Protein classification on the basis sequence
Protein classification on the basis structure
Why classify proteins according to the structure?

Polypeptides are most often characterized on the basis of their biological activity/function (e.g. catalytic proteins, transport proteins). An alternative categorization of proteins into groupings or families may be made on the basis of polypeptide sequence similarities, which imply similar structural and/or functional attributes. However, it is now clear that there exists a far greater degree of sequence diversity as opposed to structural diversity in the protein world. There appears to be no more than 1000–1500 different protein folds in existence, which form the building blocks of the tens of millions of proteins in existence. It follows that various different sequences, which in themselves display little or no sequence similarity, can in fact yield very similar higher-order structural elements in proteins. One consequence of this is that sequence-based approaches such as multiple alignments will not identify all proteins displaying homology/functional similarity.

Protein classification on the basis of structure

Two best known protein structural classification databases are including:
- ❖SCOP (Structural Classification of Proteins) database
- ❖CATH database

SCOP http://scop2.mrc-lmb.cam.ac.uk/

A motivation for this classification is to determine the evolutionary relationship between proteins.

Class is determined from the overall composition of secondary structure elements in a domain.
1. all-α, those whose structure is essentially formed by α-helices;
2. all-β, those whose structure is essentially formed by β-sheets;
3. α/β, those with α-helices and β-strands;
4. α+β, those in which α-helices and β-strands are largely segregated;

A **fold** describes the number, arrangement, and connections of these secondary structure elements.

A **superfamily** includes domains of similar folds and usually similar functions, thus suggesting a common evolutionary ancestry. Families whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, the variable and constant domains of immunoglobulins.

A **family** usually includes domains with closely related amino acid sequences (in addition to folding similarities). Proteins are clustered together into families on the basis of one of two criteria that imply their having a common evolutionary origin: first, all proteins that have residue identities of 30% and greater; second, proteins with lower sequence identities but whose functions and structures are very similar; for example, globins with sequence identities of 15%.

Although the numbers of unique folds, superfamilies, and families increase as more genomes are known and analyzed, it has become apparent that the number of protein domains in nature is large but limited.

Proteins displaying significant similarity in primary sequence and tertiary structure and/or function are classified as belonging to the same protein family. Family members generally display a strong evolutionary relationship. Members of two or more protein families, although displaying little direct sequence similarity, may share considerable higher-order structural and functional similarities. Such families are grouped into superfamilies, and are likely to share an evolutionary relationship, albeit a distant one.

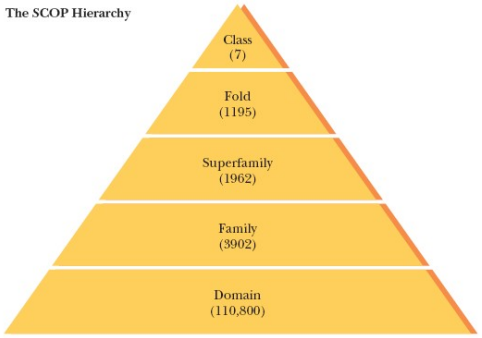
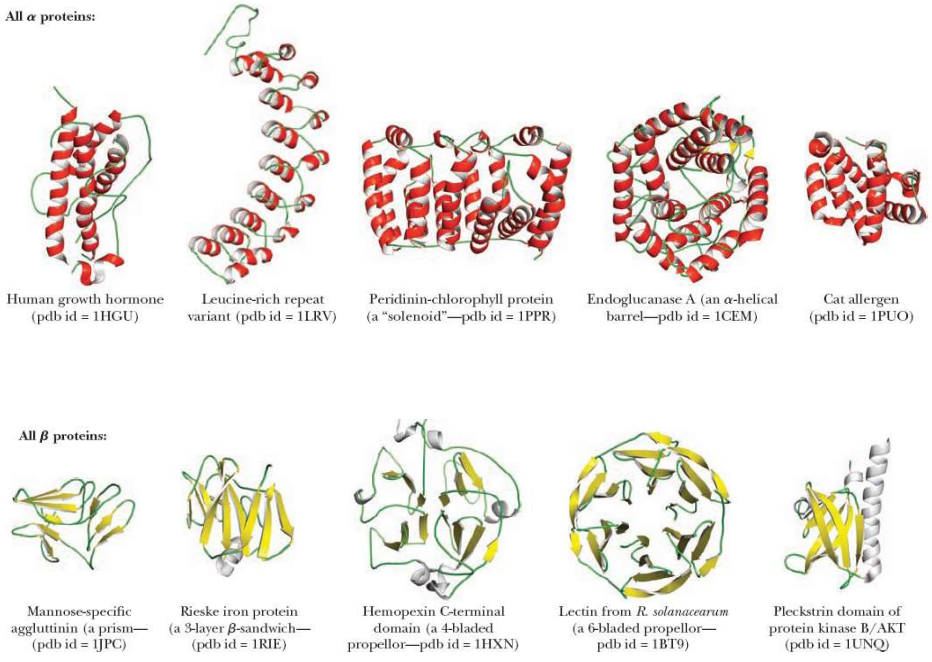
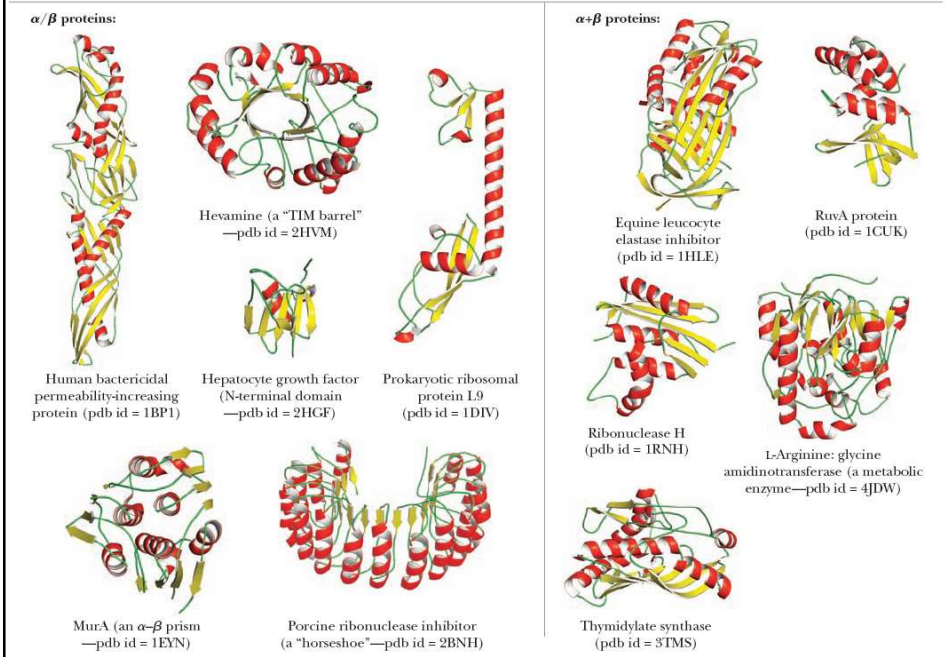


FIGURE 6.27 SCOP and CATH are hierarchical classification systems for the known proteins. Proteins are classified in SCOP by a manual process, whereas CATH combines manual and automated procedures. Numbers indicate the population of each category.





Higher structure determination

X-ray diffraction

NMR (Nuclear magnetic resonance)

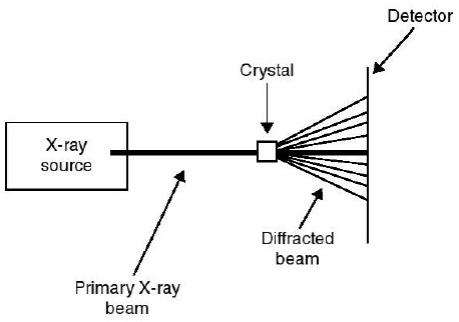


Figure 2.12 Overview of the principles of X-ray diffraction. Refer to text for details.

Prerequisite for protein x-ray diffraction: The generation of protein crystals

Why is difficult crystallize of globular (large) proteins?

Which methods are employed for protein crystallize?

Vapour diffusion or dialysis

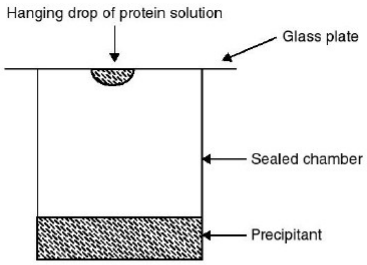


Figure 2.13 Growth of protein crystals by the vapour diffusion (hanging drop) method. A small (20-μL) drop of a concentrated purified protein solution containing a suitable precipitant (e.g. polyethylene glycol or ammonium sulfate) is placed on a glass surface. This is subsequently inverted and sealed (e.g. with vacuum grease) to the top of a chamber containing a reservoir of the precipitant. The apparatus is then incubated at a temperature of the order of 22°C, resulting in slow evaporation of water from the protein-containing hanging drop. A supersaturated solution is slowly generated, which is conducive to crystal growth.

X-ray diffraction:
Difficulties in inducing many proteins to crystallize
Do not use to determine the structure of protein in free solution (One conformation can be determined)

NMR:
The solution based nature (generates a range of closely related conformational structures)
Used for relatively small proteins

PDB :
database for three dimensional structural information

Working with proteins

Protein extraction: SDS-PAGE, 2D, IEF, Chromatography, HPLC

Protein sequencing methods: Edman and MS

Secondary structure determination: CD

Three-dimenstion structure determination: NMR, X-ray

Protein classification

Protein databases: PDB and UniProt

Protein structural stability

Protein biosynthesis → Folding (native conformation) → Functionally active protein

The final conformation depend on the polypeptide's amino acid sequence

The major stabilizing forces of a polypeptide's overall conformation are:

- ❑ Hydrophobic interactions (most important stabilizing forces)
- ❑ Electrostatic attractions (Hydrogen bond, ionic interactions,..)
- ❑ Covalent linkages (Disulfide bonds)

Polypeptides have extensive networks of **intramolecular hydrogen bonds**, but such bonds don not contribute very significantly to overall conformational stability?

Disulfide bond can **help** stabilize a polypeptide's native three-dimentional structure.

Disulfide bond as a lock

Disulfide bond in intracellular and extracellular proteins

- ❑ Free energy difference between folded and denatured form of a polypeptide (200 a.a.) is about 80-100 kJ/mol which is equal to a few hydrogen bonds. Why?
- ❑ Marginal Stability

The term "protein marginal stability" is used to give account of the low values found for protein unfolding free energies (in the order of the energy needed for breaking a few hydrogen bonds). This implies that the native state is as a thermodynamic state close to the edge with "unfolded states"

Table 2.3 Approximate bond energies associated with various (non-covalent) electrostatic interactions, as compared with a carbon–carbon single bond.

Bond type	Bond strength (kJ/mol)
Van der Waals' forces	10
Hydrogen bond	20
Ionic interactions	86
Carbon–carbon bond	350

Breathing: Allowing small molecules to diffuse in or out of the protein's interior

In addition to breathing, some proteins may undergo more marked (usually reversible) **conformational changes** (such as binding of a substrate to an enzyme or antigen binding to an antibody).

How do proteins shift efficiently and precisely from one conformation to another?

Nuclear magnetic resonance measurements by Dorothee Kern and coworkers have shown that transient hydrogen bonds are made in the conversion from one conformation to another in NtrC, a nitrogen regulatory protein.

(Gardino, A., et al., 2010. Transient non-native hydrogen bonds promote activation of a signaling protein. Cell 139:1109–1118.)

➢**Marginal Stability of the Tertiary Structure Makes Proteins Flexible**

➢**A protein's constituent atoms are constantly in motion** and groups ranging from individual amino acid side chains to entire domains can be displaced via random motion by up to about 0.2nm.

➢A protein's conformation displays a limited degree of flexibility and such movement is termed "**breathing**".

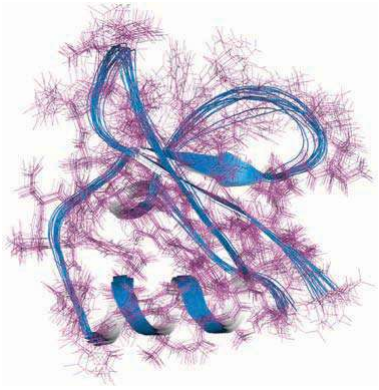


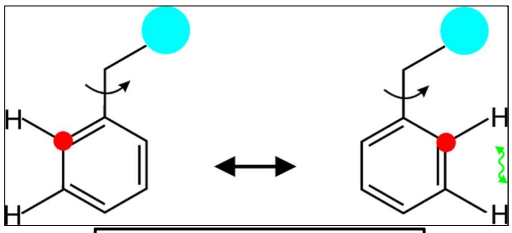
FIGURE 6.35 Proteins are dynamic structures. The marginal stability of a tertiary structure leads to flexibility and motion in the protein. Determination of structures of proteins (such as the SH3 domain of the α-chain of spectrin, shown here) by nuclear magnetic resonance produces a variety of stable tertiary structures that fit the data. Such structural ensembles provide a glimpse into the range of structures that may be accessible to a flexible, dynamic protein (pdb id = 1M8M).

Motion in Globular Proteins

TABLE 6.2 Motion and Fluctuations in Proteins

Type of Motion	Spatial Displacement (Å)	Characteristic Time (sec)	Source of Energy
Atomic vibrations	0.01–1	10^{-15}–10^{-11}	Kinetic energy
Collective motions 1. Fast: Tyr ring flips; methyl group rotations 2. Slow: hinge bending between domains	0.01–5 or more	10^{-12}–10^{-3}	Kinetic energy
Triggered conformation changes	0.5–10 or more	10^{-9}–10^{3}	Interactions with triggering agent
Proline *cis–trans* isomerization	3–10	10^{1}–10^{4}	Kinetic energy or enzyme driven

Adapted from Petsko, G. A., and Ringe, D., 1984. Fluctuations in protein structure from X-ray diffraction. *Annual Review of Biophysics and Bioengineering* **13**:331–371.



Aromatic Ring Flips

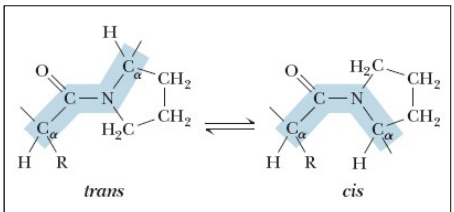


FIGURE 6.36 The *cis* and *trans* configurations of proline residues in peptide chains are almost equally stable. Proline *cis-trans* isomerizations, often occurring over relatively long time scales, can alter protein structure significantly.

Increased thermal stability is generally related to one or more of the following structural adaptations:

- an increase in the number of intramolecular polypeptide hydrogen bonds;
- an increase in the number of salt bridges;
- increased polypeptide compactness (improved packing of the hydrophobic core);
- extended helical regions.

Conversely, enhanced stability/functional flexibility of proteins derived from psychrophiles appears to be achieved by one or more of the following adaptations:

- fewer salt links;
- reduced aromatic interactions within the hydrophobic core (reduction in hydrophobicity);
- increased hydrogen bonding between the protein surface and the surrounding solvent;
- occurrence of extended surface loops.

2.5.1 Secondary structure prediction

Over 20 different methods of secondary structure prediction have been reported (Table 2.4). Traditionally these approaches fall into two main categories.

1. Empirical statistical methods based on data generated from studying proteins of known three-dimensional structure and correlation of primary amino acid sequence of such proteins with structural features.
2. Methods based on physicochemical criteria such as fold compactness (i.e. the generation of a folded form displaying a tightly packed hydrophobic core and a polar surface).

Table 2.5 Conformational preferences and assignments of amino acid residues with regard to stretches of α-helix and β structure.

α-helix			β strand		
Residue	**Pα**	**Assignment**	**Residue**	**Pβ**	**Assignment**
Glu	1.44	Hα	Val	1.64	Hβ
Ala	1.39	Hα	Ile	1.57	Hβ
Met	1.32	Hα	Thr	1.33	hβ
Leu	1.30	Hα	Tyr	1.31	hβ
Lys	1.21	hα	Trp	1.24	hβ
His	1.12	hα	Phe	1.23	hβ
Gln	1.12	hα	Leu	1.17	hβ
Phe	1.11	hα	Cys	1.07	hβ
Asp	1.06	hα	Met	1.01	Iβ
Trp	1.03	Iα	Gln	1.00	Iβ
Arg	1.00	Iα	Ser	0.94	iβ
Ile	0.99	iα	Arg	0.94	iβ
Val	0.97	iα	Gly	0.87	iβ
Cys	0.95	iα	His	0.83	iβ
Thr	0.78	iα	Ala	0.79	iβ
Asn	0.78	iα	Lys	0.73	bβ
Tyr	0.73	bα	Asp	0.66	bβ
Ser	0.72	bα	Asn	0.66	bβ
Gly	0.63	Bα	Pro	0.62	Bβ
Pro	0.55	Bα	Glu	0.51	Bβ

Pα, propensity to form α-helical regions; Pβ, propensity to form β stretches; Hα, strong helix former; hα, helix former; Iα, weak helix former; iα, indifferent; bα, helix breaker; Bα, strong helix breaker. Similar designations are used in the case of β formers, with 'β' replacing 'α'.
Source: reproduced from *Current Protocols in Protein Science* with kind permission of the publisher, John Wiley & Sons, Ltd.

The analysis carried out by Chou and Fasman also allowed the following observations to be made.

- An α-helical stretch is usually initiated by a six-residue sequence containing at least four Hα or hα residues (Table 2.5).
- Proline residues, if present, are located at the amino terminus of the helix.
- Any group of four successive residues present in an α-helix will have an average Pα value greater than 1.0 (Table 2.5).
- A β stretch is usually initiated by a five-residue sequence containing at least three Hβ or hβ residues.
- Any group of four successive residues present in a β stretch will have an average Pβ value greater than 1.0.

Most such traditional predictive methods are at best 50–70% accurate.

Some of the more recently developed programs also take into consideration multiple sequence alignment data but even the most modern programs usually achieve at best 70–75% accuracy. A range of such programs (e.g. APSSP, CFSSP, GOR, J Pred, Prof and SOPMA) are available via the ExPASy home page (see Box 1.1) and can be accessed by following the links pathway: ExPASY home page > proteomics > protein structure.

2.5.2 *Tertiary structure prediction*

Accurate prediction of a protein's three-dimensional structure is a still more complex problem. However, the fact that the architecture of all proteins is largely based on a limited number of building blocks (protein folds) helps in the development of such predictive tools. Moreover, as the number of proteins whose three-dimensional structure is resolved increases, associated bioinformatic analysis will continue to build a better picture of the range of amino acid sequences that can ultimately support the formation of specific protein folds.

Currently, three different approaches may be adopted in an attempt to predict the three-dimensional structure of a polypeptide from primary sequence data:

- comparative modelling;
- fold recognition approaches;
- *ab initio* structural prediction.

Homology modelling (comparative modelling) is applied when the target protein shares substantial sequence similarity to proteins whose three-dimensional structure has already been experimentally established. In this approach initial homology searches are undertaken using tools such as BLAST. Resolved structural details of homologous proteins can then be identified using structural databases such as PDB and CATH, allowing identification of conserved structural regions, as well as more variable regions. These provide a structural template with which the query sequence can be aligned, allowing a model of the target protein to be built. The accuracy of the predicted structure is closely related to the percentage amino acid identity shared by the query protein and its template. If sequence identity stands at 50% or greater, the predicted structure is usually quite accurate. Accuracy declines with decreasing percentage identity, particularly if it falls below about 30%.

Fold recognition approaches (also called threading) are based on the fact that proteins can share characteristic folds even if they are not homologous. Essentially the process entails 'threading' the target sequence (or subsets thereof) onto different known folds, while using software tools to evaluate likely compatibility of the sequence to the fold in question.

Threading is an approach to fold recognition which used a detailed 3-D representation of protein structure.
The idea was to physically "thread" a sequence of amino acid side chains onto a backbone structure (a fold) and to evaluate this proposed 3-D structure using a set of pair potentials and (importantly) a separate solvation potential.

Ab initio (*de novo*) structure prediction is, understandably, the most high-risk approach to structure prediction and is applied in cases where the target sequence lacks detectable homology to any protein of known structure. One common approach to *ab initio* prediction entails comparing short (nine amino acid) sequence fragments of the target protein to resolved protein structures.

A range of protein structural prediction tools (e.g. CPHmodels, ESYPred3D, HHpred and Phyre2) are available via the ExPASy home page (see Box 1.1), and can be accessed by following the links pathway: ExPASY home page > proteomics > protein structure.