

Comments

Structural motif v sequence motif

polyproline (“PXXP”) motif for SH3 binding

“RGD” motif for integrin binding

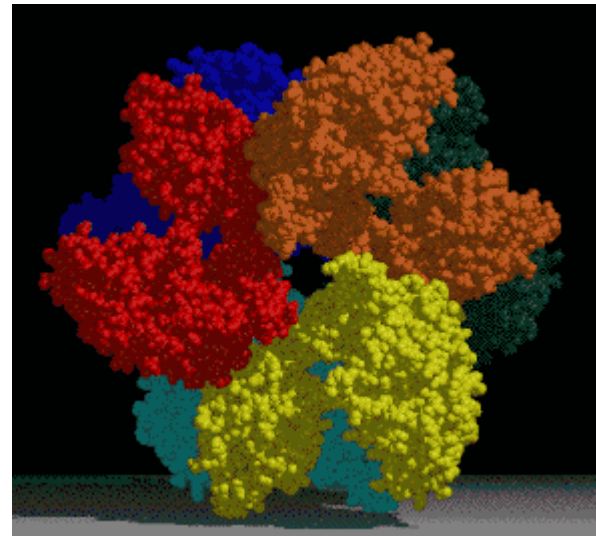
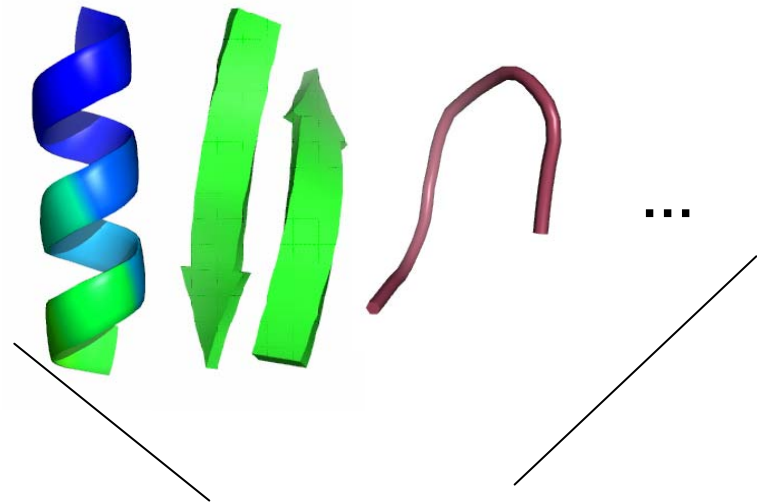
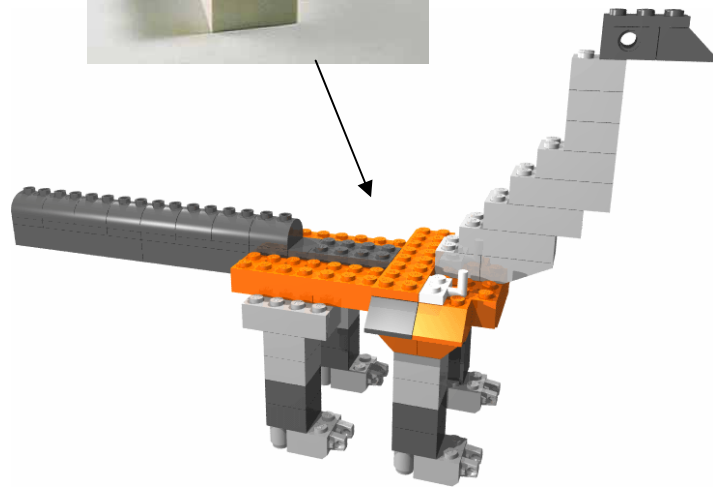
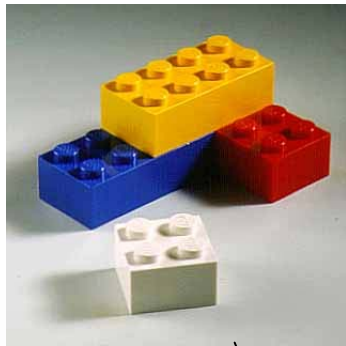
“GXXXG” motif within the TM domain of membrane protein

Most common type I' beta turn sequences: X – (N/D/G)G – X

Most common type II' beta turn sequences: X – G(S/T) – X

Putting it together

Alpha helices and beta sheets are not proteins—only marginally stable by themselves



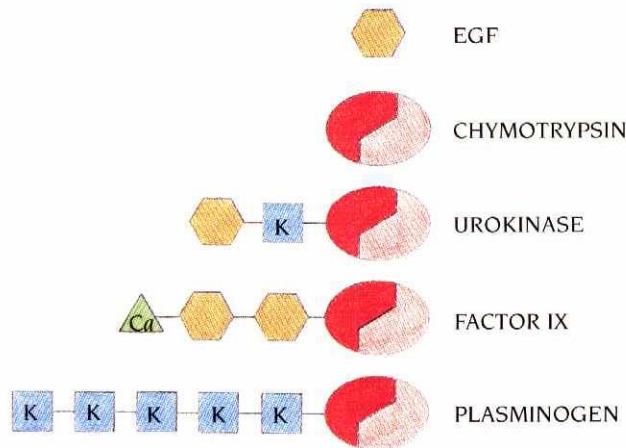
Extremely small “proteins” can’t do much

Tertiary structure

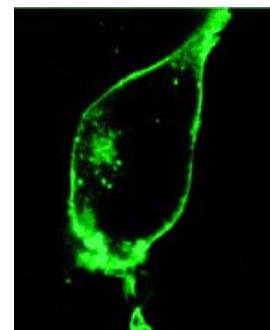
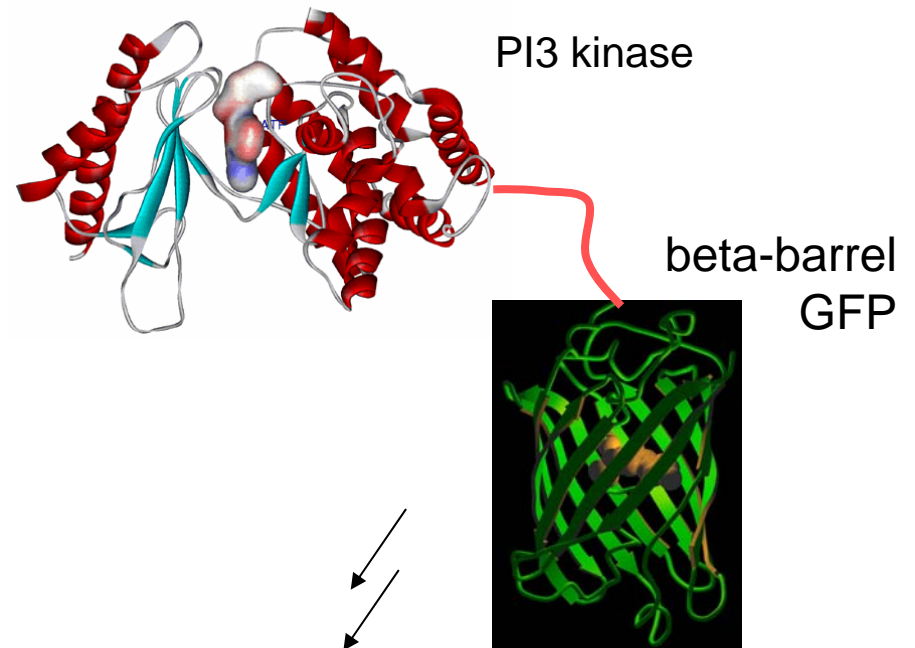
- Concerns with how the secondary structure units within a **single** polypeptide chain associate with each other to give a three-dimensional structure
- Secondary structure, super secondary structure, and loops come together to form “**domains**”, the smallest tertiary structural unit
- Structural domains (“domains”) usually contain 100 – 200 amino acids and fold stably.
- Domains may be considered to be connected units which are to varying extents independent in terms of their structure, function and folding behavior. Each domain can be described by its fold, i.e. how the secondary structural elements are arranged.
- Tertiary structure also includes the way domains fit together

Domains are modular

- Because they are self-stabilizing, domains can be swapped both in nature and in the laboratory



- Domains that are homologous to the epidermal growth factor, EGF, which is a small polypeptide chain of 53 amino acids.
- Serine proteinase domains that are homologous to chymotrypsin, which has about 245 amino acids arranged in two domains.
- Kringle domains, which have a characteristic pattern of three internal disulfide bridges within a region of about 85 amino acid residues.
- Calcium-binding domain (see Figure 2.13).



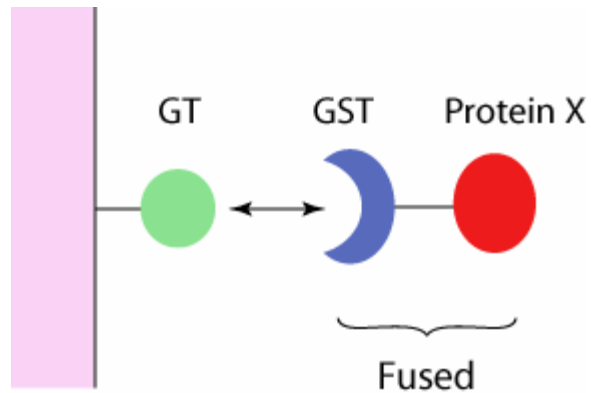
fluorescence localization experiment

Branden & Tooze

Chimeras

Recombinant proteins are often expressed and purified as fusion proteins (“chimeras”) with

- glutathione S-transferase
- maltose binding protein
- or peptide tags, e.g. hexa-histidine, FLAG epitope



helps with solubility, stability, and purification

Structural Classification

All classifications are done at the domain level

In many cases, structural similarity implies a common evolutionary origin

- structural similarity without evolutionary relationship is possible
- but no structural similarity means no evolutionary relationship

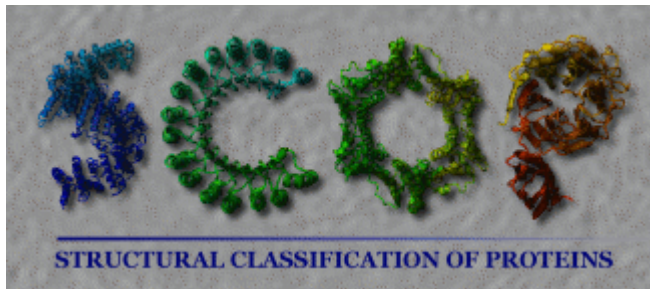
Each domain has its corresponding “fold”, i.e. the identity and connectivity of secondary structural elements

There appears to be a limited number of actual folds (~ 1000) utilized by naturally occurring proteins

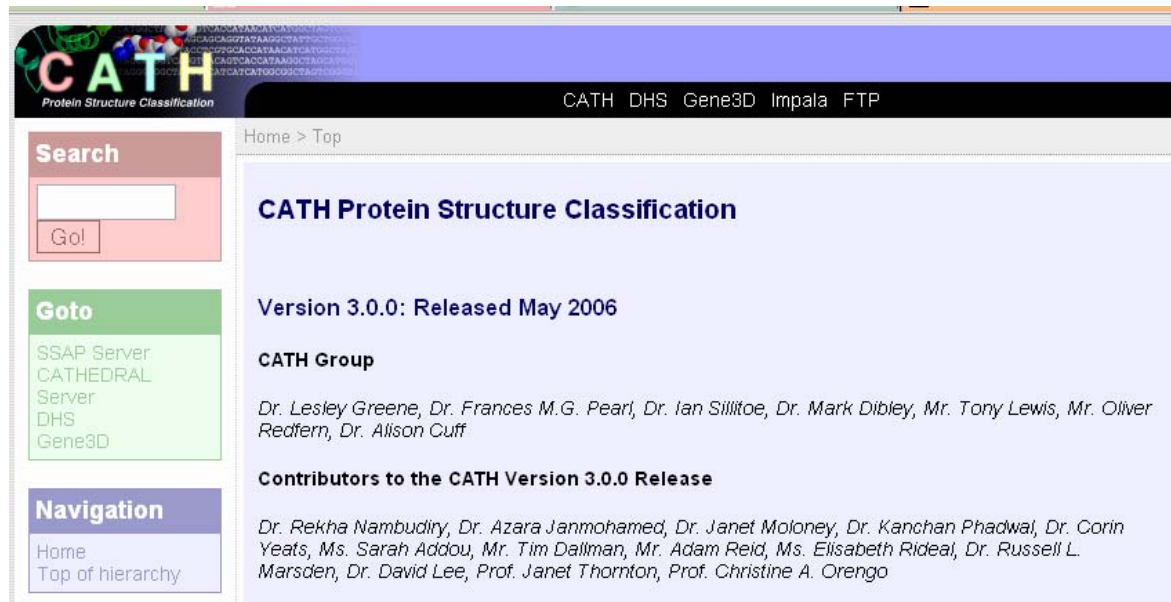
- Chothia, Nature 357, 543 (1992)

Nearly all proteins have structural similarities with some other proteins

- Two widely used protein structure classification systems are **CATH** (class, architecture, topology, and homology) and **SCOP** (structural classification of proteins)



<http://scop.berkeley.edu/>

The screenshot shows the CATH Protein Structure Classification website. The header features the CATH logo and navigation links for CATH, DHS, Gene3D, Impala, and FTP. The main content area includes a search bar, a "Goto" section with links to SSAP Server, CATHEDRAL Server, DHS, and Gene3D, and a "Navigation" section with links to Home and Top of hierarchy. The main heading is "CATH Protein Structure Classification" and the version information is "Version 3.0.0: Released May 2006". The "CATH Group" section lists contributors: Dr. Lesley Greene, Dr. Frances M.G. Pearl, Dr. Ian Sillitoe, Dr. Mark Dibley, Mr. Tony Lewis, Mr. Oliver Redfern, and Dr. Alison Cuff. The "Contributors to the CATH Version 3.0.0 Release" section lists: Dr. Rekha Nambudiry, Dr. Azara Janmohamed, Dr. Janet Moloney, Dr. Kanchan Phadwal, Dr. Corin Yeats, Ms. Sarah Addou, Mr. Tim Dailman, Mr. Adam Reid, Ms. Elisabeth Rideal, Dr. Russell L. Marsden, Dr. David Lee, Prof. Janet Thornton, and Prof. Christine A. Orengo.

<http://www.cathdb.info/latest/index.html>

SCOP Classification Statistics

SCOP: Structural Classification of Proteins. 1.73 release (Nov 2007)

34494 PDB Entries, 97178 Domains
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

- detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known

CATH

- Levitt and Thornton, 261, 552 (1976)
- Automated and manual classification based on sequence and structural similarity
 - If a given domain has sufficiently high sequence and structural similarity (e.g. 35% sequence identity) with a domain that has been previously classified in CATH, the classification is automatically inherited from the other domain
 - Otherwise, the protein is manually classified

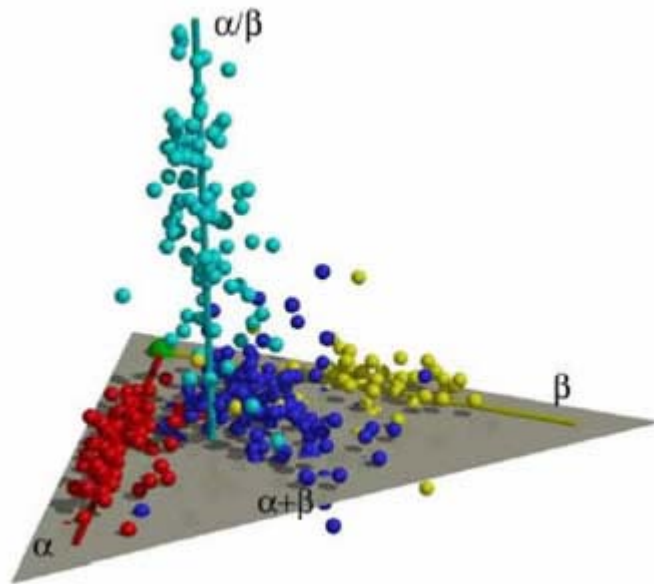
Class: mainly alpha, mainly beta, alpha+beta, little secondary structure

Architecture: overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures

Topology (Fold): both the overall shape and connectivity of the secondary structures

Homology: protein domains which are thought to share a common ancestor and can therefore be described as homologous

Visual representation of the protein structure space



three-dimensional map
of the protein universe

- Plot structural related proteins are placed close in 3D space
- Representative proteins from each of the SCOP family are compared to one another
- Scoring matrix S_{ij} is computed for proteins i, j
- Use of three different folds (alpha, beta, alpha/beta) is sufficient to describe all known folds

Shape of the fold space and the overall distribution of the folds are influenced by three factors

- Secondary structure class
- Chain topology
- Protein domain size

Hou, et al PNAS 100, 2386 (2003)

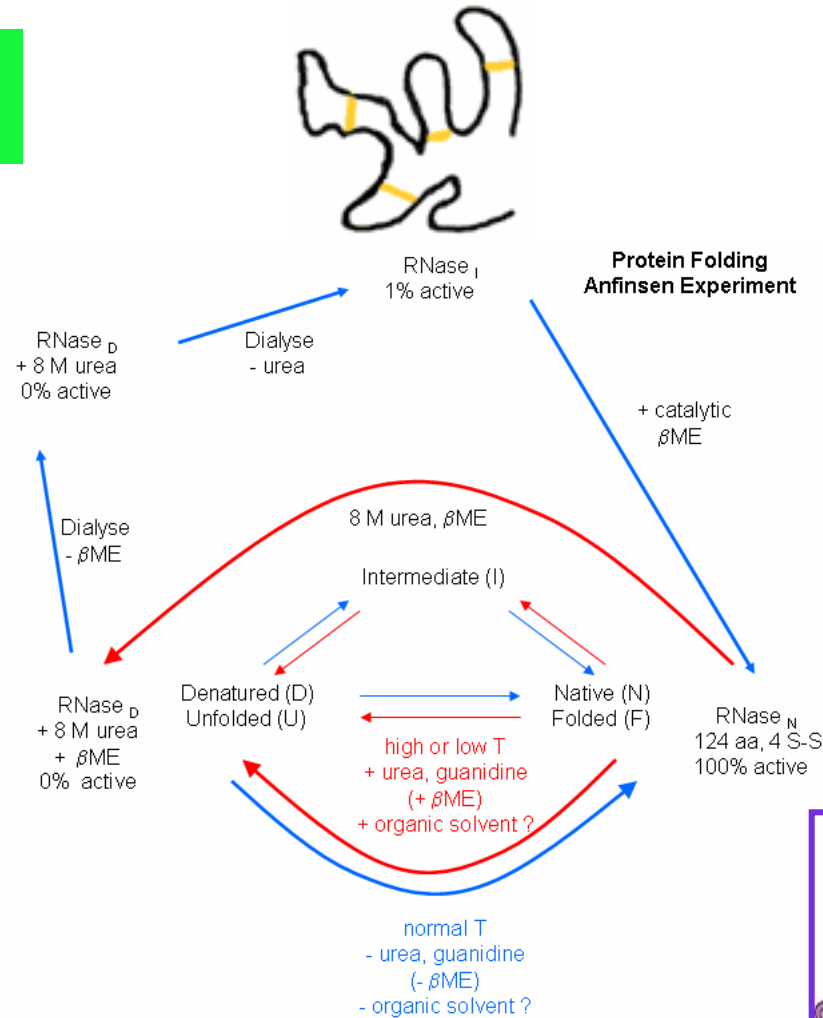
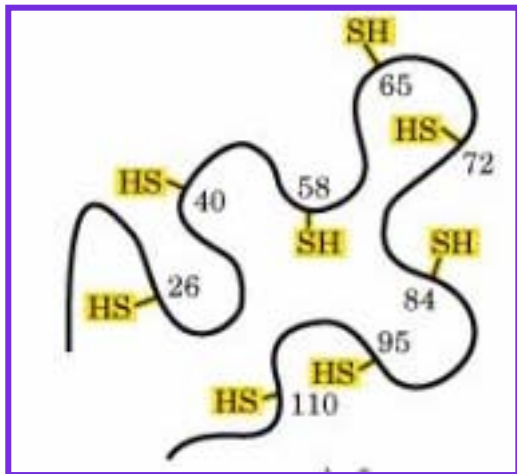


Anfinsen, Nobel prize in Chemistry 1972

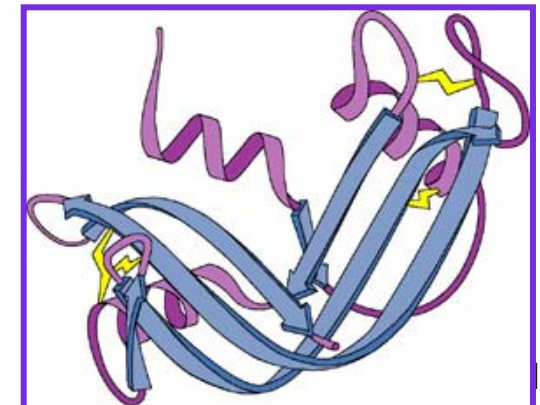
Proteins fold autonomously



denatured protein



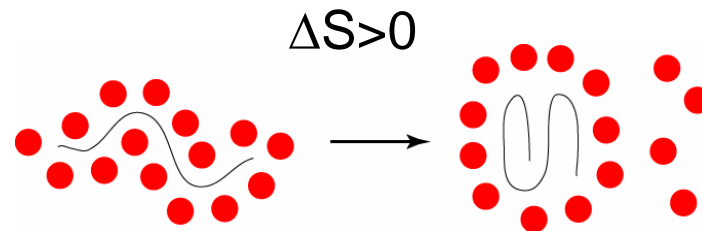
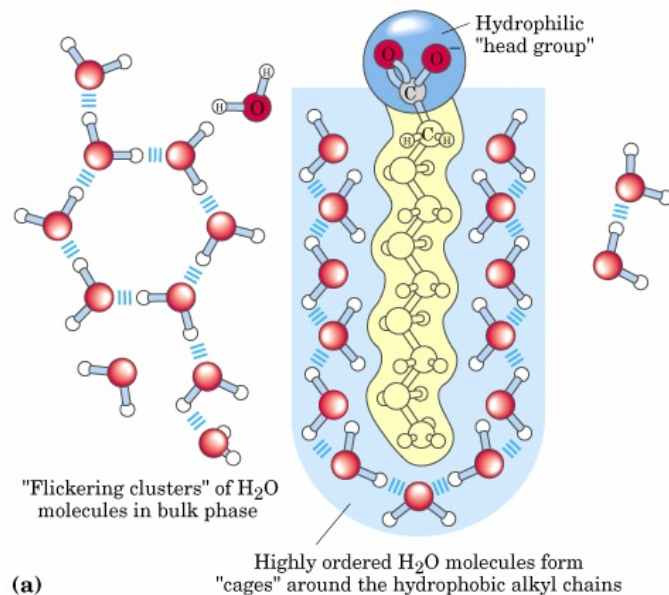
ribonuclease A



The information required to fold a protein into a 3D structure is stored in its sequence

Why do proteins fold

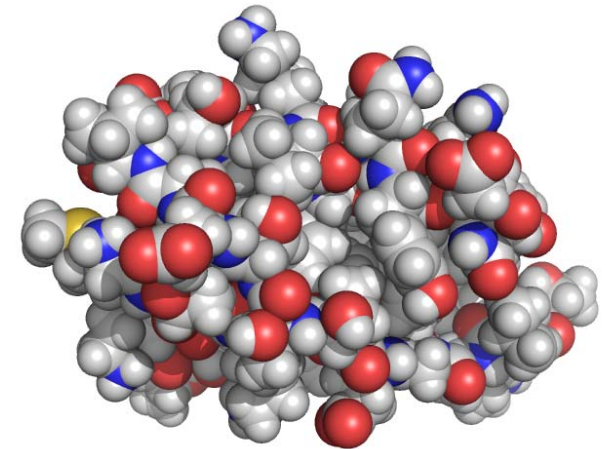
- Unfolded (“denatured”) polypeptides have a large hydrophobic surface exposed to the solvent
- Water molecules in the vicinity of a hydrophobic patch are highly ordered
- When protein folds, these water molecules are released from the hydrophobic surface, vastly increasing the solvent entropy



Corollary: Proteins fold to bury hydrophobic residues in the protein core, inaccessible to solvent molecules

(-) Energetic contributions to drive folding

- desolvation of hydrophobic surface
- intramolecular hydrogen bonds
- van der Waals interactions
- on the order of ~ -200 kcal/mol (free energy change for $U \rightarrow N$)

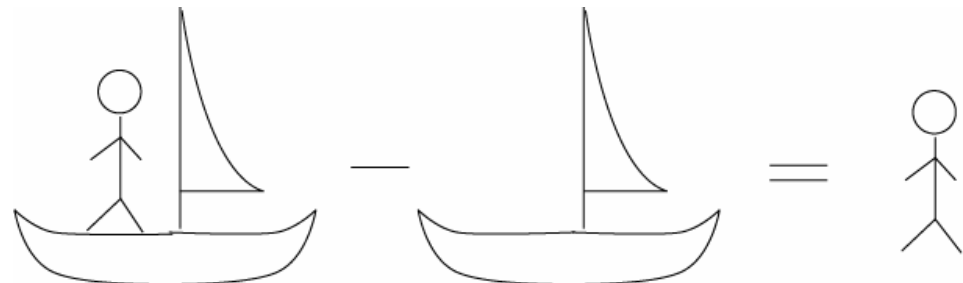


(+) Contributions to oppose folding

- loss of hydrogen bonds to water molecules
- entropy loss due to restricted backbone and side chain movements
- on the order of $+190$ kcal/mol

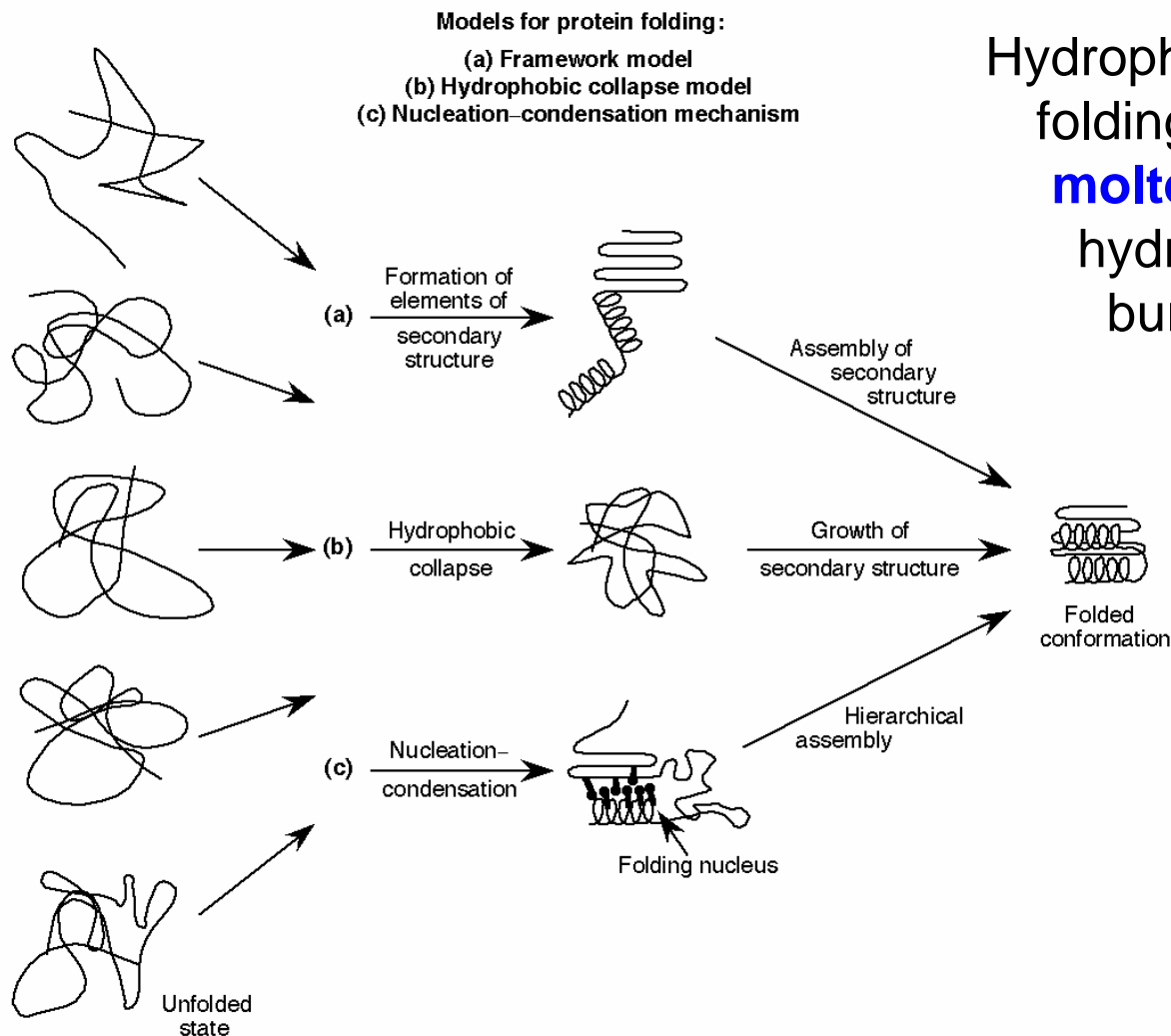
Net result:

stabilization by ~ -10 kcal/mol



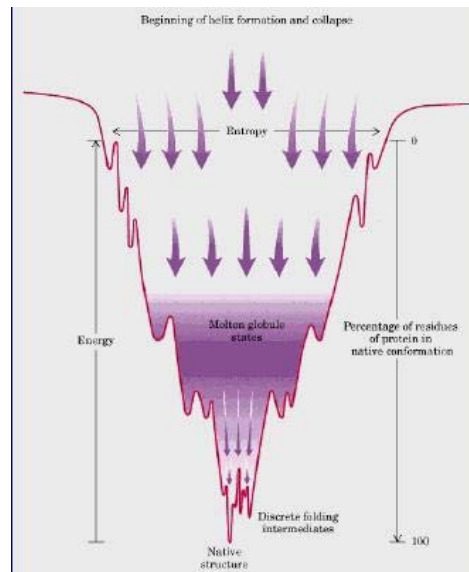
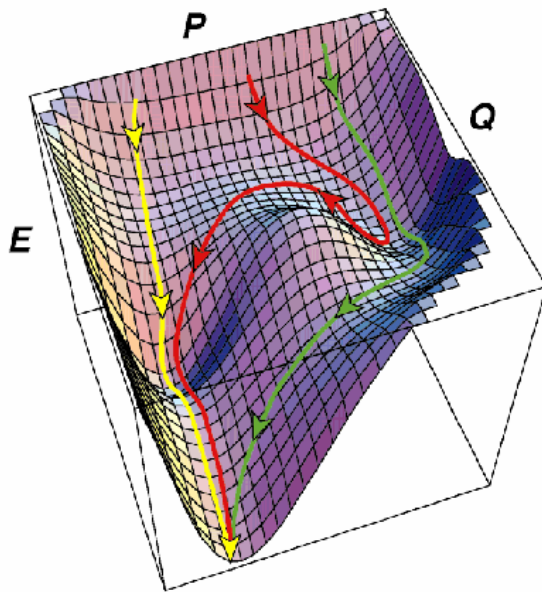
weight of captain and boat – weight of boat = weight of captain

How do proteins fold

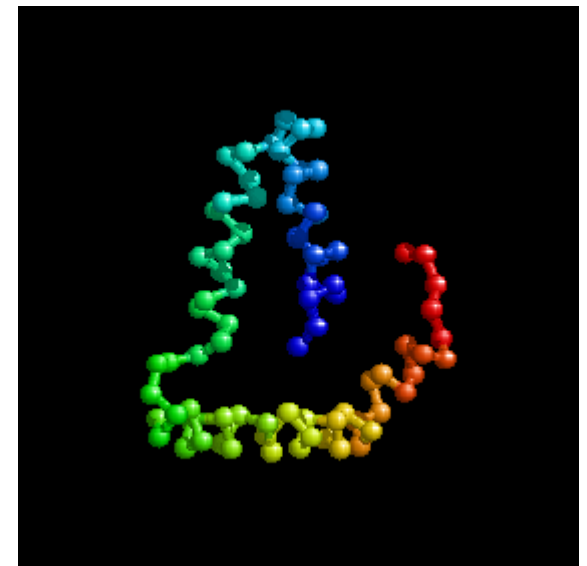


Hydrophobic collapse leads to a folding intermediate known as **molten globule**, in which the hydrophobic side chains are buried and hydrophilic side chains are exposed

- Protein fold much more rapidly than one might expect (often in μs to ms)
- Protein does not sample every possible conformation in order to reach the native state (i.e. the folded state)
- Otherwise, there are simply too many possibilities—e.g. 10^{100}
 - **Levinthal paradox**: if a protein samples every possible conformation, it'll never fold



folding funnel



Simulation of the folding of NK-lysin

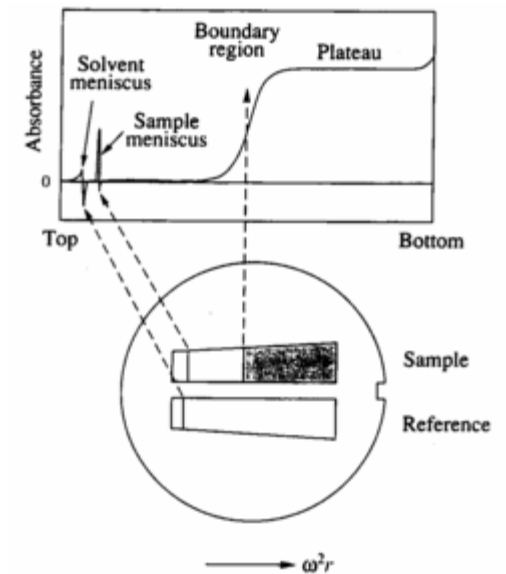
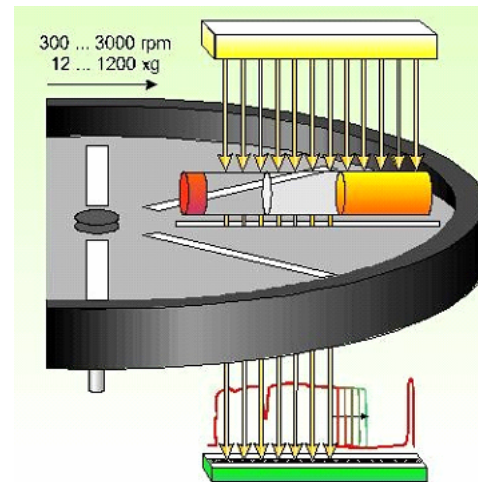
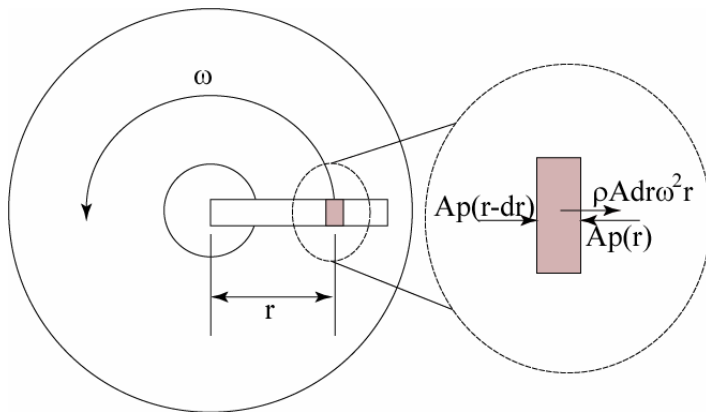
Jones, PROTEINS. Suppl. 1, 185-191 (1997)

Structural characterization

Molecular weight

- analytical ultracentrifugation: larger molecules sediment more quickly
- analytical gel filtration: larger molecules take **shorter** time to travel

Ultra (or analytical) centrifugation



$$P(r) = P(r_0) \exp\left(-\frac{m\omega^2}{2RT}(r^2 - r_0^2)\right)$$

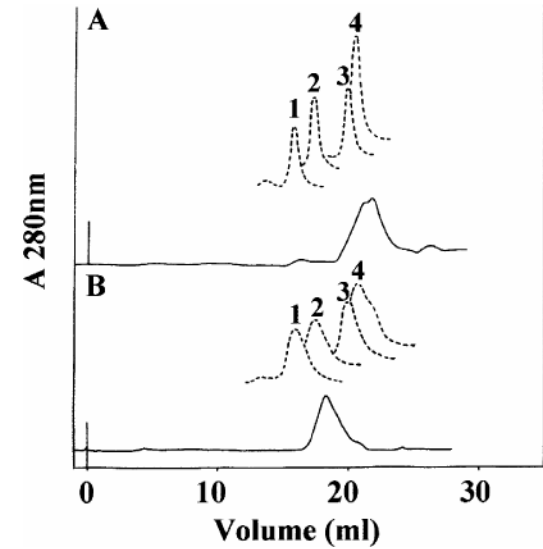
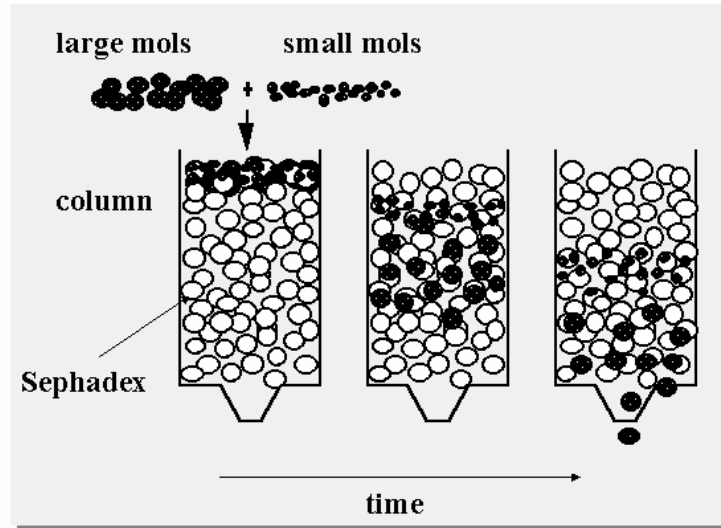
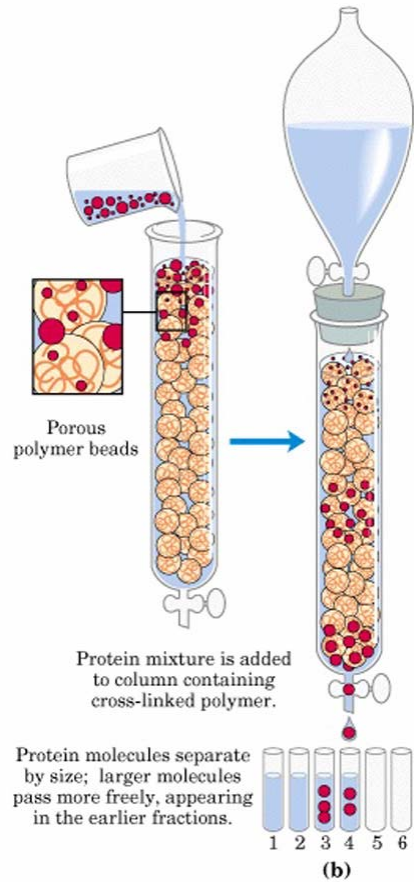
$$n(r) = n(r_0) \exp\left(-\frac{m\omega^2}{2RT}(r^2 - r_0^2)\right)$$

$$\log(n(r)) = \log(n(r_0)) - \frac{m\omega^2}{2RT}(r^2 - r_0^2)$$

$$= a + b \cdot r^2$$

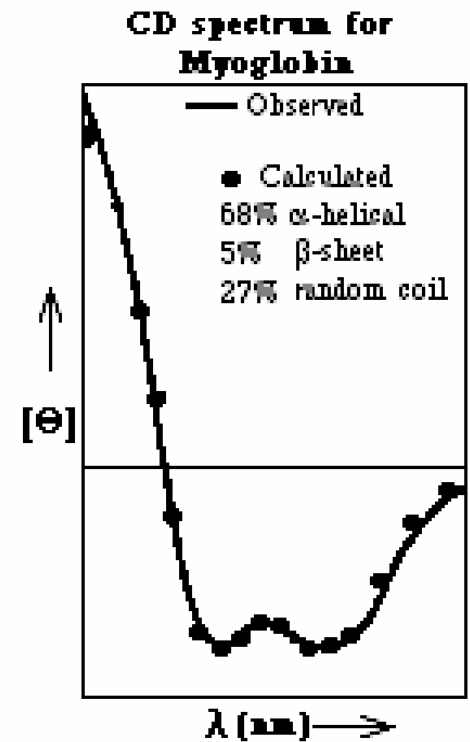
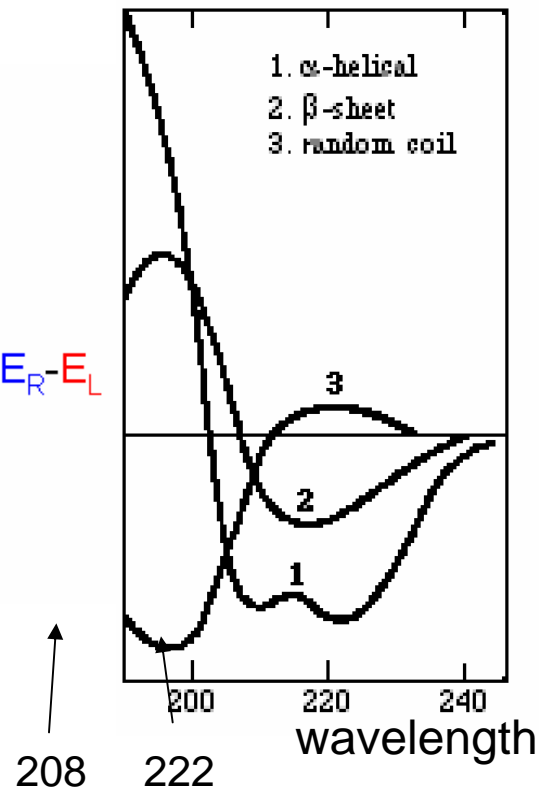
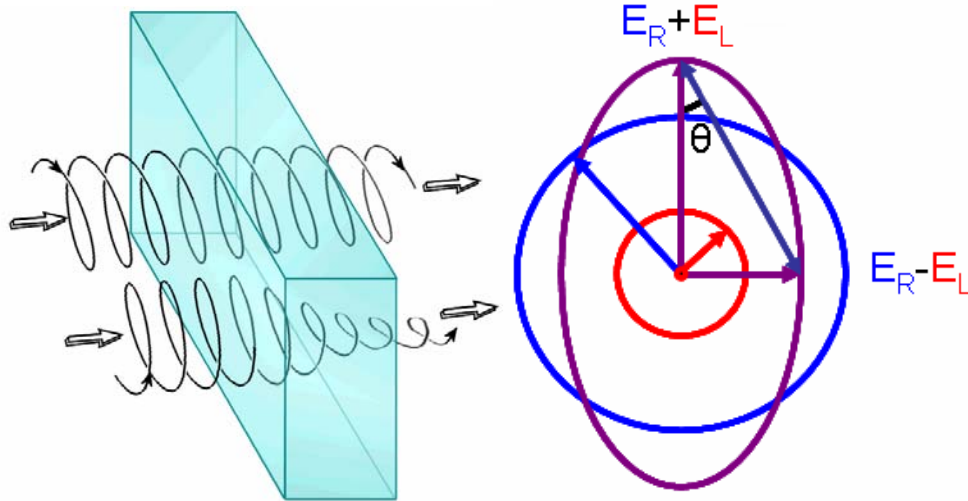
Gel filtration

also known as size exclusion column



Optical activity of protein

- Chiral compounds (including alpha helix and beta strand) absorb right-handed and left-handed polarized light to different degrees
- Circular dichroism (CD) spectroscopy measures the extent to which two circularly polarized lights get absorbed to measure secondary structure content



High resolution structure determination

- Nuclear magnetic resonance (NMR) spectroscopy



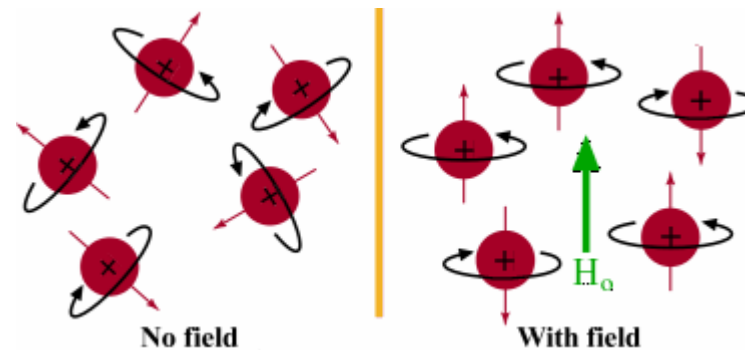
- ⊕ When placed in a magnetic field, NMR active nuclei (e.g. ^1H , ^{13}C , ^{15}N) absorb energy at a specific frequency, dependent on strength of the magnetic field
- ⊕ The resonance frequency depends critically on the local electronic structure
- ⊕ By measuring these frequency shifts (called **chemical shifts** after normalized for the strength of the magnetic field), it is possible to determine:
 - dihedral angle of a bond (hence, conformation)
 - distance between two atoms (hence, 3D distance constraint)



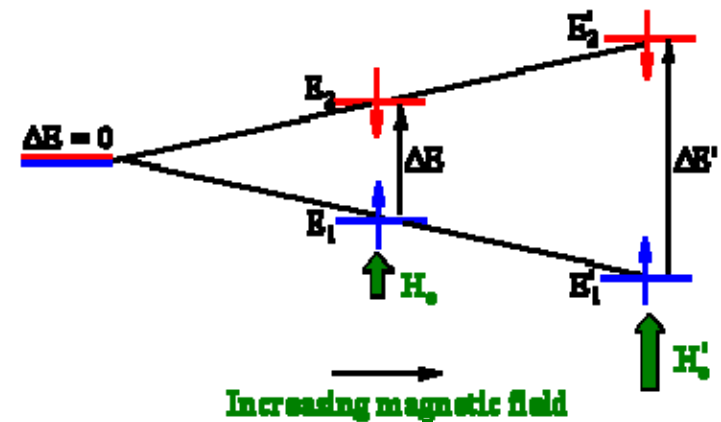
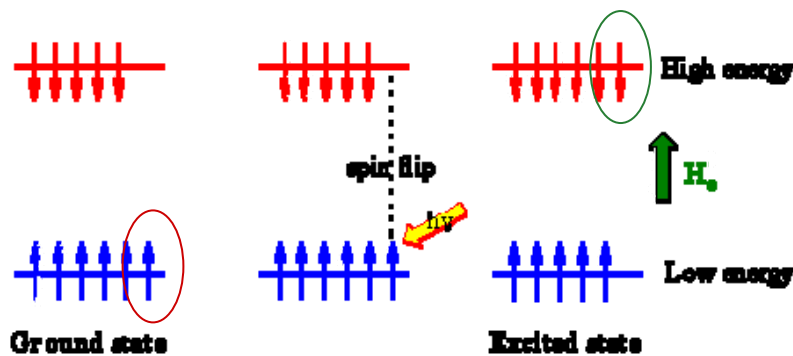
Chemistry Nobel prize 1995 and 2002

Principles of NMR

- When an atom with non-zero spin is placed in magnetic field, the up and down states have two different energies
- The transition from one to the other is accompanied by an absorption or emission of a photon

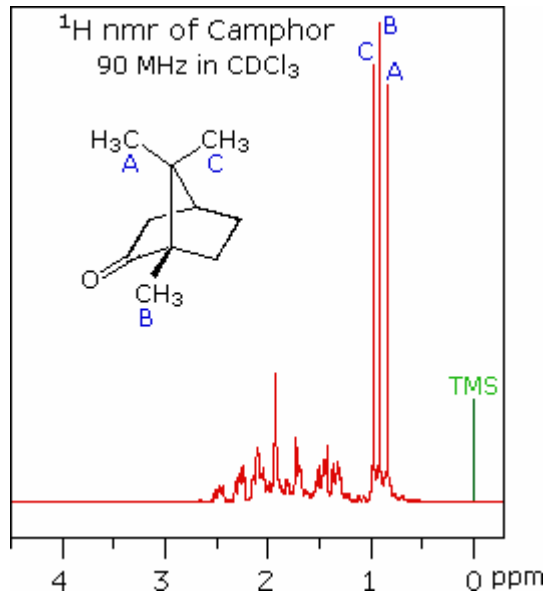


$$\Delta E = -2\vec{\mu} \cdot \vec{H}$$

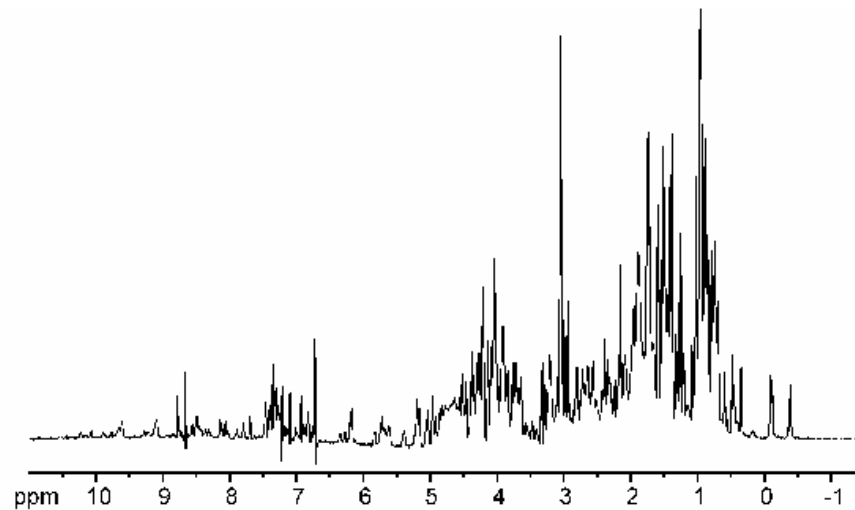


1D NMR

There are many more protons in a protein than in small molecules, leading to crowding and degeneracy of chemical shifts



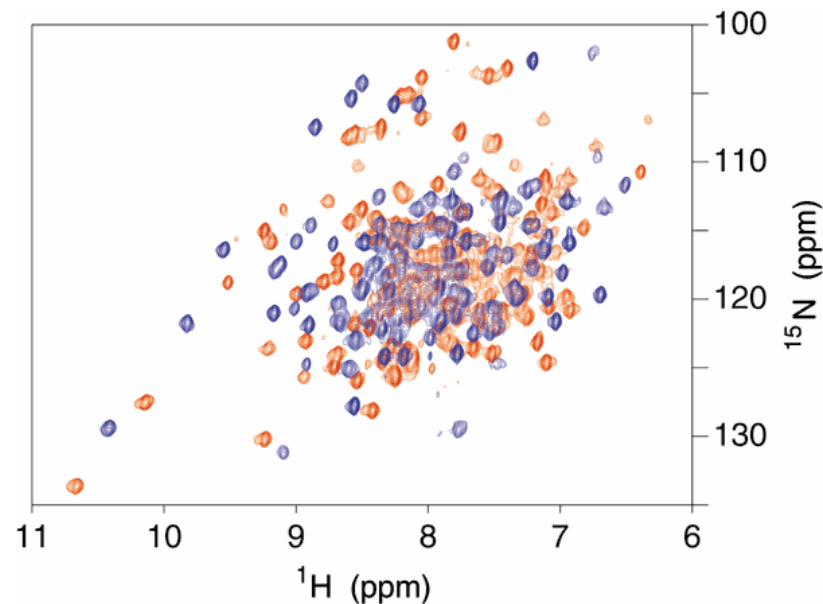
Typical protein 1D NMR



Multidimensional NMR

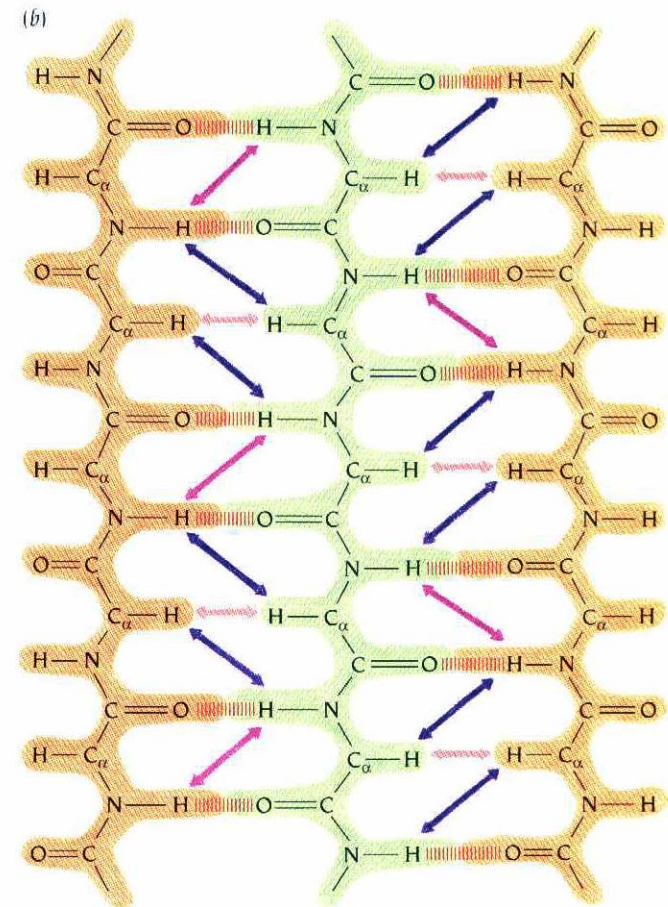
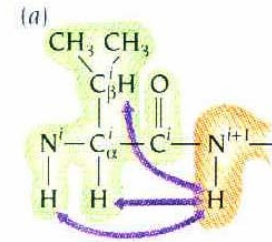
Spin labeling of protein with NMR active isotopes (e.g. **^{15}N** , **^{13}C**) makes these atoms “visible” and allows their chemical couplings to neighboring atoms to be examined

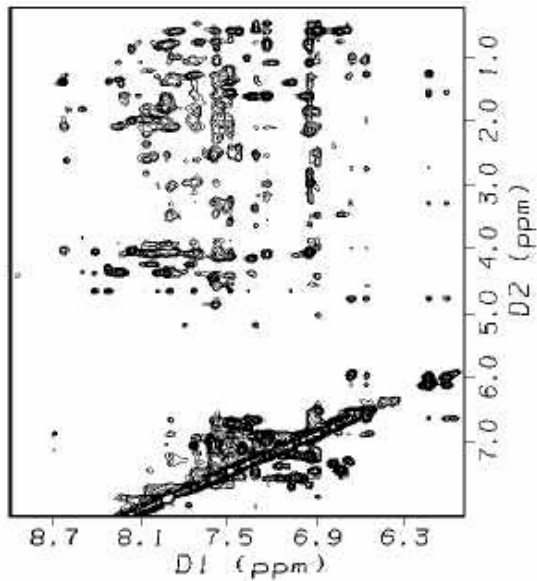
Backbone amide region of the 600-MHz (^1H , ^{15}N)-HSQC spectra of chain-selectively labeled carbon monoxide-bound hemoglobin A in water at pH 6.5 and 29 deg. Cross-peaks of a tetrameric hemoglobin



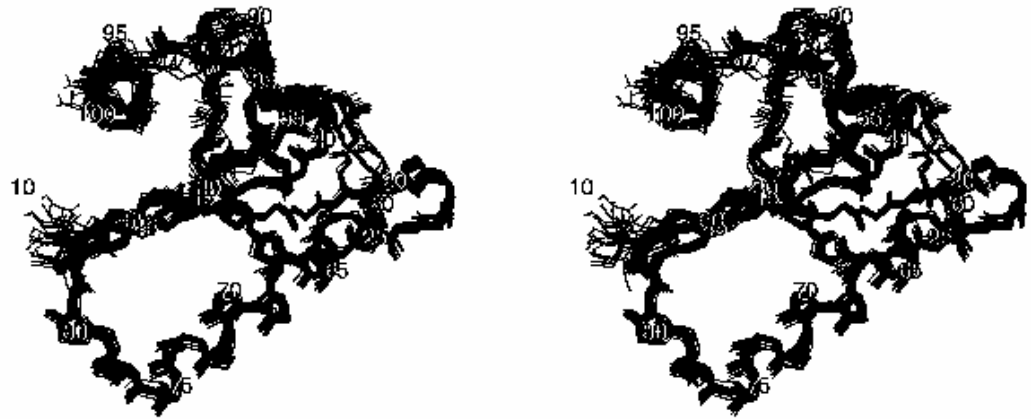
2D NOESY

- Spin labeled particles can interact with each other through space
- The intensity of the interaction (nuclear Overhauser effect, “**NOE**”) is $\sim 1/r^6$, where r is the distance between two nuclei
- Gives detectable signal up to $\sim 5 \text{ \AA}$
- NOE thus provides distance constraints between all pairs of protons that are important to construct the 3D structure of a protein
- Spatial arrangements of hydrogens are then computationally determined

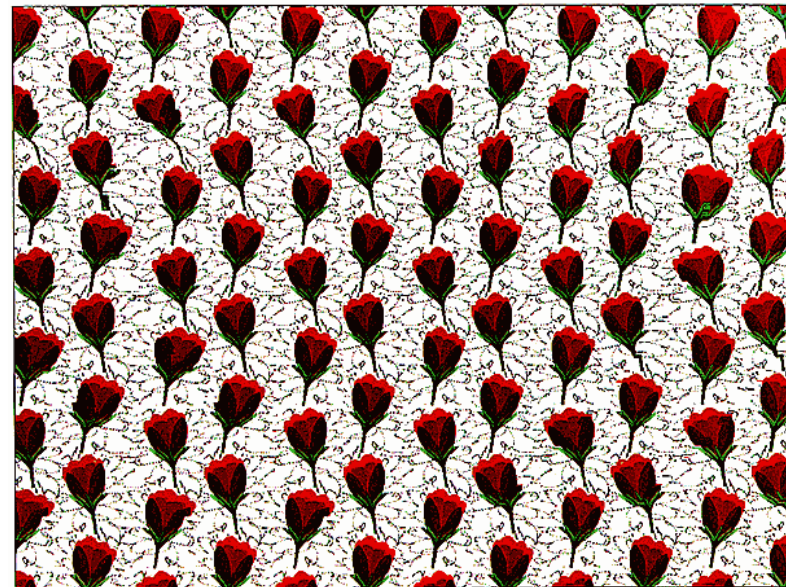




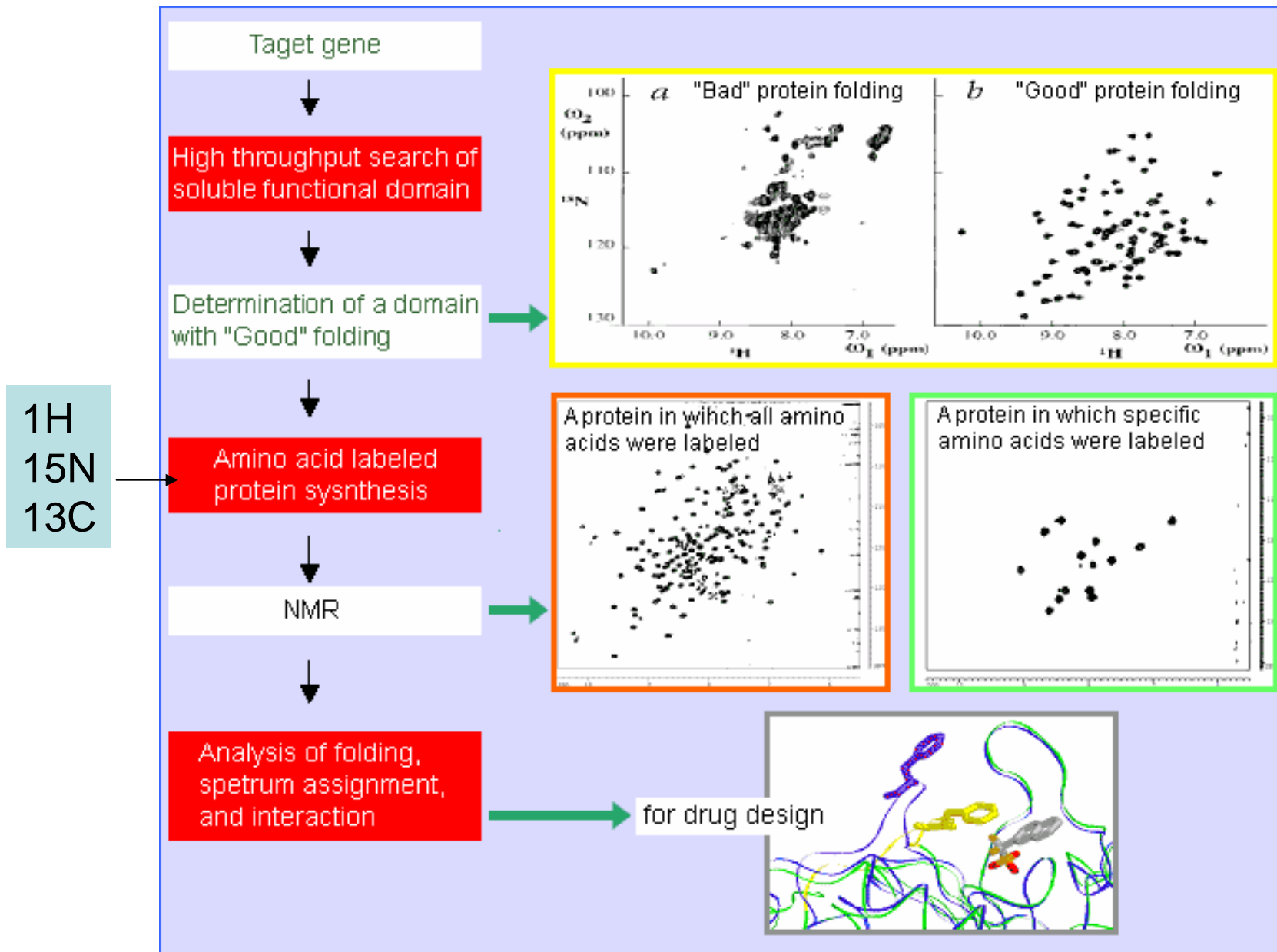
**2D NOESY spectrum of a BIR domain.
Cross-peaks show through-space
interactions between BIR domain protons.**



Stereoview of 10 lysozyme structures

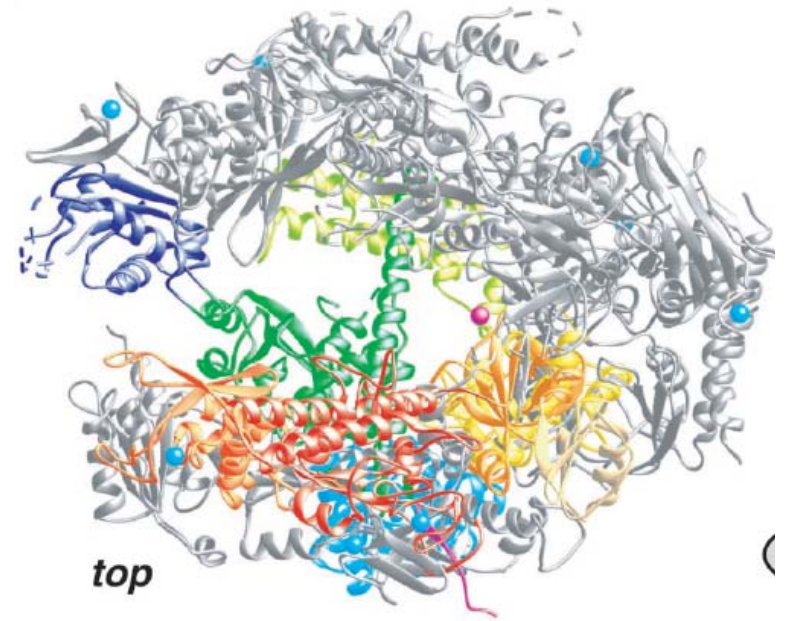


stereogram of a heart or a ghost of a tombstone



X-ray crystallography

- Atomic resolution structural information
- No size limit on how large the protein and protein complex can be
 - Compare with the upper limit on the molecular weight of the protein of ~50 kDa for NMR (although new techniques are being developed constantly)
- Amenable to high throughput approach
 - Structural GenomiX (SGX Pharmaceuticals) seeks high throughput structure determination of well characterized proteins
- High quality crystals are absolute musts

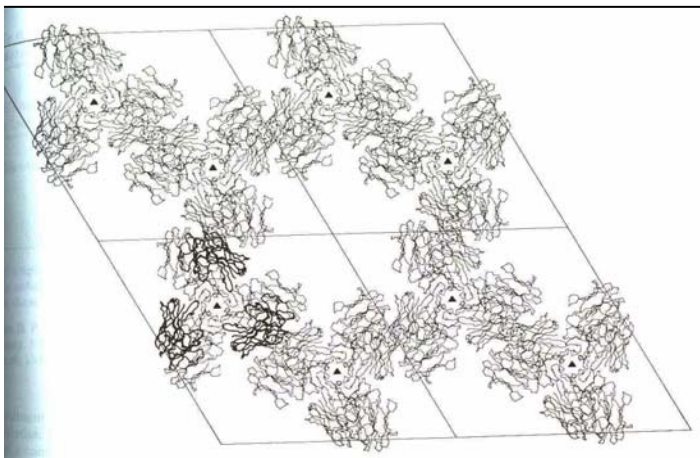


RNA polymerase II
3500 amino acids or ~ 400 kDa

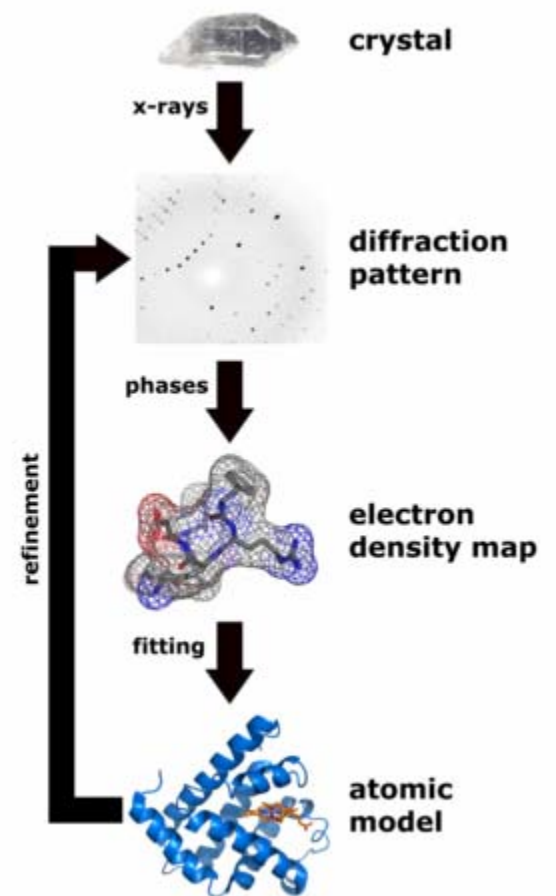
Cramer et al Science 292, 1863 (2001)
Chemistry Nobel prize 2006



- In a good protein crystal, each molecule of protein is laid out in precisely the same way in each **unit cell**
- An orderly arrangement of like atoms in space scatters (“**diffracts**”) light in geometric patterns
- The diffraction pattern is used to compute the corresponding electron density map
- Need to determine the phase
- Protein backbone and side chains can be fitted into the measured electron density
- Refinement of the structure results in 3D coordinates of the individual atoms of a protein

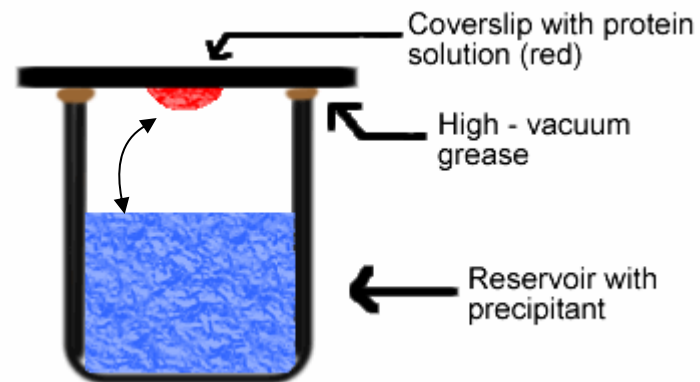
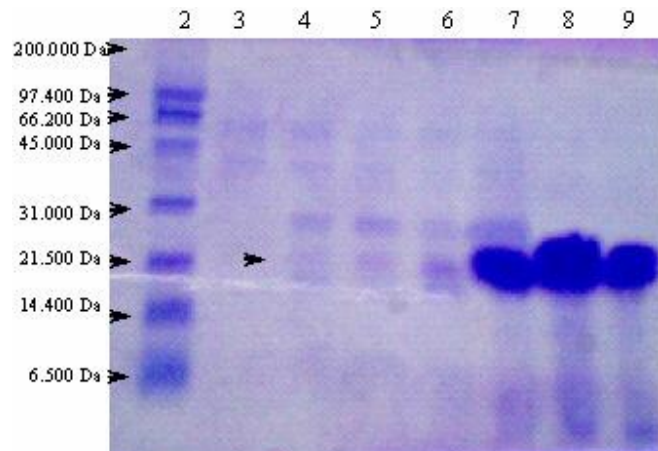


Creighton, Protein Structure



Growing protein crystal

- First the protein of interest needs to be purified to near complete homogeneity (~ 99%)
- Protein crystals can be grown by gradually increasing the protein concentration in solution through vapor diffusion against a reservoir



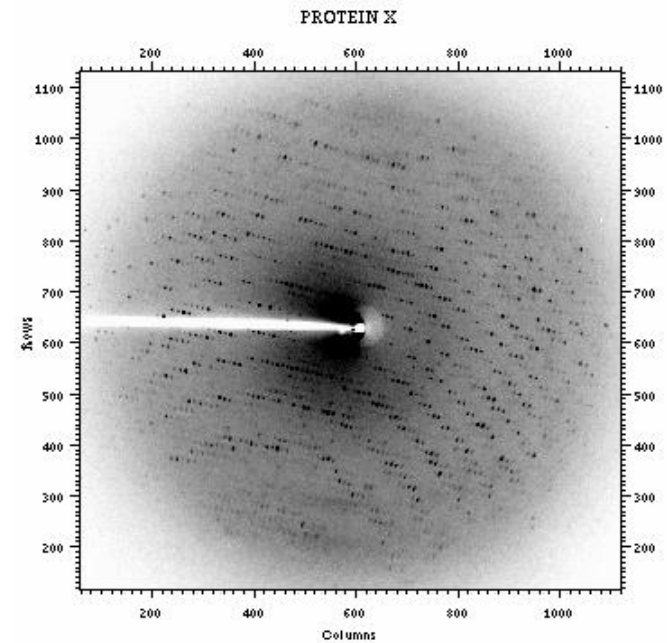
X-ray diffraction

X-ray can be generated using an in-house machine or at a synchrotron

Synchrotron beams are much more intense and the wavelength can be fine-tuned



Argonne National Lab synchrotron

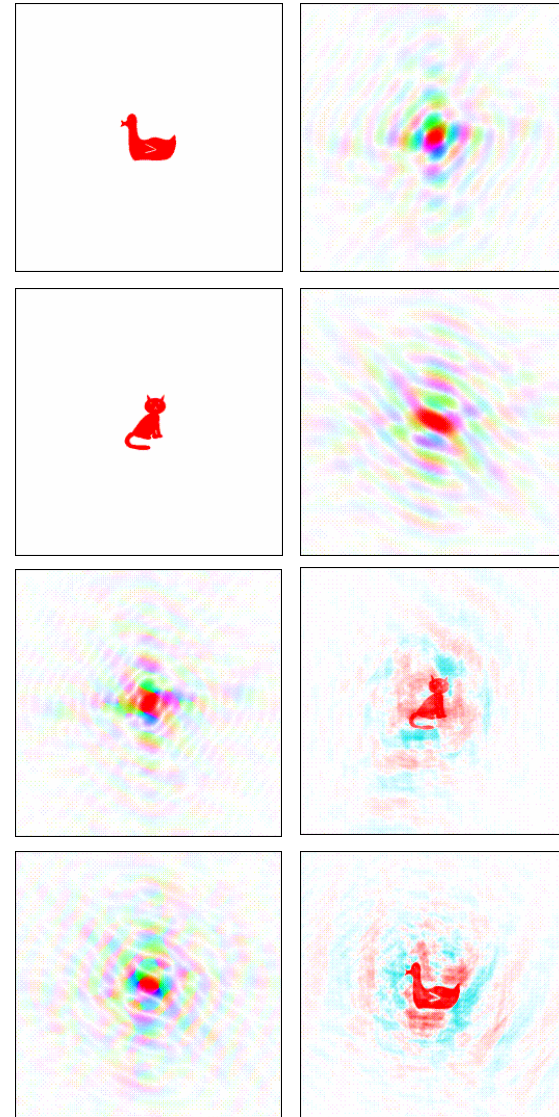


Solving the crystal structure

- Diffraction pattern corresponds to the Fourier transformation of the electron density
- In order to get the electron density back, we need to perform inverse Fourier transformation
- However, the phase information is required
 - molecular replacement
 - heavy atom soaking
 - multiple wavelength anomalous dispersion (MAD)
 - direct method

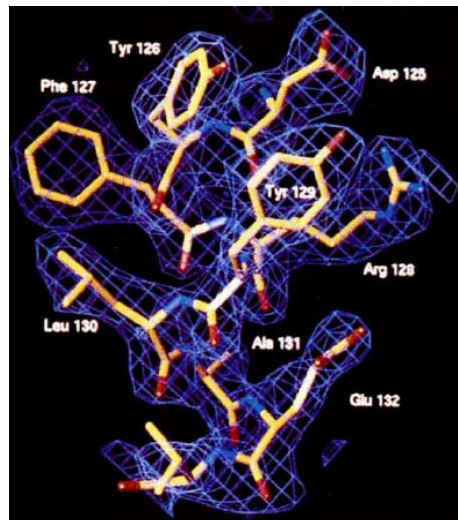


Hauptman, SUNY Buffalo
1985 Chemistry Nobel prize

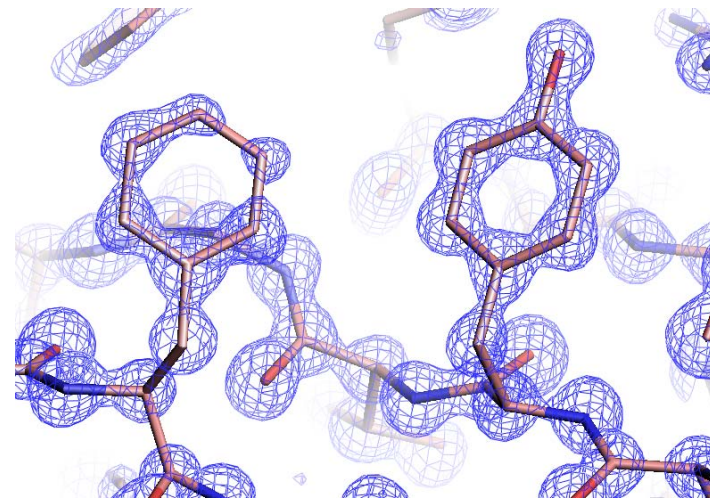


Fitting the peptide chain

- Once the electron density has been computed, the peptide chain can be threaded (computationally) into the density.
- This is easier if the resolution is higher (i.e. $< 2.0 \text{ \AA}$)
- Amino acids have known connectivities, bond lengths, angles, and stereochemistry, which aid in fitting



2.9 Å



1.2 Å

Structural classification (cont'd)

- Alpha
- Beta
- Alpha/Beta
- Small proteins

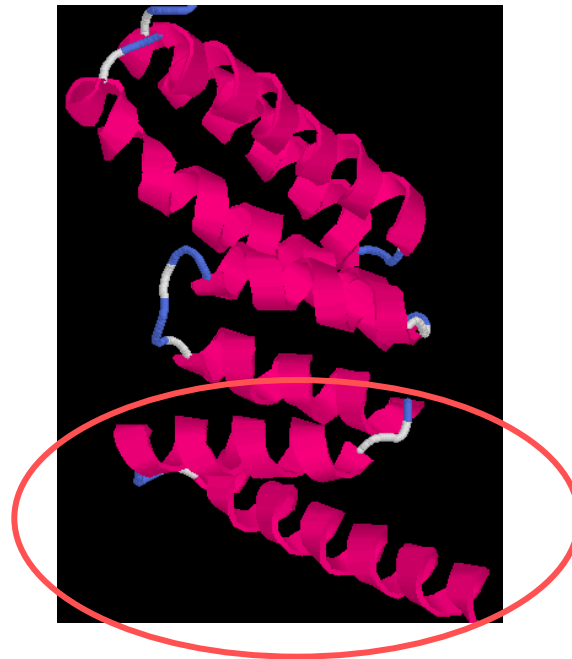
Alpha domain proteins

May or may not contain any motifs

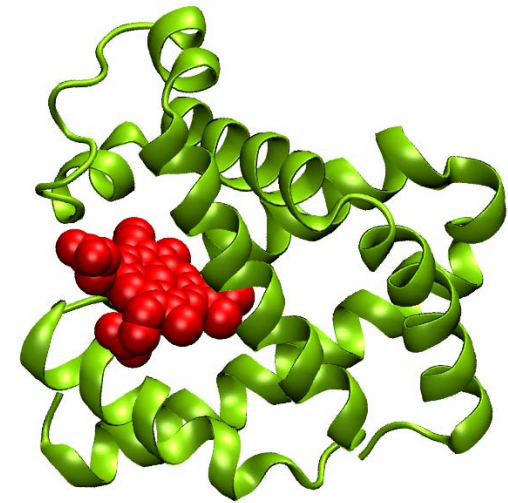
Small and large



hemerytherin
(four helix
bundle domain)

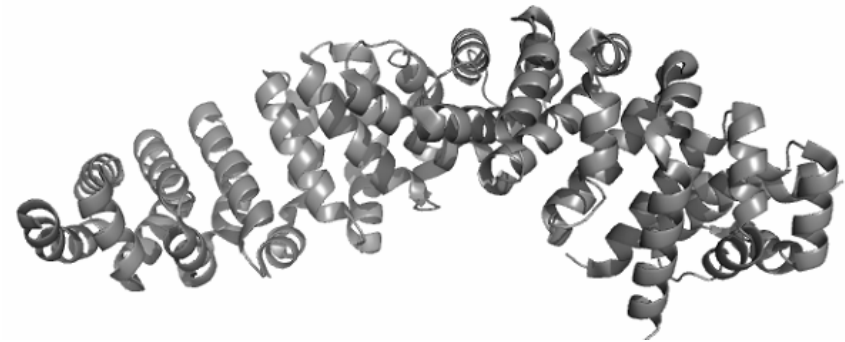
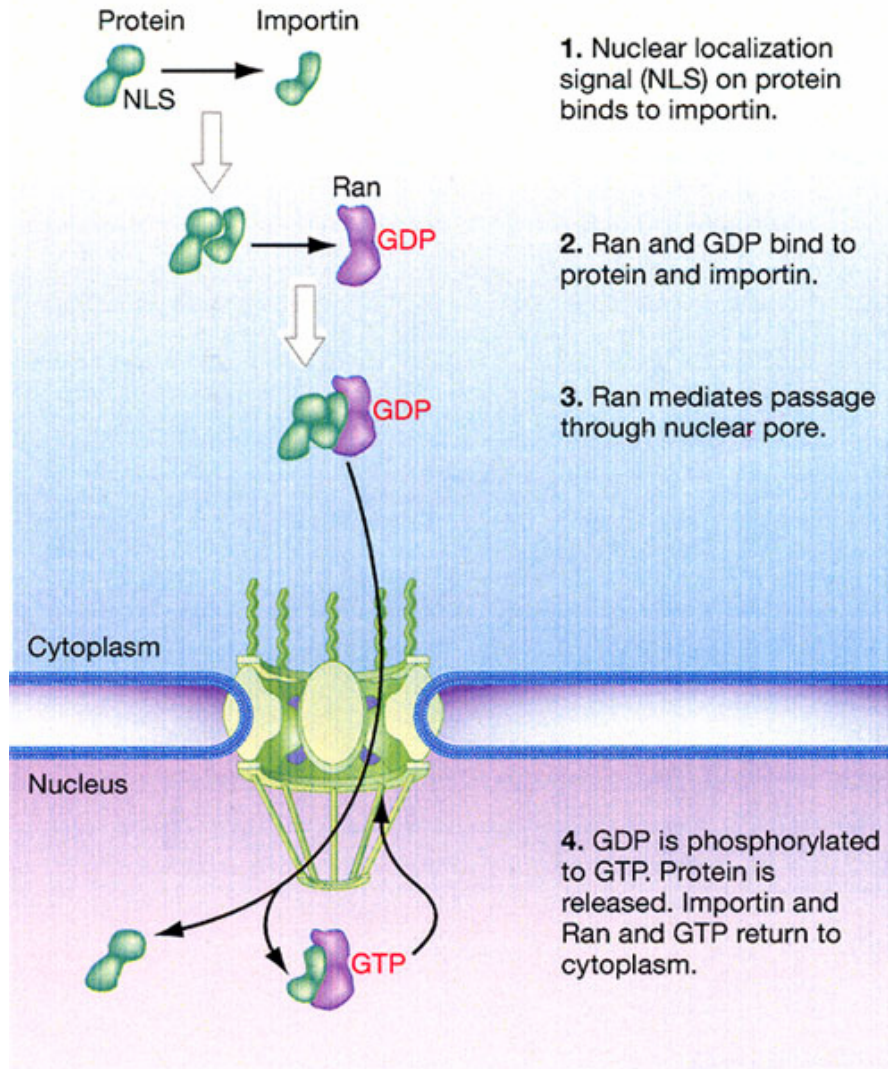


bovine cyclophilin
(tetratricopeptide
repeat domain)



Myoglobin (globin domain)

HOW PROTEINS ARE IMPORTED INTO THE NUCLEUS



importin
(Armadillo repeat)

FIGURE 7.24 An Importin, Ran, and GDP Are Required to Import Proteins into the Nucleus

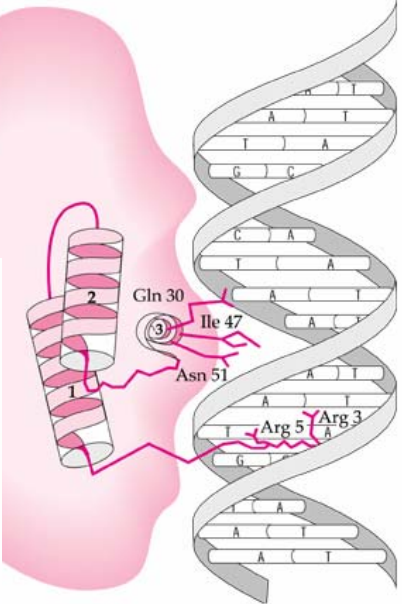
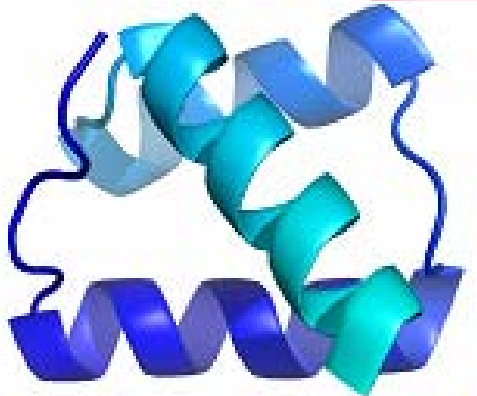
The current model for how proteins are imported into the nucleus. Importin, Ran, and GDP are recycled to the cytoplasm after they deliver cargo to the nucleus.

Functional diversity

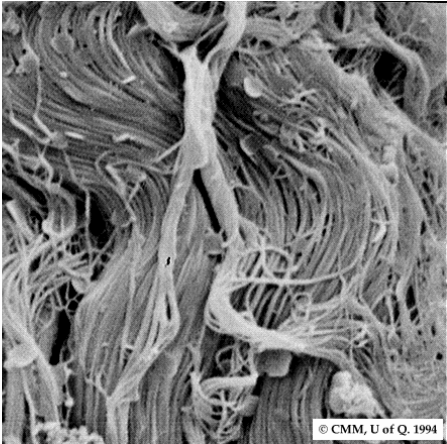
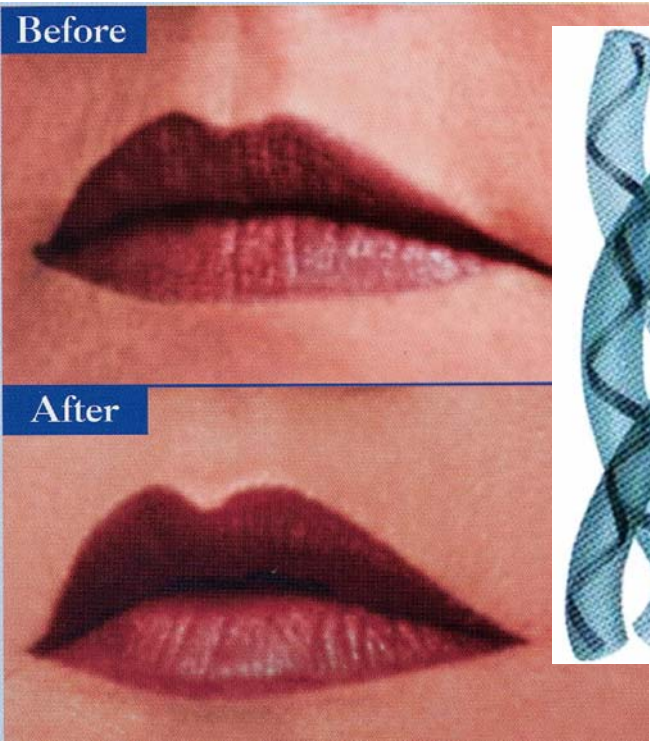
Structural : collagen

DNA binding : engrailed homeodomain

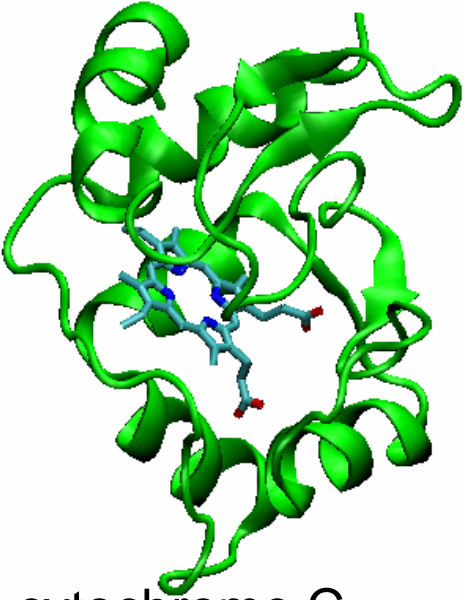
Catalysis : cytochrome c



engrailed homeodomain



collagen



cytochrome C

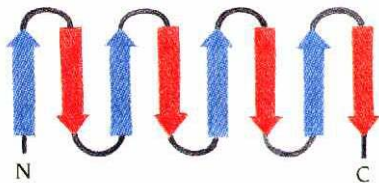
Beta proteins

Structurally and functionally diverse

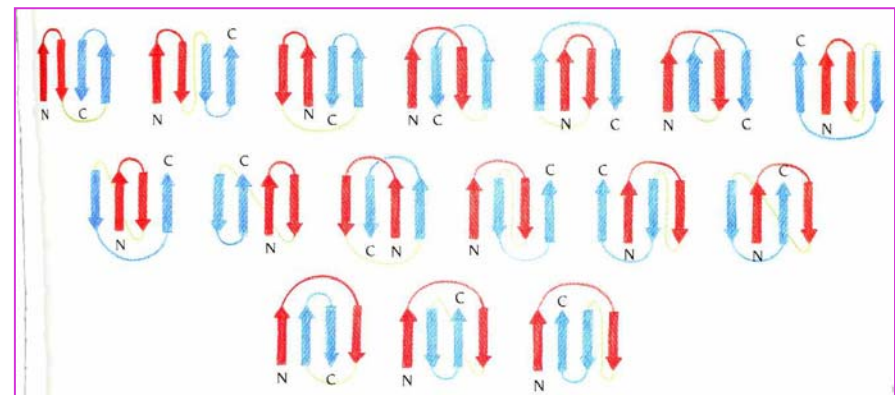
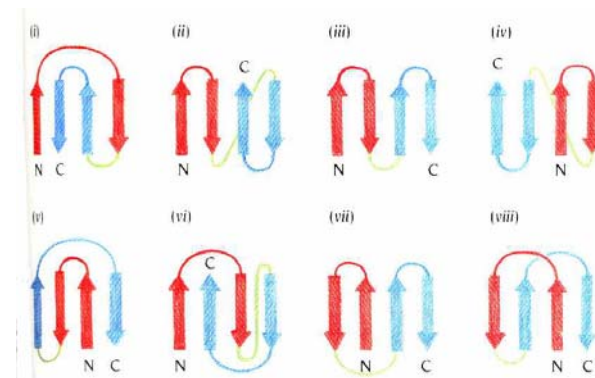
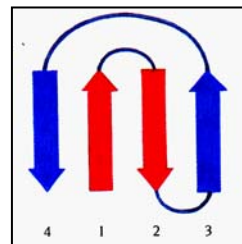
- enzymes, transport proteins, antibodies
- cell surface proteins, viral coat proteins

Strands are often arranged in an antiparallel fashion

up-down



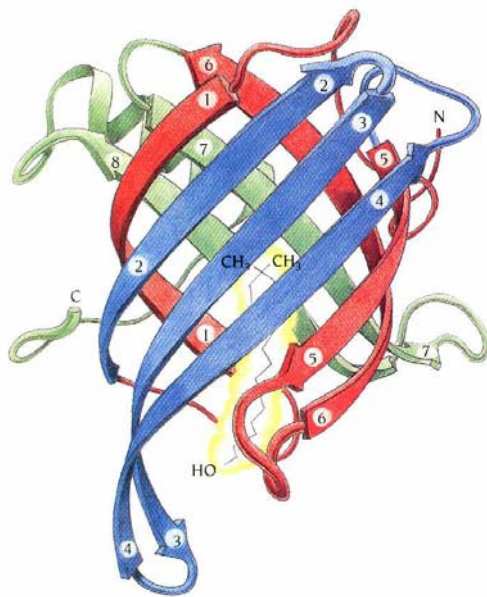
Greek key



not seen in nature

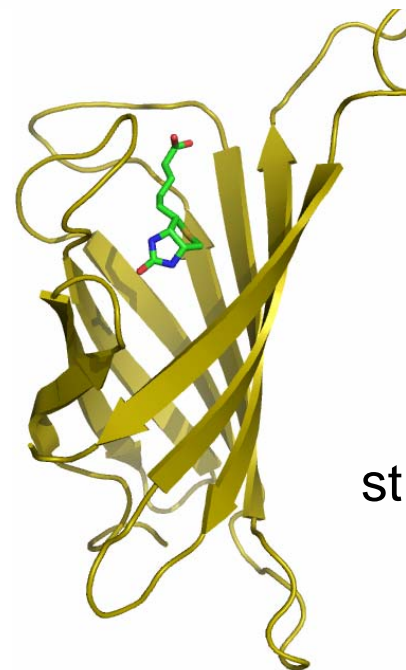
Beta barrels

- Intrinsic twist in beta sheets often leads to a barrel-like structure when two sheets are packed against each other
- Amino acid sequence reflects beta structure
 - alternating hydrophobic and hydrophilic residues



retinol binding protein

		↓		↓		↓		↓										
2	41–48	-	Ile	-	Val	-	Ala	-	Glu	-	Phe	-	Ser	-	Val	-	Asp	-
3	53–60	-	Met	-	Ser	-	Ala	-	Thr	-	Ala	-	Lys	-	Gly	-	Arg	-
4	71–78	-	Ala	-	Asp	-	Met	-	Val	-	Gly	-	Thr	-	Phe	-	Thr	-



streptavidin

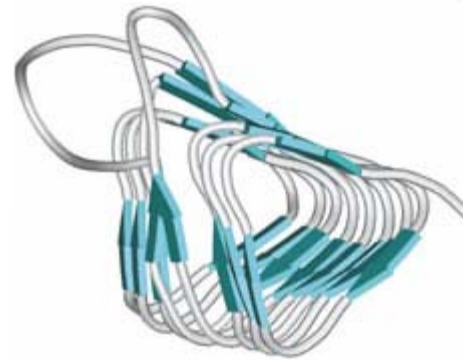
Include enzymes, antibodies, cell surface receptors, signaling molecules, viral coat proteins



immunoglobulin
(beta sandwiches)



SH-3

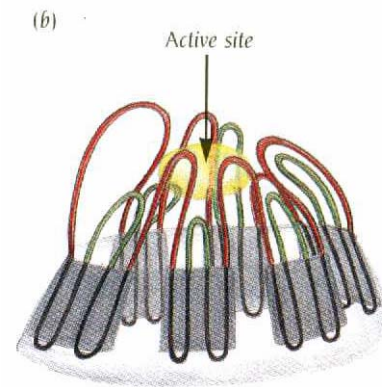


beta helix



collagenase
(trefoil)

In beta barrel enzymes, the loop regions often contain residues involved in catalysis



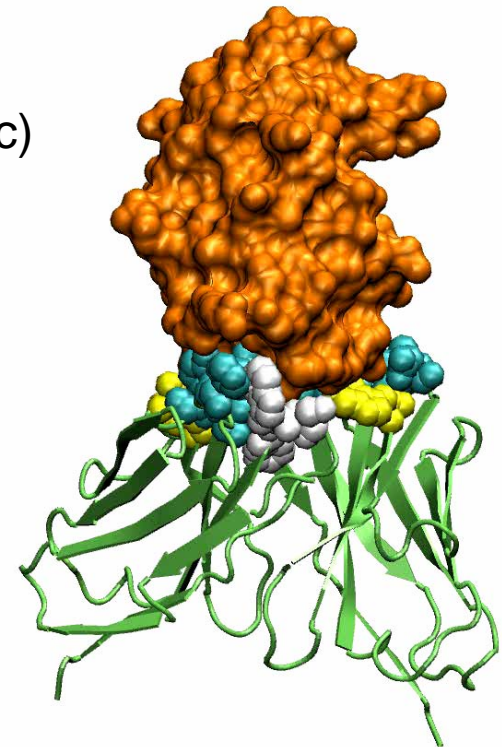
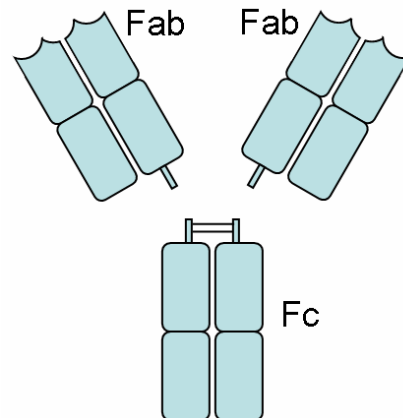
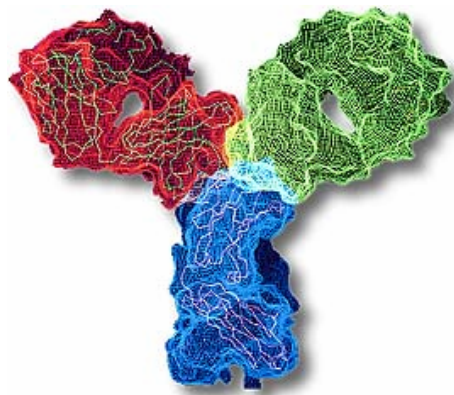
Immunoglobulin (Ig) – beta sandwich

Cell surface receptors

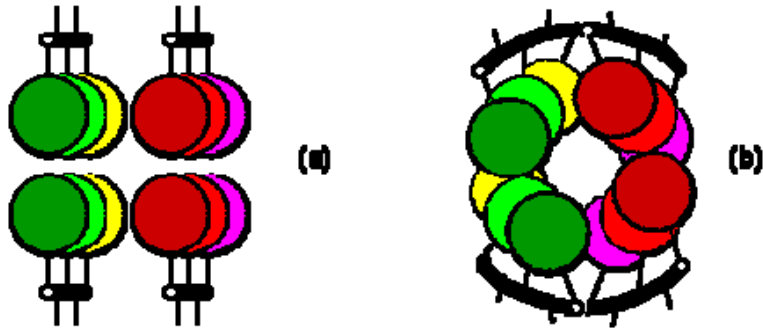
- Tumor necrosis factor receptor
- Interferon receptor

Antibodies

- “Y” shaped protein involved in (acquired) immune response
- Five different types (A, D, E, G, M)
- Comprises two heavy chains (4 Ig domains each) and two light chains (2 Ig domains each)
- Each Ig domain is about 120 amino acids
- Variable domain (Fab) and constant domain (Fc)
- Binding specificity resides within loop residues

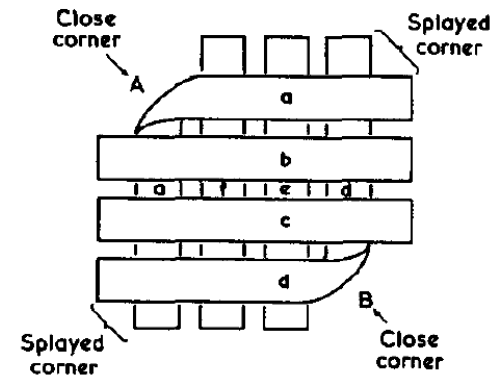
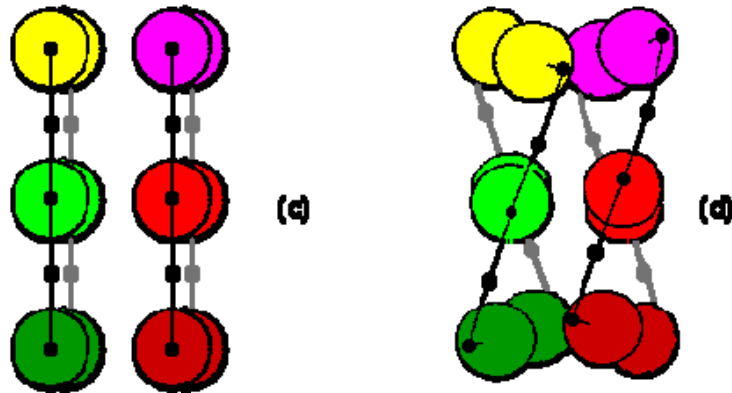


Sheet packing



(a, c) End and side views of two untwisted beta sheets. (b, d) The intrinsic twist of the beta sheet results in the sheets forming an angle of approximately 30° with each other

Chothia, Annu Rev Biochem 1984



Orthogonal beta sheet packing consist of beta sheets folded on themselves
The corner strands continues from one layer to the other

Alpha/beta proteins

Most frequent domain structure

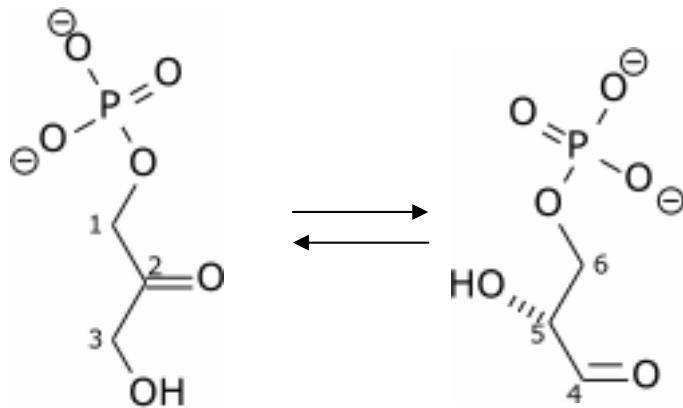
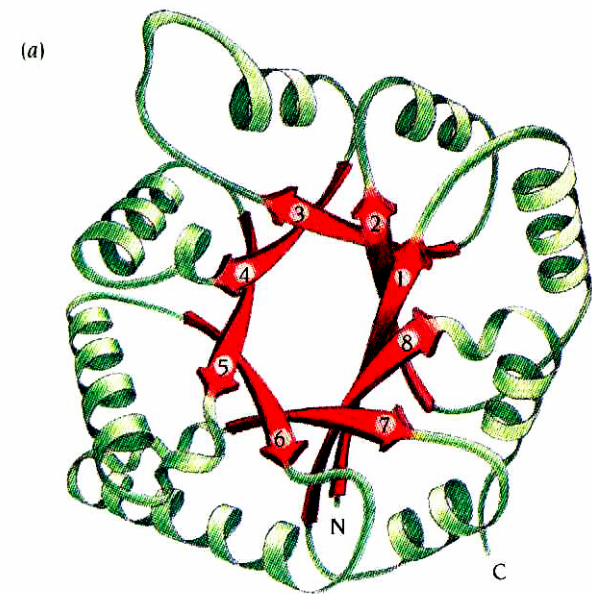
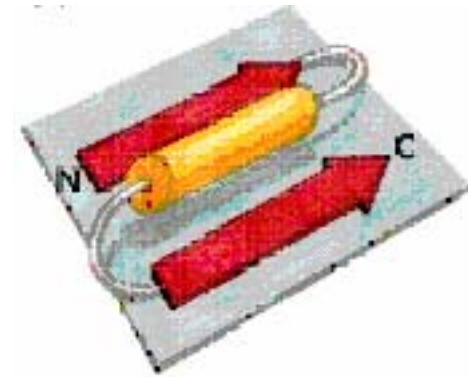
Central beta sheet (parallel or mixed) surrounded by alpha helices

Used in enzymes and transporter molecules

1. Triose phosphate isomerase (TIM) barrel
2. Open beta sheet (e.g. Rossmann fold)
3. Leucine-rich motif

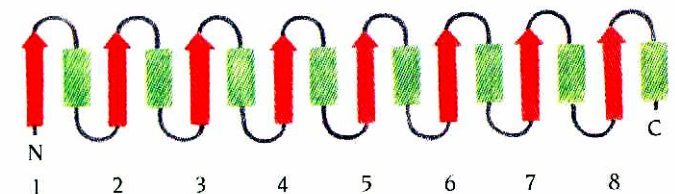
TIM barrel

- Triose phosphate isomerase is involved in glycolysis
- Eight copies of beta-alpha-beta motif joined in the same orientation
 - strands 1 and 8 hydrogen are bonded to each other
 - second strand of i -th motif and first strand of $(i+1)$ -th motif are shared



dihydroxyacetone phosphate

D-glyceraldehyde-3-phosphate



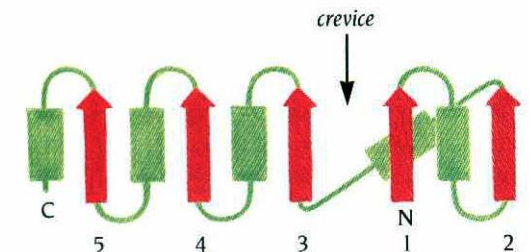
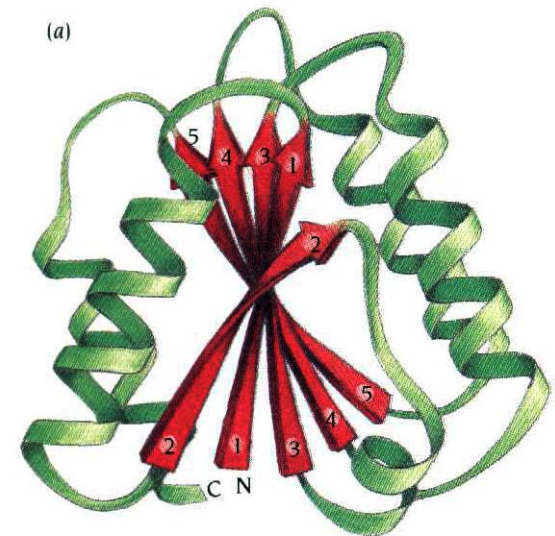
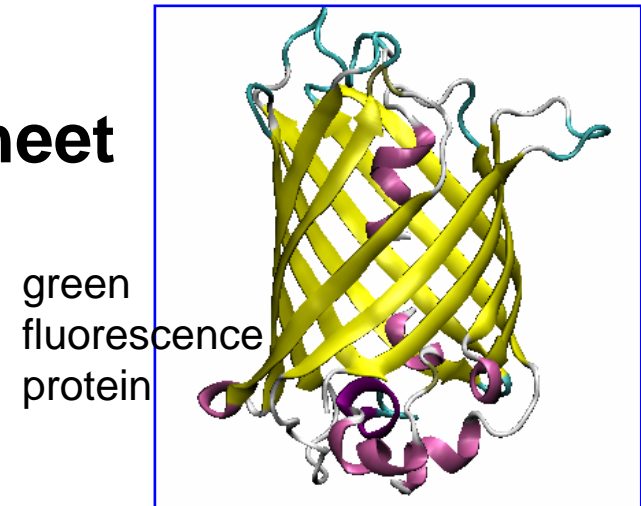
- Rediscovered at least 21 times during evolution
 - Nagano et al, JMB, 321, 741 (2002)
 - Entire database devoted to TIM barrel enzymes (“**DATE**”)
 - <http://www.mrc-lmb.cam.ac.uk/genomes/date/>
- Minimum of 200 residues are required to construct the entire protein
- TIM barrel proteins are all enzymes and make up ~10% of all enzymes
- Beta strands and alpha helices (~ 160 amino acids) provide structural framework
- Loop residues do not contribute to structural stability but instead are involved in catalytic activity
- Branched hydrophobic side chains in the core (tightly packed)
 - Val, Ile, Leu together make up 40%



Open twisted beta sheet

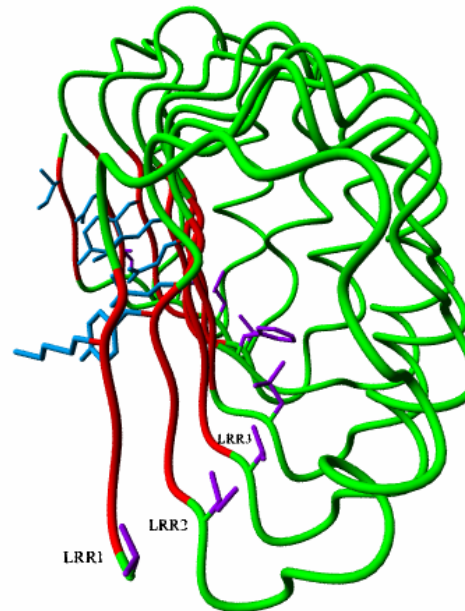
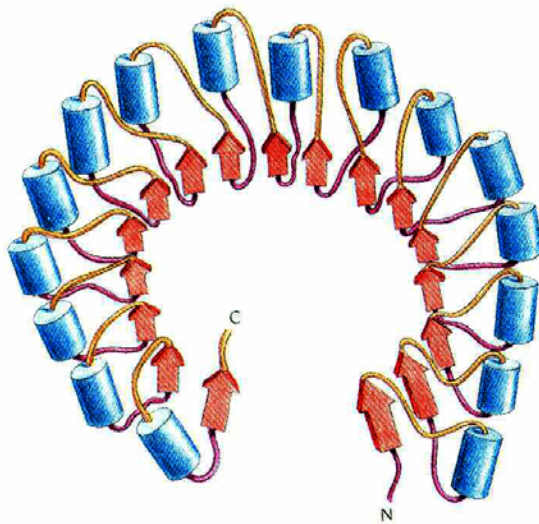
- Alpha helices on both sides of a beta sheet
- Does not form a barrel (see figure of GFP)
- Always contains two adjacent beta strands whose connections to the next strand are on opposite sides of the beta sheet, creating a crevice
- Active site always at the carboxy end of the beta sheet
- Functional residues come from the **loop regions**

Examples: carboxypeptidase, arabinose-binding protein, tyrosyl-tRNA synthetase, phosphoglycerate mutase



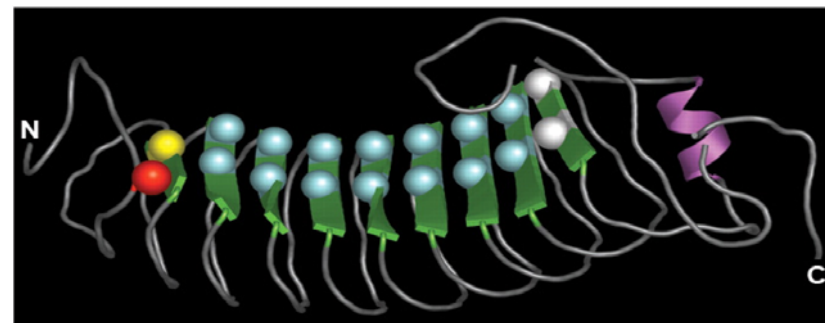
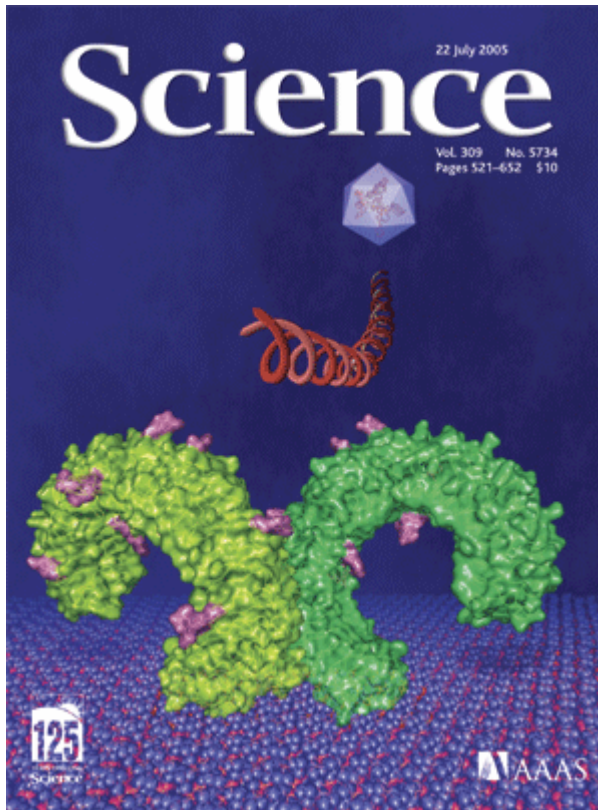
Leucine rich repeat

- Right-handed beta-alpha superhelix
- Composed of repeating 20-30 amino acid stretches
- The region between the helices and sheets is hydrophobic and is tightly packed with recurring leucine residues
- One face of the beta sheet and one side of the helix array are exposed to solvent and are therefore dominated by hydrophilic residues
- Diversity may be generated by mutating residues on the solvent exposed side of beta strands



LRR proteins in nature

- Toll-like receptor—innate immunity in mammalian cells
- Adaptive immune system in jawless fish lamprey
- RNase inhibitor



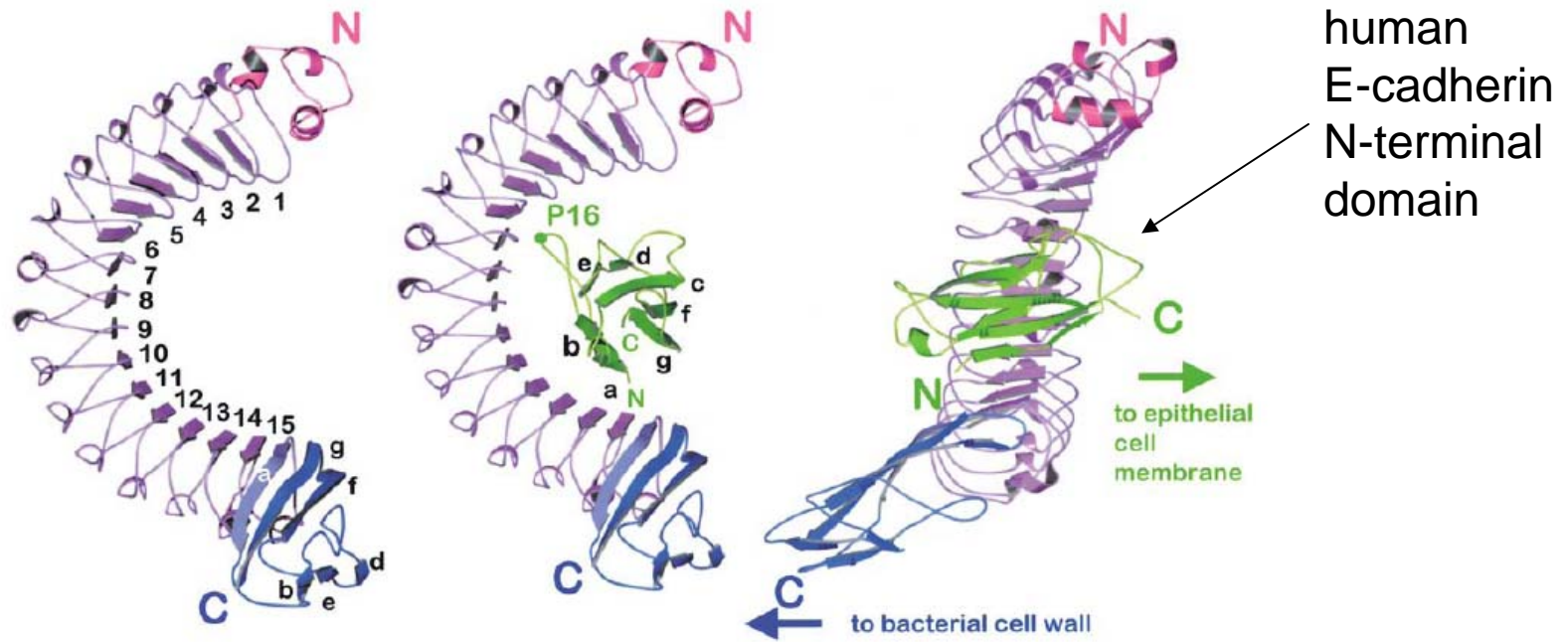
Alder et al, Science 310, 1970 (2005)



Ribonuclease inhibitor

- *Listeria* internalin

- Schubert et al, Cell 111, 825 (2002)

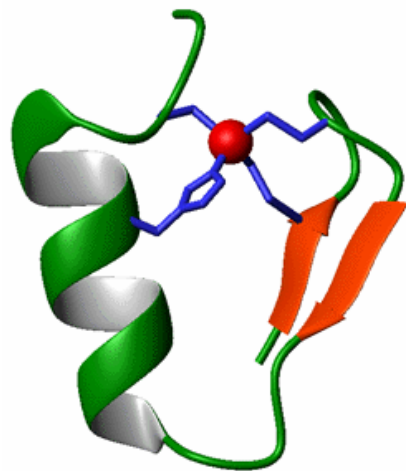


Small proteins

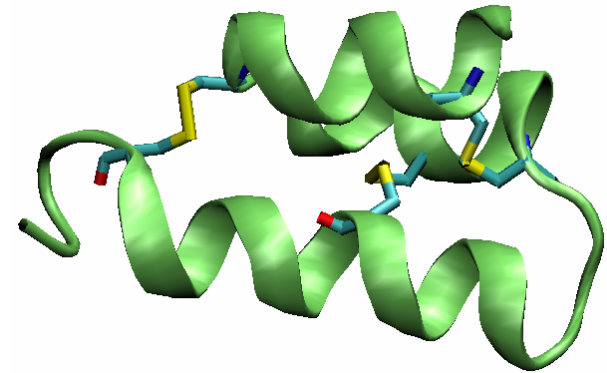
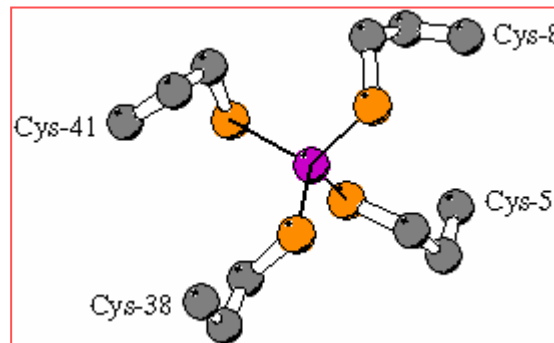
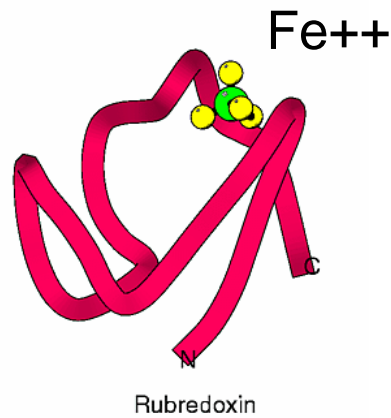
Often require additional interactions to make up for the absence of a significant hydrophobic core

metals, e.g. Zn^{++} , Fe^{++} , Ca^{++}

disulfide bonds



zinc finger



protozoan pheromone