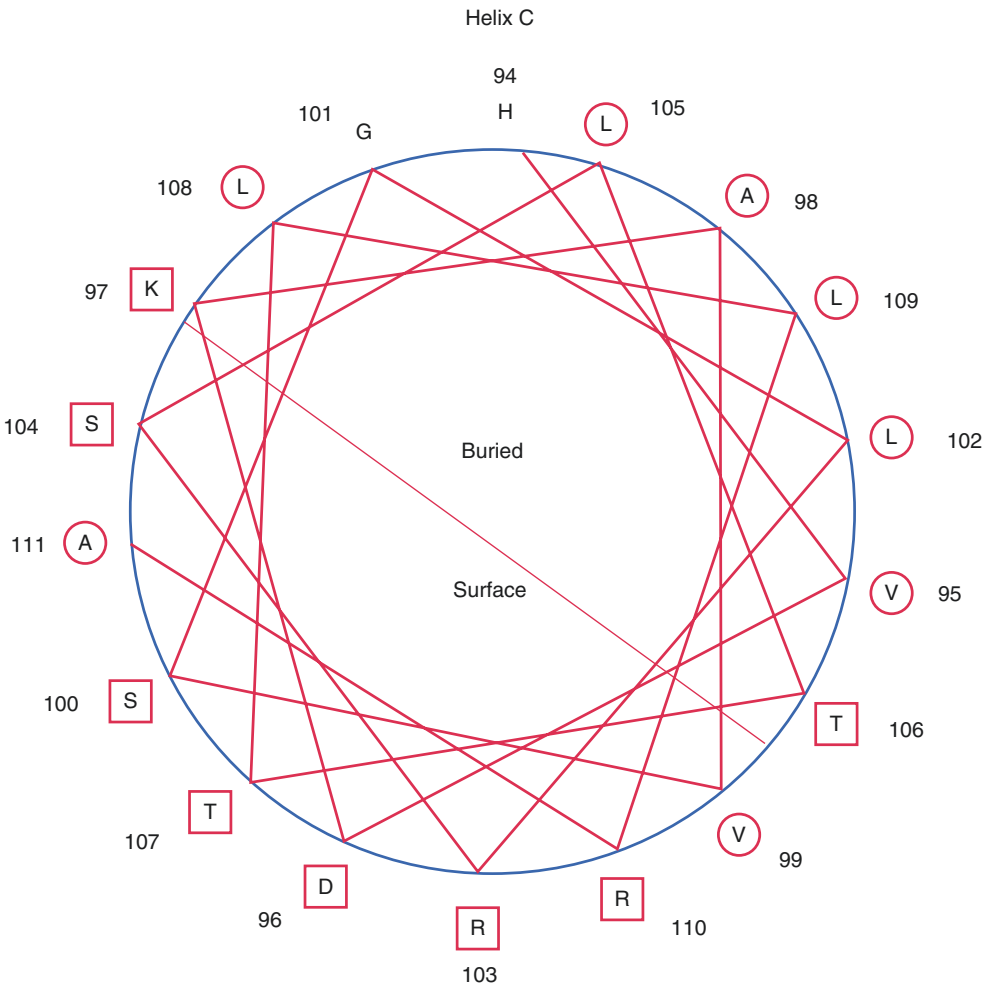

Figure 2.7 ■ Helical wheel analysis of erythropoietin sequence, from His94 to Ala111, with amino acid residues indicated in one-letter code; open circle = hydrophobic side chain, open rectangle = hydrophilic side chains) (Elliott S, personal communication, 1990)

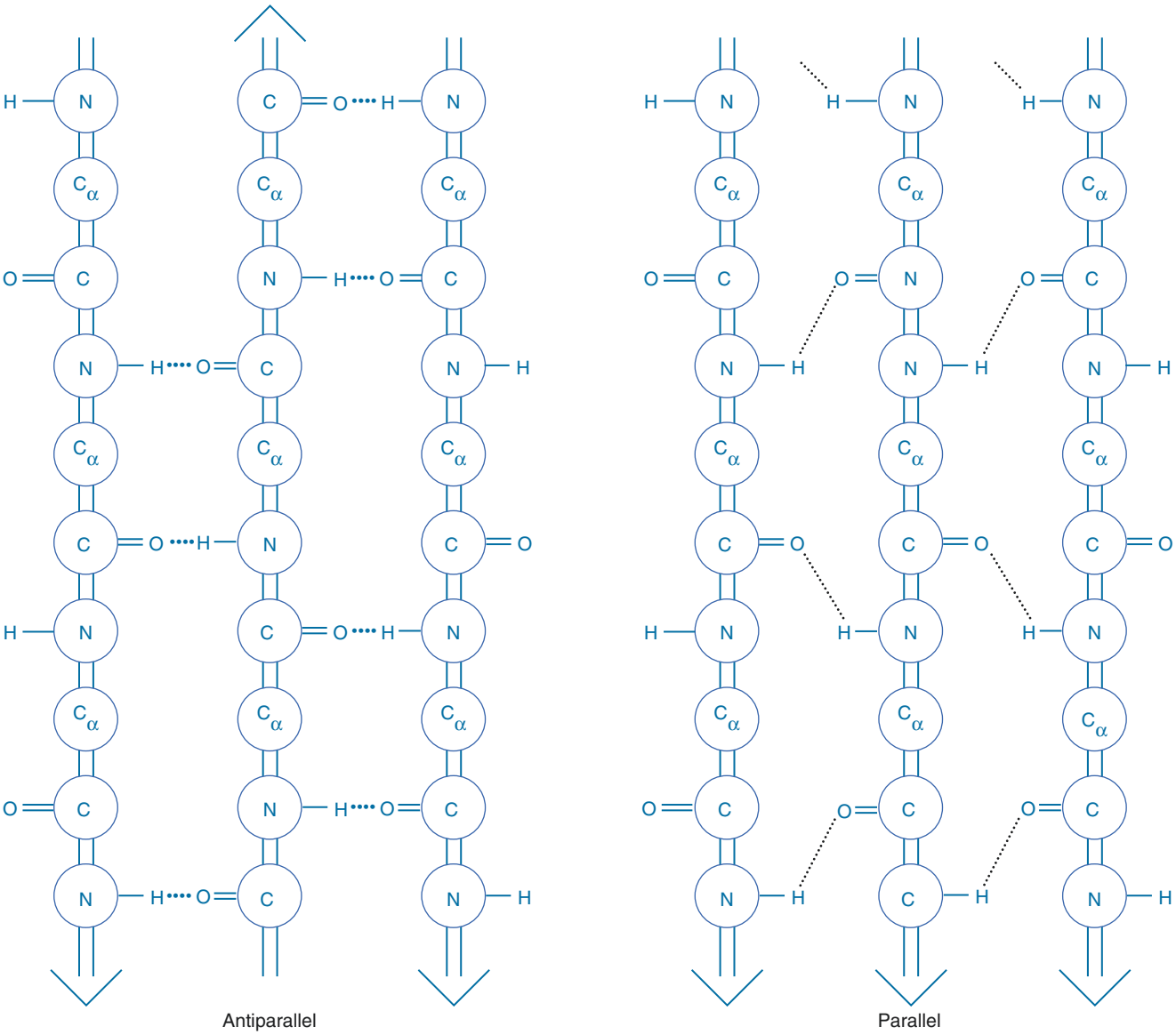

Figure 2.8 ■ Schematic illustration of the structure of an antiparallel (*left side*) and a parallel (*right side*) β-sheet. *Arrow* indicates the direction of amino acid sequence from the N-terminus to C-terminus

These regions consist of β-turns (reverse turns), short hairpin loops, and long loops. Loops and turns usually cause a change in direction of the polypeptide chain, allowing it to fold back to create a compact three-dimensional structure. Turns are short, typically consisting of four amino acid residues, and are stabilized by hydrogen bonds. Loops can be longer and are typically unstructured. Many hairpin loops are formed to connect two antiparallel β-strands.

The amino acid sequences which form β-turns are relatively easy to predict, since turns must be present periodically to fold a linear sequence into a globular structure. Amino acids found most frequently in the β-turn are usually not found in α-helical or β-sheet structures. Thus, proline and glycine represent the least-observed amino acids in these typical secondary structures. However, proline has an extremely high frequency of occurrence at the second position in the β-turn, while glycine has a high preference at the third and fourth position of a β-turn.

Although loops are not as predictable as β-turns, amino acids with high frequency for β-turns also can

form a long loop. Loops are an important secondary structure, since they form a highly solvent-exposed region of the protein molecules and allow the protein to fold onto itself.

■ Tertiary Structure and Quaternary Structure

The spatial arrangement of the various secondary structures in a protein results in its three-dimensional structure. Many proteins fold into a fairly compact, globular structure. Larger proteins usually are folded into several structural domains. Examples are immunoglobulins and Factor VIII.

The folding of a protein molecule into a distinct three-dimensional structure determines its function. Enzyme activity requires the exact coordination of catalytically important residues in the three-dimensional space. Binding of antibodies to antigens and binding of growth factors and cytokines to their receptors all require a distinct, specific surface for high-affinity binding. These interactions do not occur if the tertiary structures of antibodies, growth factors, and cytokines are altered.

A unique tertiary structure of a protein can often result in the assembly of the protein into a distinct quaternary structure consisting of a fixed stoichiometry of protein chains within the complex. Assembly can occur between identical or between different polypeptide chains. Each molecule in the complex is called a subunit. For instance, actin and tubulin self-associate into F-actin and microtubule; hemoglobin is a tetramer consisting of two α- and two β-subunits; among the cytokines and growth factors, interferon-γ is a homodimer, while platelet-derived growth factor is a homodimer of either A or B chains or a heterodimer of the A and B chain. The formation of a quaternary structure occurs via non-covalent interactions and may be stabilized through disulfide bonds between the subunits, such as in the case of the two heavy chains and two light chains of immunoglobulins.

■ Forces

Interactions occurring between chemical groups in proteins are responsible for formation of their specific secondary, tertiary, and quaternary structures. Either repulsive or attractive interactions can occur between different groups. Repulsive interactions consist of steric hindrance and electrostatic effects. Like charges repel each other and bulky side chains, although they do not repel each other, cannot occupy the same space. Folding is also against the natural tendency to move toward randomness, i.e., increasing entropy. Folding leads to a fixed position of each atom and hence a decrease in entropy. For folding to occur, this decrease in entropy, as well as the repulsive interactions, must be overcome by attractive interactions, i.e., hydrophobic interactions, hydrogen bonds, electrostatic attraction, and van der Waals interactions. Hydration of proteins, discussed in the next section, also plays an important role in protein folding.

These interactions are all relatively weak and can be easily broken and formed. Hence, each folded protein structure arises from a fine balance between these repulsive and attractive interactions. The stability of the folded structure is a fundamental concern in developing protein therapeutics.

Hydrophobic Interactions

The hydrophobic interaction reflects a summation of the van der Waals attractive forces among nonpolar groups in the protein interior, which change the surrounding water structure necessary to accommodate these groups if they become exposed. The transfer of nonpolar groups from the interior to the surface requires a large decrease in entropy, so that hydrophobic interactions are essentially entropically driven. The resulting large positive free energy change prevents the transfer of nonpolar groups from the largely sheltered interior to the more solvent-exposed exterior of the protein molecule. Thus, nonpolar groups preferentially reside in the protein interior, while the more polar groups are exposed to the surface and surrounding environment. The partitioning of different amino acyl residues between the inside and outside of a protein correlates well with the hydration energy of their side chains, that is, their relative affinity for water.

Hydrogen Bonds

The hydrogen bond is ionic in character since it depends strongly on the sharing of a proton between two electronegative atoms (generally oxygen and nitrogen atoms). Hydrogen bonds may form either between a protein atom and a water molecule or exclusively as protein intramolecular hydrogen bonds. Intramolecular interactions can have significantly more favorable free energies (because of entropic considerations) than intermolecular hydrogen bonds, so the contribution of all hydrogen bonds in the protein molecule to the stability of protein structures can be substantial. In addition, when the hydrogen bonds occur in the interior of protein molecules, the bonds become stronger due to the hydrophobic environment.

Electrostatic Interactions

Electrostatic interactions occur between any two charged groups. According to Coulomb's law, if the charges are of the same sign, the interaction is repulsive with an increase in energy, but if they are opposite in sign, it is attractive, with a lowering of energy. Electrostatic interactions are strongly dependent upon distance, according to Coulomb's law, and inversely

related to the dielectric constant of the medium. Electrostatic interactions are much stronger in the interior of the protein molecule because of a lower dielectric constant. The numerous charged groups present on protein molecules can provide overall stability by the electrostatic attraction of opposite charges, for example, between negatively charged carboxyl groups and positively charged amino groups. However, the net effects of all possible pairs of charged groups must be considered. Thus, the free energy derived from electrostatic interactions is actually a property of the whole structure, not just of any single amino acid residue or cluster.

Van der Waals Interactions

Weak van der Waals interactions exist between atoms (except the bare proton), whether they are polar or nonpolar. They arise from net attractive interactions between permanent dipoles and/or induced (temporary and fluctuating) dipoles. However, when two atoms approach each other too closely, the repulsion between their electron clouds becomes strong and counterbalances the attractive forces.

■ Hydration

Water molecules are bound to proteins internally and externally. Some water molecules occasionally occupy small internal cavities in the protein structure and are hydrogen bonded to peptide bonds and side chains of the protein and often to a prosthetic group, or cofactor, within the protein. The protein surface is large and consists of a mosaic of polar and nonpolar amino acids, and it binds a large number of water molecules, from the surrounding environment, i.e., it is hydrated. As described in the previous section, water molecules trapped in the interior of protein molecules are bound more tightly to hydrogen-bonding donors and acceptors because of a lower dielectric constant.

Solvent around the protein surface clearly has a general role in hydrating peptide and side chains but might be expected to be rather mobile and nonspecific in its interactions. Well-ordered water molecules can make significant contributions to protein stability. One water molecule can hydrogen bond to two groups distant in the primary structure on a protein molecule, acting as a bridge between these groups. Such a water molecule may be highly restricted in motion and can contribute to the stability, at least locally, of the protein, since such tight binding may exist only when these groups assume the proper configuration to accommodate a water molecule that is present only in the native state of the protein. Such hydration can also decrease the flexibility of the groups involved.

There is also evidence for solvation over hydrophobic groups on the protein surface. So-called hydrophobic hydration occurs because of the unfavorable nature of the interaction between water molecules and hydrophobic surfaces, resulting in the clustering of water molecules. Since this clustering is energetically unfavorable, such hydrophobic hydration does not contribute to the protein stability. However, this hydrophobic hydration facilitates hydrophobic interaction. This unfavorable hydration is diminished as the various hydrophobic groups come in contact either intramolecularly or intermolecularly, leading to the folding of intrachain structures or to protein-protein interactions.

Both the loosely and strongly bound water molecules can have an important impact, not only on protein stability but also on protein function. For example, certain enzymes function in nonaqueous solvent provided that a small amount of water, just enough to cover the protein surface, is present. Bound water can modulate the dynamics of surface groups, and such dynamics may be critical for enzyme function. Dried enzymes are, in general, inactive and become active after they absorb 0.2 g water per g protein. This amount of water is only sufficient to cover surface polar groups, yet may give sufficient flexibility for function.

Evidence that water bound to protein molecules has a different property from bulk water can be demonstrated by the presence of non-freezable water. Thus, when a protein solution is cooled below −40 °C, a fraction of water, ~0.3 g water/g protein, does not freeze and can be detected by high-resolution nuclear magnetic resonance (NMR). Several other techniques also detect a similar amount of bound water. This unfreezable water reflects the unique property of bound water that prevents it from adopting an ice structure.

Proteins are dissolved under physiological conditions or in test tubes in aqueous solutions containing not only water but also other solution components, e.g., salts, metals, amino acids, sugars, and many other minor components (cf. Chap. 5). These components also interact with the protein surface and affect protein folding and stability. For examples, sugars and amino acids are known to enhance folding and stability of the proteins, as described later.

■ Post-translational Modifications

In eukaryotic cells the amino acid sequence of a protein is synthesized in the ribosomes. Subsequently, so-called post-translational modification processes in the endoplasmatic reticulum and the Golgi body of the cell may change the 'amino-acid-only' structure. For instance, sugar groups (glycosylation), phosphate groups (phosphorylation), sulfate groups (sulfation) can be enzymatically attached to the pri-

mary amino acid structure of the protein. Disulfide bridge formation is a post-translational modification as well. For therapeutic proteins that undergo post-translational modifications, glycosylation and disulfide bridging are the most relevant ones. An important family of glycosylated and disulfide bridges-carrying molecules is the monoclonal antibody family (Cf. Chap. 8). Other examples of highly glycosylated proteins (also called glycoproteins) are follicle stimulating hormone (Chap. 19), erythopoeitin (Chap. 24) and Factor VIII (Chap. 21). On the other hand, proteins such as insulins, human growth hormone and interferon alfa lack sugar chains. But, all those proteins contain disulfide bonds.

Protein disulfide bonds are essential for the proper folding, functioning and stability of proteins. Oxidative enzymatic steps in the endoplasmatic reticulum lead to cystine bridge formation. When a protein product goes through red-ox cycling, disulfide bonds may open up and form again. Then, disulfide bond shuffling may result in improper protein structures and covalent aggregates.

Glycosylation is a process where many different enzymes generate often complex sugar structures. During glycosylation various sugars can be attached via glycosidic linkages in chains of different lengths and complexity. An important prerequisite for glycosylation is that the protein has sites for N-linked (via asparagine) or O-linked (usually via serine or threonine) glycosylation. Some sugar chains are unbranched, others are branched chains. Common units found in sugar chains of glycoproteins are mannose, galactose, xylose, sialic acid (derivatives of N-acetylneuraminic acid, negatively charged) and N-acetylgalactosamine, N-acetylglucosamine. Figure 2.9 shows examples of glycosylation patterns found for the monoclonal antibody rituximab.

Glycosylation plays a prominent role in the folding process of proteins and their stability, e.g., against aggregation. Moreover, therapeutic activity may be affected as, e.g., with monoclonal antibodies, where the antibody dependent cellular cytotoxicity (ADCC) depends on the glycosylation pattern. Finally, glycosylation, in particular the presence of sialic acid affects PK profiles. High sialic acids densities cause prolonged circulation times (cf. Fig. 19.3).

Standard prokaryotic cells such as *E. coli* don't perform glycosylation reactions. That may be different for engineered prokaryotes (cf. Chap. 4). Proteins expressed in eukaryotic cells can attach sugar chains to their primary structure. They undergo varying degrees of glycosylation depending on the host cell used and cell culture conditions. Moreover, downstream processing conditions may affect (and are sometimes exploited to adjust) the glycosylation profile of the final

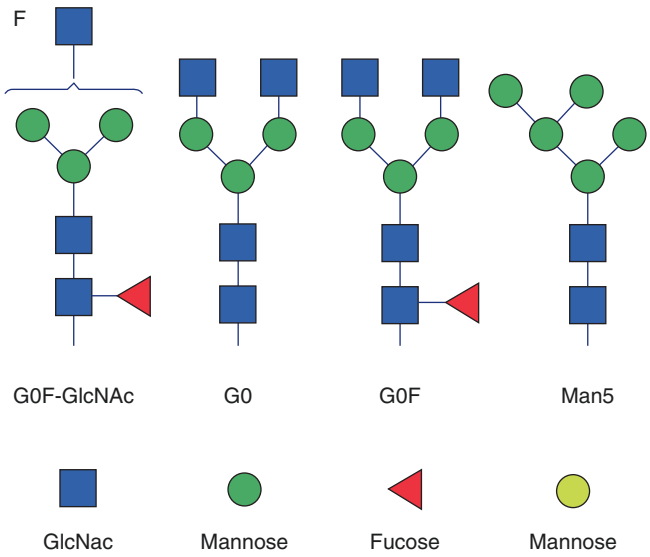

Figure 2.9 ■ Glycosylation pattern of the monoclonal antibody rituximab (Rituxan®/Mabthera®). From: Schiestl et al. (2011) Acceptable changes in quality attributes of glycosylated biopharmaceuticals. Nature Biotechnology 29: 310–312

drug substance. Mammalian cells have an advantage over yeast or plant production cells. Their sugar chain composition closest resembles the 'natural' one. The sugar chains produced in yeast are rich in mannose; their sialic acid content is low. This will increase the clearance rate of these glycoproteins upon injection. Plant derived glycoproteins are rich in fucose and xylose levels. Those structures are believed to increase immunogenicity. As with yeast-derived glycoproteins, they have a relatively low sialic acid content, which affects the pharmacokinetic profile, i.e., accelerated clearance from the blood.

It is important to realize that, unlike the synthesis of a protein's unique primary structure, post-translational modifications result in multiple chemical species, i.e., they introduce heterogeneity in the chemical structure of a protein. For instance, glycosylation variants within one batch of a glycoprotein (e.g., follicle stimulating hormone, FSH, monoclonal antibodies, cf. Fig. 2.9), yield a mixture of various molecular species within one product. This heterogeneity complicates protein characterization but can be identified by analytical techniques such as isoelectric focusing and mass spectrometry (see Chap. 3).

PROTEIN FOLDING

Proteins become functional only when they assume a distinct tertiary structure. Many physiologically and therapeutically important proteins present their surface for recognition by interacting with molecules such as substrates, receptors, signaling proteins, and cell-

surface adhesion macromolecules. Upon biosynthesis in vivo, proteins fold in a spatially organized environment that is comprised of ribosomes, ribosome-associated enzymes, chaperones and a highly concentrated macromolecule solution (200–400 mg/mL). The nascent peptide chain often may fold co-translationally. The interplay of codon usage and tRNA abundance plays an important role in introducing appropriate translation pauses in order to regulate the in vivo speed of protein synthesis to achieve correct folding.

The majority of the current therapeutic proteins require several post-translational modifications, such as disulfide formation and glycosylation, in order to be functional (see above). This is why mammalian cell line expression platforms such as CHO and HEK293 are best suited for expression of proteins that require post-translational modifications for their activity. Several *E. coli* strains have been engineered to translocate recombinant protein to the periplasmic space around the *E. coli* cell membrane, which harbors enzymes that can introduce or break disulfide bonds. However, when recombinant proteins are produced in *E. coli*, they often form inclusion bodies into which they are deposited as insoluble proteins. Therefore, an in vitro process is required to refold insoluble recombinant proteins into their native, physiologically active state. This is usually accomplished by solubilizing the insoluble proteins with detergents or denaturants, followed by the purification and removal of these reagents concurrent with refolding the proteins (see Chap. 4).

Unfolded states of proteins are usually highly stable and soluble in the presence of denaturing agents. Once the proteins are folded correctly, they are also relatively stable. During the transition from the unfolded form to the native state, the protein must go through a multitude of other transition states in which it is not fully folded, and denaturants or solubilizing agents are at low concentrations or even absent.

The refolding of proteins can be achieved in various ways. The dilution of proteins at high denaturant concentration into aqueous buffer will decrease both denaturant and protein concentration simultaneously. The addition of an aqueous buffer to a protein-denaturant solution also causes a decrease in concentrations of both denaturant and protein. The difference in these procedures is that, in the first case, both denaturant and protein concentrations are the lowest at the beginning of dilution and gradually increase as the process continues. In the second case, both denaturant and protein concentrations are highest at the beginning of dilution and gradually decrease as the dilution proceeds. Dialysis or the diafiltration of protein in the denaturant against an aqueous buffer resembles the second case, since the denaturant concentration decreases as the procedure continues. In this case, however, the protein concentration remains practically unchanged. Refolding can also be achieved by first binding the protein in denaturants to a solid phase, i.e., to a column matrix, and then equilibrating it with an aqueous buffer. In this case, protein concentrations are not well defined. Each procedure has advantages and disadvantages and may be applicable to one protein, but not to another.

If proteins in the native state have disulfide bonds, cysteines must be correctly oxidized. Such oxidation may be done in various ways, e.g., air oxidation, glutathione-catalyzed disulfide exchange, or mixed-disulfide formation followed by reduction and oxidation or by disulfide reshuffling.

Protein folding has been a topic of intensive research since Anfinsen's demonstration that ribonuclease can be refolded from the fully reduced and denatured state in in vitro experiments. This folding can be achieved only if the amino acid sequence itself contains all information necessary for folding into the native structure. This is the case, at least partially, for many proteins. However, a lot of other proteins do not refold in a simple one-step process. Rather, they refold via various intermediates which are relatively compact and possess varying degrees of secondary structures, but which lack a rigid tertiary structure. Intrachain interactions of these preformed secondary structures eventually lead to the native state. However, the absence of a rigid structure in these preformed secondary structures can also expose a cluster of hydrophobic groups to those of other polypeptide chains, rather than to their own polypeptide segments, resulting in intermolecular aggregation. High efficiency in the recovery of native protein depends to a large extent on how this aggregation of intermediate forms is minimized. The use of chaperones or polyethylene glycol has been found quite effective for this purpose. The former are proteins, which aid in the proper folding of other proteins by stabilizing intermediates in the folding process and the latter serves to solvate the protein during folding and diminishes interchain aggregation events.

Protein folding is often facilitated by cosolvents, such as polyethylene glycol or glycerol. As described above, proteins are functional and highly hydrated in aqueous solutions. True physiological solutions, however, contain not only water but also various ions and low- and high-molecular-weight solutes, often at very high concentrations. These ions and other solutes play a critical role in maintaining the functional structure of the proteins. When isolated from their natural environment, the protein molecules may lose these stabilizing factors and hence must be stabilized by certain compounds, often at high concentrations. These solutes are

also used in vitro to assist in protein folding and to help stabilize proteins during large-scale purification and production as well as for long-term storage. These solutes encompass sugars, amino acids, inorganic and organic salts, and polyols. They may not strongly bind to proteins, but instead typically interact weakly with the protein surface to provide significant stabilizing energy without interfering with their functional structure.

SELF-ASSESSMENT QUESTIONS

■ Questions

1. What is the net charge of granulocyte-colony-stimulating factor at pH 2.0, assuming that all the carboxyl groups are protonated?
2. Based on the above calculation, do you expect the protein to unfold at pH 2.0?
3. Are hydrophilic and hydrophobic amino acids ad random distributed in an alfa-helix primary sequence of a folded protein?
4. (A) What are the different types of forces that stabilize the secondary, tertiary and, eventually, quaternary structure of a protein? (B) Why are solutes changing these folded structures?

■ Answers

1. Based on the assumption that glutamyl and aspartyl residues are uncharged at this pH, all the charges come from protonated histidyl, lysyl, arginyl residues, and the amino terminus, i.e., 5 His + 4 Lys + 5 Arg + N-terminal = +15.
2. Whether a protein unfolds or remains folded depends on the balance between the stabilizing and destabilizing forces. At pH 2.0, extensive positive charges destabilize the protein, but whether such destabilization is sufficient or insufficient to unfold the protein depends on how stable the protein is in the native state. The charged state alone cannot predict whether a protein will unfold.
3. An α-helix serves as a building block for the three-dimensional structure of globular proteins by bringing hydrophobic side chains to one side of a helix and hydrophilic side chains to the opposite side of the same helix.
4. (A) These forces are covalent forces (disulfide bonds), hydrophobic interactions, hydrogen bonds, electrostatic interactions, van der Waals interactions and hydration forces. (B) These components interact with the protein surface and affect protein folding and stability. For example, sugars and amino acids are known to enhance folding and stability of the proteins, as described below. Another example is buffers that change the pH and by that the charge on the proteins and by that electrostatic intereactions, or (high) concentrations of surfactants that denature (unfold) proteins.

FURTHER READING

Buxbaum E (2015) Fundamentals of protein structure and function, 2nd edn. Springer, New York

Creighton TE (ed) (1989) Protein structure: a practical approach. IRL Press, Oxford

Gregory RB (ed) (1994) Protein-solvent interactions. Marcel Dekker, New York

Schulz GE, Schirmer RH (eds) (1979) Principles of protein structure. Springer, New York

Shirley BA (ed) (1995) Protein stability and folding. Humana Press, Totowa

Whitford D (2005) Proteins: structure and function. Wiley, Hoboken, NJ