Protein Classification

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 (Class Architecture Topology Homologous) database

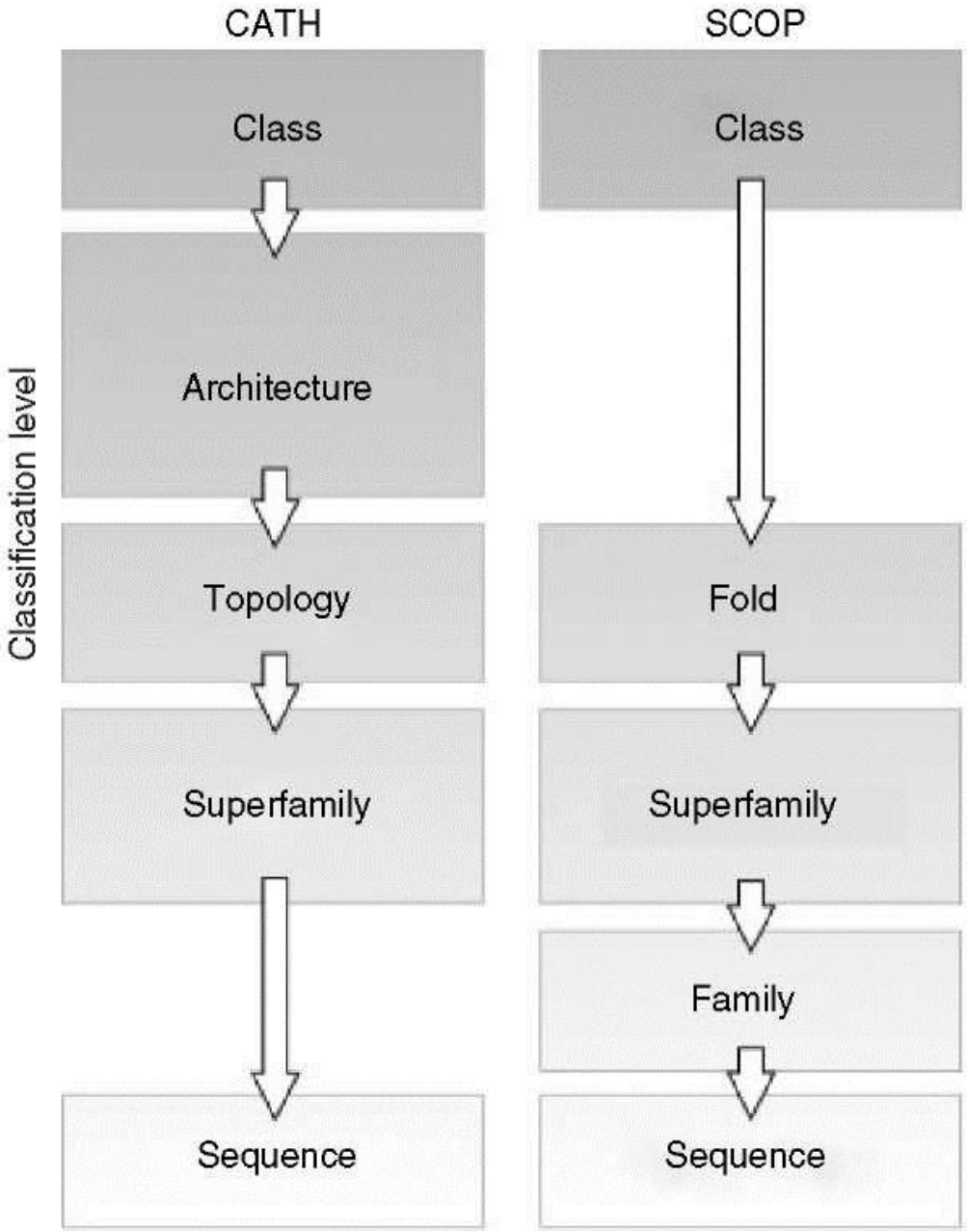
SCOP http://scop2.mrc-lmb.cam.ac.uk/

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.


CATH
SCOP
Class
Class
Architecture
Topology
Fold
Superfamily
Superfamily
Family
Sequence
Sequence
Classification level

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.

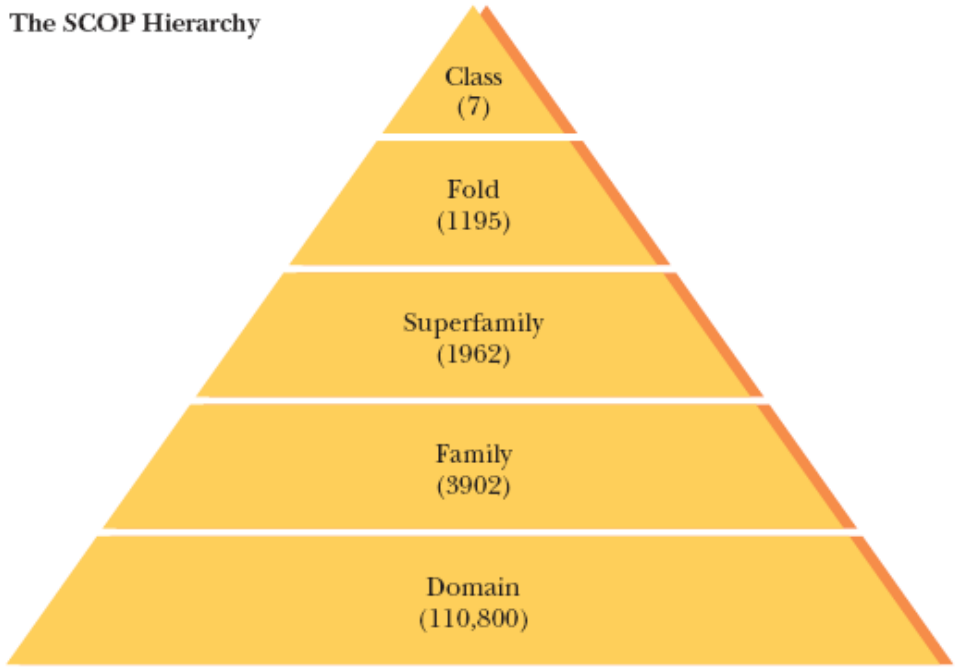
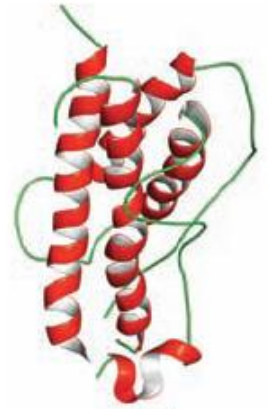
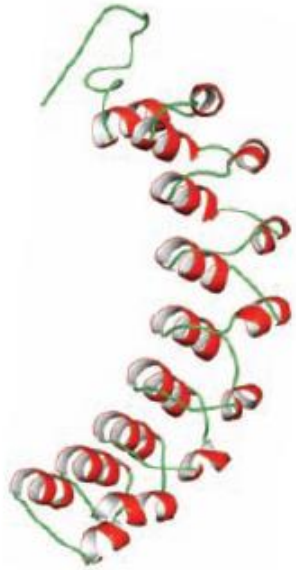


FIGURE 6.27 SCOP and CATH are hierarchical classification systems for the known proteins. Proteins are classified in SCOP by a manual process, whereas CATH combines manual and automated procedures. Numbers indicate the population of each category.

All α proteins:



Human growth hormone (pdb id = 1HGU)



Leucine-rich repeat variant (pdb id = 1LRV)

Peridinin-chlorophyll protein (a "solenoid"—pdb id = 1PPR)

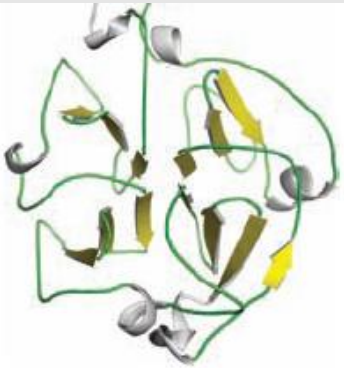
Endoglucanase A (an α-helical barrel—pdb id = 1CEM)

Cat allergen (pdb id = 1PUO)

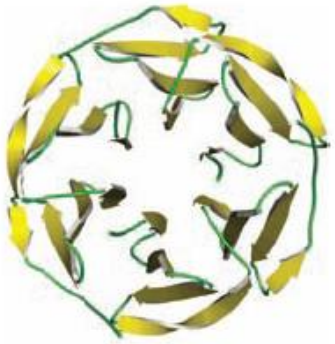
All β proteins:

Mannose-specific agglutinin (a prism—(pdb id = 1JPC)

Rieske iron protein (a 3-layer β-sandwich—(pdb id = 1RIE)



Hemopexin C-terminal domain (a 4-bladed propellor—pdb id = 1HXN)

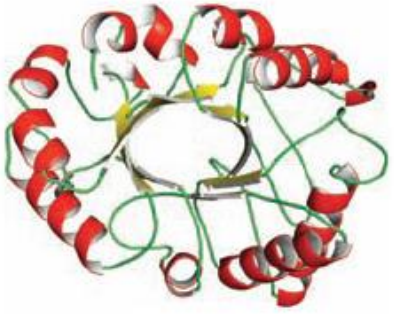


Lectin from *R. solanacearum* (a 6-bladed propellor—pdb id = 1BT9)

Pleckstrin domain of protein kinase B/AKT (pdb id = 1UNQ)

α/β proteins:

Human bactericidal permeability-increasing protein (pdb id = 1BP1)

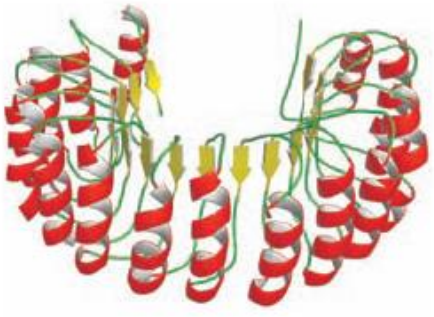


Hevamine (a "TIM barrel" —pdb id = 2HVM)

Hepatocyte growth factor (N-terminal domain —pdb id = 2HGF)

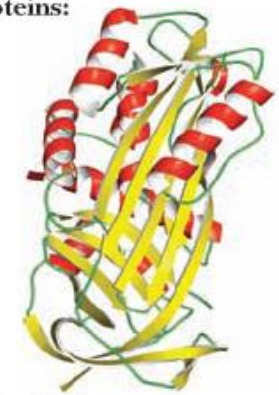
Prokaryotic ribosomal protein L9 (pdb id = 1DIV)

MurA (an α–β prism —pdb id = 1EYN)



Porcine ribonuclease inhibitor (a "horseshoe"—pdb id = 2BNH)

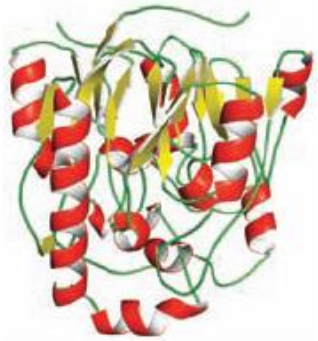
α+β proteins:



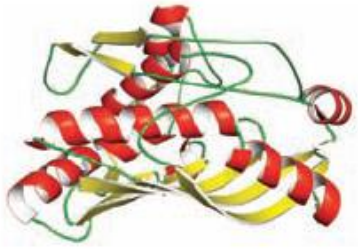
Equine leucocyte elastase inhibitor (pdb id = 1HLE)

RuvA protein (pdb id = 1CUK)

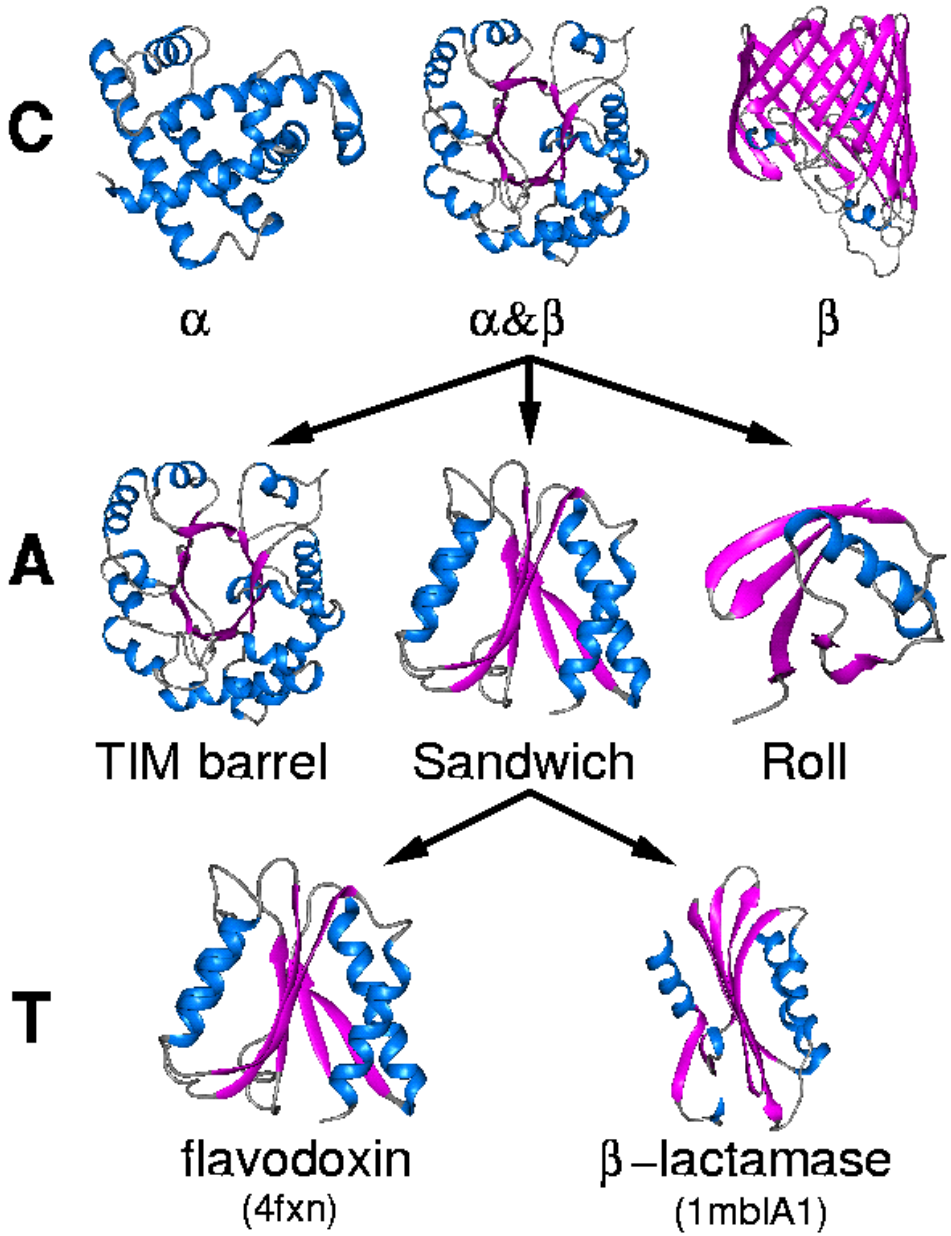
Ribonuclease H (pdb id = 1RNH)



L-Arginine: glycine amidinotransferase (a metabolic enzyme—pdb id = 4JDW)



Thymidylate synthase (pdb id = 3TMS)



The four major hierarchical levels in CATH are class, architecture, topology (or fold) and homologous family. There are currently three major classes recognised in CATH (mainly-alpha, mainly-beta and alpha-beta). Below class, the architecture level simply describes the orientations of the secondary structures in 3D without regard to their connectivity. We currently recognize 28 well-defined architectures and within each architecture, the topology or fold is determined by the connectivity of the secondary structures (figure 1).

Three-dimensional structure determination of Proteins

Higher structure determination

X-ray diffraction

NMR (Nuclear magnetic resonance)

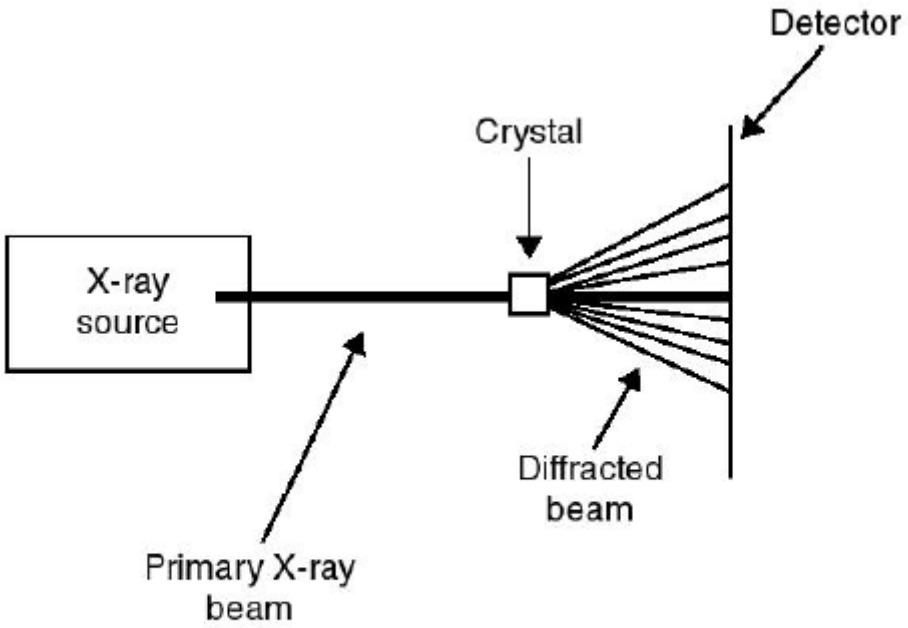


Figure 2.12 Overview of the principles of X-ray diffraction. Refer to text for details.

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

Why is difficult crystallize of globular (large) proteins?

Why is difficult crystallize of globular (large) proteins?

- some proteins, especially **membrane proteins**, are exceedingly difficult to crystallize, but some researchers have solved this problem by genetically re-engineering the protein. This would involve deleting regions that are known to be disordered, etc. Thus when interpreting 3D crystal structures of re-engineered proteins, one should be very careful to consider the artificial changes that have been made to them. Membrane proteins, which have predominantly hydrophobic surfaces, are hard to crystallize because they tend to aggregate in aqueous solutions.
- Proteins which have **post-translational modifications** are also hard to crystallize because the PTMs are usually not uniform among the protein molecules..
- some of the features which make proteins hard to form good crystals (good enough to study them by X-ray diffraction) include the **presence of very flexible regions**, or **the presence of carbohydrate moieties**. Some proteins have inherently disordered regions (IDRs) and those are also hard to crystallize. Most intrinsically disordered proteins (IDP) are impossible to crystallise.
- **Large**, **unstable**, **un-soluble proteins** or **complexes** are hard to prepare and will impede extensive sparse-screening of crystallization conditions. **Structural heterogeneity** causes difficulties in crystallization. Dynamic parts of a protein or subunits exchange in a complex, will prevent the formation of regular crystals.
- Proteins with **high surface hydrophobicity and high surface charge** are not good candidates for forming crystals in solutions.

Which methods are employed for protein crystallize?

Vapour diffusion or dialysis

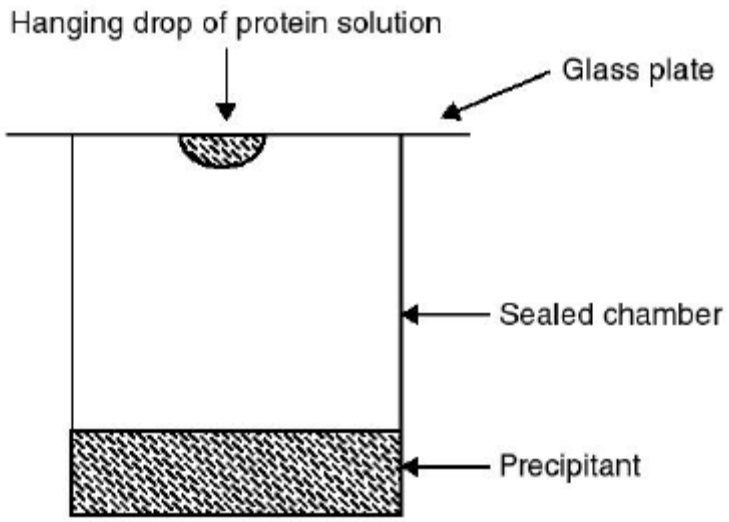


Figure 2.13 Growth of protein crystals by the vapour diffusion (hanging drop) method. A small (20-μL) drop of a concentrated purified protein solution containing a suitable precipitant (e.g. polyethylene glycol or ammonium sulfate) is placed on a glass surface. This is subsequently inverted and sealed (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

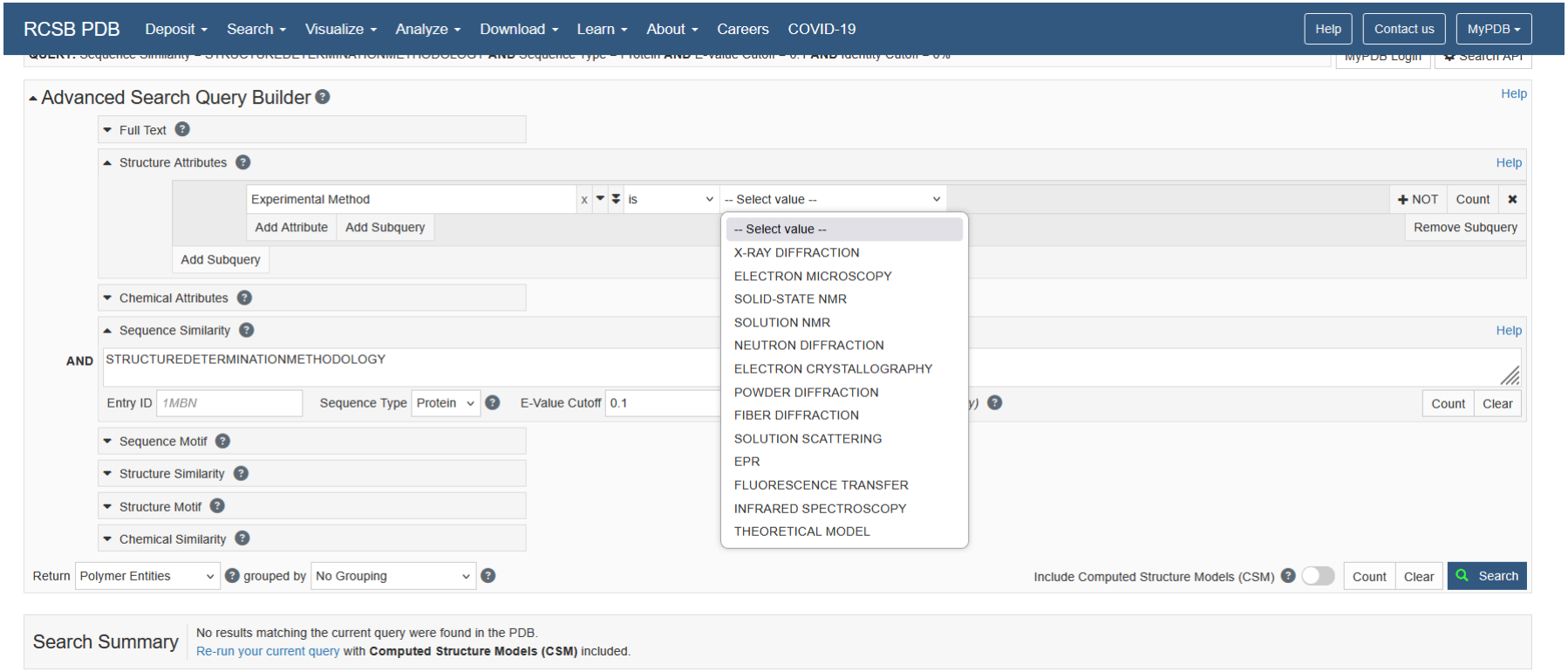
Secondary structure determination: CD

Three-dimenstion structure determination: NMR, X-ray

Protein classification

Protein databases: PDB and UniProt

RCSB PDB
(Research Collaboratory for Structural Bioinformatics PDB)


RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
About
Careers
COVID-19
Help
Contact us
MyPDB
Advanced Search Query Builder
Help
Full Text
Structure Attributes
Help
Experimental Method
is
-- Select value --
+ NOT
Count
Add Attribute
Add Subquery
Remove Subquery
Add Subquery
Chemical Attributes
Sequence Similarity
Help
AND
STRUCTUREDETERMINATIONMETHODOLOGY
Entry ID
1MBN
Sequence Type
Protein
E-Value Cutoff
0.1
Count
Clear
Sequence Motif
Structure Similarity
Structure Motif
Chemical Similarity
-- Select value --
X-RAY DIFFRACTION
ELECTRON MICROSCOPY
SOLID-STATE NMR
SOLUTION NMR
NEUTRON DIFFRACTION
ELECTRON CRYSTALLOGRAPHY
POWDER DIFFRACTION
FIBER DIFFRACTION
SOLUTION SCATTERING
EPR
FLUORESCENCE TRANSFER
INFRARED SPECTROSCOPY
THEORETICAL MODEL
Return
Polymer Entities
grouped by
No Grouping
Include Computed Structure Models (CSM)
Count
Clear
Search
Search Summary
No results matching the current query were found in the PDB.
Re-run your current query with Computed Structure Models (CSM) included.

Feature	X-ray Diffraction	NMR Spectroscopy
Protein Sample State	**Crystalline solid.** The protein must be arranged in a rigid, repeating crystal lattice, providing a static view.	**Free solution.** The protein is studied in a more natural state, allowing for the observation of its flexibility.
Fundamental Principle	A beam of X-rays is passed through a protein crystal, and the resulting diffraction pattern is analyzed.	A strong magnetic field and radio waves are used to probe the magnetic properties of specific atomic nuclei in the protein.
Type of Information	A single, high-resolution 3D structure that represents an average conformation of the protein in the crystal.	An ensemble of closely related 3D structures that reflects the protein's conformational flexibility or "breathing."
Primary Limitation	**The difficulty of crystallization.** Inducing most large, globular proteins to form suitable crystals is a major bottleneck.	**The size of the protein.** The complexity of the data generally limits the technique to relatively small proteins (up to 40-50 kDa).

1. X-ray Diffraction (پراش اشعه ایکس)

Feature	English (مزایا/Pros)	فارسی (مزایا/Pros)
Resolution	**Provides high-resolution protein structural information** and **atomic level structural resolution** (typically 1–1.5 Å).	**اطلاعات ساختاری پروتئین با وضوح بالا** را فراهم می‌کند و **وضوح ساختاری در سطح اتمی** (معمولاً ۱ تا ۱.۵ آنگستروم) را ارائه می‌دهد.
Wavelength	The wavelength of X-rays approximates the dimensions of proteins (ångströms), making it appropriate for visualization.	طول موج اشعه ایکس تقریباً با ابعاد پروتئین‌ها (آنگستروم) برابر است، که آن را برای تجسم مناسب می‌سازد.

Feature	English (معایب/Cons)	فارسی (معایب/Cons)
Sample State	**Requires the protein to be in crystalline form**.	**مستلزم این است که پروتئین به شکل بلوری (کریستالی) باشد.**
Crystallization	Crystallization is a **major bottleneck**, as the vast majority of globular proteins are extremely large, display irregular surfaces, and are difficult to crystallize.	کریستالیزاسیون یک **مانع اصلی** است، زیرا اکثریت قریب به اتفاق پروتئین‌های کروی بسیار بزرگ هستند، سطوح نامنظمی از خود نشان می‌دهند و تبلور آن‌ها دشوار است.
Purity/Conditions	The protein must generally be very pure, and optimal crystallization conditions (pH, salts, etc.) must usually be determined by direct experimentation.	پروتئین عموماً باید بسیار خالص باشد و شرایط بهینه تبلور (pH ، نمک‌ها و غیره) معمولاً باید از طریق آزمایش مستقیم تعیین شوند.
Crystal Properties	Protein crystals are soft and relatively easily destroyed. They contain significant solvent-filled channels/pores, typically occupying 30–80% of the crystal volume.	بلورهای پروتئینی نرم هستند و نسبتاً به راحتی تخریب می‌شوند. آنها حاوی کانال‌ها یا منافذ قابل توجهی پر از حلال هستند که معمولاً ۳۰ تا ۸۰ درصد از حجم بلور را اشغال می‌کنند.

2. Nuclear Magnetic Resonance (NMR)

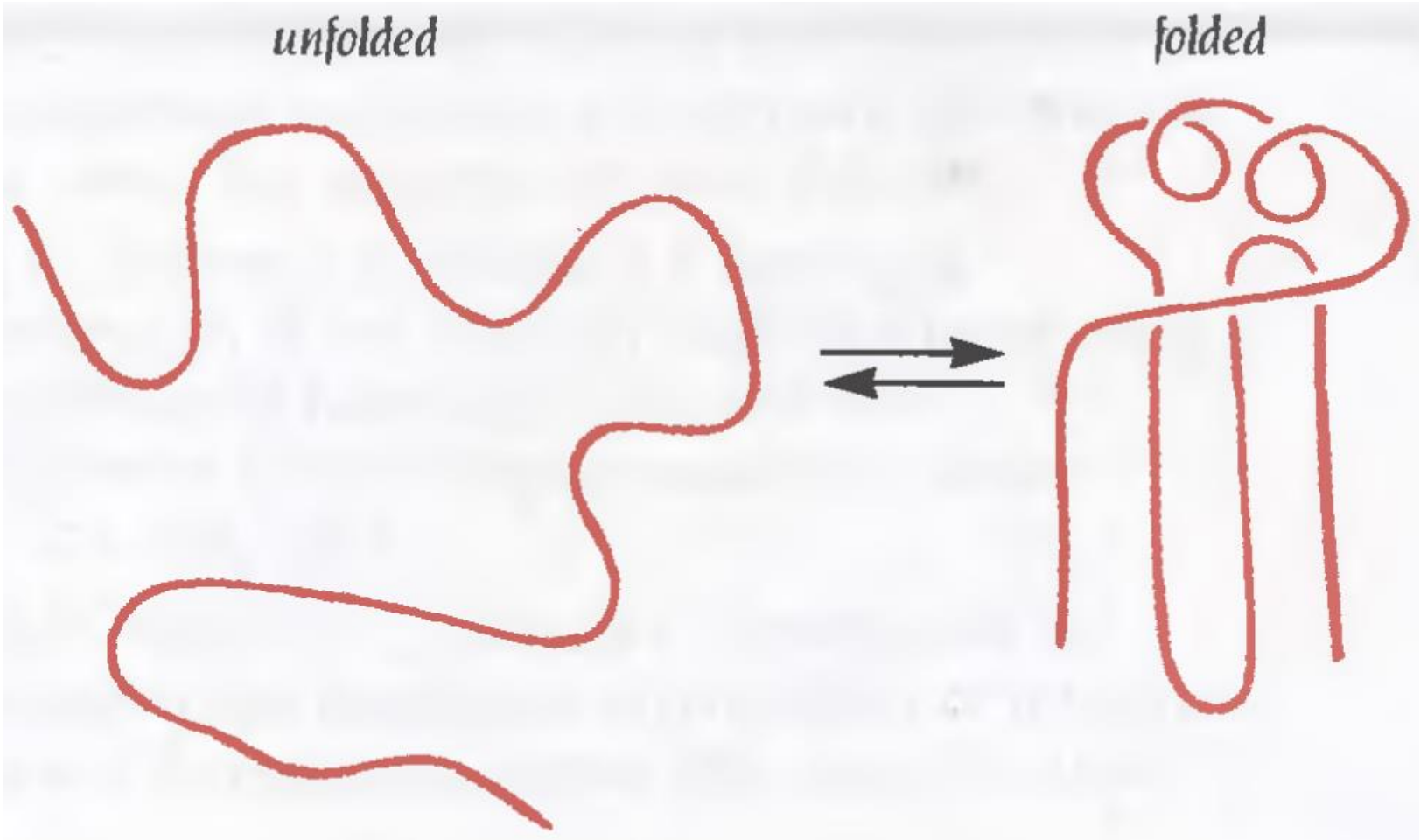
Feature	English (مزایا/Pros)	فارسی (مزایا/Pros)
Resolution	A technique widely used to obtain **high-resolution protein structural information**.	تکنیکی است که به طور گسترده برای به دست آوردن **اطلاعات ساختاری پروتئین با وضوح بالا** استفاده می‌شود.
Sample State	May be used to determine the structure of **proteins in free solution** (solution-based nature).	می‌توان از آن برای تعیین ساختار **پروتئین‌ها در محلول آزاد** (ماهیت مبتنی بر محلول) استفاده کرد.
Conformational Data	Generates a **range of closely related conformational structures**, which reflects the fact that protein conformation can flex or 'breathe' in solution.	**مجموعه‌ای از ساختارهای کانفورماسیونی نزدیک به هم** را ایجاد می‌کند که نشان‌دهنده این واقعیت است که کانفورماسیون پروتئین می‌تواند در محلول انعطاف‌پذیر باشد یا "نفس بکشد."

Feature	English (معایب/Cons)	فارسی (معایب/Cons)
Complexity	The **complexity of the technique**, and particularly the data generated, is a drawback.	**پیچیدگی تکنیک** و به ویژه داده‌های تولید شده توسط آن، یک نقطه ضعف است.
Size Limitation	The complexity often **limits the use of this approach to relatively small proteins**. However, recent technical advances now render practicable the analysis of proteins of 40–50 kDa or more.	پیچیدگی اغلب **استفاده از این روش را به پروتئین‌های نسبتاً کوچک محدود می‌کند** .با این حال، پیشرفت‌های فنی اخیر اکنون تحلیل پروتئین‌های ۴۰ تا ۵۰ کیلو دالتون یا بیشتر را عملی کرده است.

این دو روش را می‌توان مانند عکاسی و فیلمبرداری از یک سازه مقایسه کرد: پراش اشعه ایکس (X-ray) مانند یک عکس فوری با وضوح بسیار بالا از سازه در حالت ثابت (بلوری) است، در حالی که NMR به مثابه گرفتن یک سری عکس‌های متحرک یا یک فیلم از آن سازه در حال حرکت و انعطاف‌پذیری (در محلول) است.

Protein structural stability

Folding and Flexibility



- The process by which a polypeptide chain acquires its correct three-dimensional structure to achieve the biologically active native state is called protein folding.
- Some polypeptide chains spontaneously fold into the native state, others require the assistance of enzymes for example to catalyze the formation and exchange of disulfide bonds; and many require the assistance of a class of proteins called chaperones.
- After a polypeptide has acquired most of its correct secondary structure, with the α-helices and β-sheets formed, it has a looser tertiary structure than the native state and is said to be in the molten globular state. The compaction that is necessary to go from the molten globular state to the final native state occurs spontaneously.

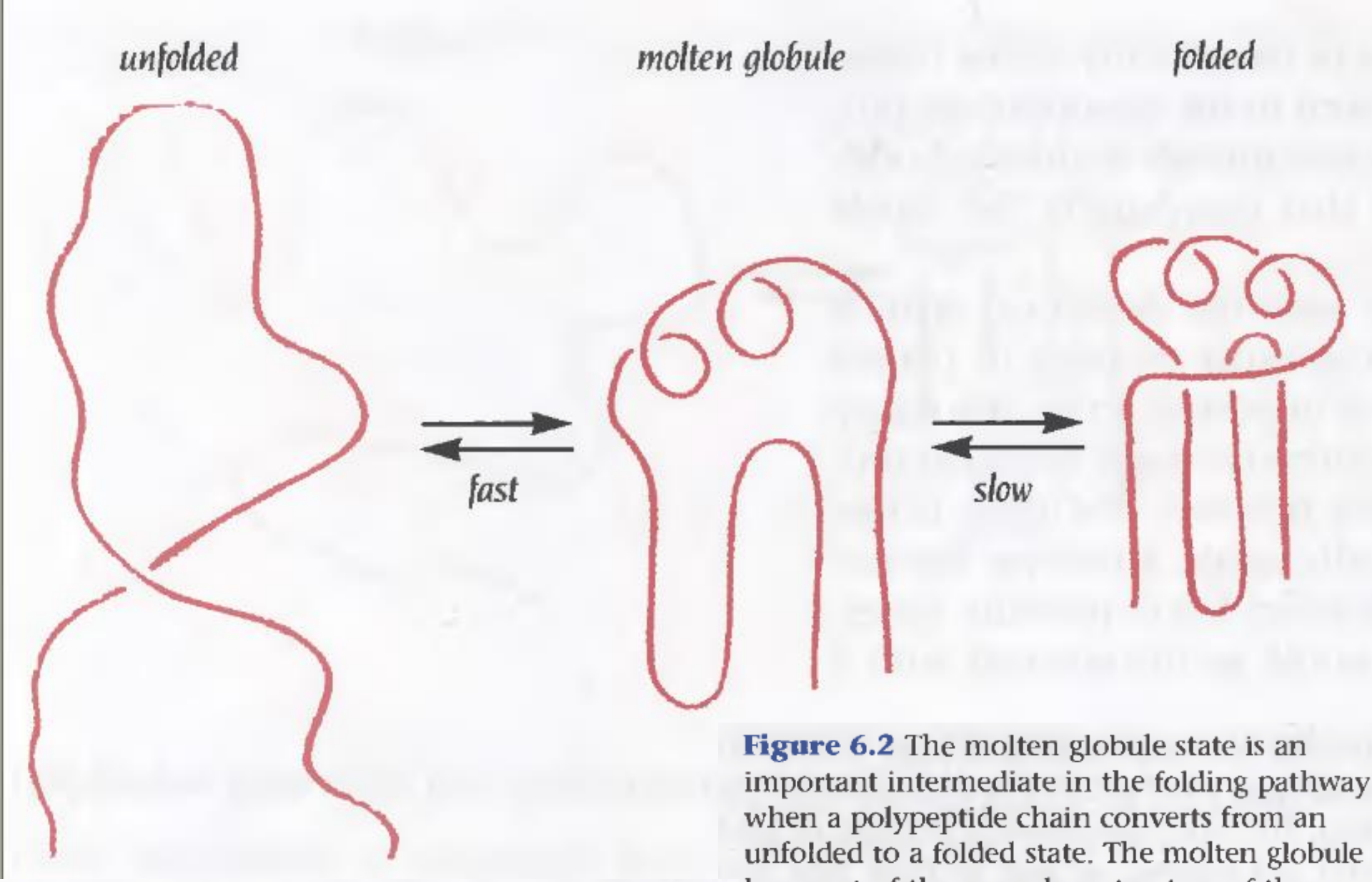


Figure 6.2 The molten globule state is an important intermediate in the folding pathway when a polypeptide chain converts from an unfolded to a folded state. The molten globule has most of the secondary structure of the native state but it is less compact and the proper packing interactions in the interior of the protein have not been formed.

Molten globules are intermediates in folding

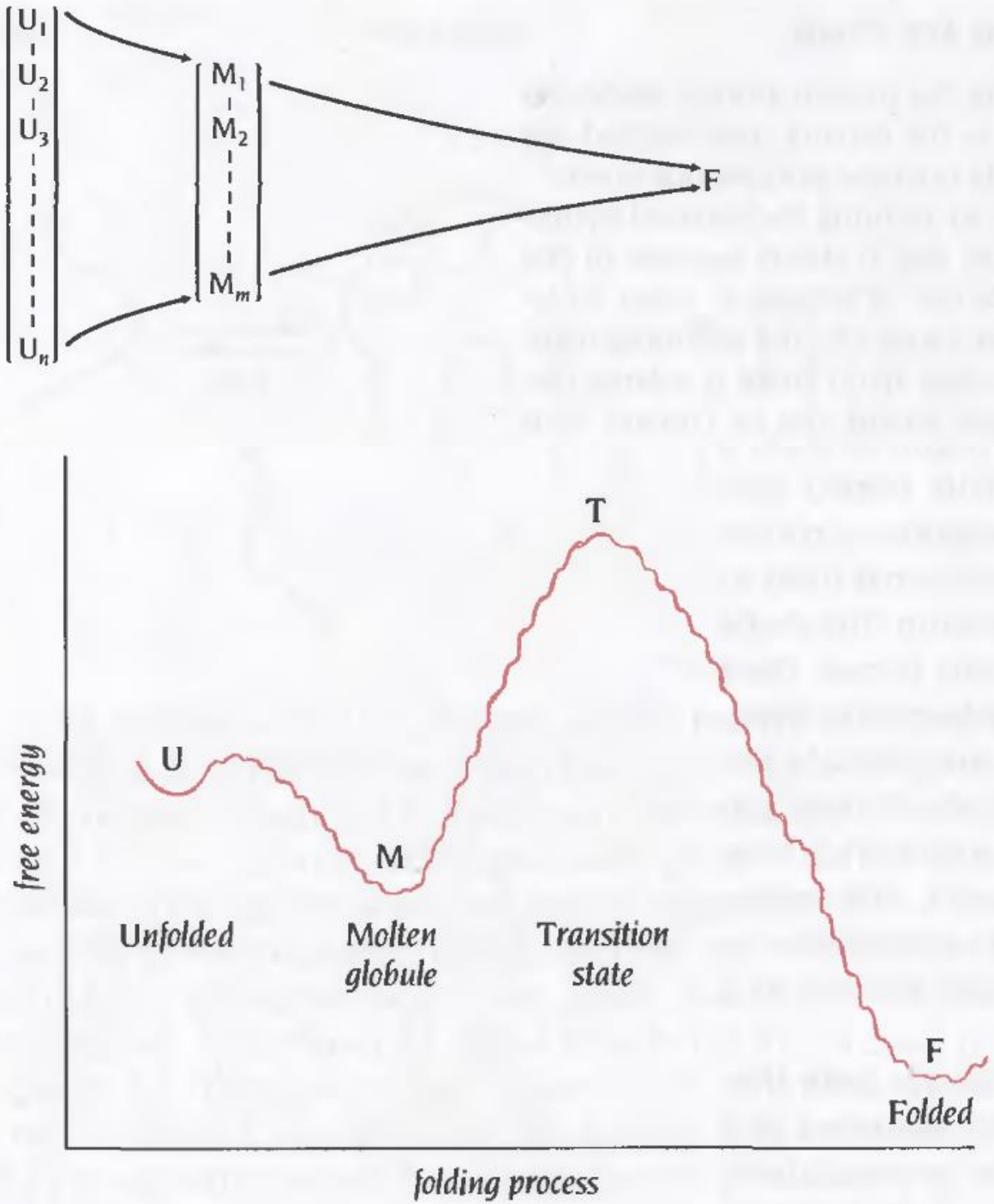


Figure 6.3 The unfolded state is an ensemble of a large number of conformationally different molecules, $U_1 \ldots U_n$, which undergo rapid interconversions. The molten globule is an ensemble of structurally related molecules, $M_1 \ldots M_m$, which are rapidly interconverting and which slowly change to a single unique conformation, the folded state F. During the folding process the protein proceeds from a high energy unfolded state to a low energy native state. The conversion from the molten globule state to the folded state is slow and passes through a high energy transition state, T.

سد بالای انرژی در گذار molten globule → folded ناشی از:

- کاهش شدید آنتروپی با تنظیم اتمی زنجیر جانبی
- بسته‌بندی دقیق هسته آب‌گریز
- خروج کامل آب از محیط داخلی پروتئین (desolvation)
ایزومریزاسیون کند پرولین
- تشکیل یا اصلاح پیوندهای دی‌سولفید
- گیر افتادن در چاه‌های انرژی محلی

به همین دلیل این مرحله **کندترین و پرانرژی‌ترین** قسمت فرایند تاخوردگی است.

FIGURE 6.34 A model for the steps involved in the folding of globular proteins. The funnel represents a free energy surface or energy landscape for the folding process. The protein folding process is highly cooperative. Rapid and reversible formation of local secondary structures is followed by a slower phase in which establishment of partially folded intermediates leads to the final tertiary structure. Substantial exclusion of water occurs very early in the folding process.

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?

❑ 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?

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

There are two major contributors to the energy difference between the folded and the denatured state: enthalpy and entropy. **Enthalpy** derives from the energy of the noncovalent interactions within the polypeptide chain—the hydrophobic interactions, hydrogen bonds and ionic bonds. The covalent bonds within and between the amino acid residues in the polypeptide chain are the same in the native and denatured states, with the exceptions of disulfide bonds in those proteins where these form between cysteine residues. The noncovalent interactions on the other hand differ significantly between the two states. In the native state these interactions are maximized to produce a compact globular molecule with a tightly packed hydrophobic core whereas the denatured state is more open and the side chains are more loosely packed (Figure 6.1). These noncovalent interactions are therefore stronger and more frequent in the native state and hence their energy contribution, enthalpy, is much larger. The enthalpy difference between native and denatured states can reach several hundred kcal/mol.

Entropy derives from the second law of thermodynamics which states that energy is required to create order. Proteins in the native state are highly ordered in one main conformation whereas the denatured state is highly disordered, with the protein molecules in many different conformations. A typical experimental preparation of unfolded protein (a solution in 6 M guandinium chloride or 8 M urea) contains 10^{15}–10^{20} protein molecules, each of which will have a unique conformation. In the absence of compensating factors it would therefore be entropically much more favorable for the protein to be in the disordered denatured state. The energy difference due to entropy between the native ordered state and the denatured state can also reach several hundred kcal/mole but in the opposite direction to the enthalpy difference. The total energy difference between the native and the denatured state of 5–15 kcal/mol, which is called the **free energy** difference, is thus a difference between two large numbers, the enthalpy difference and the entropy difference. The fact that this difference is very small is a severe complicating factor both for predictions of possible native states and for interpretation of factors responsible for the stability or instability of protein molecules, because our knowledge about the denatured state is very incomplete.

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

Disulfide bond as a lock

Disulfide bond in intracellular and extracellular proteins

Globular proteins are only marginally stable

❑ 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”

Slight changes in pH or temperature can convert a solution of biologically active protein molecules in the **native state** to a biologically inactive **denatured state**. The energy difference between these two states in physiological conditions is quite small, about 5-15 kcal/mol.

Breathing: A protein's conformation displays a limited degree of flexibility and such movement is termed "**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).

- **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.

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.)

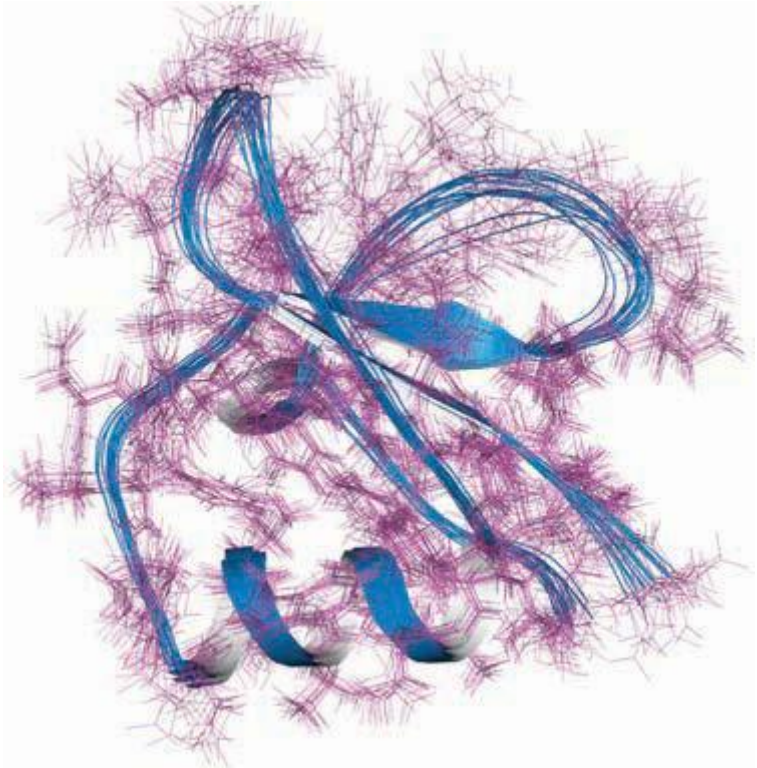
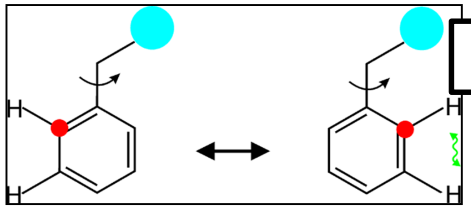


FIGURE 6.35 Proteins are dynamic structures. The marginal stability of a tertiary structure leads to flexibility and motion in the protein. Determination of structures of proteins (such as the SH3 domain of the α-chain of spectrin, shown here) by nuclear magnetic resonance produces a variety of stable 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

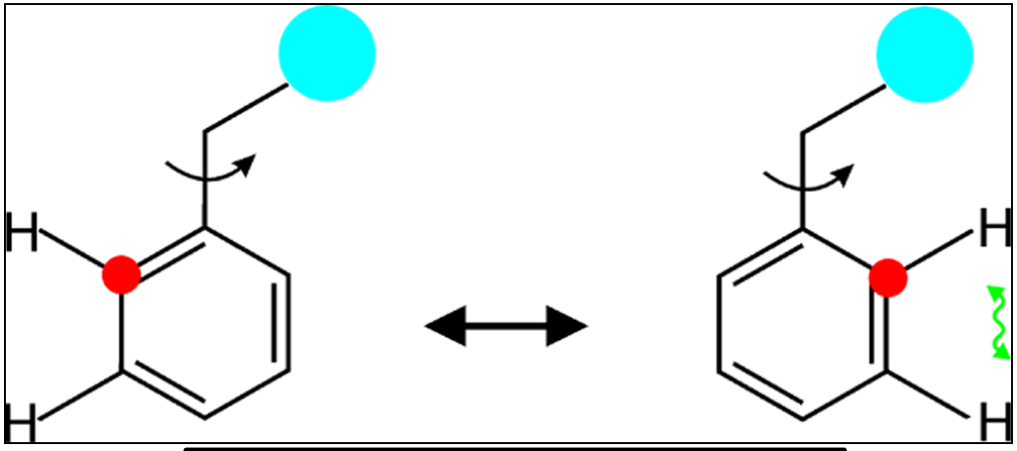
TABLE 6.2 Motion and Fluctuations in Proteins

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



Aromatic Ring Flips

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

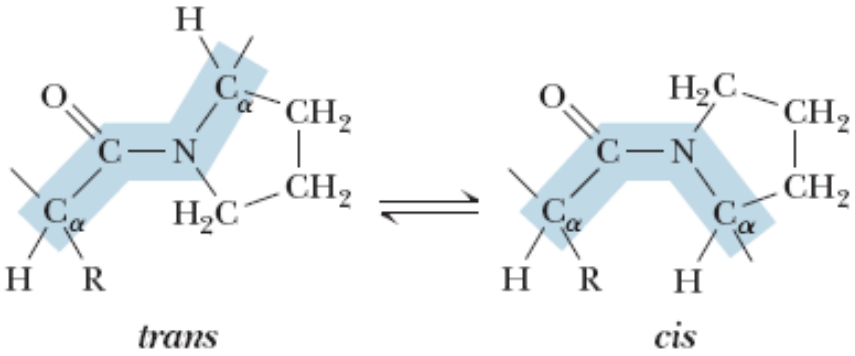


FIGURE 6.36 The *cis* and *trans* configurations of proline residues in peptide chains are almost equally stable. Proline *cis-trans* isomerizations, often occurring over relatively long time scales, can alter protein structure significantly.

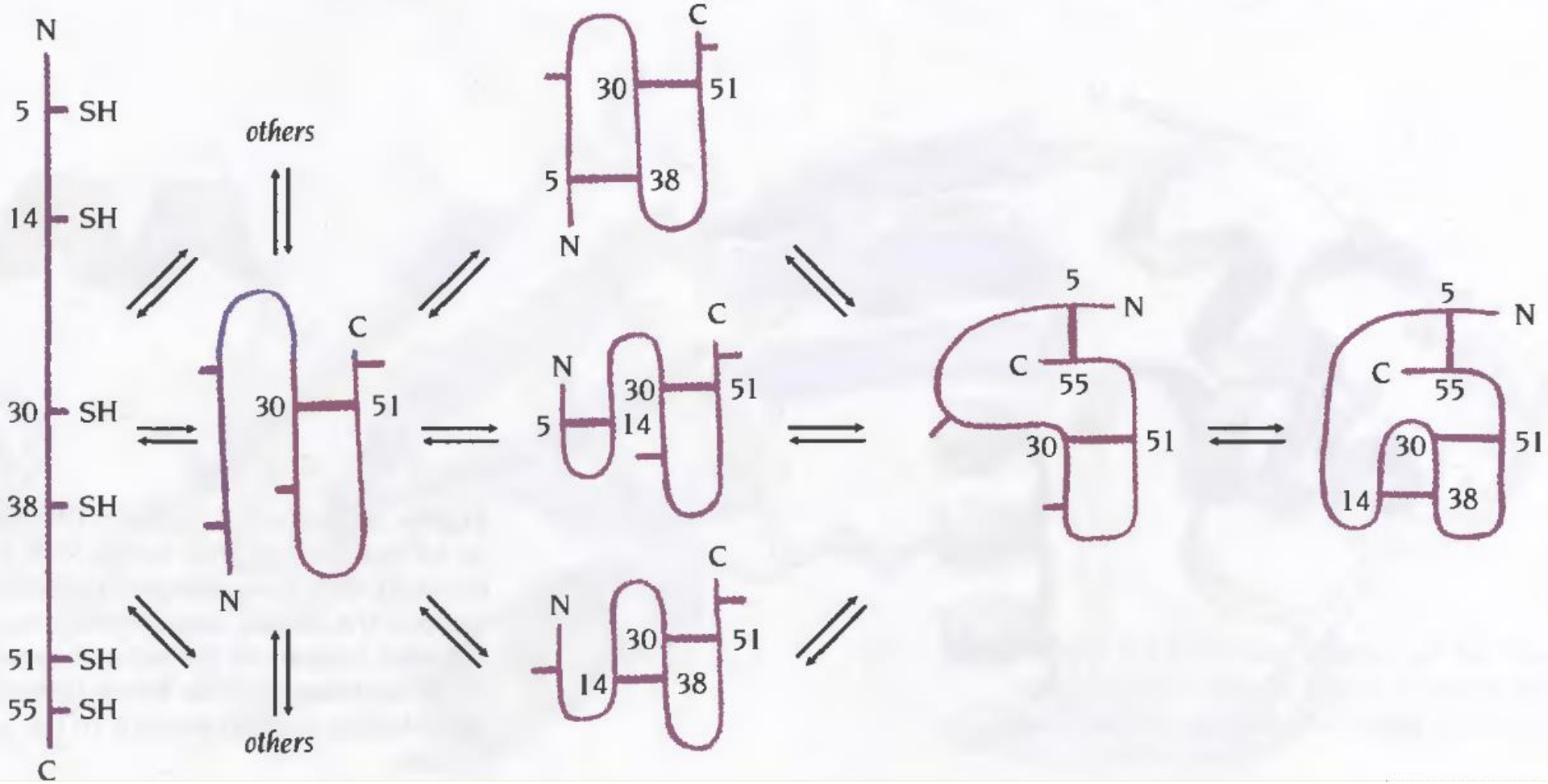
Increased thermal stability is generally related to one or more of the following structural adaptations:

- 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.

Kinetic factors are important for folding


N
5 SH
14 SH
30 SH
38 SH
51 SH
55 SH
C
others
others
30
51
N
C
5
38
N
C
N
30
51
5
14
C
5
N
C
55
30
51
5
N
C
55
30
51
14
38

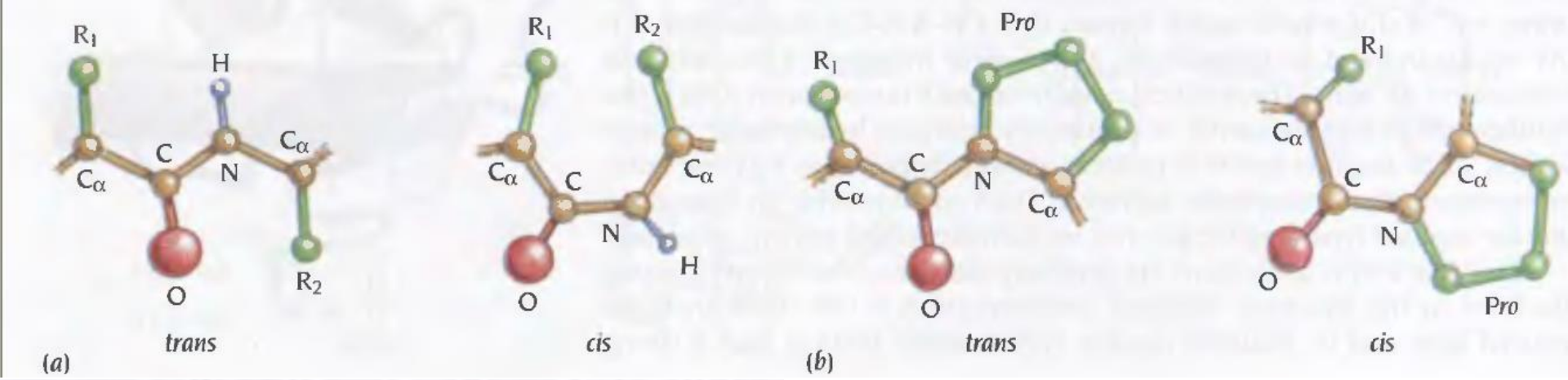


Figure 6.9 (a) Peptide units can adopt two different conformations, *trans* and *cis*. In the *trans*-form the C=O and the N-H groups point in opposite directions whereas in the *cis*-form they point in the same direction. For most peptides the *trans*-form is about 1000 times more stable than the *cis*-form. (b) When the second residue in a peptide is proline the *trans*-form is only about four times more stable than the *cis*-form. *Cis*-proline peptides are found in many proteins.

Isomerization of proline residues can be a rate-limiting step in protein folding

Cis–trans isomerization of proline peptides is intrinsically a slow process and *in vitro* it is frequently the rate-limiting step in folding for those molecules that have been trapped in a folding intermediate with the wrong isomer.

Many proteins can fold spontaneously *in vitro*, although some appear to fold more slowly/less accurately than they do *in vivo*. Although the primary sequence ultimately dictates tertiary structure, several obstacles to correct folding exist, including:

- aggregation of partially folded intermediates via intermolecular hydrophobic interactions;
- isomerization of proline residues;
- formation of disulfide linkages between incorrect pairs of cysteine residues.

Protein Structure Prediction
-Secondary Structure
-Tertiary Structure

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.

Quaternary Structure

How Do Protein Subunits Interact at the Quaternary Level of Protein Structure?

- Many proteins exist in nature as oligomers, complexes composed of (often symmetric) noncovalent assemblies of two or more monomer subunits. In fact, subunit association is a common feature of macromolecular organization in biology. Most intracellular enzymes are oligomeric and may be composed either of a single type of monomer subunit (homomultimers) or of several different kinds of subunits (heteromultimers). The simplest case is a protein composed of identical subunits.
- Oxygen is carried in the blood by hemoglobin, which contains two each of two different subunits (heterotetramer). A counterpoint to these small clusters is made by the proteins that form large polymeric aggregates.

How Do Protein Subunits Interact at the Quaternary Level of Protein Structure?

- Proteins are synthesized on large complexes of many protein units and several RNA molecules called ribosomes.
- Muscle contraction depends on large polymer clusters of the protein myosin sliding along filamentous polymers of another protein, actin. The way in which separate folded monomeric protein subunits associate to form the oligomeric protein constitutes the quaternary structure of that protein. Table 6.3 lists several proteins and their subunit compositions. Proteins with two to four subunits predominate in nature, but many cases of higher numbers exist.

How Do Protein Subunits Interact at the Quaternary Level of Protein Structure?

- The subunits of an oligomeric protein typically fold independently and then interact with other subunits. The surfaces at which subunits interact are similar in nature to the interiors of the individual subunits—closely packed with both polar and hydrophobic interactions. Interacting surfaces must therefore possess complementary arrangements of polar and hydrophobic groups.
- Oligomeric associations of protein subunits can be divided into those between identical subunits and those between nonidentical subunits. Interactions among identical subunits can be further distinguished as either isologous or heterologous. In isologous interactions, the interacting surfaces are identical and the resulting structure is necessarily dimeric and closed, with a twofold axis of symmetry (Figure 6.42). If any additional interactions occur to form a trimer or tetramer, these must use different interfaces on the protein’s surface.
- Many proteins, such as transthyretin, form tetramers by means of two sets of isologous interactions (Figure 6.43). Such structures possess three different twofold axes of symmetry. In contrast, heterologous associations among subunits involve nonidentical interfaces. These surfaces must be complementary, but they are generally not symmetric.

TABLE 6.3 Aggregation Symmetries of Globular Proteins

Protein	Number of Subunits
Alcohol dehydrogenase	2
Malate dehydrogenase	2
Superoxide dismutase	2
Triose phosphate isomerase	2
Glycogen phosphorylase	2
Aldolase	3
Bacteriochlorophyll protein	3
Concanavalin A	4
Glyceraldehyde-3-phosphate dehydrogenase	4
Immunoglobulin	4
Lactate dehydrogenase	4
Prealbumin	4
Pyruvate kinase	4
Phosphoglycerate mutase	4
Hemoglobin	2 + 2
Insulin	6
Aspartate transcarbamoylase	6 + 6
Glutamine synthetase	12
TMV protein disc	17
Apoferritin	24
Coat of tomato bushy stunt virus	180

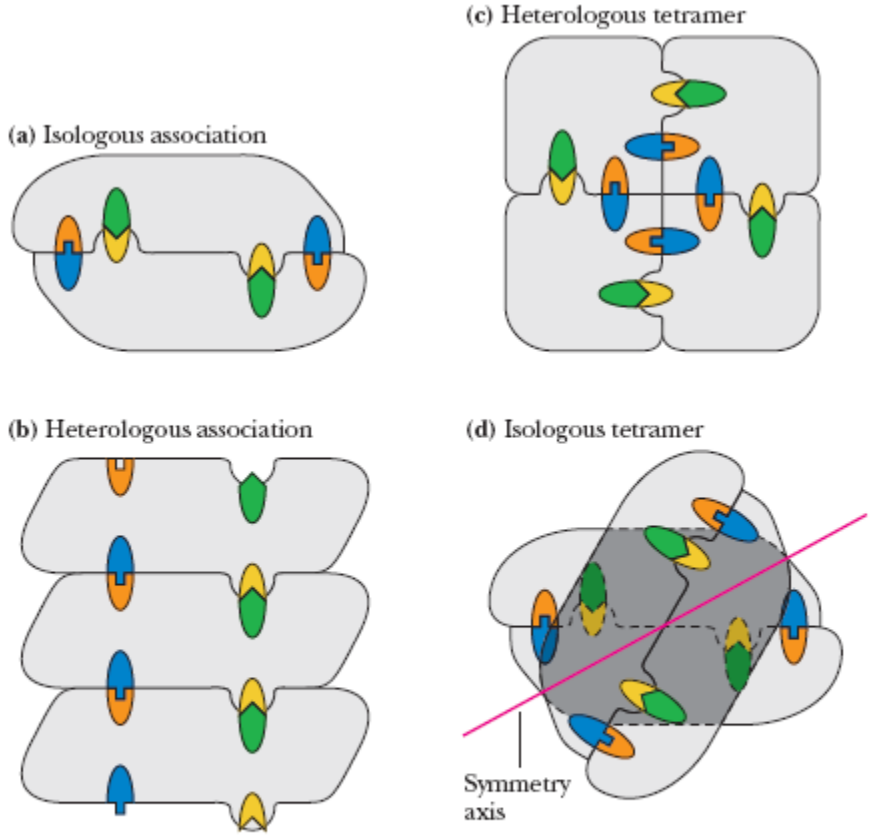
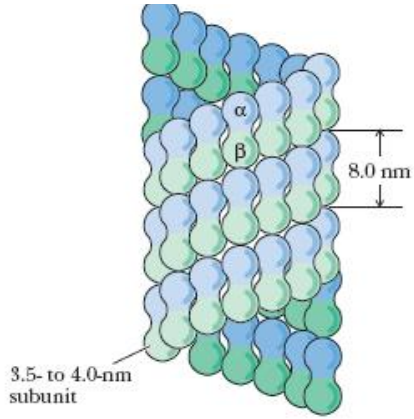


FIGURE 6.42 Isologous and heterologous associations between protein subunits. **(a)** An isologous interaction between two subunits with a twofold axis of symmetry perpendicular to the plane of the page. **(b)** A heterologous interaction that could lead to the formation of a long polymer. **(c)** A heterologous interaction leading to a closed structure–a tetramer. **(d)** A tetramer formed by two sets of isologous interactions.



There Are Structural and Functional Advantages to Quaternary Association

There are several important consequences when protein subunits associate in oligomeric structures.

- **Stability** One general benefit of subunit association is a favorable reduction of the protein's surface-to-volume ratio. The surface-to-volume ratio becomes smaller as the radius of any particle or object becomes larger. (This is because surface area is a function of the radius squared and volume is a function of the radius cubed.) Because interactions within the protein usually tend to stabilize the protein and because the interaction of the protein surface with solvent water is often energetically unfavorable, decreased surface-to-volume ratios usually result in more stable proteins. Subunit association may also serve to shield hydrophobic residues from solvent water. Subunits that recognize either themselves or other subunits avoid any errors arising in genetic translation by binding mutant forms of the subunits less tightly.

There Are Structural and Functional Advantages to Quaternary Association

Genetic Economy and Efficiency Oligomeric association of protein monomers is genetically economical for an organism. Less DNA is required to code for a monomer that assembles into a homomultimer than for a large polypeptide of the same molecular mass. Another way to look at this is to realize that virtually all of the information that determines oligomer assembly and subunit–subunit interaction is contained in the genetic material needed to code for the monomer. For example, HIV protease, an enzyme that is a dimer of identical subunits, performs a catalytic function similar to homologous cellular enzymes that are single polypeptide chains of twice the molecular mass (see Chapter 14).

There Are Structural and Functional Advantages to Quaternary Association

Bringing Catalytic Sites Together Many enzymes (see Chapters 13 to 15) derive at least some of their catalytic power from oligomeric associations of monomer subunits. This can happen in several ways. The monomer may not constitute a complete enzyme active site. Formation of the oligomer may bring all the necessary catalytic groups together to form an active enzyme. For example, the active sites of bacterial glutamine synthetase are formed from pairs of adjacent subunits. The dissociated monomers are inactive.

Oligomeric enzymes may also carry out different but related reactions on different subunits. Thus, tryptophan synthase is a tetramer consisting of pairs of different subunits, $\alpha_2\beta_2$. Purified α-subunits catalyze the following reaction:

$$\text{Indoleglycerol phosphate} \rightleftharpoons \text{indole} + \text{glyceraldehyde-3-phosphate}$$

and the β-subunits catalyze this reaction:

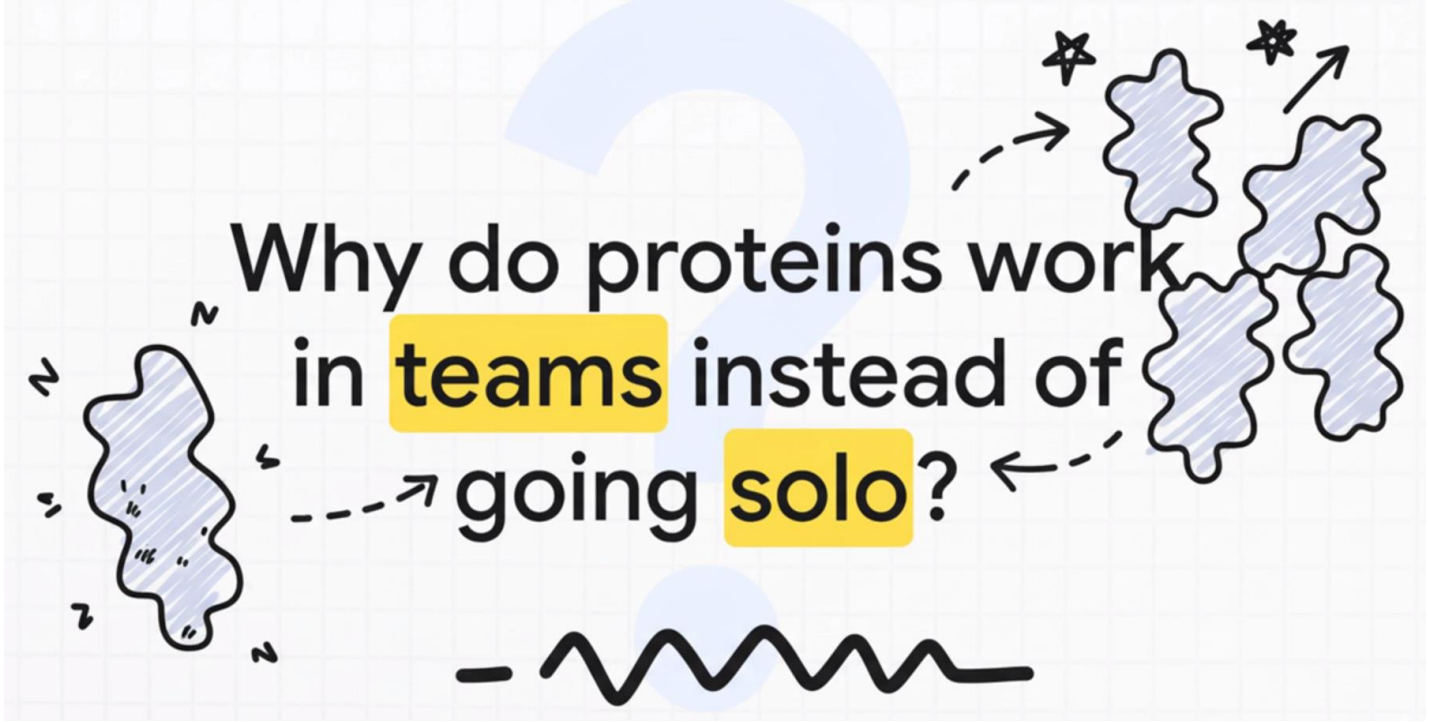
$$\text{Indole} + \text{L-serine} \rightleftharpoons \text{L-tryptophan}$$

Indole, the product of the α-reaction and the reactant for the β-reaction, is passed directly from the α-subunit to the β-subunit and cannot be detected as a free intermediate.

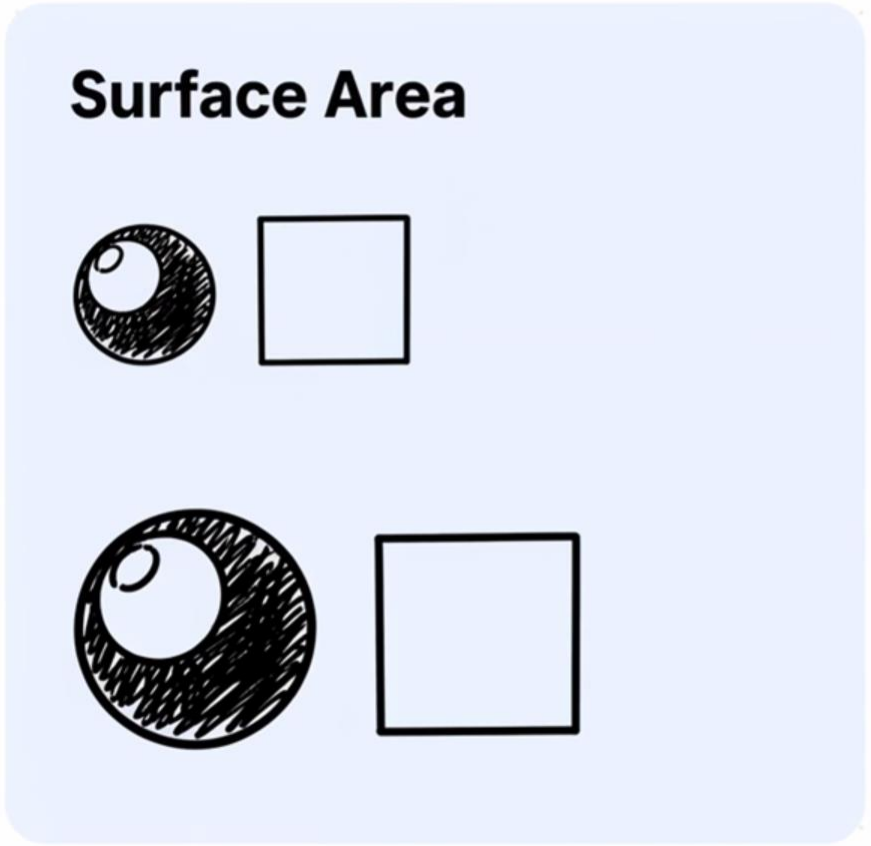
There Are Structural and Functional Advantages to Quaternary Association

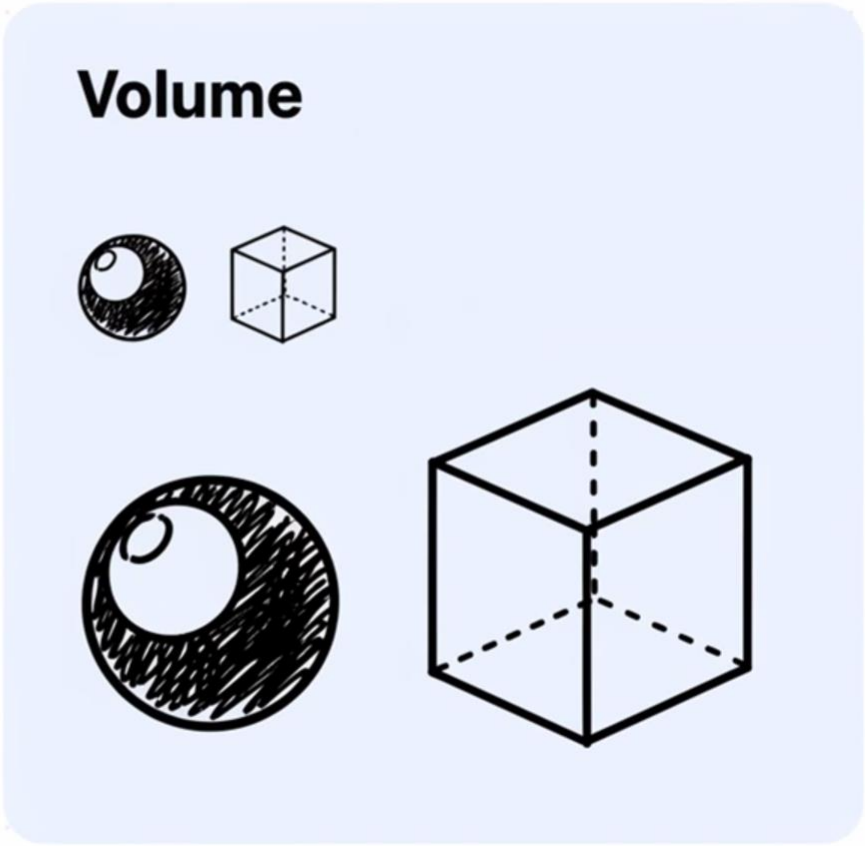
Cooperativity There is another, more important consequence when monomer subunits associate into oligomeric complexes. Most oligomeric enzymes regulate catalytic activity by means of subunit interactions, which may give rise to cooperative phenomena.

Multisubunit proteins typically possess multiple binding sites for a given ligand. If the binding of ligand at one site changes the affinity of the protein for ligand at the other binding sites, the binding is said to be **cooperative.** Information transfer in this manner across long distances in proteins is termed **allostery,** literally action "at another site." Increases in affinity at subsequent sites represent positive cooperativity, whereas decreases in affinity correspond to negative cooperativity. The points of contact between protein subunits provide a mechanism for this signal transduction through the protein structure and for communication between the subunits. This in turn provides a way in which the binding of ligand to one subunit can influence the binding behavior at the other subunits. Such cooperative behavior, discussed in greater depth in Chapter 15, is the underlying mechanism for regulation of many biological processes.



- Functional Imperatives of Quaternary Protein Structure
  - Stability
  - Genetic Economy and Efficiency
  - Bringing Catalytic Sites Together
  - Cooperativity (Regulation)





Surface $\sim r^2$

$r=2 \rightarrow r^2 = 4$

$r=4 \rightarrow r^2 = 16$

Volume $\sim r^3$

$r=2 \rightarrow r^3 = 8$

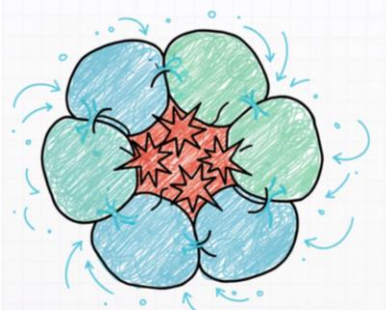
$r=4 \rightarrow r^3 = 64$

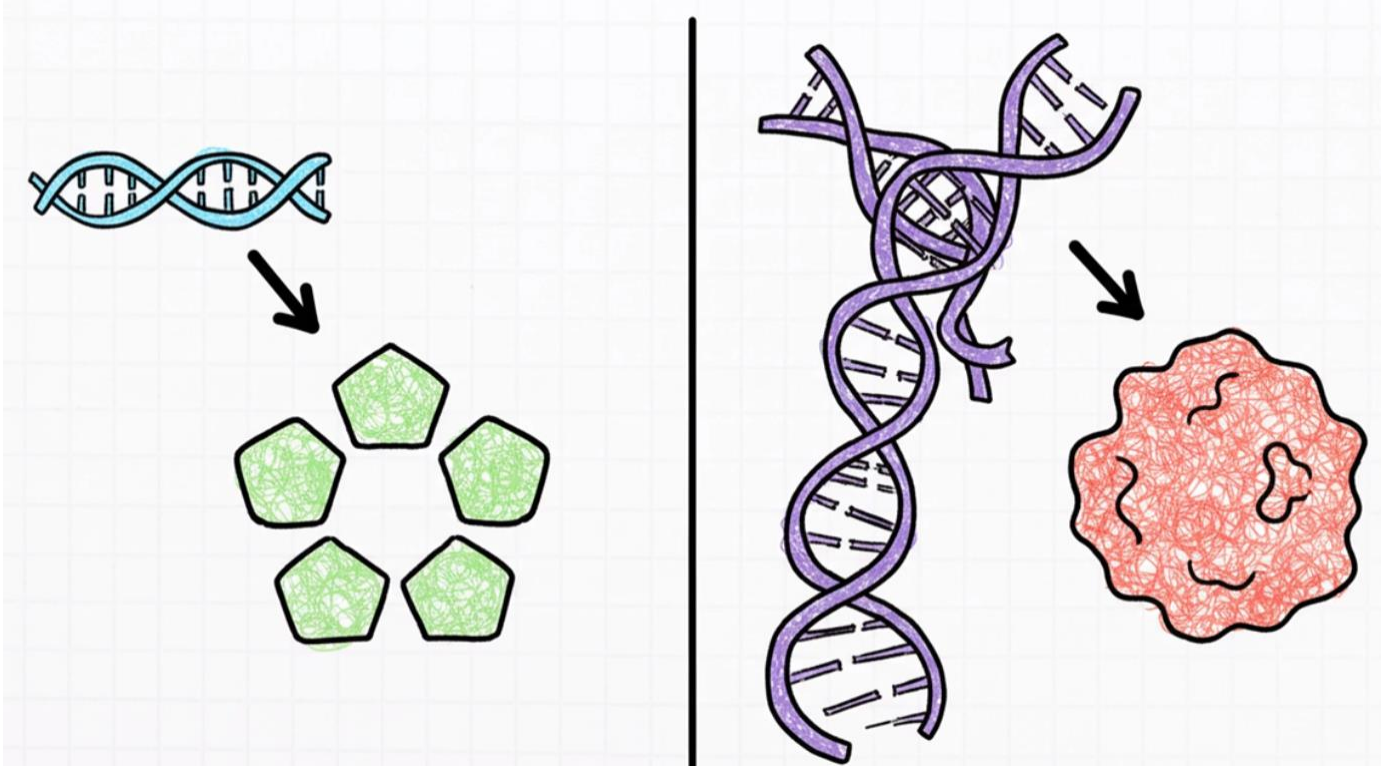
Surface/Volume:

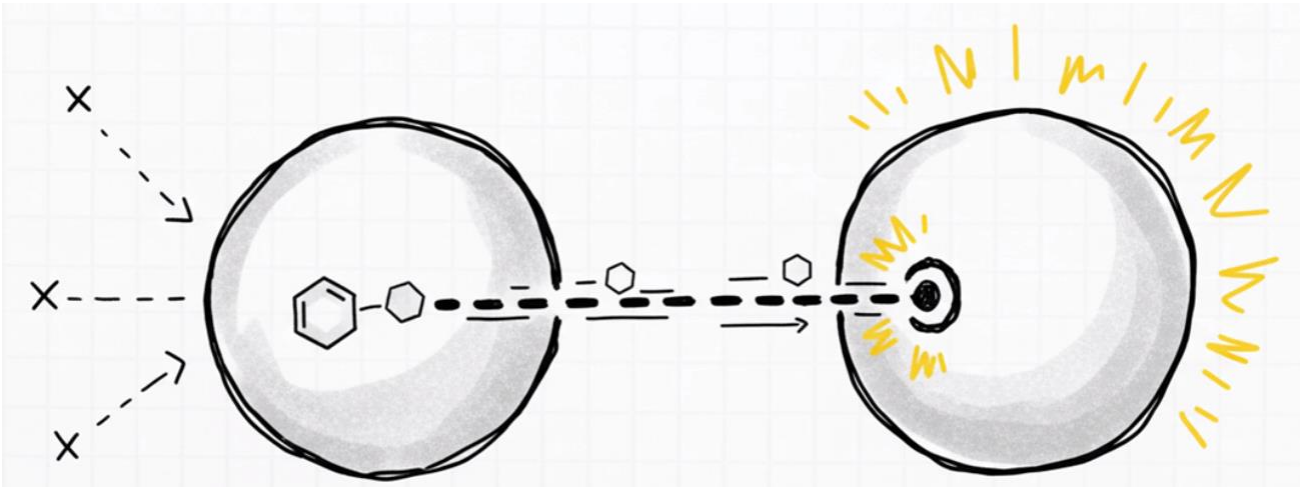
4/8=1/2=0.5

16/64=1/4=0.25

بخش‌های سطحی کمتر در تماس با محیط (معمولاً) آبدوست قرار می‌گیرند.






α-subunit
β-subunit

Faster-Acting Insulin: Genetic Engineering Solves a Quaternary Structure Problem

Insulin is a peptide hormone secreted by the pancreas that regulates glucose metabolism in the body. Insufficient production of insulin or failure of insulin to stimulate target sites in liver, muscle, and adipose tissue leads to the serious metabolic disorder known as diabetes mellitus. Diabetes afflicts millions of people worldwide. Diabetic individuals typically exhibit high levels of glucose in the blood, but insulin injection therapy allows these individuals to maintain normal levels of blood glucose.

Insulin is composed of two peptide chains covalently linked by disulfide bonds. This “monomer” of insulin is the active form that binds to receptors in target cells. However, in solution, insulin spontaneously forms dimers, which themselves aggregate to form hexamers. The surface of the insulin molecule that self-associates to form hexamers is also the surface that binds to insulin receptors in target cells. Thus, hexamers of insulin are inactive. Insulin released from the pancreas is monomeric and acts rapidly at target tissues. However, when insulin is administered (by injection) to a diabetic patient, the insulin hexamers dissociate slowly and the patient’s blood glucose levels typically drop slowly (over several hours). In 1988, G. Dodson showed that insulin could be genetically engineered to prefer the monomeric (active) state. Dodson and his colleagues used recombinant DNA technology to produce insulin with an **aspartate** residue replacing a **proline** at the contact interface between adjacent subunits. The negative charge on the Asp side chain creates electrostatic repulsion between subunits and increases the dissociation constant for the hexamer ↔ monomer equilibrium. Injection of this mutant insulin into test animals produced more rapid decreases in blood glucose than did ordinary insulin. This mutant insulin, known as insulin aspart, marketed by the Danish pharmaceutical company Novo as NovoLog in the United States and as NovoRapid in Europe, has several advantages over ordinary insulin, in the treatment of diabetes. NovoLog has a faster rate of absorption, a faster onset of action, and a shorter duration of action than regular human insulin. It is particularly suited for mealtime dosing to control postprandial glycemia, the rise in blood sugar following consumption of food. Regular human insulin acts more slowly, so patients must usually administer it 30 minutes before eating.

Some Proteins Are Intrinsically Unstructured

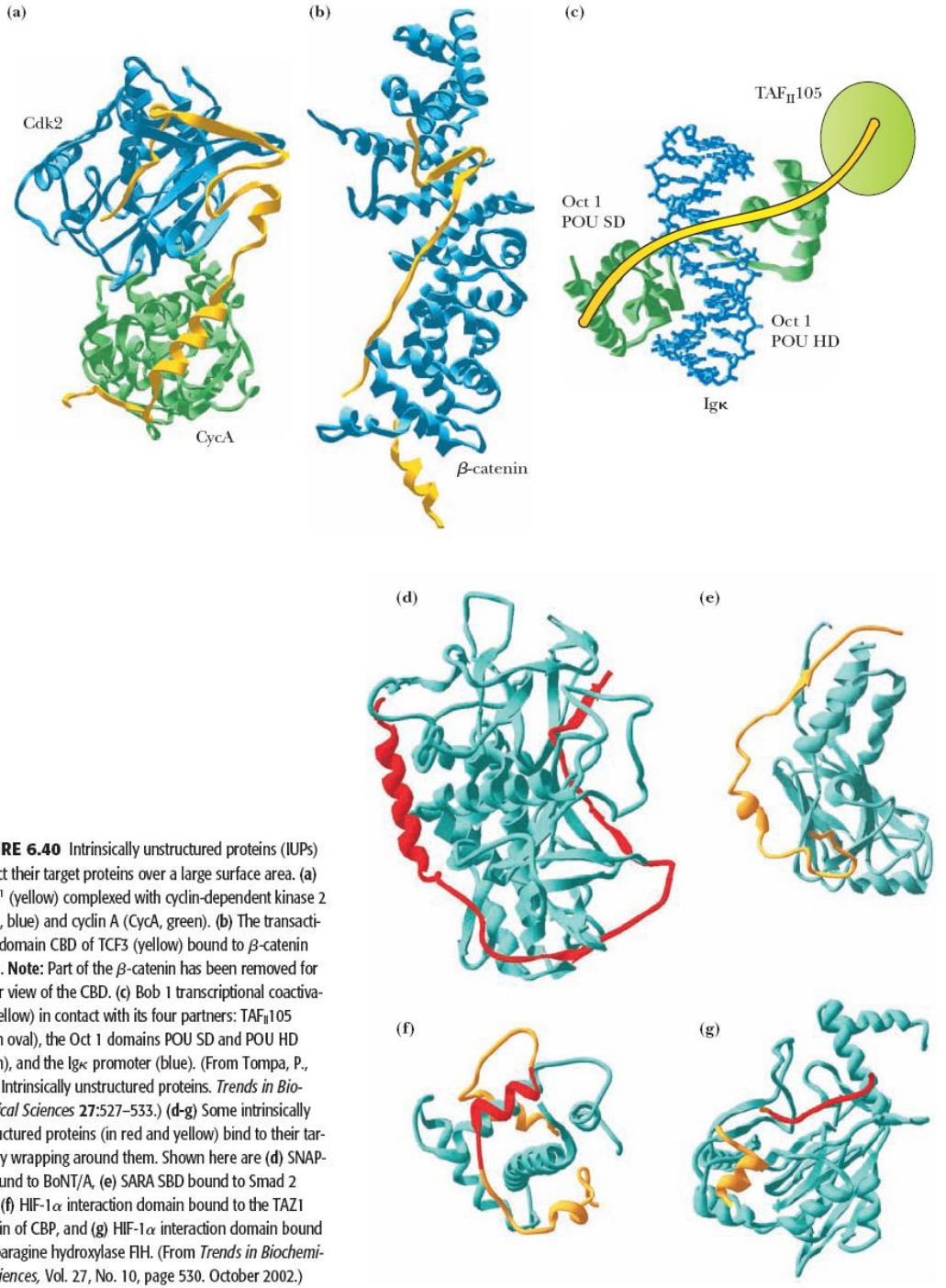


FIGURE 6.40 Intrinsically unstructured proteins (IUPs) contact their target proteins over a large surface area. **(a)** $p27^{Kip1}$ (yellow) complexed with cyclin-dependent kinase 2 (Cdk2, blue) and cyclin A (CycA, green). **(b)** The transactivator domain CBD of TCF3 (yellow) bound to β-catenin (blue). **Note:** Part of the β-catenin has been removed for a clear view of the CBD. **(c)** Bob 1 transcriptional coactivator (yellow) in contact with its four partners: $TAF_{II}105$ (green oval), the Oct 1 domains POU SD and POU HD (green), and the Igκ promoter (blue). (From Tompa, P., 2002. Intrinsically unstructured proteins. *Trends in Biochemical Sciences* **27**:527–533.) **(d-g)** Some intrinsically unstructured proteins (in red and yellow) bind to their targets by wrapping around them. Shown here are **(d)** SNAP-25 bound to BoNT/A, **(e)** SARA SBD bound to Smad 2 MH2, **(f)** HIF-1α interaction domain bound to the TAZ1 domain of CBP, and **(g)** HIF-1α interaction domain bound to asparagine hydroxylase FIH. (From *Trends in Biochemical Sciences,* Vol. 27, No. 10, page 530. October 2002.)