The Structure of Biopolymers

Tudor I. Oprea
Division of Biocomputing
University of New Mexico School of Medicine
MSC08 4560, 1 University of New Mexico
Albuquerque, NM 87131-0001

There is no life on Earth without proteins or nucleic acids. Viruses that lack their own metabolic machinery are still assembled in viral particles that exhibit a protein coating and a nucleus (genetic material). Proteins and nucleic acids are the most diverse biopolymers found in a living cell. Each cell type contains a structurally distinct group of proteins, which forms the basis of cellular diversity. Genetic information, encoded in genes, gives the uniqueness of cells, organs and species.

Understanding the structure of proteins and nucleic acids is essential, if one considers these biopolymers as targets for therapeutic manipulation.

1. The Structure of Proteins

Proteins have a wide variety of functional and biological roles, raging from enzymes (biomolecular catalysts), receptors (small molecule recognition), and antibodies (macromolecule recognition) to structural proteins (cellular endoskeleton) and protectants (e.g., outer skin layer). Their importance was recognized as early as 1838, when Danish scientist Gerardus J. Mulder first used the term *protein*, from the Greek *proteis*, meaning first in importance.

Regardless of their mass and function, all proteins are polymers made of (alpha) amino acids linked by peptide bonds (CO-NH). It is the number, order and chemical nature of these amino acids that determines their primary/secondary/tertiary structure, as well as the chemical character and function of each protein. Thus, we take a look at amino acids first.

1.1. The general properties of amino acids in relationship to protein structure

Small molecules (molecular weight, MW, around 100 atomic mass units), all amino acids (AAs) include an amino ($R\text{-}NH_3^+$) and a carboxyl (R-COOH) moiety, respectively, both substituted at the same methyl group (R-CH). Although two additional AAs have been discovered in proteins, we frequently refer to the set of **twenty** AAs encoded by the standard genetic code as *proteinogenic*. Their chemical structures are rendered in **Figure 1**. Except Proline, all AAs have a free amino and carboxyl moiety substituted at the alpha carbon, the difference among them being given by the *side chains*. Fig. 1 also includes the three-letter and one-letter acronym for each AA. Some of these amino acids are also "essential" since they require an external (dietary) supply, i.e., they cannot be produced from other resources in the body. **The essential AAs are: His, Ile, Leu, Lys, Met (or Cys), Phe (or Tyr), Thr, Trp and Val.** Vegetarians must ensure that Lys and Trp are consumed in adequate quantities, as plants are typically a poor source for these two amino acids.

The **primary structure** (or primary sequence) refers to AA positions in the polypeptide chain (i.e., sequence), as well as their chemical bonds – not unlike those found in small molecules and (short) peptides. As the number of AAs increases, paralleled by MW, new types of bonds occur, e.g., hydrogen bonds; these play a major role in determining the **secondary structure**, which includes structural motifs characteristic of proteins: α-helices, β-sheets, etc. Stable, well-defined secondary structures interact with each other (e.g., hydrophobic interactions, salt-bridges, disulfide bonds, etc.), determining the **tertiary structure** of proteins. Tertiary structures allow the packing of protein domains (compact units). Further packing among tertiary structures, based on complex surface interactions as well as specific interactions among peptidic monomers or protein subunits determines the **quaternary structure** of proteins.

Disulfide bridges are S-S bonds that occur between two Cysteine side-chains, playing an important role in tertiary and quaternary structures. Peptide bonds are rather rigid in character, the most frequent being the *trans*-amide bonds (the N-hydrogen and C-oxygen in opposite directions), as opposed to the *cis*-amide bond (the N-hydrogen and C-oxygen on the same side). Most other bonds (except guanidines and double bonds) are rotatable (sigma) bonds, contributing to protein flexibility. Hydrophobic and electrostatic interactions play major roles in the tertiary and quaternary structures. We usually find peptides with secondary (2°) and tertiary (3°) structures, whereas proteins have mostly tertiary (3°) and (4°) quaternary structures. **Figure 2** shows a portion of a peptide chain, its one-letter code structure, and further levels of protein structures. *By convention, the numbering scheme in a protein sequence starts at the N-terminus.*

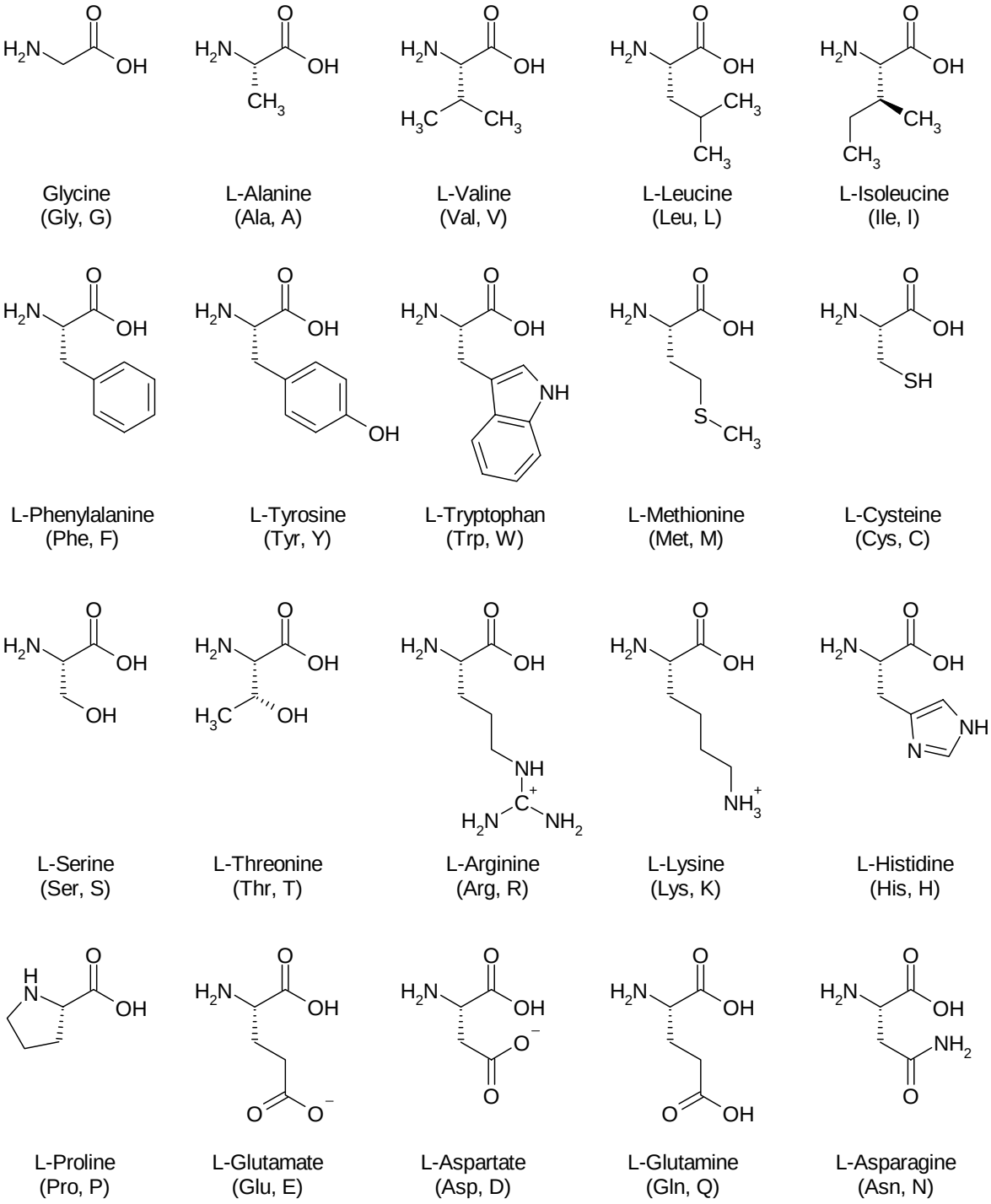


Figure 1. Chemical structures of the twenty alpha amino acids found in proteins (L-isomers).
Read a bedtime story about each amino acid on-line.

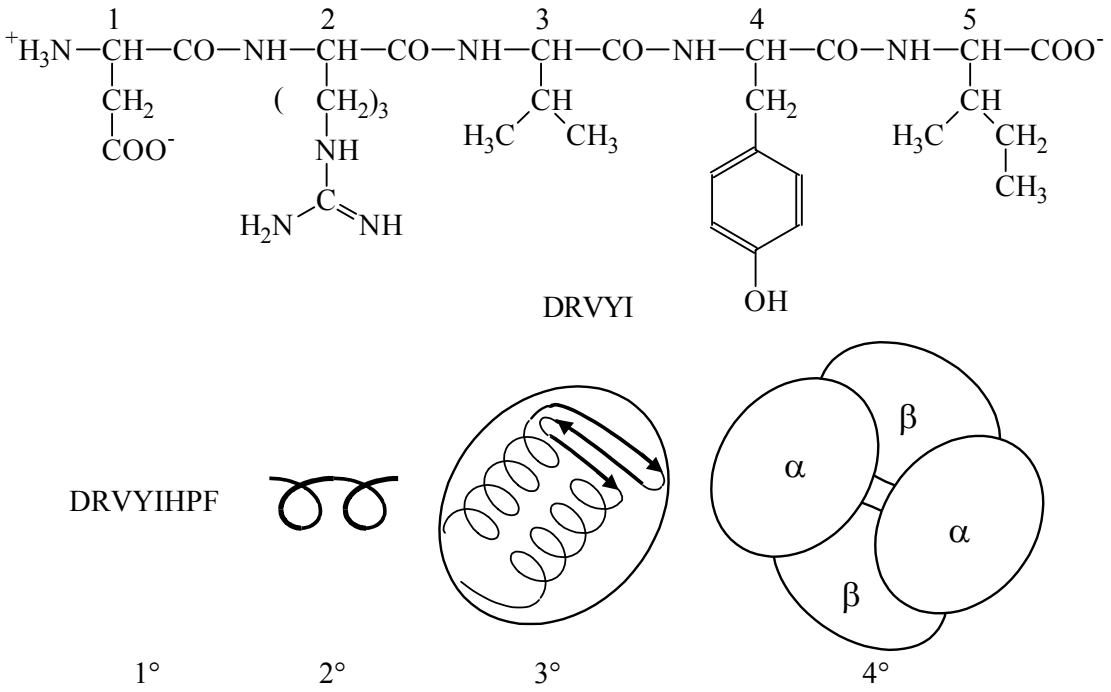


Figure 2. The primary sequence and the four levels of protein structure.

Except Glycine (R = H in **Fig. 3**), all AAs are optically active. All natural amino acids from proteins have the **L** stereochemistry (**Fig. 3**). D amino acids are also present in the cellular walls of certain micro-organisms, or in certain antibiotic peptides (e.g., Gramycidin, Actinomycin D).

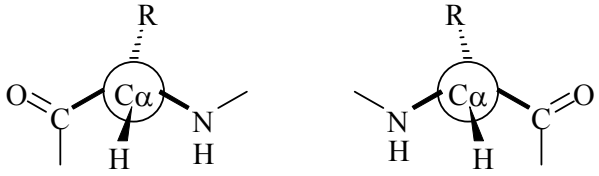


Figure 3. Amino acid chirality.

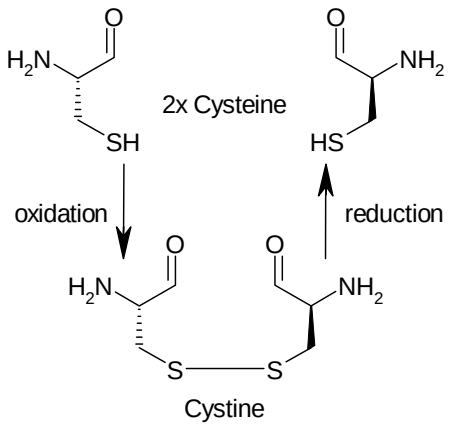


Figure 4. Cysteine, cystine and the disulfide bridge.

Because of the thiol (R-SH) moiety, Cysteine is weakly polar. Frequently, two cysteinyl residues bind covalently under oxidative conditions, forming a disulfide bridge (**Figure 4**). The disulfide bond gives Cystine a weak absorption at 240 nm. S-S bonds play an important role in the

formation of secondary and tertiary structures. Such bonds can be deliberately formed when one attempts to determine the tertiary/quaternary structure of proteins via *site-directed mutagenesis* (by mutating key amino acids with Cys).

The chemical nature of the polypeptide chain is the same for all peptides and proteins, as all amino acids* contribute the same chemical fragment to this chain: -CO-C(**R**)H-NH-. Thus, the side-chain structures (**R**-groups) are the major determinants of the polypeptide secondary/tertiary/quaternary structure. Their properties are summarized in **Table 1**: measured pK_a values for α-carboxyls and α-amino moieties as well as the pK_a of ionisable **R**-groups at physiological pH are given. ***Hydrophobicity*** (measuring the propensity of a molecule to avoid aqueous solution and migrate into the lipid phase) is expressed by the π constant, derived by measuring the *n*-octanol/water partition coefficient for the corresponding N-acetyl-C-amides at pH 7.1.

Table 1. Some physico-chemical characteristics of the proteinogenic amino acids.

Amino acid	pK_a (α-COOH)	pK_a (α-NH_3^+)	pK_a (**R**)	π (**R**)	**R** polarity	Comments (side chain character)
Ala (A)	2.35	9.69	-	0.31	non-polar	A little hydrophobic
Arg (R)	2.17	9.04	12.48	-1.01	positive charge	hydrophilic
Asn (N)	2.02	8.80	-	-0.60	polar, no charge	marked polar
Asp (D)	2.09	9.82	3.86	-2.57	negative charge	hydrophilic
Cys(C)	1.71	10.78	8.33	1.54	polar, no charge	less polar as Cystine
Gln (Q)	2.17	9.13	-	-0.22	polar, no charge	marked polar
Glu (E)	2.19	9.67	4.25	-2.29	negative charge	hydrophilic
Gly (G)	2.34	9.60	-	0.00	non-polar	no side-chain effect
His (H)	1.82	9.17	6.0	0.13	positive charge	only 10% protonated
Ile (I)	2.36	9.68	-	1.80	non-polar	hydrophobic
Leu (L)	2.36	9.60	-	1.70	non-polar	hydrophobic
Lys (K)	2.18	8.95	10.53	-0.99	positive charge	hydrophilic
Met (M)	2.28	9.21	-	1.23	non-polar	hydrophobic
Phe (F)	1.83	9.13	-	1.79	non-polar	marked hydrophobic
Pro (P)	1.99	10.60	-	0.72	non-polar	hydrophobic
Ser (S)	2.21	9.15	-	-0.04	polar, no charge	hydrogen-bonds
Thr (T)	2.63	10.43	-	0.26	polar, no charge	hydrogen-bonds
Trp (W)	2.38	9.39	-	2.25	non-polar	marked hydrophobic
Tyr (Y)	2.20	9.11	10.07	0.96	polar, no charge	hydrogen-bonds
Val (V)	2.32	9.62	-	1.22	non-polar	hydrophobic

* Except Pro and 4-hydroxy-Pro (a modified AA found in collagen)

1.2. Peptide units: Structure and conformation.

The spatial structure of the peptide chain is characterized by mapping two dihedral angles, Φ (C-N-Cα-C) and Ψ (N-Cα-C-N) – see **Fig. 5**. The rigid character of the amide bond implies that these are the only freely rotating chemical bonds in the polypeptide chain.

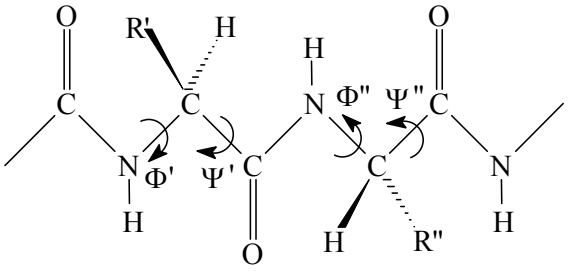


Figure 5. Conformationally flexible bonds in a peptide chain.

Computational studies for dipeptides show that Φ/Ψ minima are determined by the **R**-group of the AA side-chain, and are less influenced by other residues. Tripeptides, further studies show, have an additional minimum determined by the formation of a hydrogen bond – a ten-atom U turn [U10] being formed – see **Fig. 6**.

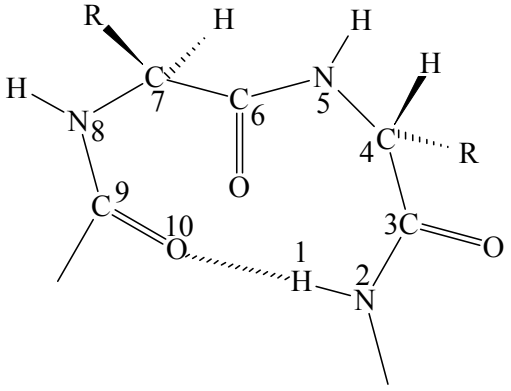


Figure 6. The 10-atoms "U-turn".

Energy estimates using non-bonded interactions (homo-polypeptides, isolated side-chains) show five conformational minima for the peptide skeleton, as shown in **Table 2** and **Figure 7** (the angle values are rounded to the nearest integer). The minima correspond to typical values for the right-turn (R_α) or left-turn (L_α) alpha-helices, for the beta-sheet (β) and for the U-turn.

Table 2. Typical Φ/Ψ angles for secondary structures.

Structure	Φ	Ψ
Cis	0	0
R_α	-60	-50
β	-150	150
U_{10}	-60	-30
L_α	60	50

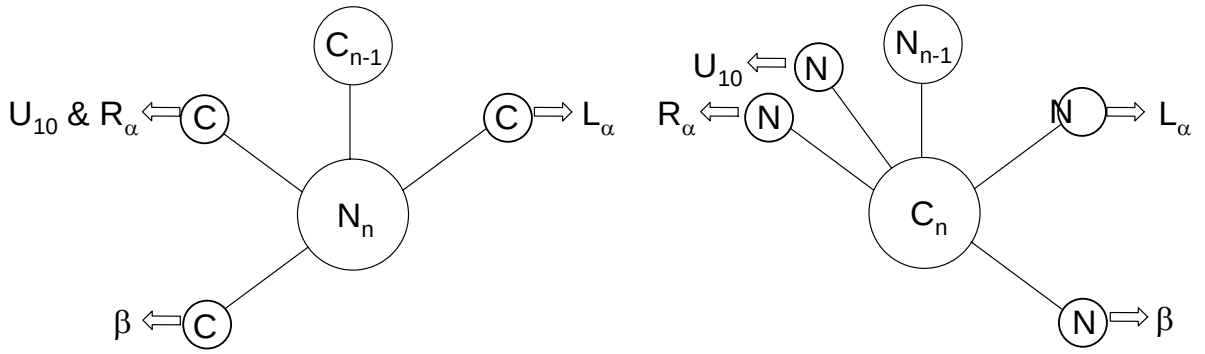


Figure 7. Newman projections for stable peptide skeleton conformations.

Side-chain conformations are described by the dihedral angles χ_1 (N-Cα-Cβ-Cγ), χ_2 (Cα-Cβ-Cγ-Cδ), χ_3, etc, **Figure 8**. Energy estimates indicate three "normal" minima for χ = -60°, 180° and +60°. However, actual values for these angles differ because of steric repulsions (due to packing), in particular when tertiary and quaternary structures are present. Studies that examined the "allowed" conformations for side-chains are available (also called rotamer libraries) and are frequently used when modeling proteins.

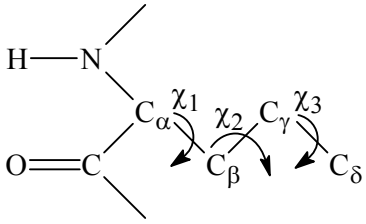


Figure 8. Side-chain dihedral angle definitions.

Not all biologically relevant biopolymers based on amino acids are large. Some of these naturally occurring peptides are shown in **Table 3**. Leu- and Met-enkephalin are naturally-occurring pentapeptides that act on the pain-mediating opioid receptors (endorphins). Read on-line about endorphins and **chilli peppers**. Morphine and other alkaloids also act on these receptors. Angiotensin II is a very powerful vasoconstrictor that increases blood pressure. Blocking its formation from Angiotensin I (via Angiotensin converting enzyme, ACE) lead to a very successful category of anti-hypertensive agents known as ACE-inhibitors (e.g., captopril, enalapril). Vasopressin (also known as the anti-diuretic hormone, ADH) reduces the daily urine volume (diuresis). Oxytocin is a hormone predominantly involved with lactation and uterine contraction (in particular at birth). These two hormones, while having a rather similar peptide structure, have very different physiological roles. Gramycidin S, a cyclic decapeptide from *Bacillus brevis* is a typical peptide antibiotic containing D-amino acids. It also contains Ornithine (Orn), a peptide not present in proteins (see **Figure 9**).

Table 3. Naturally-occurring peptides (examples)

Name	Structure
Leu-enkephalin	Tyr-Gly-Gly-Phe-Leu
Met-enkephalin	Tyr-Gly-Gly-Phe-Met
Angiotensin II	Asp-Arg-Val-Tyr-Ile-His-Pro-Phe
Vasopressin	Cys-Tyr-**Phe**-Gln-Asn-Cys-Pro-**Arg**-Gly
Oxytocin	Cys-Tyr-**Ile**-Gln-Asn-Cys-Pro-**Leu** -Gly
Gramycidin S (cyclic)	[D-Phe-L-Pro-L-Val-L-Orn-L-Leu- L-Leu-L-Orn-L-Val-L-Pro-D-Phe]$_2$

Ornithine is also present in the urea cycle of the biochemical pathways. Other amino acids, not present in proteins, are given in Figure 9.

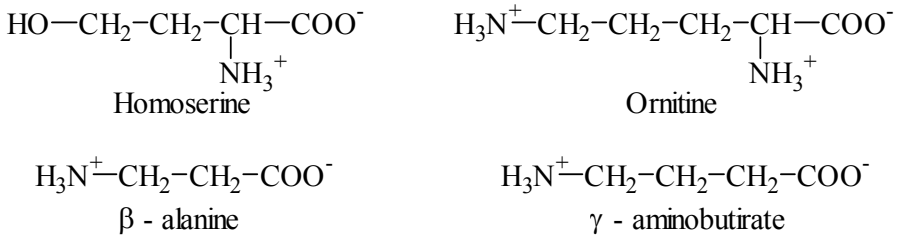


Figure 9. Some natural amino acids not present in proteins.

1.3. Conformational aspects of polypeptides and proteins.

Understanding the primary structure (or sequence) is essential for understanding proteins. By recognizing the conserved sequences, we can classify proteins and perform *primary sequence alignments* (see Part I of BioMed505) and predictions (e.g., secondary structure predictions). The primary sequence further serves as input for predicting the tertiary structure (e.g., threading, folding, homology modeling). Based on H-bonds and the study of peptide bond reactivity, Pauling and Corey found that the most stable **secondary structures** of proteins are alpha-helices (H-bonds between AAs that are close to each other in the primary sequence) and beta-sheets (H-bonds between AAs that are distant in the primary sequence).

The **Ramachandran plot** shows the "allowed" and "forbidden" conformations for the polypeptide chains. As seen in Fig. 5, each third C-N bond from the polypeptide chain has a partially double bond character and cannot rotate freely. The other bonds, namely Cα-N and Cα-C are – in theory – free to rotate since they are true sigma bonds. However, the **R**-group may hinder this rotation, if it has a considerable size. Moreover, the possible overlap of the H and O atoms of the peptide bonds can also cause steric hindrance. The first to map the Φ and Ψ angles, Ramachandran plotted the allowed vs. forbidden angles and found a rather limited number of stable Cα-C & Cα-N values (see also Table 2). For example, the α-helix is allowed according to the Ramachandran plot because all pairs of dihedral angles in this spatial arrangement are found in a favorable area: The most favorable right-hand turn α-helix have Φ = -57° & Ψ = -47°. If the Φ value increases without modifying Ψ, we move away from the conformationally allowed space; if the Ψ varies significantly without altering Φ, we are also moving into forbidden regions. For any given pair of Φ/Ψ values, we can predict which combination is sterically favored. All naturally occurring combinations for native (folded) proteins fall into the allowed Φ/Ψ space (**Fig. 10**).

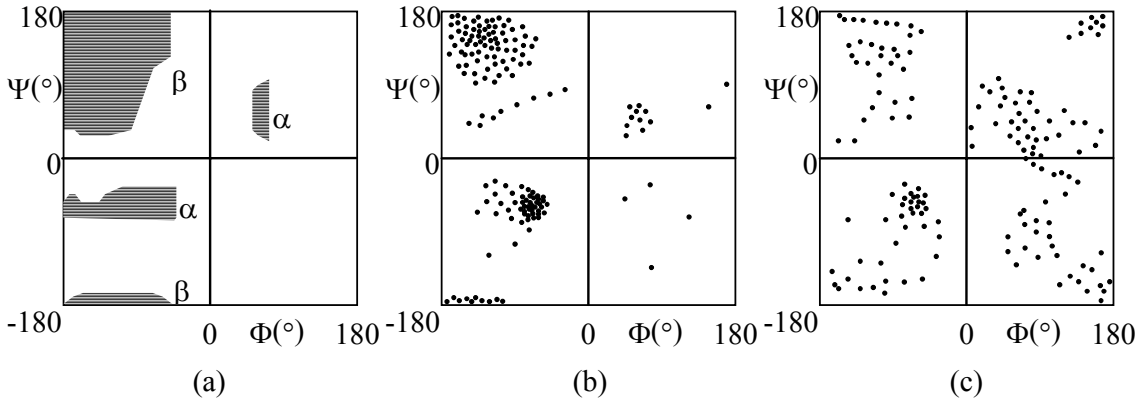


Figure 10. The Ramachandran plot. (a): Sterically-allowed values for the right-handed α-helix (R_α ; the most common), for the β-strand and the left-handed α-helix (L_α). (b): Observed values

for the 19 proteinogenic AAs (except Gly). Each dot represents the values for an amino acid found in high-resolution X-ray structures. (c): Observed values for Gly (which can sample regions that are "forbidden" to other AAs). Gly and Pro (restricted conformationally) play key roles in determining the tertiary and quaternary structures. Do you have access to software that can output Ramachandran plots? Then follow this on-line tutorial from UCSF (Midas optional).

In the **α-helix** structure the peptide chain forms a helical spiral having ~3.6 amino acids per helical step. The AA side-chains face the outer region of the helix (**Figure 11**). The repeating unit of this structure, making a single complete turn of the helix, has ca. 0.54 nm along its longitudinal axis – as confirmed by studying naturally occurring keratins via X-ray crystallography. Such an α-helix allows the formation of intra-molecular H-bonds that parallel its longitudinal axis, between the N-H atom of the first AA's peptide moiety and the 4th consecutive AA carbonyl oxygen of the peptide unit (*i to i+4*) – see **Fig. 11**. Both D- and L- amino acids can form α-helices, although never in a polypeptide chain that contains a mixture of the two stereoisomers. As shown in **Fig. 10**, both left-turn and right-turn helices can be formed – the latter ones being more stable.

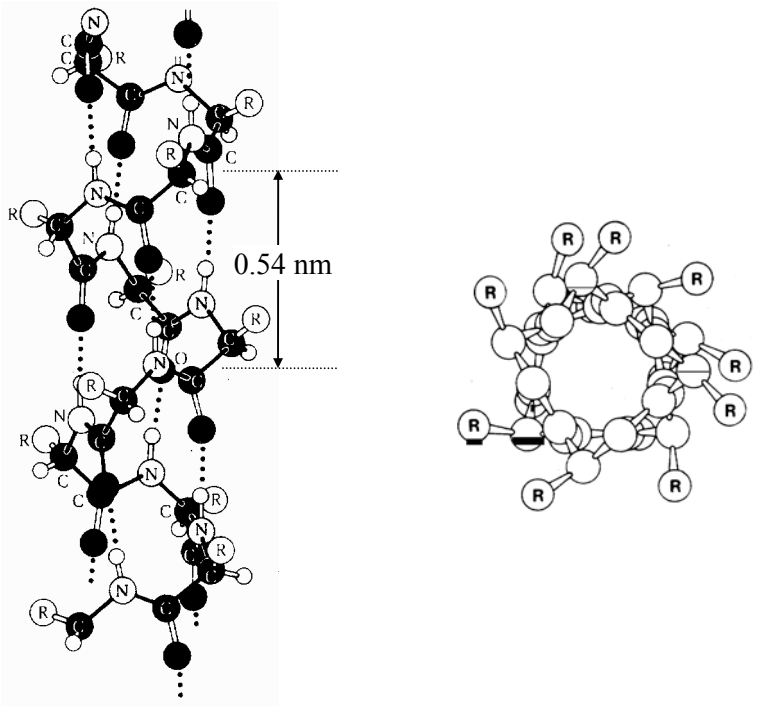


Figure 11. The alpha helix: Side-view (left) and top-view. The intra-molecular hydrogen bonds formed between the peptidic (N)-H and O(=C) moieties are shown with dotted lines.

For an α-helix: Φ, Ψ ≈ -60°, -50° , the number of residues varying from 4 to 80. An average α-helix has 10 AAs, thus being ≈15 Å long. The dipole moment of an α-helix plays an important role in stabilizing the protein structure, an AA contributing ≈3 D (Debye units) to the helix, e.g., a total of ≈30 D for 10 AAs. Barnase is a microbial ribonuclease containing a 3-turns α-helix at the N-terminus. Introducing a His (positively charged side-chain) at the C-terminus via point mutation causes a 5°C increase in the barnase melting point, as determined by reversible thermal denaturation, compared to wild-type. This increased protein stability of ≈2 kcal/mol cam be attributed to the His -α_1-helix ion-dipole interaction. This interaction is spatially favored by the tertiary structure adopted by barnase (**Fig. 12**).

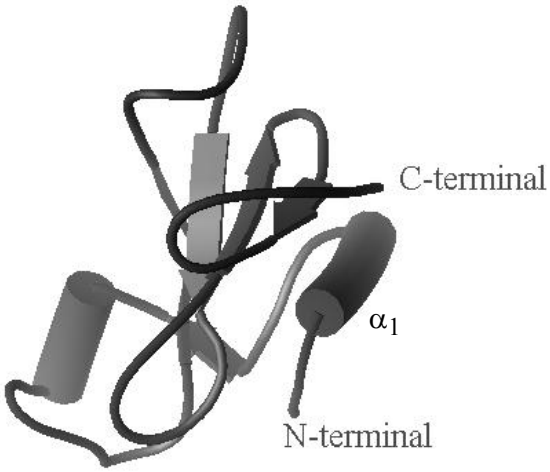


Figure 12. The 3D structure of barnase, rendered using a convention for secondary structural motifs of proteins: α-helices are shown as cylinders, the flat arrows indicate β-sheets, while the thinner tubes indicate a lack of organized secondary structures. The figure indicates the vicinity of the α_1-helix to the C-terminus.

The α-helix propensity is great for A, E, L, M and F, and decreased for P, G, Y and S: If any four of these "helix-making"AAs are in sequence (with maximum one Gly), then an α-helix is likely to occur. The α-helix continues until 2 "helix-breaking" AAs (or 1 Pro) follow in sequence. Other studies have shown that there are other factors favoring α-helix formation, namely the interaction between the preceding 4-5 AAs (N – terminal: N^{I}, N^{II}, N^{III}, N^{IV}, etc.) and the following 4-5 AAs (C – terminal: C^{I}, C^{II}, C^{III}, C^{IV}, … etc.) with the first 3 (N^{1}, N^{2}, N^{3}) and last 3 (C^{1}, C^{2}, C^{3}) AAs from the α-helix itself. Key in this interaction are the border (capping) AAs as well, N^{cap} & C^{cap}:

$$\cdots N^{III}, N^{II}, N^{I}, N^{cap}, N^{1}, N^{2}, N^{3}, \cdots, C^{3}, C^{2}, C^{1}, C^{cap}, C^{I}, C^{II}, C^{III}, C^{IV} \ldots$$

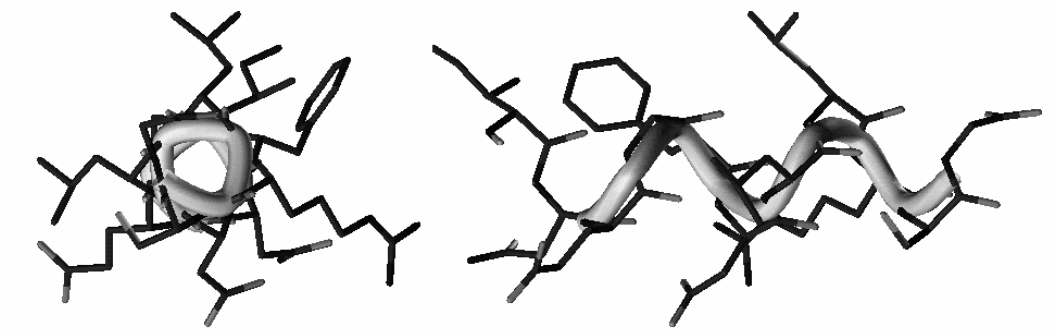
Favoring the α-helix formation are **hydrophobic interactions** between N^{III}, N^{II}, N^{I} & N^{3}/N^{4} and C^{II}, C^{III}, C^{IV}, C^{V} & C^{3}, respectively, and **H-bonds** between the N^{cap} side-chain & the amido N^{3} moiety, and the amido H & carbonyl O between α-helix residues and the residues following C^{cap}, as well as between C^{cap} & C^{3}/C^{4}, respectively. C^{I} is either Gly or Pro, usually. Besides the hydrogen bonds shown in **Fig. 11**, other important interactions are shown in **Table 4**. Based on the above, we can conclude that *the formation of an α-helix depends significantly on the local primary sequence.* An average α-helix, together with its side-chains, is depicted in **Fig. 13**.

Table 4. Interactions stabilizing α-helix formation (^{O} indicates the carbonyl-peptide oxygen):

Charge – aromatic	Charge – H-bond	H-bond	Hydrophobic
Phe (F) – His^{+}(H)	Gln(Q) - Asp^{-}(D)	Gln(Q) - Asp^{O}	Leu(L) - Ile(I)
Trp(W) – His^{+}(H)	Glu^{O}(E) - Lys^{+}(K)	Gln(Q) - Glu^{O}	Leu(L) - Leu(L)
	His^{+}(H) - Asp^{O}(D)		Leu(L) - Val(V)

The nature of the side-chain **R**-groups is essential in determining the interactions of an α-helix with the (micro)environment – see also **Table 1**. Three α-helices from different proteins, illustrated in **Figure 14** (citrate-synthase, alcohol-dehydrogenase and troponin C), have remarkably different configurations. The first one is predominantly hydrophobic, the 2nd one has a hydrophobic, as well as a solvent-exposed, hydrophilic face (seen better from above), whereas the 3rd one is predominantly polar, being entirely exposed to solvent. While not all proteins contain α-

helices, some have a high content: e.g., myoglobin (153 AAs) is 75% α-helical, with 8 such regions.

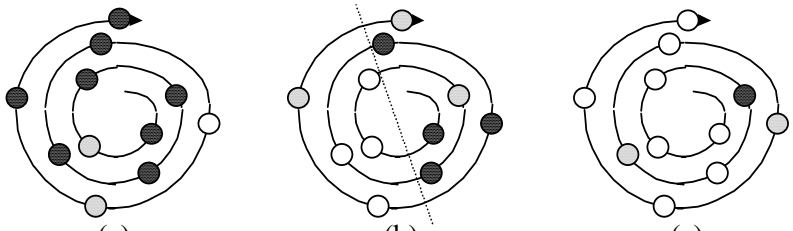


(a) **(b)**

Figure 13. An α-helix from alcohol dehydrogenase: The longitudinal axis is orthogonal **(a)** and parallel **(b)** with the paper plane. The peptide chain is indicated by tubular rendering.

Primary sequence of the α-helix

(a)	259 → L	S	F	A	A	A	M	N	G	L	A → 271
	+1.70	-0.04	+1.79	+0.31	+0.31	+0.31	+1.23	-0.60	0.00	+1.70	+0.31	
(b)	354 → I	N	E	G	F	D	L	L	R	S	G → 366
	+1.80	-0.60	-2.29	0.00	+1.79	-2.57	+1.70	+1.70	-1.01	-0.04	0.00	
(c)	86 → K	E	D	A	K	G	K	S	E	E	E → 98
	-0.99	-2.29	-2.57	0.31	-0.99	0.00	-0.99	-0.04	-2.29	-2.29	-2.29	



(a) (b) (c)

Figure 14. Three α-helices from **(a)** citrate-synthase, **(b)** alcohol-dehydrogenase and **(c)** troponin C: The primary sequence and the numbering scheme, as well as the π constant for the side-chain **R**-groups are given in the top panel (see also **Table 1**). In **(a)** the α-helix is mostly hydrophobic, and is in fact buried in the protein; in **(b)** the α-helix is exposed to solvent (left-hand side), whereas in **(c)** the entire α-helix is solvent-exposed. The lower-panel depiction of α-helices shows each AA positioned at 100° intervals, color-coded by polarity (light colors for polar residues, darker shades for hydrophobic ones).

*Are you on-line? Search the **PDB** and find out how these 3 helices really look like in 3D.*

Less frequent are the 3_{10}-helices, named this way because 3 AAs determine one helical turn. Easy to recognize because of its triangular 'look' when viewed from above, the 3_{10}-helix allows the formation of intra-molecular H-bonds that parallel its longitudinal axis, between the N-H atom of the first AA's peptide moiety and the 3rd consecutive AA carbonyl oxygen of the peptide unit (*i to i+3*). When only 10 atoms are involved, the structure is also known as a U-turn (see **Fig. 6**); its Φ/Ψ angles are given in **Table 2**. Lysozyme contains two 3_{10}-helices.

Polypeptidic chains having an extended conformation (Φ,Ψ ≈ -180°, -180°) are frequently arranged in **β-sheets**, by forming inter-chain (>C=)O···H(-N<) H-bonds; in this type of **secondary structure**, the "middle" peptide chains tend to maximize the number of H-bonds (see also **Fig. 15**). As all peptide bonds are involved in the H-bonding pattern, this gives β-sheets an additional stability; side-chains zigzag above or below the plane formed by the β-sheet. As the peptide chain is completely extended, the distance between adjacent AAs is 0.35 nm. Based on the primary sequence orientation, β-sheets are either parallel or anti-parallel (**Fig. 15**).

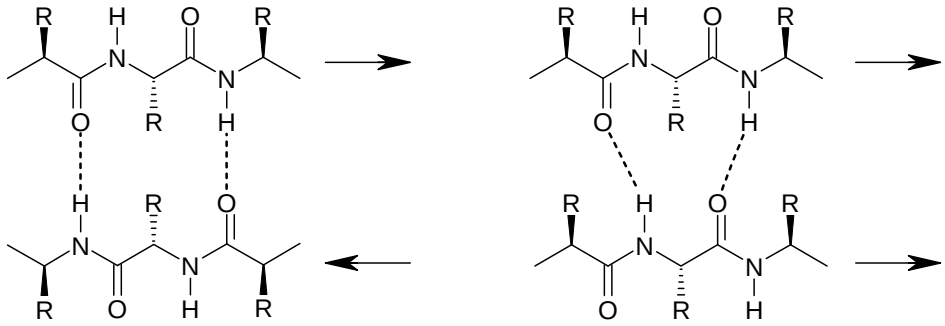


Figure 15. The β-sheet polypeptide chain. Hydrogen bonds (thick dotted lines) are shown for both the **parallel** (left) and **anti-parallel** (right) orientations.

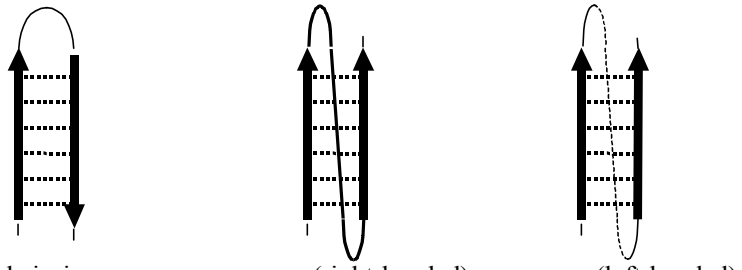


Figure 16. The topology of β-sheets. See text for details.

Topological arrangements of β-sheets are illustrated in **Fig. 16**. The middle sketch (b) shows a right-handed super-helix, since by looking form the N-terminus to the C-terminus (the direction defined by primary sequence), the peptide chain 'turns' to the right; the sketch to the right (c) shows a left-handed super-helix, since the peptide chain 'turns' to the left. **Table 5** gives the propensity ***P*** (+ or –) of each proteinogenic AAs to form β-sheets. A thermodynamic scale provides the energetic equivalent for ***P***, expressed as $\Delta\Delta G_\beta$ (with Ala as reference point).

Table 5. Amino acid propensity to form β-sheets.

Amino acid	*P*	$\Delta\Delta G_\beta$	Amino acid	*P*	$\Delta\Delta G_\beta$	Amino acid	*P*	$\Delta\Delta G_\beta$
Thr		1.1	Trp	-	0.54	**Ala**	+	0.00
Ile	+	1.0	Cys		0.52	His	-	-0.02
Tyr		0.96	Leu	+	0.51	Asn		-0.08
Phe	+	0.86	Arg	-	0.45	Asp	-	-0.94
Val	+	0.82	Lys	-	0.27	Gly		-1.20
Met	+	0.72	Gln		0.23	Pro	-	<-3
Ser		0.70	Glu	-	0.01			

Two distinct scenarios for β-sheets occur in the context of tertiary structures: Central β-sheets have β-sheets on both sides, whereas peripheral β-sheets have β-sheets on one side only. In a similar context to **Table 5**, using point mutations, melting point determinations and statistical analyses, the likelihood $\Delta\Delta G_{\beta}p$ [1] of AAs to be present in peripheral β-sheets has been investigated (**Table 6**). However, $\Delta\Delta G_{\beta}p$ does not correlate with the experimentally observed propensities ***P.*** The same **Table 6** includes the ΔΔΔG values: ΔΔΔG = $\Delta\Delta G_{\beta}p$ − $\Delta\Delta G_{\beta}c$, where $\Delta\Delta G_{\beta}c$ [2] measures the preference of AAs to occur in central β-sheets. **Table 6** includes two additional side-chain characteristics: Λ – a scale for polarity, and π-π_{Ala} – a scale for hydrophobicity. ΔΔΔG values do not correlate with any of these values. They do correlate, however, with the non-polar surface areas of the AA side-chains (not shown) and the n-octanol/water partition coefficients for the same (not shown). Based on the above, we can conclude that *the formation of a β-sheet, unlike the formation of an α-helix, depends significantly on tertiary structure information, not only on the local information.*

Table 6. Statistically determined side-chain properties for proteinogenic amino acids.

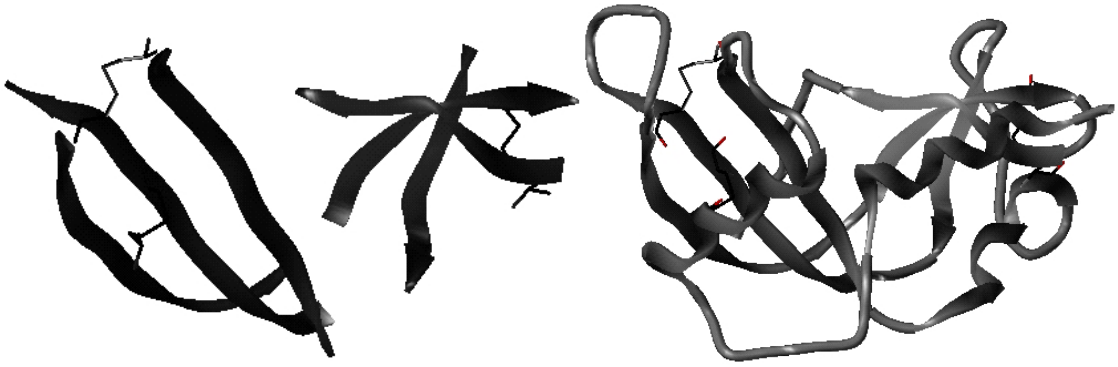
Amino acid	P	$\Delta\Delta G_{\beta}p$	Amino acid	P	ΔΔΔG	Λ	π-π_{Ala}
Thr		0.83	Ile	+	-0.98	0.3	+1.49
Ser		0.63	Arg	-	-0.88	4.0	-1.32
Glu	-	0.31	Trp	-	-0.77	1.5	+1.94
Val	+	0.17	Leu	+	-0.75	0.4	+1.39
Phe	+	0.16	Tyr		-0.75	2.1	0.65
Tyr		0.11	Met	+	-0.74	1.0	+.92
Cys		0.08	Phe	+	-0.70	1.1	+1.48
Gln		0.04	Lys	-	-0.67	3.6	-1.30
Ile	+	0.02	Val	+	-0.65	0.4	+.91
Ala	+	0.00	Cys		-0.44	0.6	+1.23
His	-	-0.01	Thr		-0.27	1.0	-0.05
Met	+	-0.02	Gln		-0.19	2.2	-0.53
Asp	-	-0.10	Asn		-0.16	2.0	-0.91
Trp	-	-0.17	Ser		-0.07	0.8	-0.35
Asn		-0.24	**Ala**	+	0.00	0.2	0.00
Leu	+	-0.24	His	-	0.01	2.1	-0.18
Lys	-	-0.40	Glu	-	0.30	2.3	-2.60
Arg	-	-0.43	Gly		0.35	0	-0.31
Gly		-0.85	Asp	-	0.84	2.0	-2.88
Pro	-	<-4	Pro	-	-	0.7	+0.41

The β-sheet secondary structure is usually associated with structural proteins, e.g., β-keratin, a protein found in silk fibers that contains *anti-parallel* (**Fig. 15**) β-sheets. Another structural protein, found in e.g., hair & wool, is α-keratin. This protein has 3 polypeptides arranged in right-turn α-helices bundled in a left-handed spiral stabilized by disulfide (**Fig. 4**)

1. 'p' for peripheral
2. 'c' for central

bridges. However, when heated under wet conditions and elongated, α-keratin transitions into β-keratin, but with *parallel* β-sheets. Both proteins are water-insoluble.

Other proteins containing β-sheets include lysozyme, carboxypeptidase A and ribonuclease. For example, bovine pancreatic ribonuclease contains 3 segments of extended peptide chains arranged in an anti-parallel β-sheet – see **Figure 17**. Its active form is stabilized by 3 inter-chain S-S bridges (e.g., Cys^{26}–Cys^{84} & Cys^{40}–Cys^{95}, **(a)**). Looking at the entire protein one can notice the undergoing twists of the β-sheet secondary structures **(b)**.



(a) (b)

Figure 17. The β-sheets in bovine pancreatic ribonuclease: Three disulfide bridges are shown in the detailed view **(a)**, with the entire enzyme rendered in **(b)**.

Around one fifth of the observed β-sheets include parallel – anti-parallel combinations, typically as 'right-handed' turns – as shown for thioredoxin (**Fig. 18**). Often, the β-sheet peptide chains (in particular the anti-parallel ones) are connected by hairpin loops that are structurally similar to the "U-turn" depicted in **Fig. 6**.

(a) (b)

Figure 18. The β-sheets in E. coli thioredoxin: **(a)** detail and **(b)** the entire protein.

*Are you on-line? Search the **PDB** and find out how these β-sheets look like in 3D.*

Connecting secondary structure elements we usually find **loops**, which are peptide chains with variable structure, size and shape. *Loops are typically found at the surface of proteins.* When exposed to solvent, protein loops contain polar side-chains. They are essential in defining antibody binding (Ab) and enzyme catalytic sites. For example, the *Fab* (Fragment – antigen binding) regions in antibodies contain 6 loops that provide the required diversity and specificity for appropriate molecular recognition of the corresponding antigens.

Most conserved regions in proteins are usually arranged as α-helices and β-sheets. Loops provide structural variety and are usually associated with *introns* in structural genes. Thus, the *core* of a protein, usually independent of the loop arrangement (length, shape, size), allows proteins with sometimes remarkably different primary sequences to have the same functional role. A good example are the cytochrome P450s – heme-thiolate proteins, that usually share less than 20% sequence homology, yet all function as mono-oxygenases, which act by inserting a single oxygen atom into the substrate. The variable regions (loops, but also α-helices) provide P450s the expected substrate specificity (or lack thereof). Characteristic of P450s at the primary sequence level is the cysteine heme-iron ligand signature:

F-[SNGH]-X-[GD]-X-[+]-X-**C**-[LIVMFAP]-[GAD]

where 'X' is any amino acid, and + can be R (90%), H, and rarely P or T. The cysteine (**bold**) sulfur atom binds to the heme iron (S[-]...Fe), ensuring the activation of the O_2 molecule, required for proper P450 function.

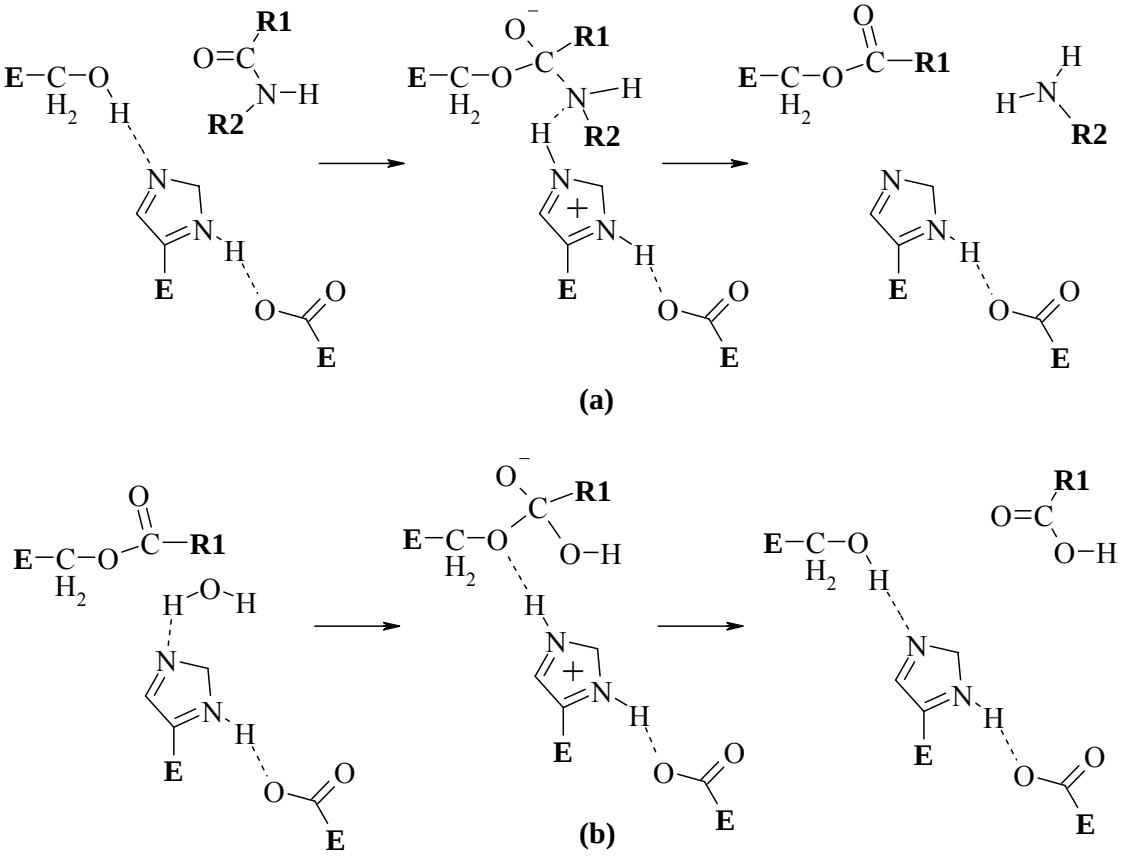


Figure 19. The catalytic mechanism of breaking an amido group in peptides by serine-proteases. The initial step leads to the formation of a free amine (R2-NH_2) and the acyl-enzyme intermediate (a); this is followed by the release **(b)** of the carboxylic acid (R1-COOH).

Serine proteases represent another example of conserved structural core that preserves protein function: All serine-proteases have 3 key residues: **serine**, **histidine** and **aspartate**, which are *structurally* conserved *in 3 dimensions* regardless of the enzyme's total primary sequence(s). This spatial arrangement is conserved in all serine proteases regardless of origin (e.g., viral or animal) and regardless of the protein's function. The catalytic mechanism, illustrated in **Fig. 19**, shows how the scissile amide bond is attacked by an oxygen anion from the serine side-chain. This

is possible because serine donates a proton to the imidazole ring of histidine. The imidazole cation is stabilized, in turn, by aspartate's carboxylate anion (**Fig. 19 a**). During the nucleophilic attack, the amide carbon changes hybridization from the sp^2 to the sp^3 state – which is in fact *the transition state* of this reaction. In the following step, the electron pair of the amide carbon migrates to the amide nitrogen, followed by a proton capture from the histidine cation, thus forming a free amine. The acyl-enzyme intermediate is relatively stable, but is rapidly hydrolised in the presence of water molecules (**Fig. 19 b**). In this step, the transition state is similar to the previous one, as it includes a nucleophilic attack and a tetrahedral (sp^3) carbon. Finally, the initial state of the enzyme is regenerated as the carboxylate moves into solvent.

1.4. Protein structural motifs

When the polypeptide chain folds over itself several times, proteins take a globular shape. Secondary structural elements thus have a tendency to interact with each other in 3 dimensions. If adjacent in the primary sequence, α-helices and β-sheets have a tendency to form **super-secondary structures** – well defined **structural** (folding) **motifs**. This unique property of polypeptide chains proves useful in understanding protein folding, as it significantly decreases the number of possible 3D-arrangements. Even in unrelated proteins, these structural motifs tend to fold in the same manner, even though the actual length of α-helices or β-sheets may differ, and even though loops may vary in length and conformation and the primary sequence may bear no resemblance. Some such super-secondary structures are depicted in **Figure 20**.

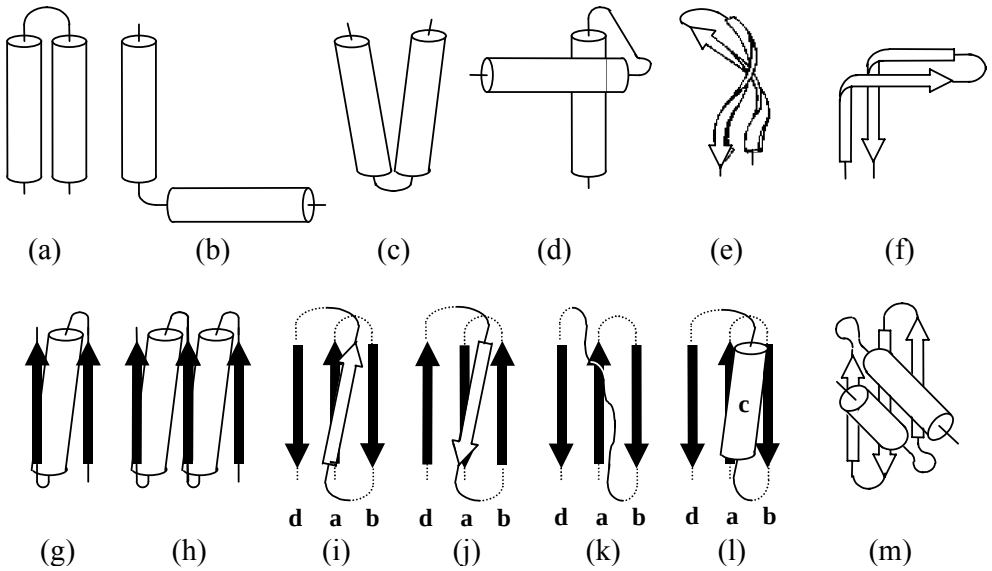


Figure 20. Protein structural motifs (α-helices rendered as cylinders, β-sheets as arrows).

The helix – loop - helix (**Figure 20, a - d**) is quite common in proteins: One usually distinguishes between anti-parallel helices (**20 a**), L- (**20 b**) or V-shaped (**20 c**) motifs, or the triangular-shaped "EF-hand" (**20 d**), which in fact is a chiral arrangement (as it does not overlap with its mirror image). The **EF-hand** (a left-handed super-helix) is found mostly in proteins that have an orthogonal folding pattern: Globulins, albumin and DNA-binding proteins are good examples. **This motif is involved in Ca^{2+}-binding**, e.g. in parvalbumin, calmodulin & troponin C.

Beta-sheets are most always right-hand twisted, if one looks from the main polypeptide chain direction. This arrangement is typical for protein β-sheets, regardless of their connectivity and 3D-assembly. Therefore, one can discuss asymmetry with respect to β-sheets: For example, the ββ hairpin (**Fig. 16 a**) can have a left-hand or right-hand loop, as the 2nd β-sheet is positioned

to the left or to the right of the 1st one (and is in fact a chiral object). When a ββ hairpin is significantly twisted and rotates into a super-helix (**Fig. 20 e**), it is always a right-handed rotation when viewed from the concave side of the super-helix. When a long β-sheet folds over itself such that its two halves are orthogonally folded into separate layers, it is also a right-handed rotation when viewed from the concave side. This implies that β-sheets "rotate" along an imaginary axis when transitioning from a layer to the next. This structural motif (**Fig. 20 f**) is right-handed in all known proteins, e.g., ribonuclease S, HIV-1 protease, serine-proteases and retinol-binding protein. Triple β-sheets can have an S- or a Z- shape (see also **20 m**).

Other examples include the βαβ (**20 g**) and βαβαβ (**20 h**) motifs, composed of alternating β-sheet/α-helix units. The polypeptide chain is always folded in a right-handed super-helix, with parallel β-sheets. The βαβ super-helix is the major structural element in α/β proteins; the right-handed rotation can be explained by the tendency of β-sheets to twist to the right, and by the fact that shorter loops are required for this, as opposed to the left-handed turn. The loops can often play a structural role or may be found in the binding site: Triosophosphate isomerase contains only βαβ repeats, and in one case 2 such units share a β-sheet (the βαβαβ — **20 h**).

The **abcd** unit (four consecutive β-sheets, **a**, **b**, **c** & **d**), is a structural motif that occurs in two-layered β-proteins having parallel axes: The **b**, **c** & **d** sheets form a right-handed super-helix, while the **a** sheet is positioned between **b** & **c**. There are two **abcd** variants which have the same overall folding pattern, but different polypeptide chain direction (**Fig. 20 i** and **20 j**). The **c** unit in **abcd** can also have an irregular conformation (**Fig. 20 k**). Such motifs are found in, e.g., the carboxypeptidase A inhibitors. Two-layered α/β proteins can also have an α-helix as the **c** element in the **abcd** unit (**Fig. 20 l**). An S-shaped triple β-sheet (**2.20 m**) flanked by 2 α-helices can also form a right-handed super-helix; this particular motif is named αSα super-helix. In fact, one or both α-helices can be replaced by a β-sheet (the βSα and βSβ super-helix, respectively). The list can go on with additional, more complex super-secondary structures in which 3 to 8 secondary structural units can form well-defined three dimensional arrangements. The fact that unrelated protein sequences fold in the same manner, and that smaller proteins are often similar to super-secondary structural motifs of larger proteins indicates that these arrangements are thermodynamically stable and can fold as such. One can regard them as "building blocks" that contribute to the tertiary and quaternary folds of proteins. The study of these structural motifs is also useful in modeling and predicting 3D structures of proteins.

1.5. Protein tertiary and quaternary structures.

Tertiary structures indicate the overall 3D arrangement of a protein peptide chain as it produces a more globular, complex structure. For example, the 153-AA myoglobin has 8 and 5 loops that produce a particular spatial arrangement, as shown in **Figure 21**. Globular proteins fold in such a manner as to create a relatively hydrophobic protein interior, while exposing polar side-chains to the surface. On the other hand, G-protein coupled receptors (GPCRs), that have 7 trans-membrane R_α helices, create a more hydrophilic protein interior (as the α-helices interface with each other), as opposed to the more hydrophobic protein exterior that interfaces with the membrane environment.

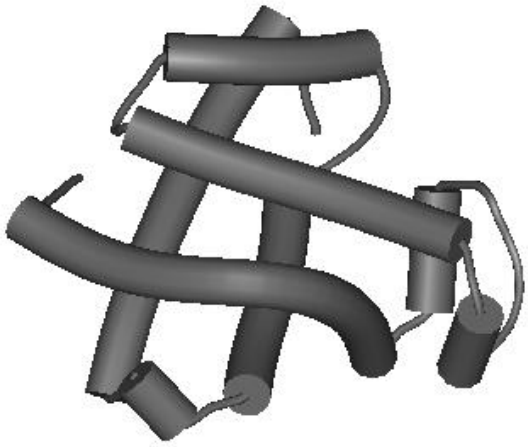


Figure 21. The tertiary structure of a protein indicates the 3D arrangement of secondary structures (in the above example, 8 α-helices).

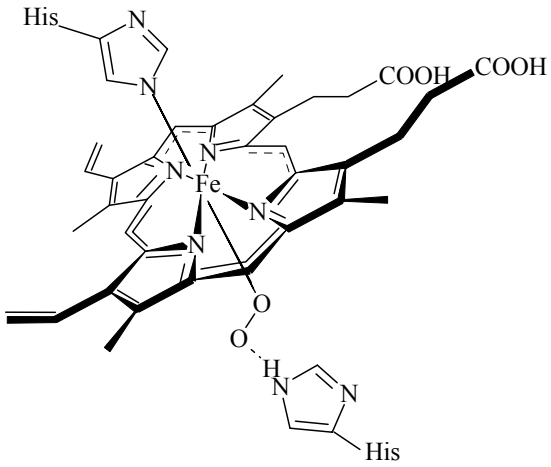


Figure 22. The heme group in myoglobin showing the helix 8 histidine bound directly to the Fe atom, and the helix 7 histidine stabilizing the O_2 molecule (itself Fe-chelated).

Myoglobin includes a heme group – a tetrapyrolic nucleus with a central Fe atom, which binds oxygen (**Fig. 22**). The heme is present in oxygen-binding proteins (hemoglobin, myoglobin), electron-transporting proteins (cytochromes), as well as cytochrome P450s (heme-thiolate mono-oxygenases). As Fig. 22 indicates, nitrogens bind four of the six valences of the Fe atom; the 5th position is typically bound to the protein (e.g., His in myoglobin and Cys in P450s), while the 6th position can bind an O_2 molecule. The heme group in myoglobin and hemoglobin has a mostly non-polar surface-exposed region. Exceptions are two His residues (from helices 7 and 8) that play a functional role, as illustrated in **Fig 22**. In cytochrome P450 enzymes, one His is replaced by Cys (heme-thiolate), whereas the 2nd one is replaced by water in the native state and by O_2 in the active state.

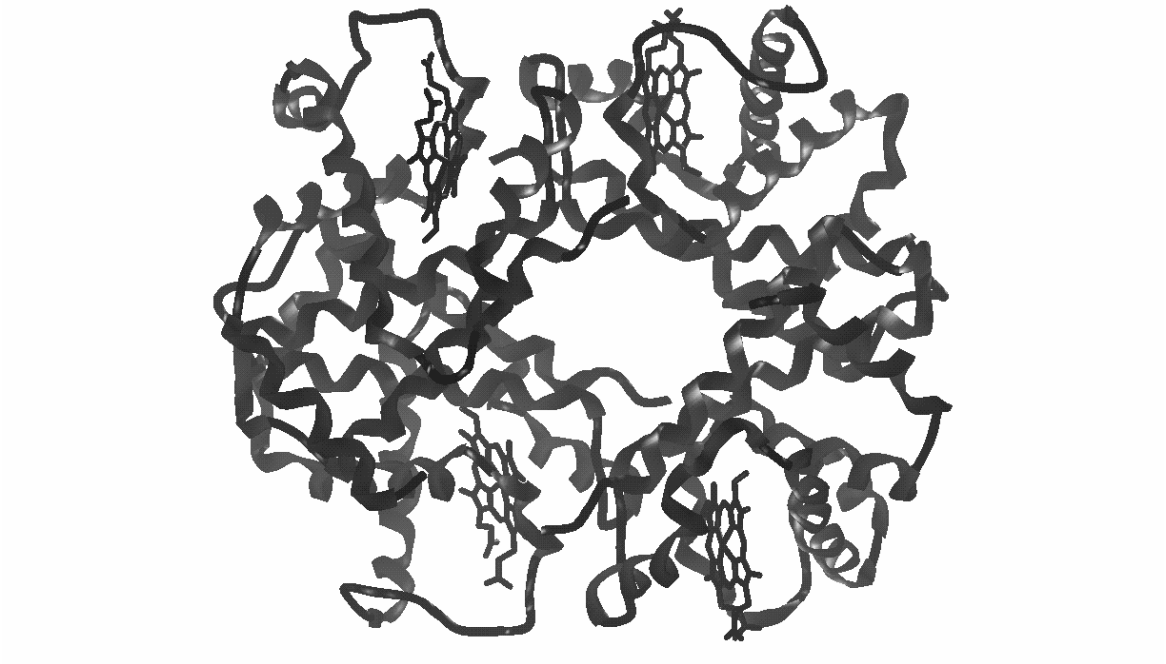


Figure 23. The quaternary structure of hemoglobin. The four heme groups are clearly visible, each heme attached to one of the four monomers. Each monomer is similar to myoglobin.

The quaternary structure reflects the highest (4^{th}) level of structural organization and relates to the pre-defined interaction between two or more polypeptide chains that define a biologically active protein. Many proteins are, in fact, oligomers composed of two or more subunits. Hemoglobin, which is the major protein component of red blood cells, is a tetramer that contains one of each of the α and β subunits (not to be confused with the α-helix & β-sheet). The four monomers fold into a nearly spherical arrangement, 55 Å in diameter – **Fig. 23**. The tertiary structure of these monomers resembles that of myoglobin , even though the primary sequence is different. Each polypeptide chain contains a heme group; most inter-monomer interactions occur between each α and β subunit, with very few inter-α or inter-β interactions.

The central cavity in hemoglobin plays a functional role, as it binds 2,3-BPG (2,3 Bisphospho-D-glycerate), an allosteric regulator of hemoglobin function. This minor product of glucose metabolism is predominantly bound by electrostatic interactions to hemoglobin – its four negative charges (at pH 7.4) salt-bridged by six positive charges from the two β subunits. As 2,3-BPG accumulates, it binds to hemoglobin and in turn decreases its O_2 affinity, resulting in improved O_2 delivery in (oxygen-deprived) tissues. O_2 delivery in tissues is a direct function of a particular tissue's metabolic needs (not only because of 2,3-BPG concentration but also because of CO2 accumulation and pH decrease in blood – known as the Bohr effect. Lungs have a relatively low concentration of 2,3-BPG, allowing up to 4 O_2 molecules to bind to one hemoglobin.

Beside the stabilizing interactions characteristic of secondary structures, we also encounter forces that stabilize tertiary and quaternary structures – mostly due to the **R**-groups found in side-chains: *(i)* hydrogen bonds; *(ii)* hydrophobic interactions; *(iii)* salt bridges between anions and cations (e.g., an aspartate's carboxylate with an Arginine guanidinium); and *(iv)* the S-S bonds (the only covalent bonds).

Protein structure prediction starting from the amino acid sequence requires, to start with, the prediction of the structural motifs (secondary structure). The overall fold (e.g., globule or fibril) needs to be established as a first step. To this end, one should examine the hydrophobic character and the formal charge (at pH 7) of the side-chain R-groups. In globular proteins, the majority of the hydrophobic R-groups would face the protein interior, forming an "oil drop" that is

not in contact with the aqueous solvent; the hydrophilic R-groups are, for the most part, exposed to solvent. If the ratio between the sum of R-groups hydrophobicity and the total protein charge is (much) greater than one, it is likely that we are looking at a globular protein; conversely, if this ratio is (much) smaller than one, a fibril protein is expected. Strongly hydrophilic R-groups, usually solvent-exposed, provide a protein's antigenic character (the 3D pattern that may, or not, be recognized by the immune system's antibodies). **Regardless of the overall fold, protein density is relatively constant – approximately 5 times higher than water.**

Read about protein folding comparison servers – reviewed on-line.

1.6. Protein Folding.

Recent studies indicate that protein folding starts as a local event, with secondary structures being formed before the tertiary and quaternary ones. This supports the hypothesis that secondary structures are by and large pre-determined according to the primary sequence *(local information).* However, the information related to folding at the 3rd and 4th levels appears to be diffuse. Mutation studies in proteins support the conclusion that conformational specificity and folding stability are not related: Often, mutations lead to reduced stability, yet have little effect on the overall fold. Predominantly helical in the wild-type form, lysozyme – a globular, soluble protein – becomes unstable and shifts toward an anti-parallel β-sheet organization as a result of mutations. As a result, mutant lysozyme becomes insoluble and tends to aggregate. A similar mechanism has been inferred for amyloid, a protein involved in Alzheimer's disease *(senile dementia).*

Kinetic studies indicate that the formation of tertiary structures requires some time. The 129 AAs lysozyme, a monomeric protein, is organized in 4 α-helices that form a 'domain' that includes the N- and C- terminal parts, as well as a 2nd 'domain' formed by a triple-chain anti-parallel β-sheet and a relatively long loop. One 3_{10}—helix each is present in both domains, which are linked by a short double-chain anti-parallel β-sheet. The structure is stabilized by four S-S bridges. *In vitro,* the transition of denatured lysozyme (having the S-S bonds intact) to its native fold takes 2 seconds.

In vivo, however, the folding of proteins appears to be assisted by specialized proteins named molecular chaperones. These proteins can interact with non-native conformations of other proteins (not-folded), possibly by recognizing solvent-exposed hydrophobic sequences that would otherwise be buried in native conformations; they do not appear to interact with non-polar residues that are solvent-exposed in the native state. Chaperones do not appear to interact with any one primary sequence in particular, thus being able to prevent incorrect folding and incorrect matching between different monomers. The function of preventing incorrect aggregation appears to continue even after the proteins are correctly folded. Key players in the protein folding events are the following chaperones: Hsp60 (Hsp – *heat-shock protein*) & Hsp70, also known as DnaK, DnaJ, GrpE (a protein that regulates DnaK), GroEL (Hsp60) and GroES (a protein that regulates GroEL); these proteins play a role under both normal conditions and under stress. DnaK – a 70 kD protein, has a C-terminal binding site for the nascent, unfolded protein (effectively, a substrate), and an N-terminal ATP-binding site. DnaJ also binds the substrate and facilitates its binding to DnaK. The 20 kD GrpE interacts with the DnaK ATP-binding site, facilitating ADP release. GroEL has 14 subunits, 57 kD each, placed in two heptameric rings that form a double thorus with a 45 Å wide central cavity. GroES appears to have a single heptameric ring that can bind to one or both of the GroEL thoroidal ends. The mechanism of action of these chaperones is depicted in **Fig. 24**.

DnaK & DnaJ are co-translationally bound to the nascent protein as during ribosomal synthesis (**24 a**). These proteins form a stable ternary complex. DnaJ stabilizes DnaK in the ADP-bound form that has a high affinity to the unfolded, *in statu nascendi*, protein (**24 b**). This complex prevents incorrect folding, as well as mis-matched aggregation to other proteins present in the cell. Appropriate folding of the nascent protein, however, requires it to be released from the ternary

complex. Enter GrpE, which interacts with DnaK's ATP binding site and facilitates ADP release. This step is rapidly followed by another ATP binding to DnaK while the ternary complex dissociates; it is the ATP binding step, not ATP hydrolysis, that is the key step in releasing the nascent substrate (**24 c**). The polypeptide chain could now – in principle – sample the available conformational space to fold in the pattern compatible with its function.

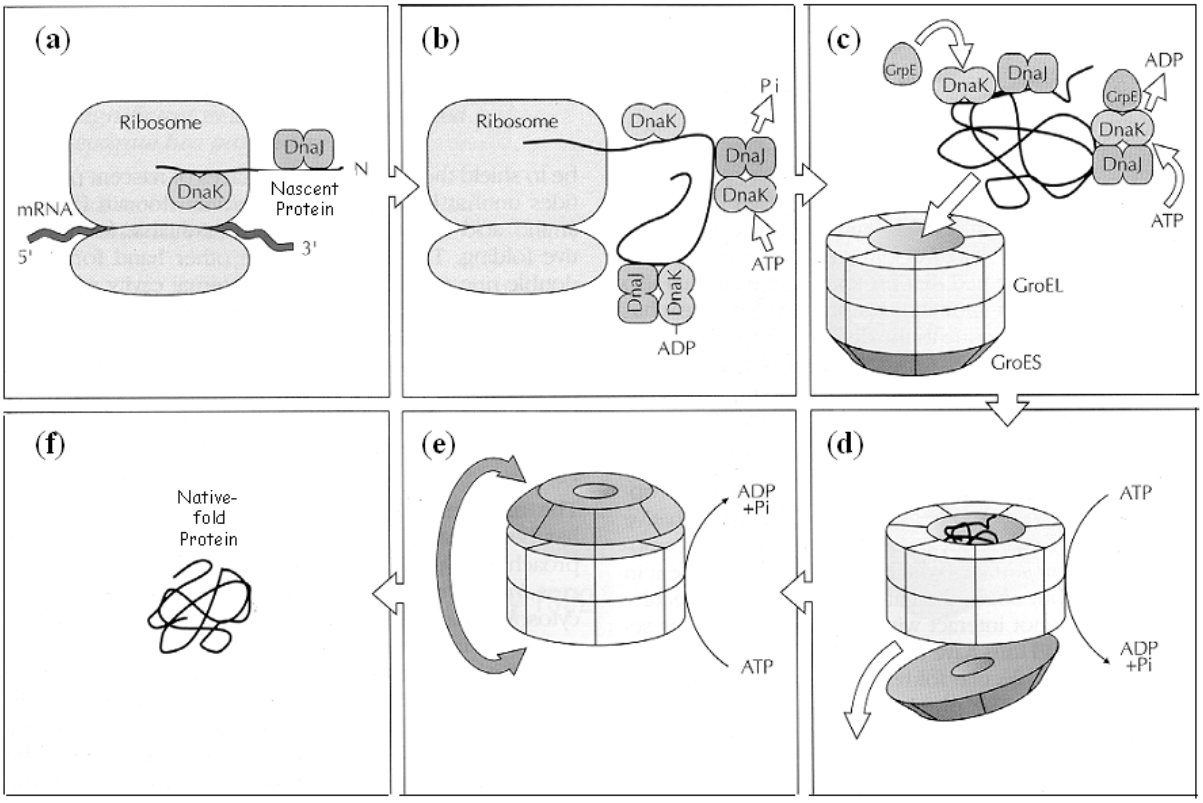


Figure 24. The mechanism of action of molecular chaperones: Pi is inorganic phosphate.

Often, efficient packing cannot occur unassisted, and the GroEL/GroES complex facilitates the process, which takes place in the GroEL central cavity. Under physiological conditions, GroEL is asymmetrically bound to GroES at one end of the thorus. The complex requires GroES-regulated ADP binding to GroEL. The nascent protein binds to the GroEL cavity using multiple interaction points – which should in effect stop its transition into the native fold; however, as this occurs, GroEL's affinity for ADP drops, and the ADP release causes GroES to dissociate from GroEL. This makes room for another ATP to bind to GroEL, in turn decreasing its affinity for the nascent protein, which is released – partially or totally – in the cavity. In some cases, when the nascent protein is strongly bound to GroEL, ATP hydrolysis appears to provide additional dissociation energy (**24 d**). GroES again binds to GroEL, even before the ATP → ADP hydrolysis. The time lapsed between the ATP-GroEL binding and the regeneration of the ADP-bound state (post-hydrolysis) is the actual time when folding occurs. If the folding process has not reached its final stage, the partially folded protein continues to be recognized by GroEL which in turn can bind again the protein with high affinity (as ADP/GroES are again present). In this manner, one or several 'folding cycles' can take place (**24 e**), as documented for some large proteins: The complete folding of one Rhodanase molecule requires 130 ATP → ADP hydrolytic steps. Once completely folded into its native state, molecular recognition sites inside the GroEL cavity no longer recognize the protein, which is now released into the environment (**24 f**).

Filename: ProteinStructure.doc
Directory: C:\TudorData\newwork\BioMed505\Proteins
Template: C:\Documents and Settings\nobody\Application Data\Microsoft\Templates\Normal.dot
Title: Proteine
Subject:
Author: TIO, LK
Keywords:
Comments:
Creation Date: 9/2/2004 11:38 PM
Change Number: 6
Last Saved On: 9/3/2004 1:06 AM
Last Saved By: nobody
Total Editing Time: 97 Minutes
Last Printed On: 9/3/2004 1:16 AM
As of Last Complete Printing
Number of Pages: 21
Number of Words: 7,218 (approx.)
Number of Characters: 41,147 (approx.)