Introduction to Proteins

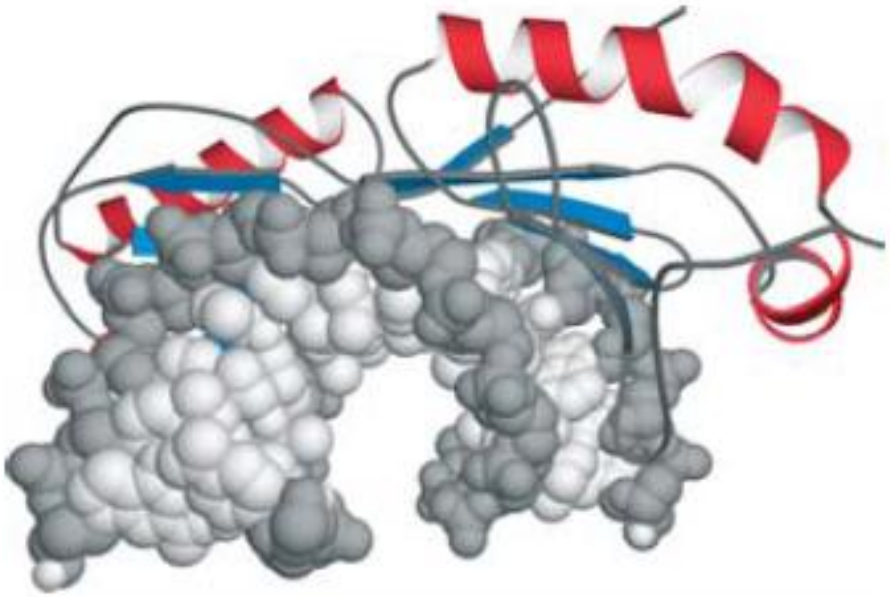
Dr. Fayyaz ul Amir Afsar Minhas
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

Importance of proteins

- 50% of the dry weight of the human body is protein
- Up to 92% of the dry weight of the red blood cell is a single protein called Hemoglobin
- Almost all cellular functions and biological functions involve proteins including
  - DNA Transcription
  - RNA Translation
  - Splicing
  - Muscular construction
  - Structural functions…

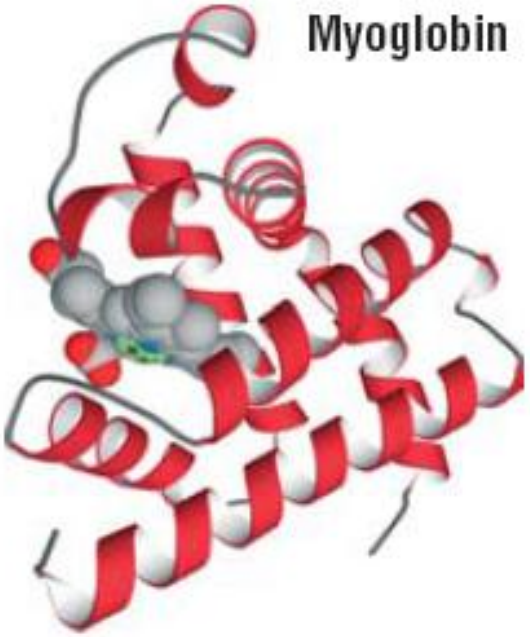
Binding

TATA binding protein



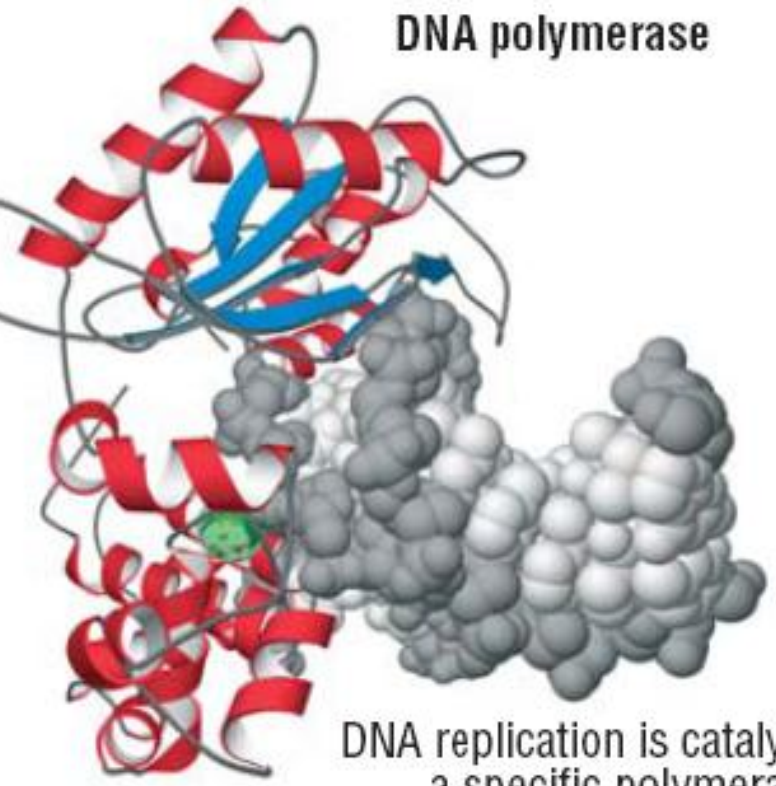
The TATA binding protein binds a specific DNA sequence and serves as the platform for a complex that initiates transcription of genetic information. (PDB 1tgh)

Myoglobin

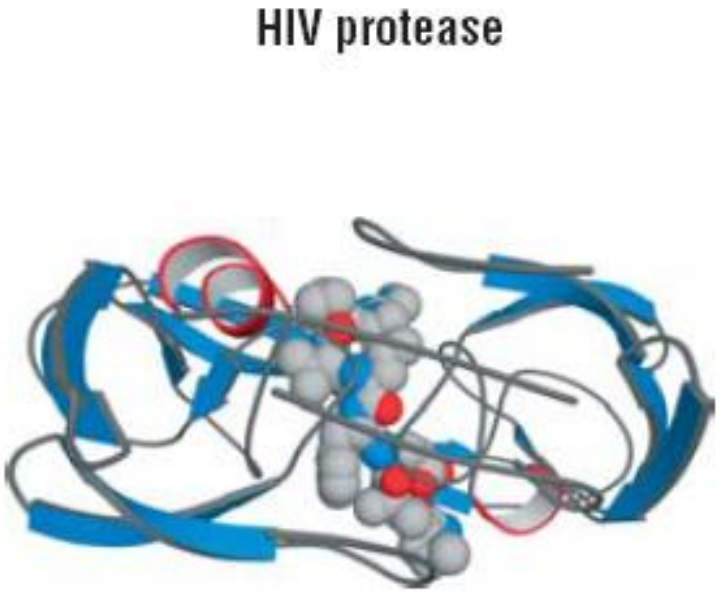


Myoglobin binds a molecule of oxygen reversibly to the iron atom in its heme group (shown in grey with the iron in green). It stores oxygen for use in muscle tissues. (PDB 1a6k)

Catalysis



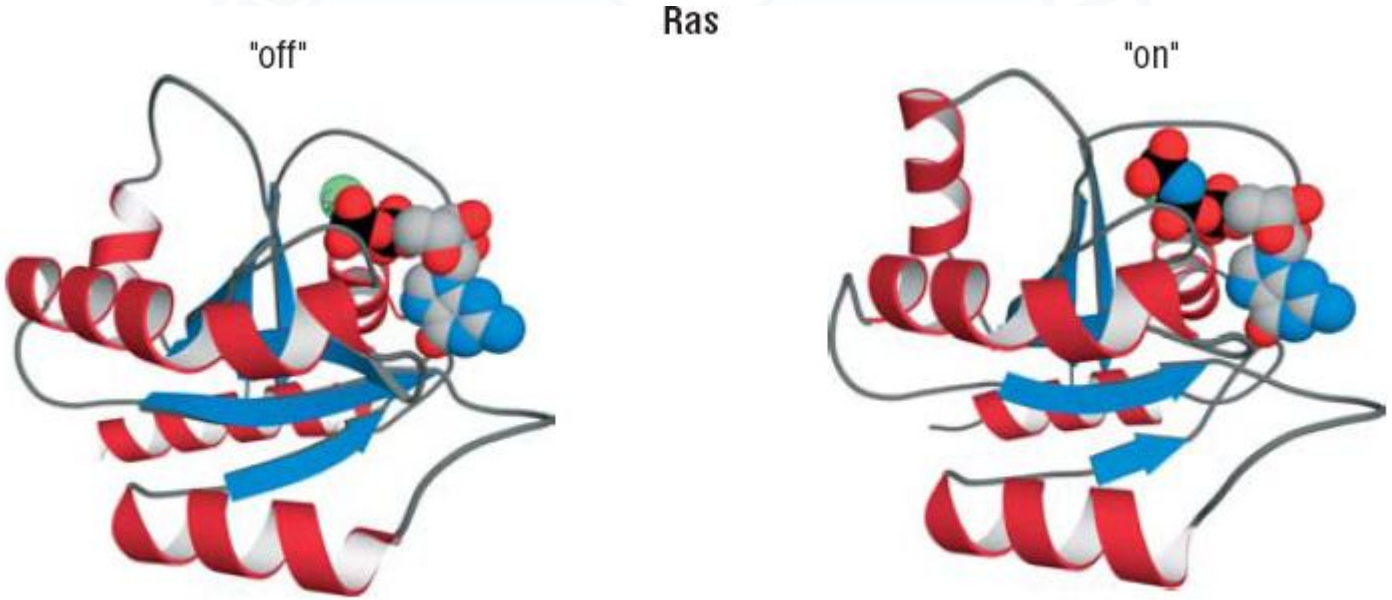
DNA replication is catalyzed by a specific polymerase that copies the genetic material and edits the product for errors in the copy. (PDB 1pbx)



Replication of the AIDS virus HIV depends on the action of a protein-cleaving enzyme called HIV protease. This enzyme is the target for protease-inhibitor drugs (shown in grey). (PDB 1a8k)

Switching

- Signal cell growth



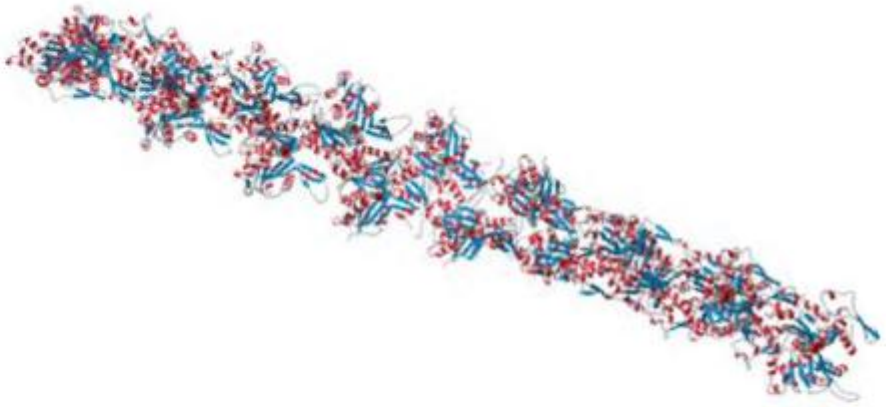
The GDP-bound ("off"; PDB 1pll) state of Ras differs significantly from the GTP-bound ("on"; PDB 121p) state. This difference causes the two states to be recognized by different proteins in signal transduction pathways.

Structural proteins

Silk

Silk derives its strength and flexibility from its structure: it is a giant stack of antiparallel beta sheets. Its strength comes from the covalent and hydrogen bonds within each sheet; the flexibility from the van der Waals interactions that hold the sheets together. (PDB 1slk)

F-actin



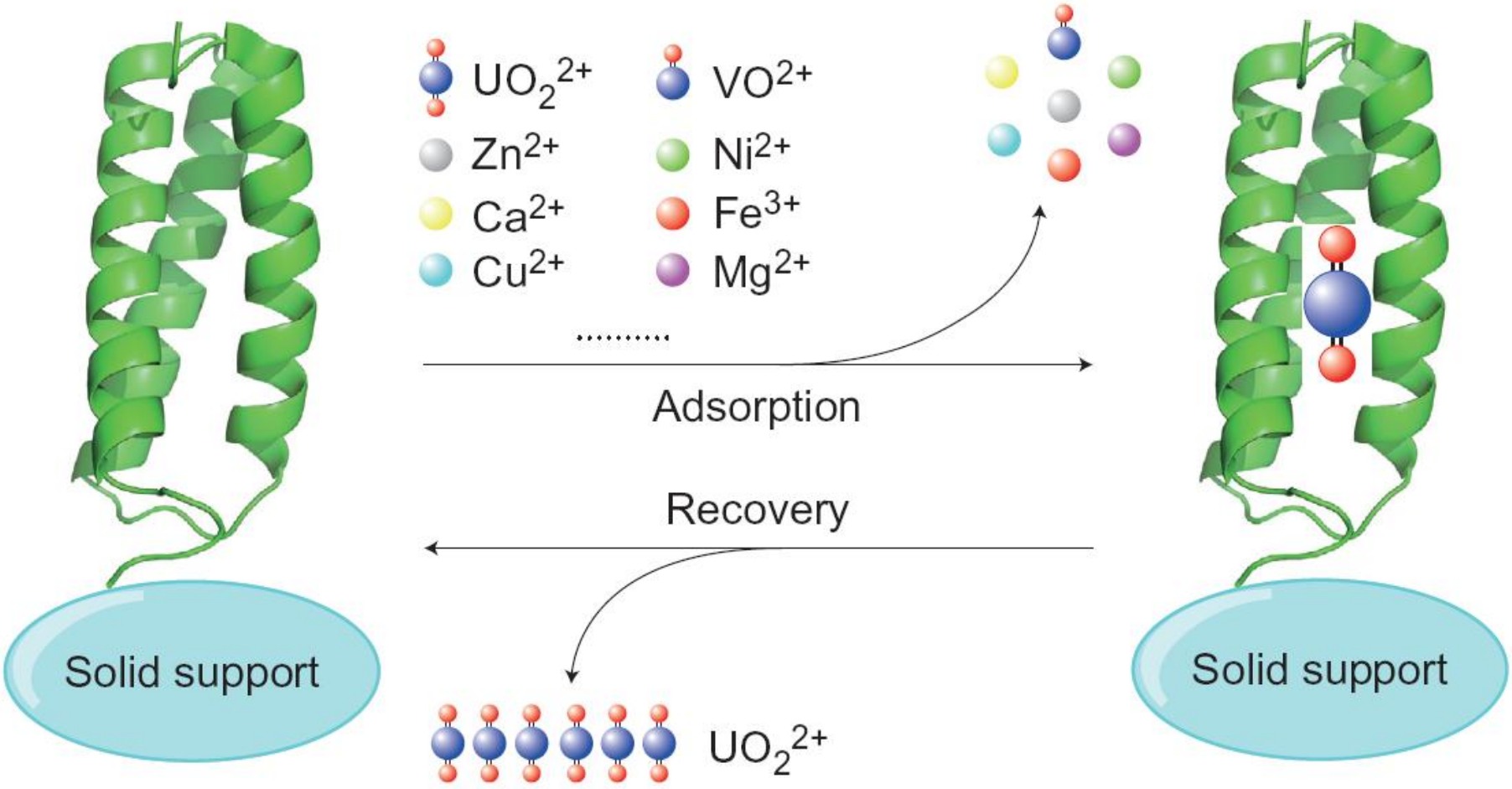
Actin fibers are important for muscle contraction and for the cytoskeleton. They are helical assemblies of actin and actin-associated proteins. (Courtesy of Ken Holmes)

The most versatile cellular molecule

- Proteins are
  - Stable
  - Flexible
  - Versatile in terms of their function

- Protein Function
  - Biochemical function of the molecule in isolation
  - Cellular function as part of larger assembly
  - Resulting Phenotype in the cell or the organism

Uranium Binding Proteins

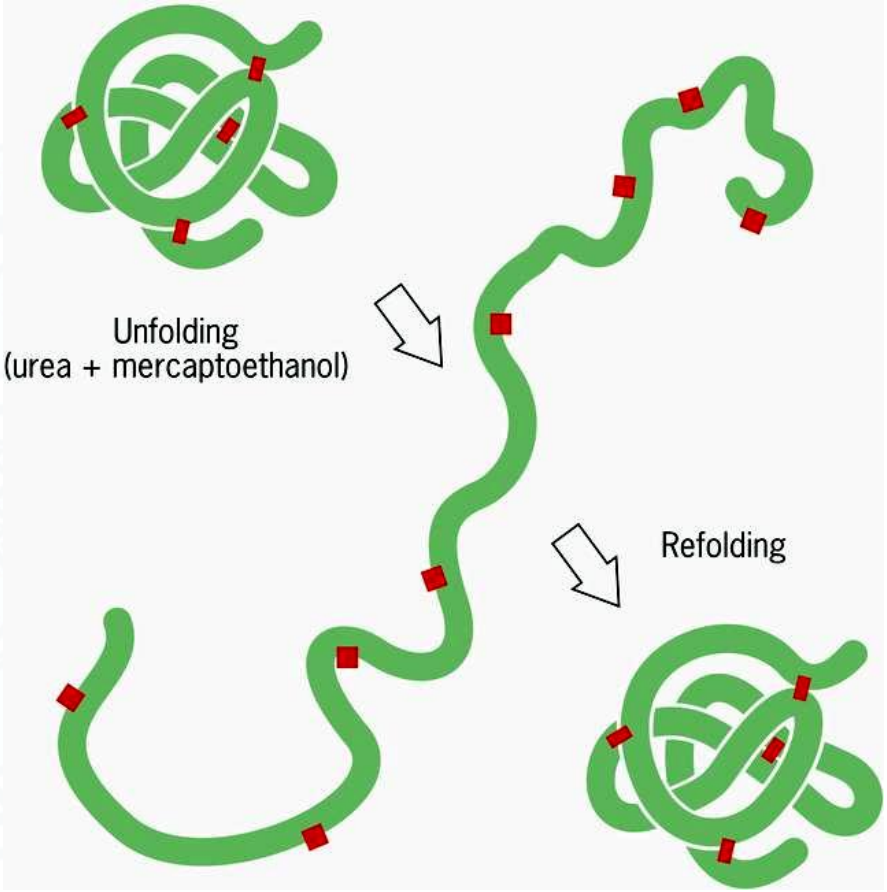


How are proteins versatile?

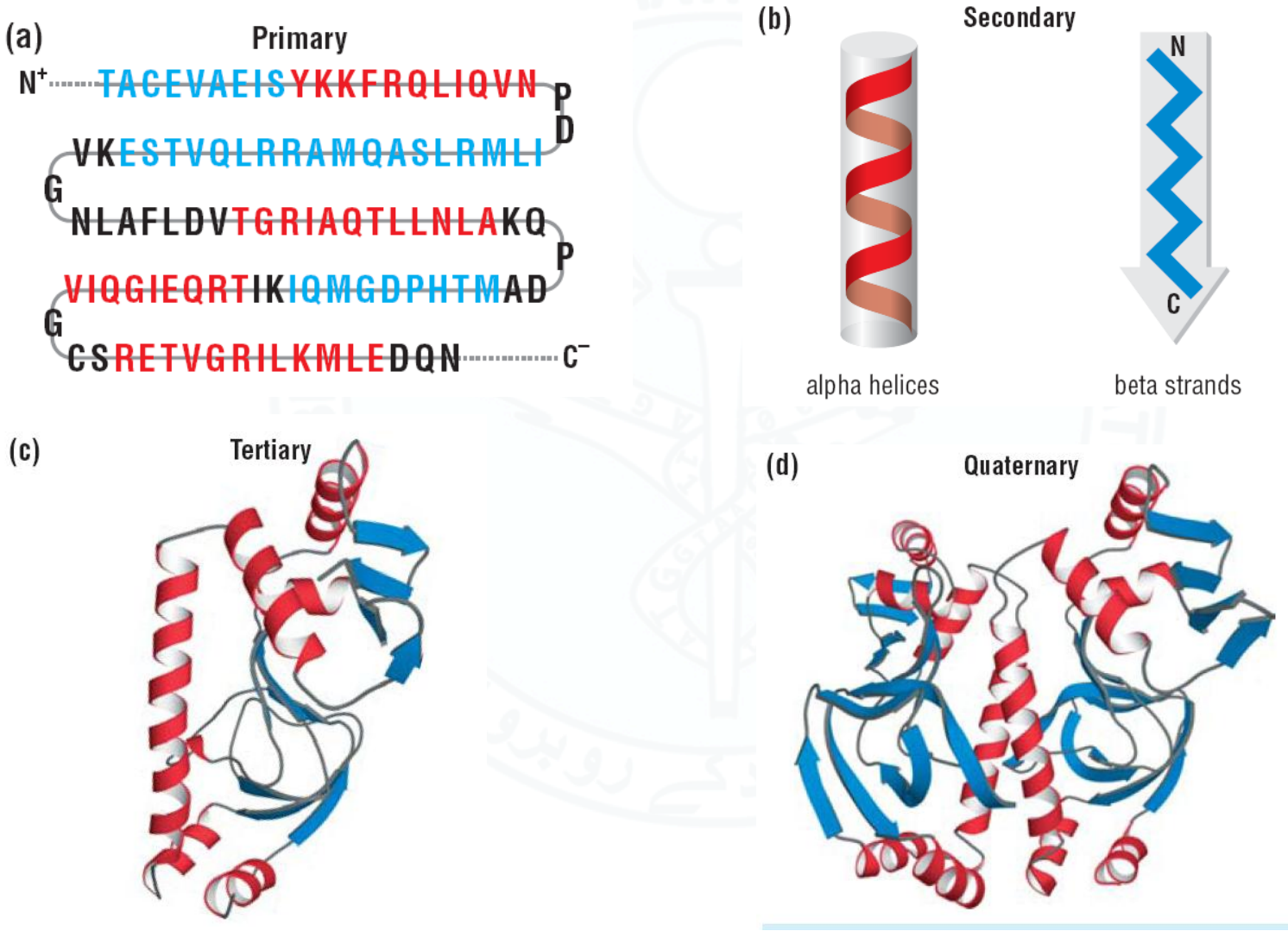
- Because of their structure!
- What determines their structure?
  - Composed of amino acids
  - Amino acids are linked to form polymer
  - Significant hydrogen bonding between amino acids
- Primary ideology
  - Sequence determines Structure determines Function
  - Proven by Anfinsen 1960

Anfinsen Experiment

- Native structure of the protein is determined by the protein's amino acid
- Principles for protein folding
- Native structure is
  - Unique
  - Stable
  - Kinetically accessible



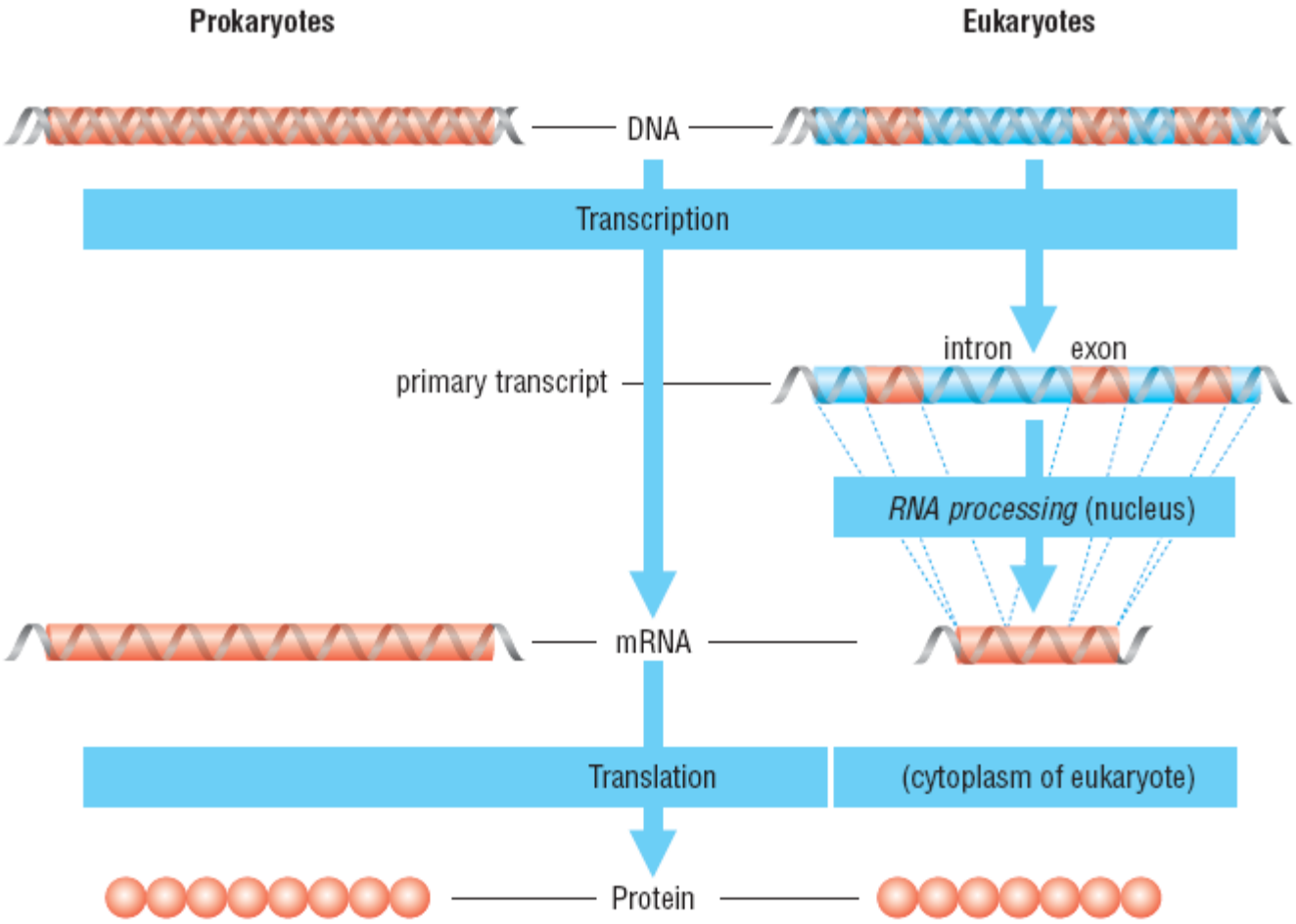
Protein Composition



Understanding protein structure

- To understand protein structure, we will try to answer the following question
  - Why do proteins take any shape at all?
    - Why do proteins fold?
  - Can proteins take on arbitrary shapes?

The Central Dogma

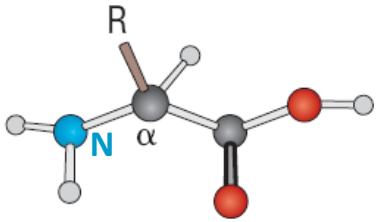


Codons

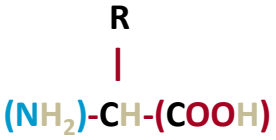
1st position (5' end)	2nd position U	C	A	G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

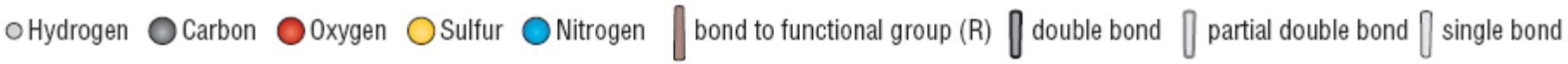
Amino acids	Abbreviations		Codons
Alanine	Ala	A	GCA GCC GCG GCU
Cysteine	Cys	C	UGC UGU
Aspartic acid	Asp	D	GAC GAU
Glutamic acid	Glu	E	GAA GAG
Phenylalanine	Phe	F	UUC UUU
Glycine	Gly	G	GGA GGC GGG GGU
Histidine	His	H	CAC CAU
Isoleucine	Ile	I	AUA AUC AUU
Lysine	Lys	K	AAA AAG
Leucine	Leu	L	UUA UUG CUA CUC CUG CUU
Methionine	Met	M	AUG
Asparagine	Asn	N	AAC AAU
Proline	Pro	P	CCA CCC CCG CCU
Glutamine	Gln	Q	CAA CAG
Arginine	Arg	R	AGA AGG CGA CGC CGG CGU
Serine	Ser	S	AGC AGU UCA UCC UCG UCU
Threonine	Thr	T	ACA ACC ACG ACU
Valine	Val	V	GUA GUC GUG GUU
Tryptophan	Trp	W	UGG
Tyrosine	Tyr	Y	UAC UAU

Amino Acids



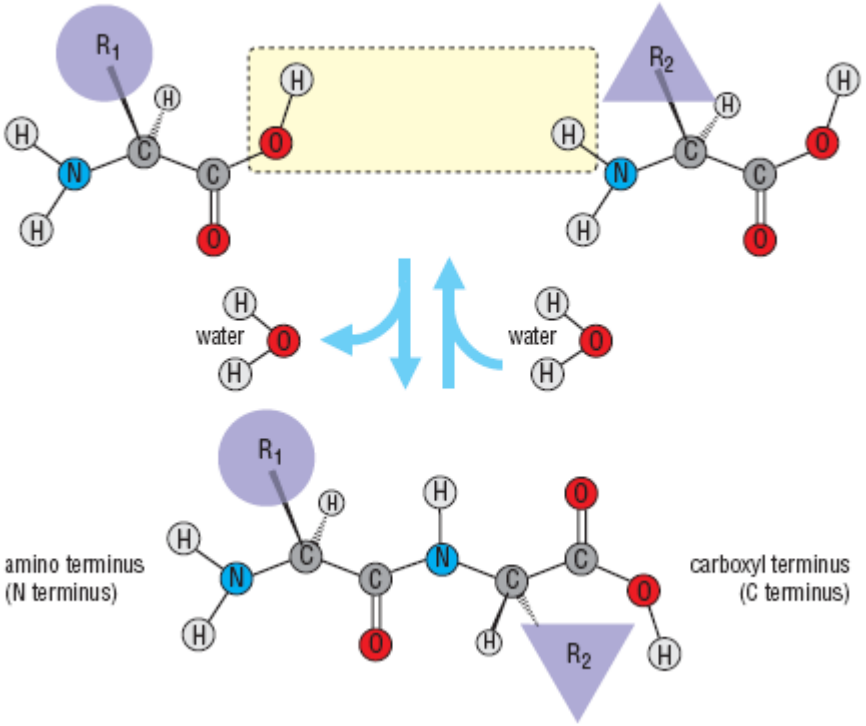
The chemical structure of an amino acid. The backbone is the same for all amino acids and consists of the amino group (NH_2), the alpha carbon and the carboxylic acid group (COOH). Different amino acids are distinguished by their different side chains, R. The neutral form of an amino acid is shown: in solution at pH 7 the amino and carboxylic acid groups ionize, to NH_3^+ and COO^-. Except for glycine, where R=H, amino acids are chiral (that is, they have a left–right asymmetry). The form shown is the L-configuration, which is most common.

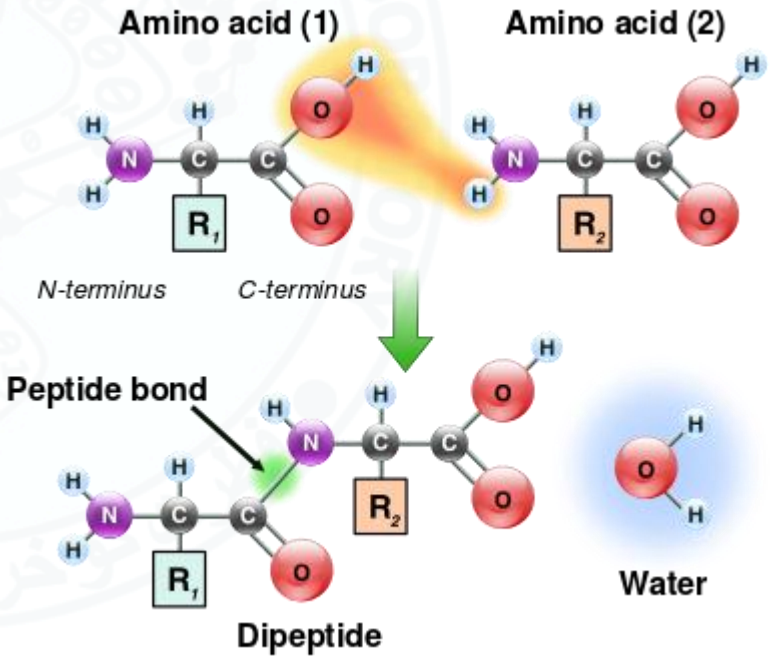




C-alpha is the carbon atom on which the functional group is attached

Peptide bond formation and hydrolysis





Extended peptide chain

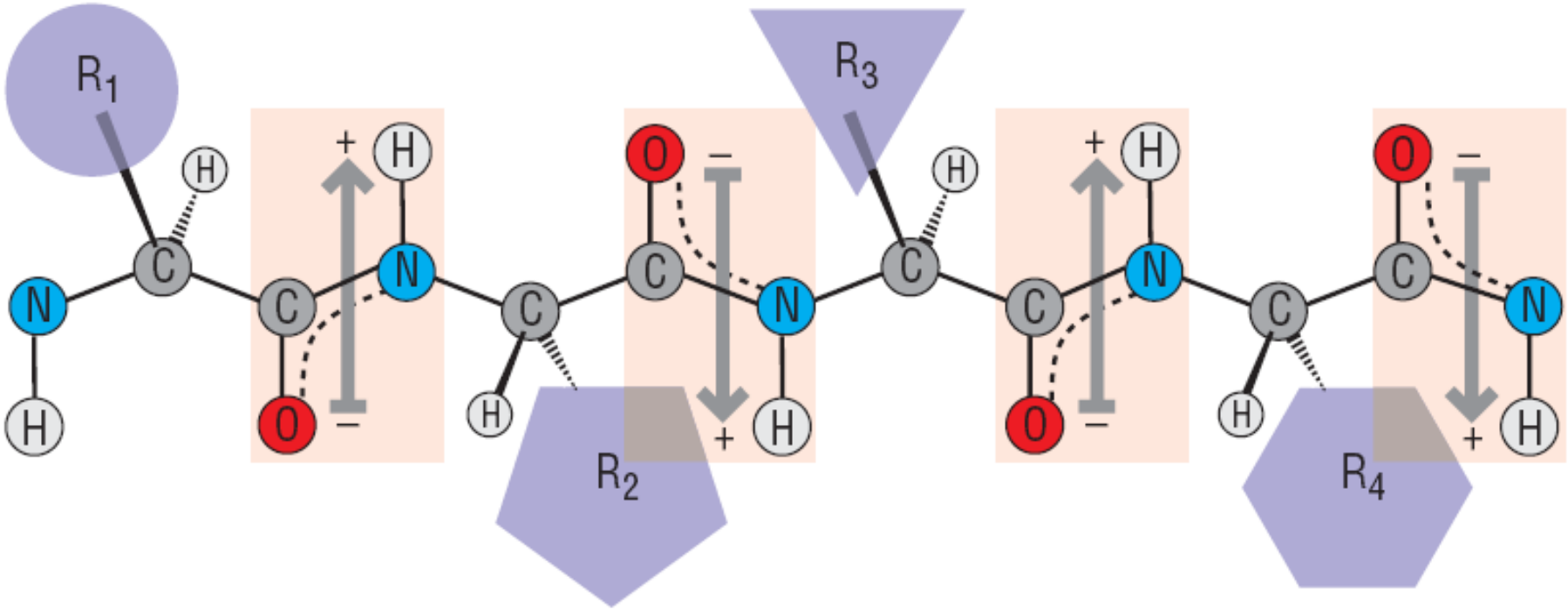
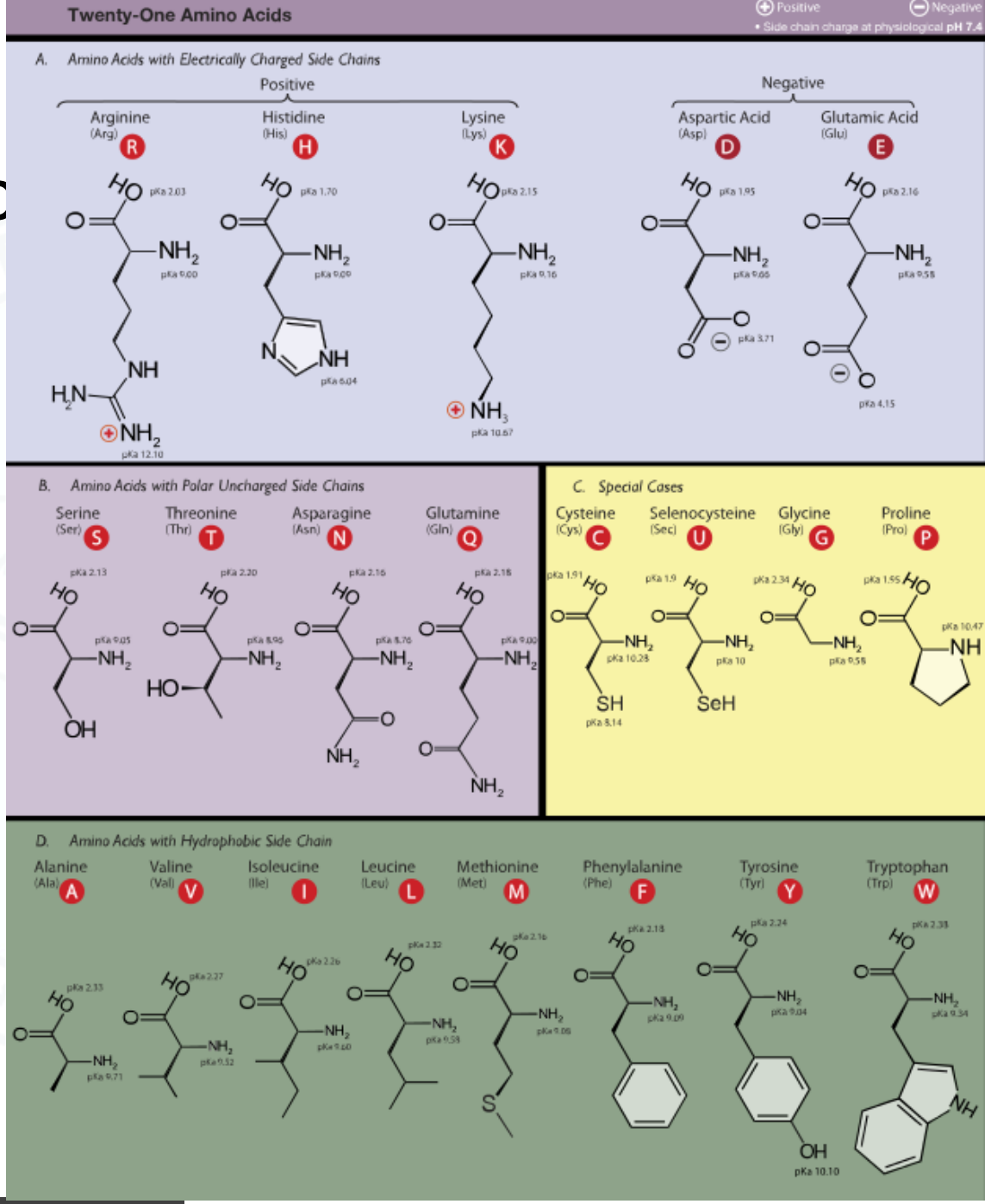


Figure 1-8 Schematic diagram of an extended polypeptide chain The repeating backbone is shown, with schematized representations of the different side chains (R_1, R_2 and so on). Each peptide bond is shown in a shaded box. Also shown are the individual dipole moments (arrows) associated with each bond. The dashed lines indicate the resonance of the peptide bond.

The carbonyl carbon, oxygen and nitrogen lie in the same plane

Types of amino acids



Amino Acid	3-Letter[116]	1-Letter[116]	Side-chain polarity[116]	Side-chain charge (pH 7.4)[116]	Hydropathy index[117]	MW(Weight)[119]
Arginine	Arg	R	Basic polar	positive	−4.5	174
Lysine	Lys	K	Basic polar	positive	−3.9	146
Asparagine	Asn	N	polar	neutral	−3.5	132
Aspartic acid	Asp	D	acidic polar	negative	−3.5	133
Glutamic acid	Glu	E	acidic polar	negative	−3.5	147
Glutamine	Gln	Q	polar	neutral	−3.5	146
Histidine	His	H	Basic polar	positive(10%) neutral(90%)	−3.2	155
Proline	Pro	P	nonpolar	neutral	−1.6	115
Tyrosine	Tyr	Y	polar	neutral	−1.3	181
Tryptophan	Trp	W	nonpolar	neutral	−0.9	204
Serine	Ser	S	polar	neutral	−0.8	105
Threonine	Thr	T	polar	neutral	−0.7	119
Glycine	Gly	G	nonpolar	neutral	−0.4	75
Alanine	Ala	A	nonpolar	neutral	1.8	89
Methionine	Met	M	nonpolar	neutral	1.9	149
Cysteine	Cys	C	nonpolar	neutral	2.5	121
Phenylalanine	Phe	F	nonpolar	neutral	2.8	165
Leucine	Leu	L	nonpolar	neutral	3.8	131
Valine	Val	V	nonpolar	neutral	4.2	117
Isoleucine	Ile	I	nonpolar	neutral	4.5	131

Properties of amino acids

- Different amino acids have different properties
- This allows all types of interactions between these amino acids
  - Charged amino acids would like to be close to oppositely charged ones and away from similarly charged ones
  - Hydrophobic ones don't like water
    - They will like to find other hydrophobic ones so that they can interact with them after displacing the water in between much like how two oil droplets in a cup of water merge to form a bigger oil droplet
      - This effect is called hydrophobic effect
- There are other types of interactions in the protein

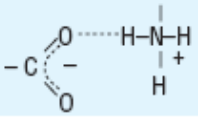
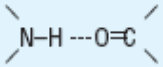
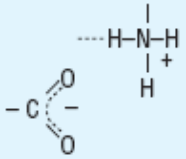
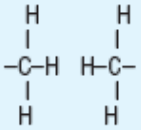
Interactions between amino acids

- Interactions between backbone atoms

- Interactions between side chains

Interactions of amino acids

Chemical Interactions that Stabilize Polypeptides

Interaction	Example	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)
Covalent bond	–Cα–C–	-	1.5 Å	356 kJ/mole (610 kJ/mole for a C=C bond)
Disulfide bond	–Cys–S–S–Cys–	-	2.2 Å	167 kJ/mole
Salt bridge		Donor (here N), and acceptor (here O) atoms <3.5 Å	2.8 Å	12.5–17 kJ/mole; may be as high as 30 kJ/mole for fully or partially buried salt bridges (see text), less if the salt bridge is external
Hydrogen bond		Donor (here N), and acceptor (here O) atoms <3.5 Å	3.0 Å	2–6 kJ/mole in water; 12.5–21 kJ/mole if either donor or acceptor is charged
Long-range electrostatic interaction		Depends on dielectric constant of medium. Screened by water. 1/r dependence	Variable	Depends on distance and environment. Can be very strong in nonpolar region but very weak in water
Van der Waals interaction		Short range. Falls off rapidly beyond 4 Å separation. $1/r^6$ dependence	3.5 Å	4 kJ/mole (4–17 in protein interior) depending on the size of the group (for comparison, the average thermal energy of molecules at room temperature is 2.5 kJ/mole)

Interactions between backbone atoms

- Carbonyl oxygen and Amide Hydrogen form nearby amino acids in the chain can form oxygen bonds
  - This is energetically favorable
  - This gives rise to regular conformations
    - Alpha helices
    - Beta sheets

Alpha helices

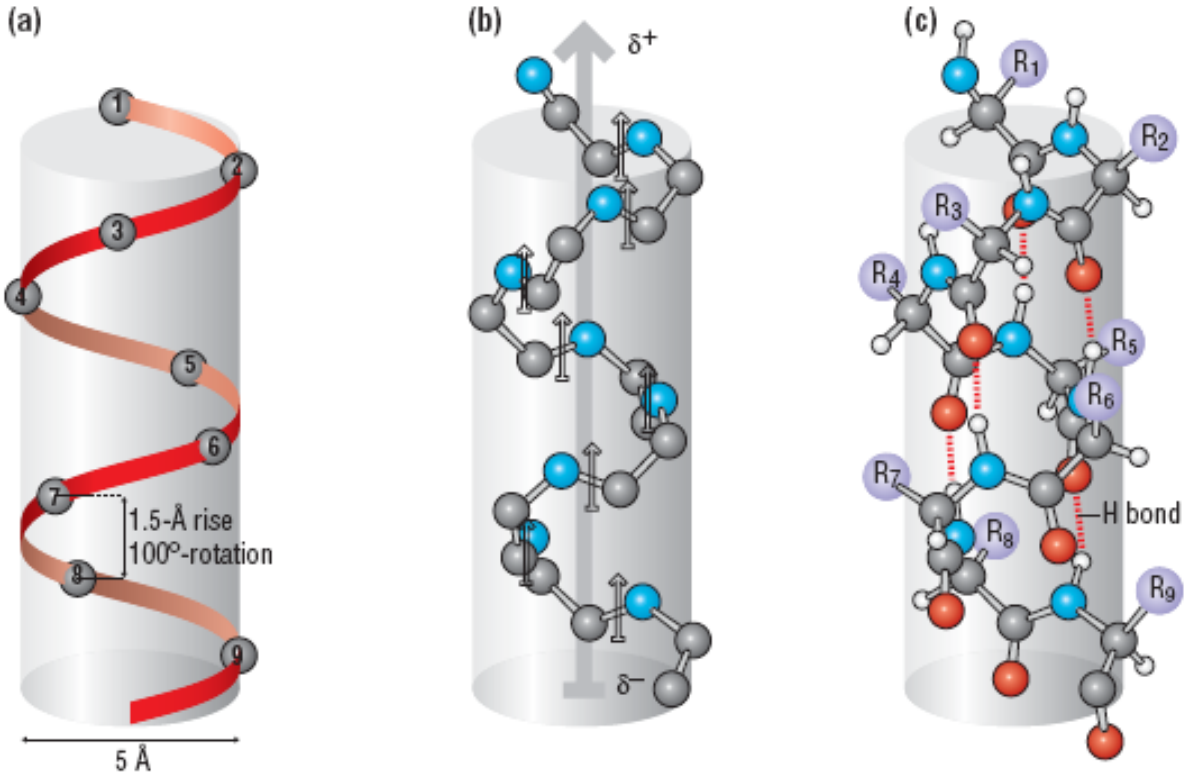
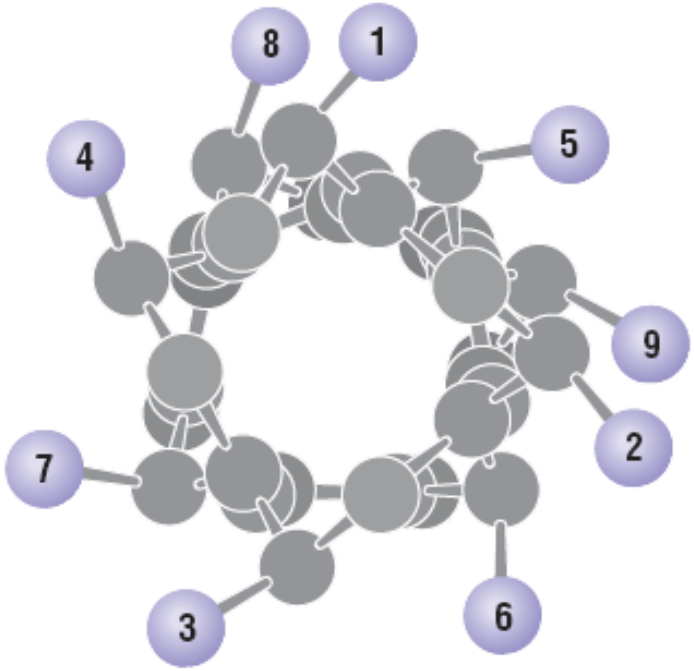


Figure 1-13 The alpha helix The chain path with average helical parameters is indicated showing **(a)** the alpha carbons only, **(b)** the backbone fold with peptide dipoles and **(c)** the full structure with backbone hydrogen bonds in red. All three chains run from top to bottom (that is, the amino-terminal end is at the top). Note that the individual peptide dipoles align to produce a macrodipole with its positive end at the amino-terminal end of the helix. Note also that the amino-terminal end has unsatisfied hydrogen-bond donors (N–H groups) whereas the carboxy-terminal end has unsatisfied hydrogen-bond acceptors (C=O groups). Usually a polar side chain is found at the end of the helix, making hydrogen bonds to these donors and acceptors; such a residue is called a helix cap.

Alpha helices from top

- 3.6 residues per turn of the helix
- Residues 3-4 amino acids apart in the sequence will project from the same face
- In some proteins the surface face will be polar and the other face will be hydrophobic



Beta sheets

- Sometimes a hydrogen bond will be formed between residues that are faraway in sequence from each other to result in beta sheets. These can also be amphipathic.

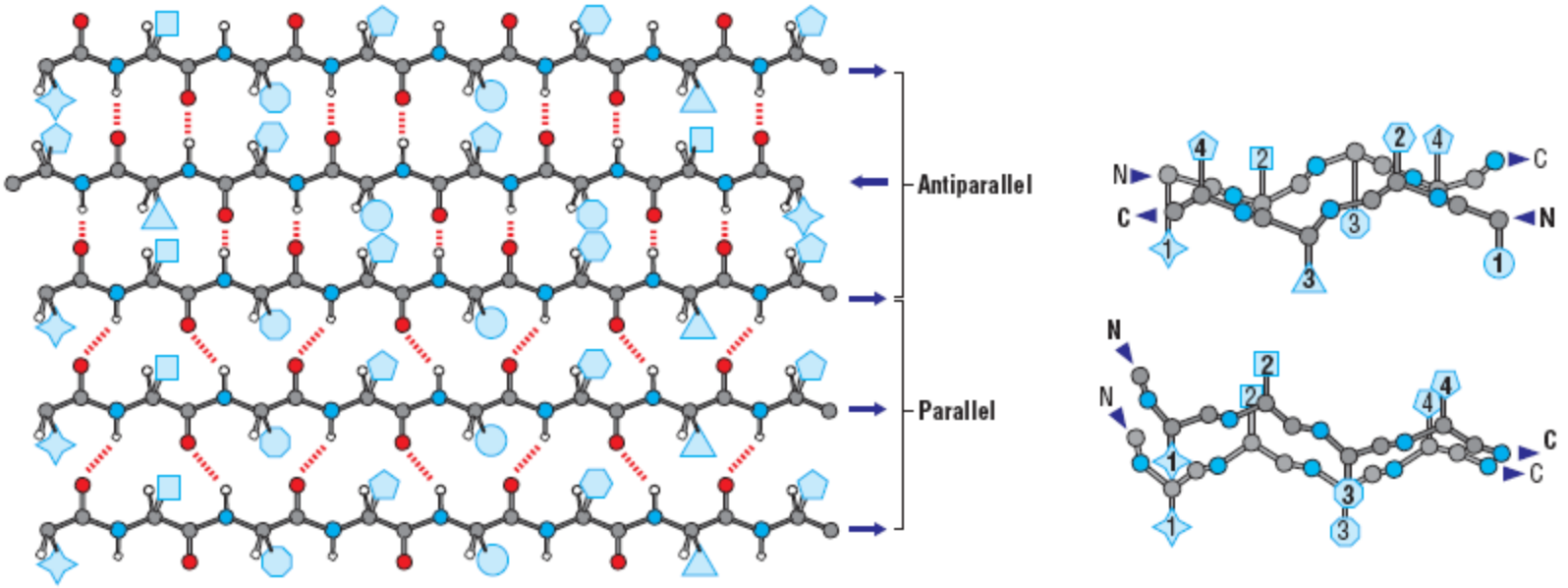
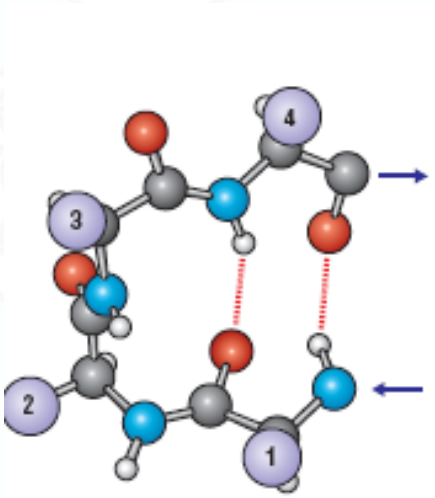


Figure 1-17 The structure of the beta sheet The left figure shows a mixed beta sheet, that is one containing both parallel and antiparallel segments. Note that the hydrogen bonds are more linear in the antiparallel sheet. On the right are edge-on views of antiparallel (top) and parallel sheets (bottom). The corrugated appearance gives rise to the name "pleated sheet" for these elements of secondary structure. Consecutive side chains, indicated here as numbered geometric symbols, point from alternate faces of both types of sheet.

Beta Hairpin



Shape of protein

- The shape of the protein is determined by competition between self-interactions and interactions with water molecules
- Condensation of multiple secondary structure elements leads to a tertiary structure

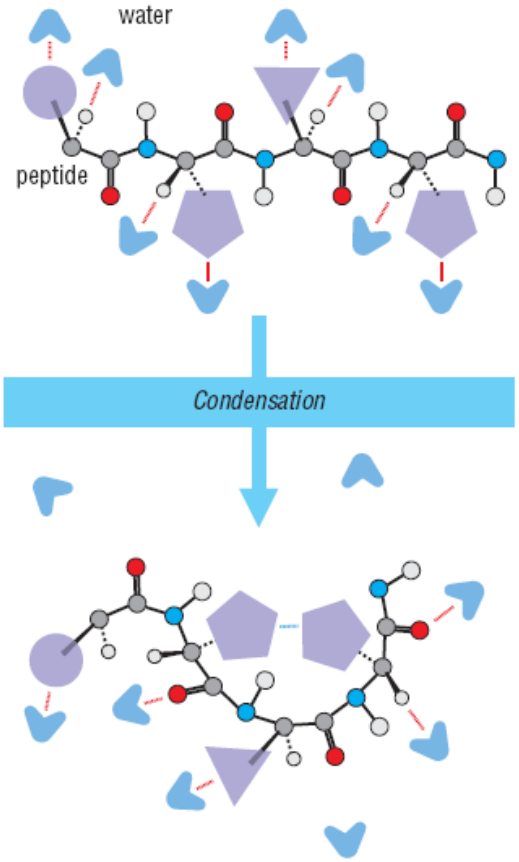
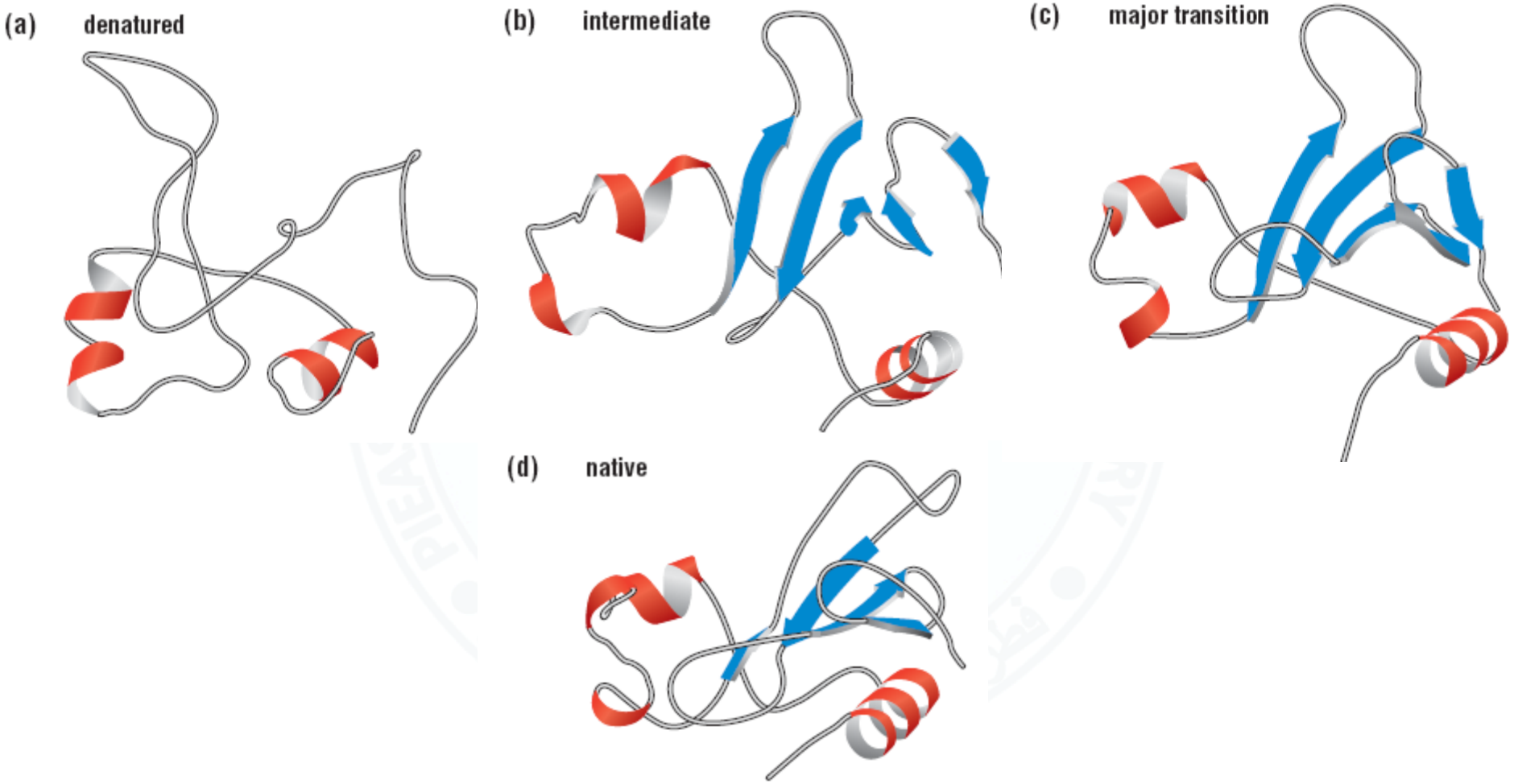


Figure 1-23 Highly simplified schematic representation of the folding of a polypeptide chain in water In the unfolded chain, side-chain and main-chain groups interact primarily with water, even if they are hydrophobic and the interaction is unfavorable. Burying the hydrophobic groups in the interior of a compact structure enables them to interact with each other (blue line), which is favorable, and leaves polar side chains on the surface where they can interact with water (red lines). The polar backbone groups that are buried along with the hydrophobic side chains must make hydrogen bonds to each other (not shown), as bulk water is no longer available.

Protein folding

- Although the long chain of amino acids is held together as a chain by strong covalent bonds, this chain folds into a three dimensional structure which is stabilized by weak non-covalent interactions
- Gulliver in Lilliput analogy



Protein folding

Protein Folding Video

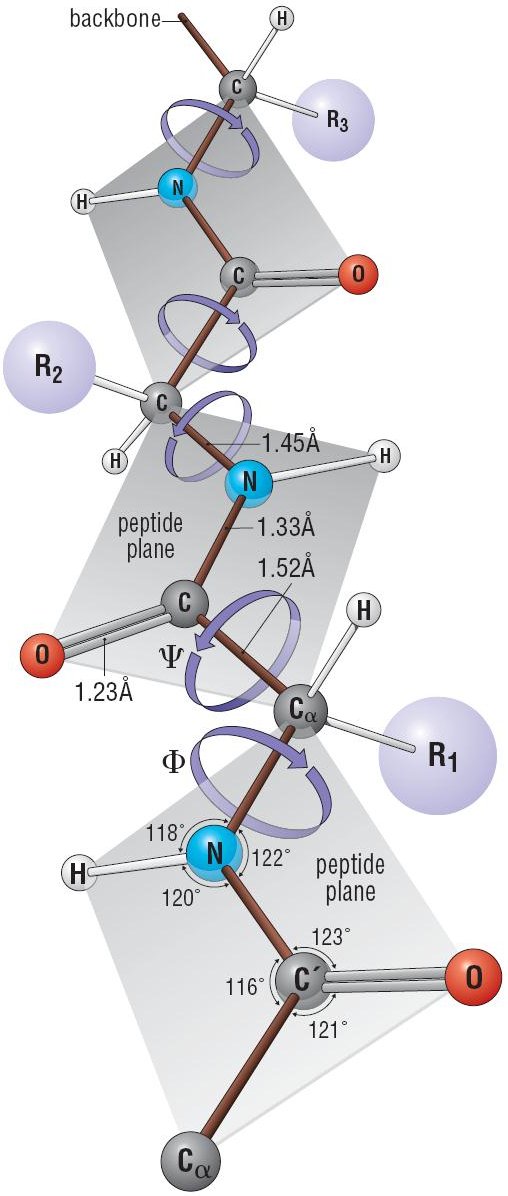
The energetics of protein folding

- A protein will spontaneously fold only if the folded state is energetically more feasible than the unfolded state
  - Entropy
  - Enthalpy

- Free energy
  - Non-bonding interactions
  - Entropy

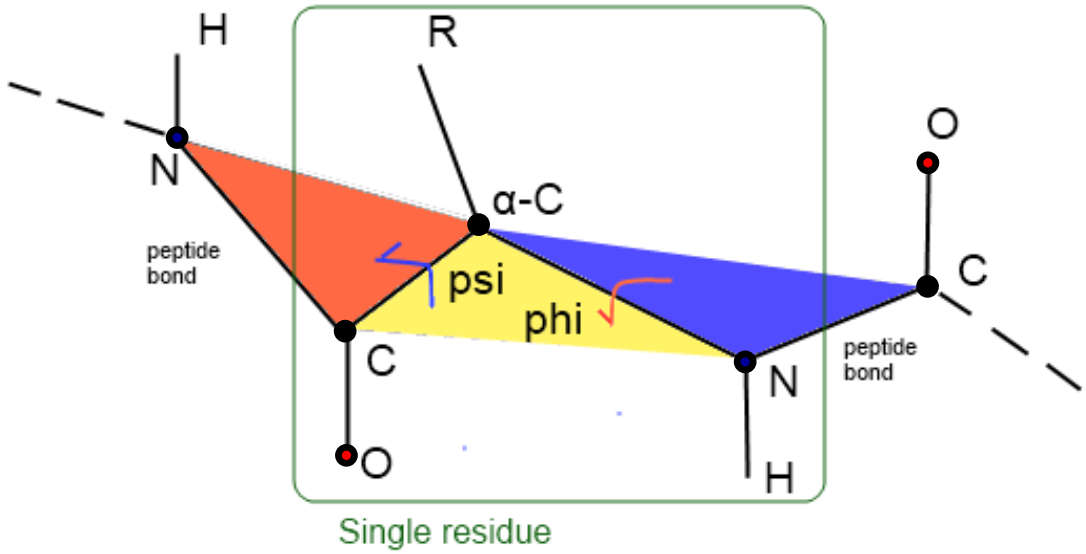
Torsion angles

- N, C, O lie in the same plane
  - Rigid Structure
- N–Ca and Ca–C bonds are rotatable and free rotation is allowed
  - Flexible
  - Can rotate as long as no steric clashes arise (i.e., no side chains collide with each other)
- Phi torsion angle
  - The angle of the N–Ca bond to the adjacent peptide bond
- Psi torsion angle
  - the angle of the C–C_a bond to the adjacent peptide bond



Torsion angles

Excellent tool: http://proteopedia.org/wiki/index.php/Psi_and_Phi_Angles



Torsion angles

- A protein has alternating rigid and rotatable flexible covalent bonds
  - This combination greatly reduces the number of possible conformations a protein can have
  - Certain torsion angles will cause clashes
  - Ranges of torsion angles are also dependent upon the type of side chains

Ramachandran plot

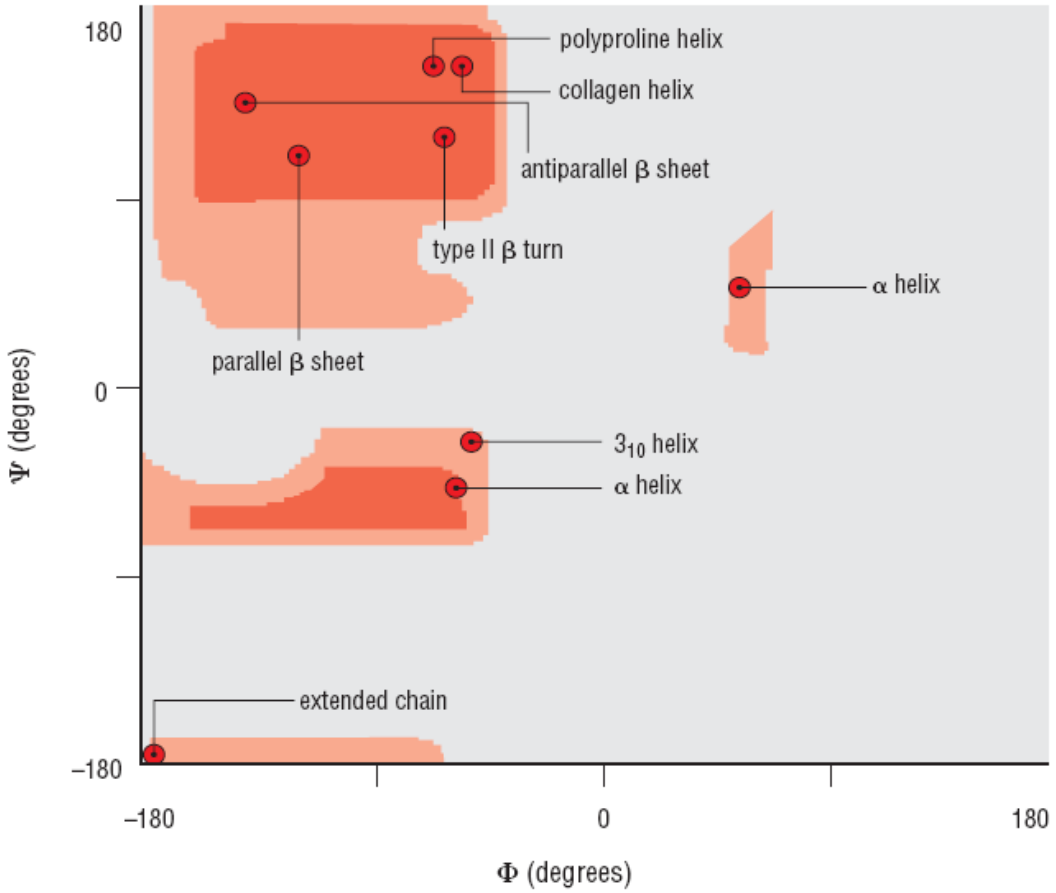


Figure 1-11 Ramachandran plot Shown in red are those combinations of the backbone torsion angles phi and psi (see Figure 1-9) that are "allowed" because they do not result in steric interference. The pink regions are allowed if some relaxation of steric hindrance is permitted. Common protein secondary structure elements are marked at the positions of their average phi, psi values. The isolated pink alpha-helical region on the right is actually for a left-handed helix, which is only rarely observed in short segments in proteins. The zero values of phi and psi are defined as the *trans* configuration.

Fold-it

- http://fold.it/portal/info/about

- Christopher B Eiben, Justin B Siegel, Jacob B Bale, Seth Cooper, Firas Khatib, Betty W Shen, Foldit Players, Barry L Stoddard, Zoran Popović, David Baker (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nature Biotechnology, 2012. [PDF] [PROJECT]
- Firas Khatib, Seth Cooper, Michael D. Tyka, Kefan Xu, Ilya Makedon, David Baker, Foldit players (2011). Algorithm discovery by protein folding game players. Proceedings of the National Academy of Sciences of the United States of America vol. 108 no. 47 18949-18953, 2011. [PDF] [PROJECT]
- Firas Khatib, Frank DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Mariusz Jaskolski, David Baker (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nature Structural and Molecular Biology 18, 1175-1177, 2011. [PDF] [PROJECT]
- Seth Cooper, Firas Khatib, Ilya Makedon, Hao Lu, Janos Barbero, David Baker, James Fogarty, Zoran Popović, Foldit players (2011). Analysis of Social Gameplay Macros in the Foldit Cookbook. Foundations of Digital Games, 2011. [PDF] [PROJECT]
- Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, Foldit players (2010). Predicting protein structures with a multiplayer online game. Nature 446 p. 756-760, 05 August 2010. [PDF] [PROJECT]
- Seth Cooper, Adrien Treuille, Janos Barbero, Andrew Leaver-Fay, Kathleen Tuite, Firas Khatib, Alex Cho Snyder, Michael Beenen, David Salesin, David Baker, Zoran Popović, >57,000 Foldit players (2010). The challenge of designing scientific discovery games. International Conference on the Foundations of Digital Games, 2010. [PDF] [PROJECT]
- Other
  Miroslaw Gilski, Maciej Kazmierczyk, Szymon Krzywda, Helena Zábranská, Seth Cooper, Zoran Popović, Firas Khatib, Frank DiMaio, James Thompson, David Baker, Iva Pichová, Mariusz Jaskolskia (2011). High-resolution structure of a retroviral protease folded as a monomer. Acta Crystallographica D67, 907-914, 2011. [PDF] [PROJECT]

How big is the protein conformation space?

- Assume that:
  - each of the torsion angles can take 2 or 3 values
    - In practice these angles are continuous
  - Side chains are fixed
  - We have a protein of length L

- Thus the number of possible "conformations" is:
  - 2L or 3L
  - If L = 100, this means 10^{30} or 10^{47} conformations

Levinthal's paradox

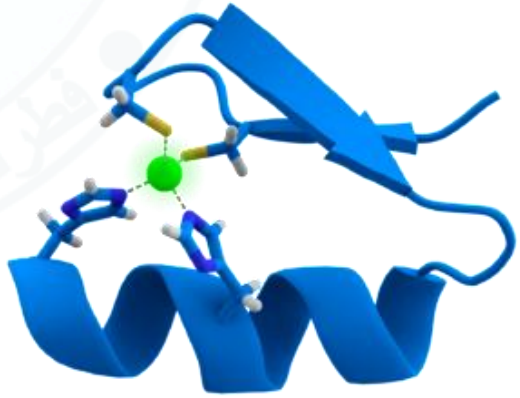
- The number of possible folding conformations of a protein is very large
- But the protein folds in microseconds or less

- How?

- Formation of local interactions

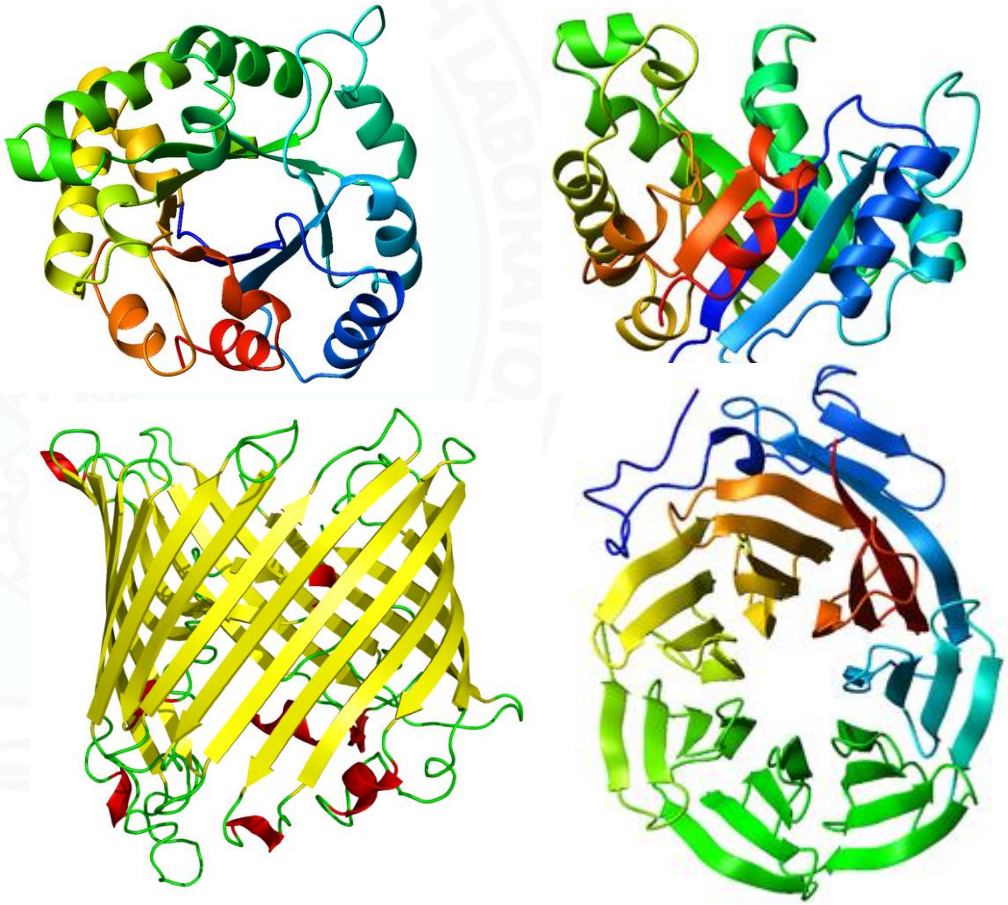
Supersecondary structures

- Motif
  - A supersecondary structure that appears in a variety of molecules
  - Can appear in proteins with dissimilar functions
  - May not be associated with a sequence motifs
- Examples
  - Helix-Turn-Helix
  - Zinc Finger
  - β-α-β
  - β-hairpin
- Databases
  - PROSITE

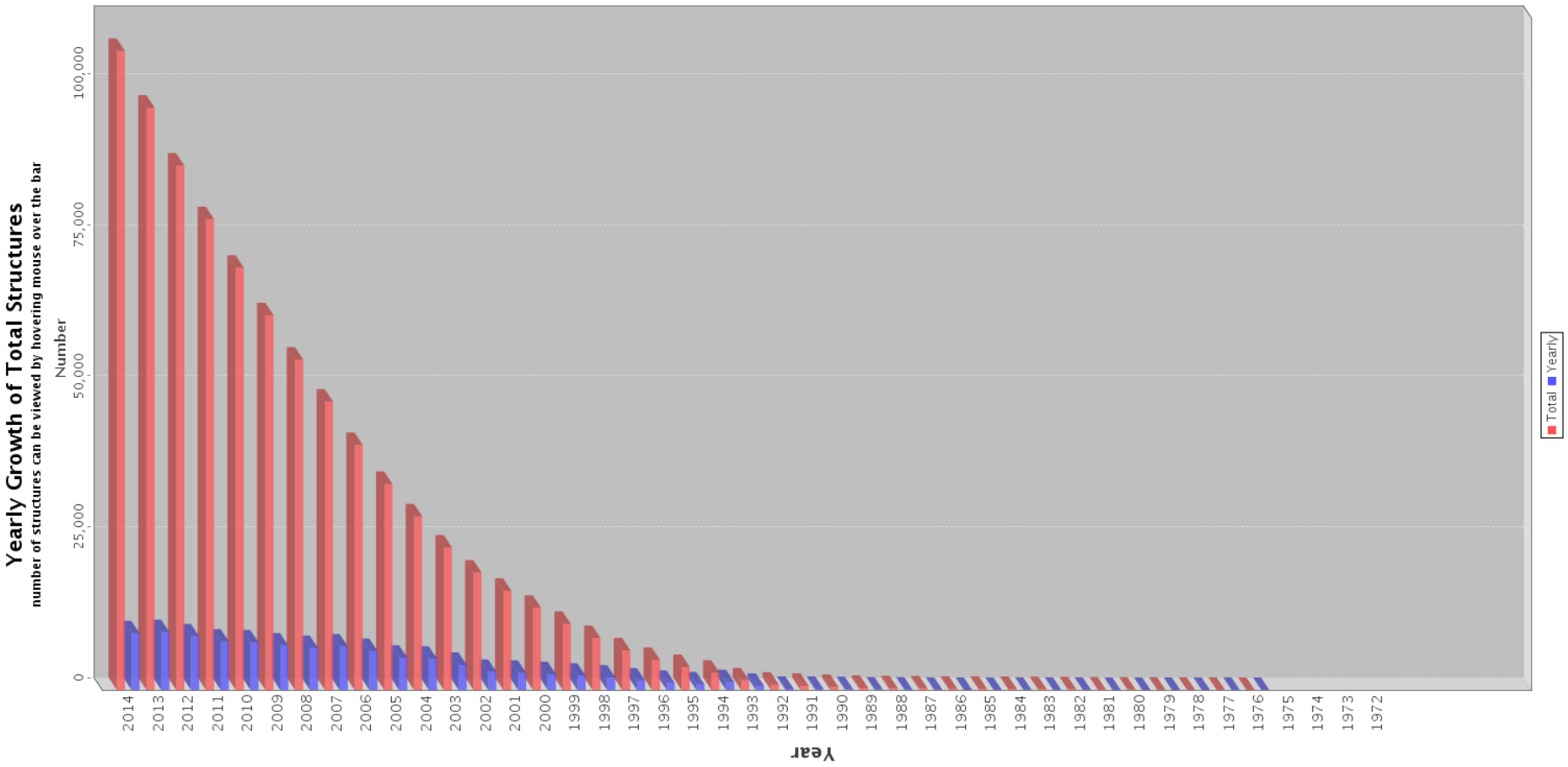


Super-secondary structures

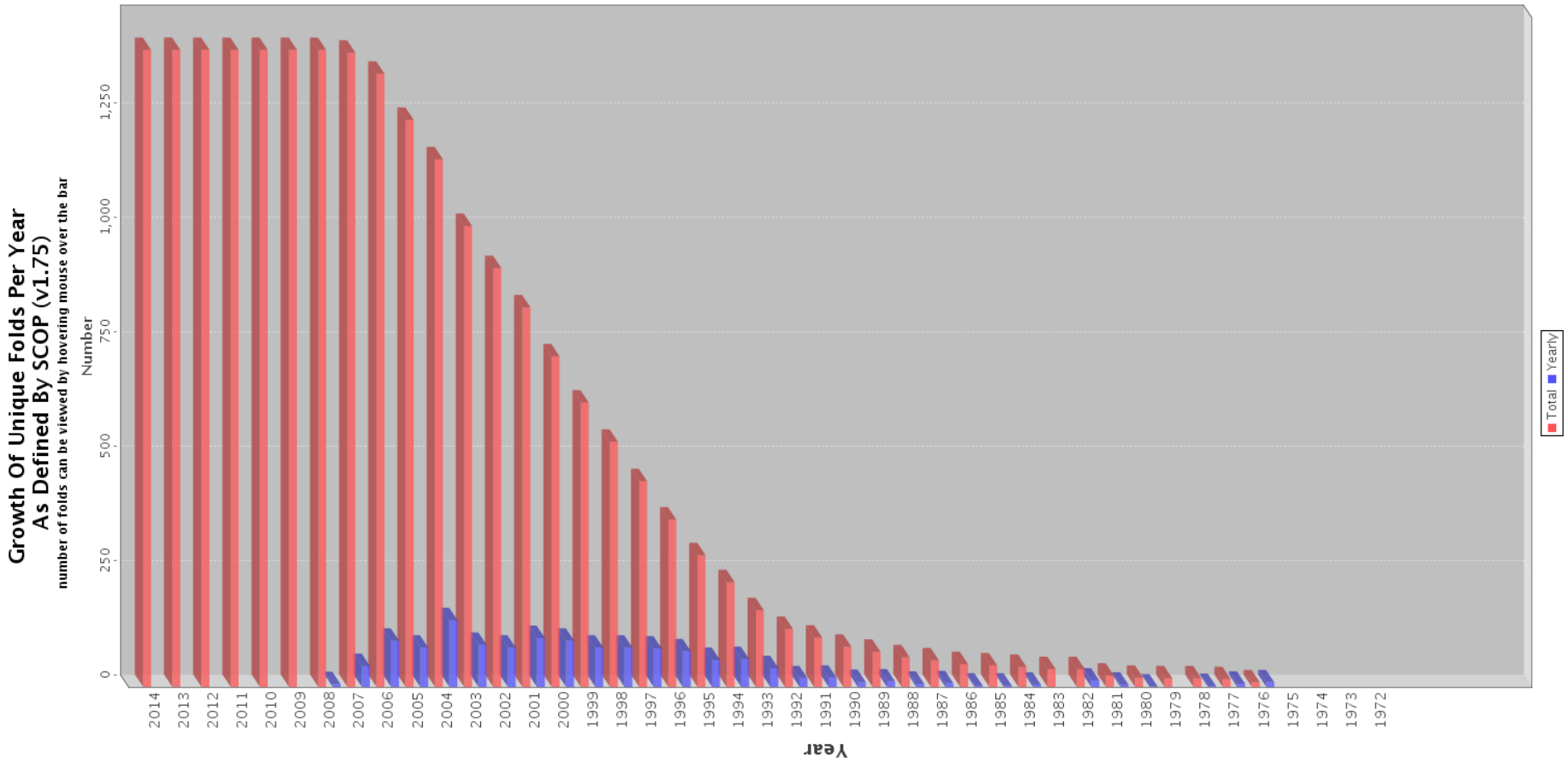
- Fold
  - A protein fold is defined by the way the secondary structure elements of the structure are arranged relative to each other in space.
  - Proteins can fold in only so many ways
- Examples
  - TIM Barrel Fold (PDB: 8TIM)
  - Beta-Barrel
  - Beta Propellor (PDB: 1ERJ)



Growth in PDB

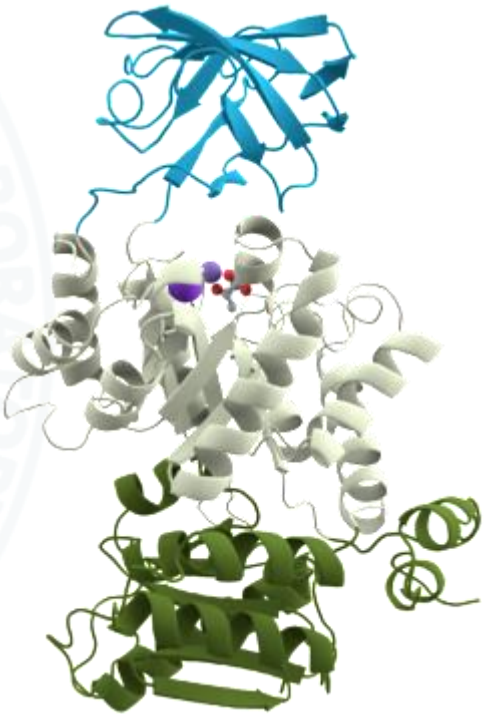


Growth of unique folds in PDB



Supersecondary structures

- Chain
  - A protein can have multiple independent chains of amino acids
- Domains
  - A protein domain is a conserved part of a given protein sequence and (tertiary) structure that can evolve, function, and exist independently of the rest of the protein chain.
  - Two proteins having the same domain are said to belong to the same family
- SCOP database



Pyruvate kinase, a protein with three domains (PDB 1PKN

Computational Problems in Protein Biology

- Visualization
- Databases
- Alignments
  - Sequence based: Local, global, homology, phylogenetics, co-evolution, motif-finding
  - Structure based alignments
- Structure prediction
  - Secondary and tertiary, contact maps, fold, domain localization, family identification
  - Prediction of structures of larger protein assemblies
  - Molecular dynamics, normal mode analysis
  - Protein folding, folding-pathways
- Properties: Surface area, trans-membrane portions, signal peptides, functional sites, post-translational modifications
- Cellular Function prediction, phenotype prediction
- Cellular localization
- Biological process prediction
- Mutation analysis
- Protein interactions: Binding, binding sites, protein, nucleic acid, water, ions, ligands
- Motif Finding
- Drug discovery and design
- **Protein design and engineering**

PyMOL Visualization

- Load 1CLL
- Visualize 1CLL
  - Show the structure as a cartoon
  - Hide lines
  - Get rid of het-atoms
  - Show the Calcium ions as spheres
  - Show lines
  - Find polar contacts of the calcium ions
  - Select all atoms in an alpha helix and show the hydrogen bonds between them
  - Do the same for a beta sheet

Visualization types

- Types:
  - Lines
  - Sticks
  - Cartoon
  - Space-fill
  - Surface
- How are secodnary structures assigned?
  - DSSP
  - STRIDE (http://en.wikipedia.org/wiki/STRIDE_%28protein%29)

PyMOL Visualization

- Load 1CFD & 1DMO
  - Is it showing hydrogens?
- Align the two structures

http://pymol.sourceforge.net/newman/user/toc.html

Protein databases

- Genbank: All DNA sequences
- Uniprot

http://en.wikipedia.org/wiki/UniProt

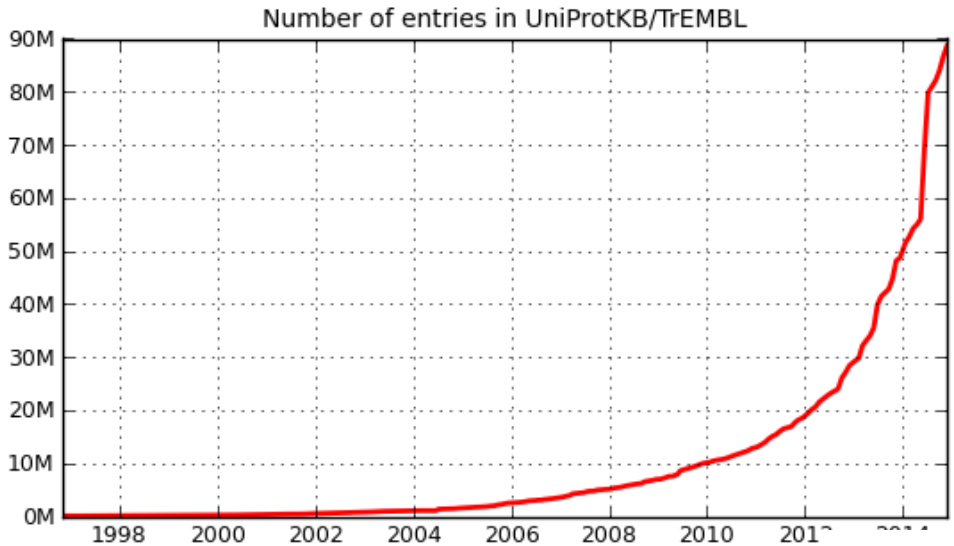
- SCOP: Structural classification of proteins
  - manual classification of protein structural domains based on similarities of their structures and amino acid sequences.

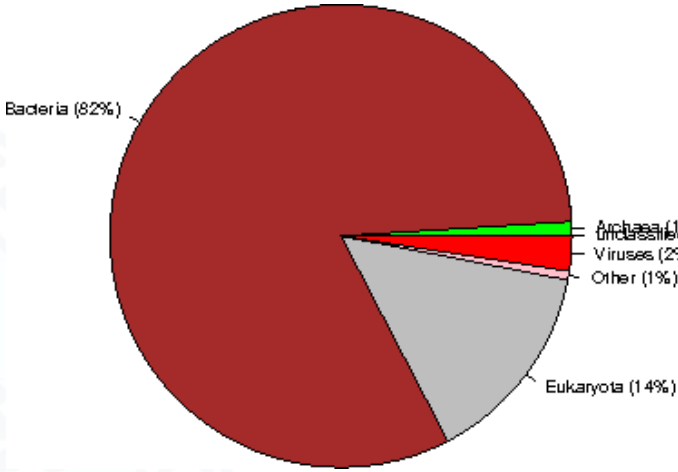
http://en.wikipedia.org/wiki/Structural_Classification_of_Proteins_database

- Pfam
  - Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models.

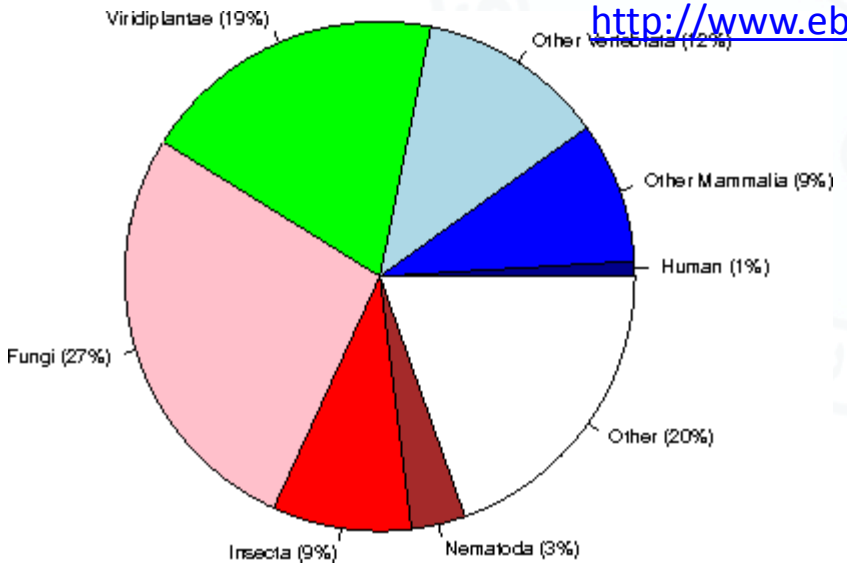
http://en.wikipedia.org/wiki/Pfam

Uniprot Statistics

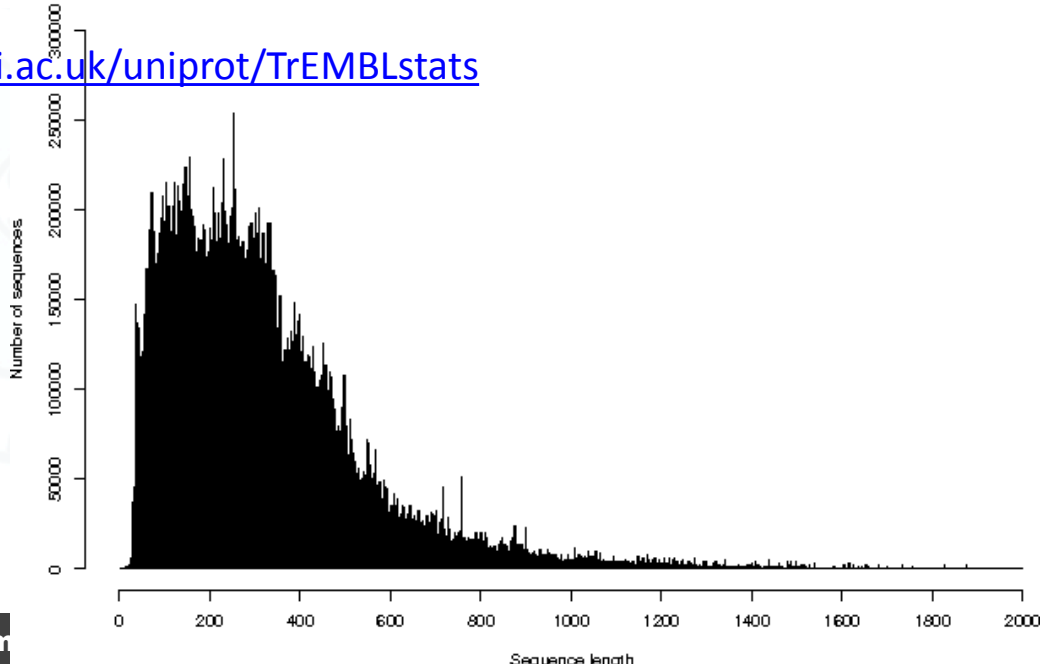




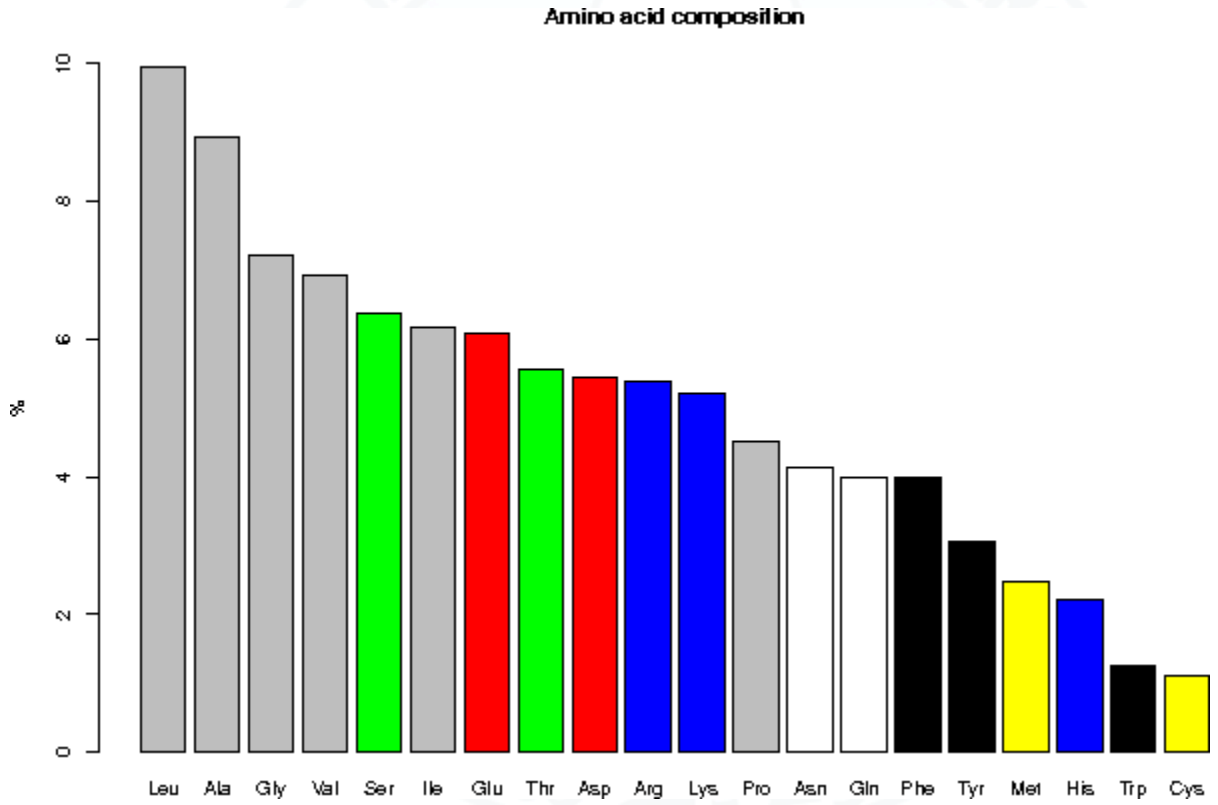
Length distribution of the sequences



http://www.ebi.ac.uk/uniprot/TrEMBLstats



Usage of amino acids



gray = aliphatic, red = acidic, green = small hydroxy, blue = basic, black = aromatic, white = amide, yellow = sulfur

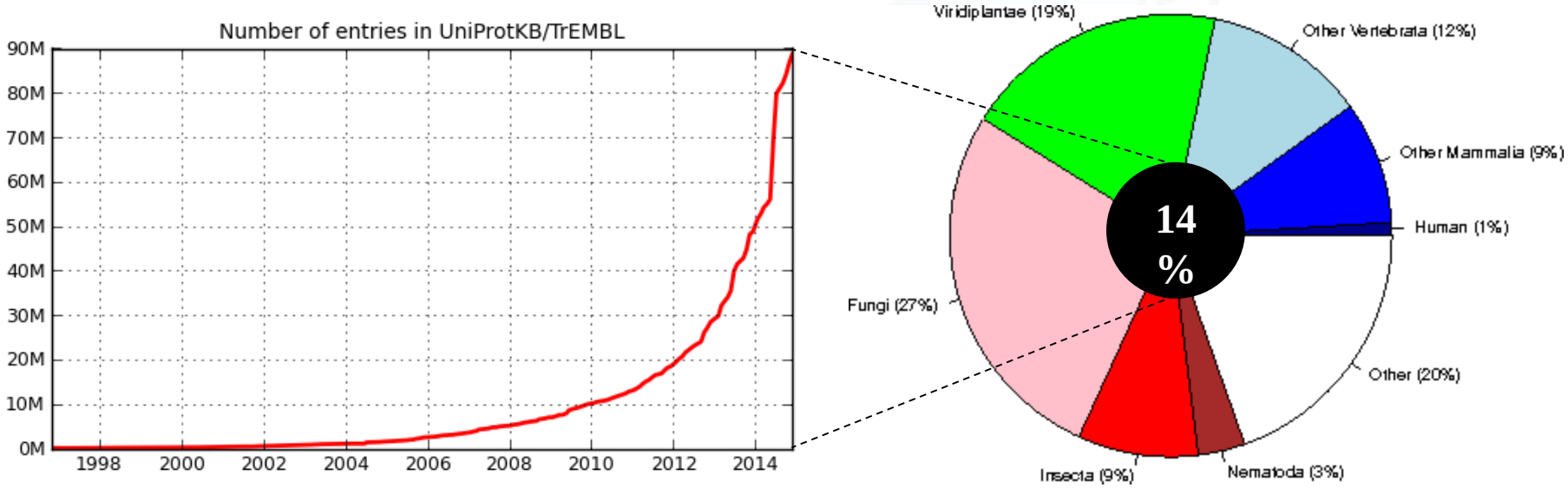
Uniprot structure

- "The primary mission of Universal Protein Resource (UniProt) is to support biological research by maintaining a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community"
- Each protein has a uniprot accession number

"Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D214–D219, Jan. 2011.

Protein databases: Sequence

- Sequence
  - UniProt
    - UniProt Knowledgebase (UniProtKB): Expertly curated annotations
    - The UniProt Archive (UniParc): Reflects history of all proteins
    - UniProt Reference Clusters (UniRef): Closely related proteins are merged
    - UniProt Metagenomic and Environmental Sequences Database (UniMES)



"Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D214–D219, Jan. 2011.

Uniprot structure

- UniProt Knowledgebase (UniProtKB)
  - Expertly curated database,
  - Central access point for integrated protein information
  - Cross-references to multiple sources.
- The UniProt Archive (UniParc)
  - is a comprehensive sequence repository
  - Reflects the history of all protein sequences
- UniProt Reference Clusters (UniRef)
  - Merge closely related sequences based on sequence identity to facilitate sequence similarity searches
  - Using CD-HIT
- UniProt Metagenomic and Environmental Sequences Database (UniMES)
  - Caters for the expanding area of metagenomic data.

Uniprot structure

- UniProtKB/Swiss-Prot
  - 542,782 sequence entries in 2014_03 release
  - Reviewed, manually annotated entries
  - Annotations include
    - Protein and gene names
    - Function
    - Enzyme-specific information such as catalytic activity, cofactors and catalytic residues
    - Subcellular location
    - Protein-protein interactions
    - Pattern of expression
    - Locations and roles of significant domains and sites
    - Ion-, substrate- and cofactor-binding sites
    - Protein variant forms produced by natural genetic variation, RNA editing, alternative splicing, proteolytic processing, and post-translational modification
- UniProtKB/TrEMBL
  - Automatic annotation

Example

- http://www.uniprot.org/uniprot/P62158

- Breuza, Lionel, Sylvain Poux, Anne Estreicher, Maria Livia Famiglietti, Michele Magrane, Michael Tognolli, Alan Bridge, Delphine Baratin, Nicole Redaschi, and The UniProt Consortium. 2016. “The UniProtKB Guide to the Human Proteome.” *Database* 2016 (January): bav120. doi:10.1093/database/bav120.

Protein Databank

- http://proteopedia.org/wiki/index.php/Believe_It_or_Not
- http://en.wikipedia.org/wiki/Protein_Data_Bank

Experimental Method	Proteins	Nucleic Acids	Protein/Nucleic Acid complexes	Other	Total
X-ray diffraction	85406	1558	4547	5	91516
NMR	9281	1092	218	7	10598
Electron microscopy	579	67	190	0	836
Hybrid	63	3	2	1	69
Other	157	4	6	13	180
Total:	95486	2724	4963	26	**103199**

Tools for PDB

Database name	Database description
DSSP	Secondary structure of proteins http://swift.cmbi.ru.nl/gv/dssp/
HSSP	Multiple sequence alignments of UniProtKB against PDB files http://swift.cmbi.ru.nl/gv/hssp/
PDBREPORT	Reports about errors and anomalies in macromolecules http://swift.cmbi.ru.nl/gv/pdbreport/
PDB_REDO	Re-refined PDB files solved by X-ray crystallography http://www.cmbi.ru.nl/pdb_redo/
PDBFINDER	Searchable summaries of PDB file information ftp://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder/
PDBFINDER2	As PDBFINDER, but with much extra information added ftp://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder2/
PDB_SELECT	Quality-sorted culled lists of protein chains in the PDB http://swift.cmbi.ru.nl/gv/select/
WHY_NOT	Explanation why entries in other databases cannot exist http://www.cmbi.ru.nl/WHY_NOT/

R. P. Joosten, T. A. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend, "A series of PDB related databases for everyday needs," *Nucleic Acids Res.*, vol. 39, no. suppl 1, pp. D411–D419, Jan. 2011.

Other tools

- OCA
- FirstGlance
- PDBSum
- ccPDB

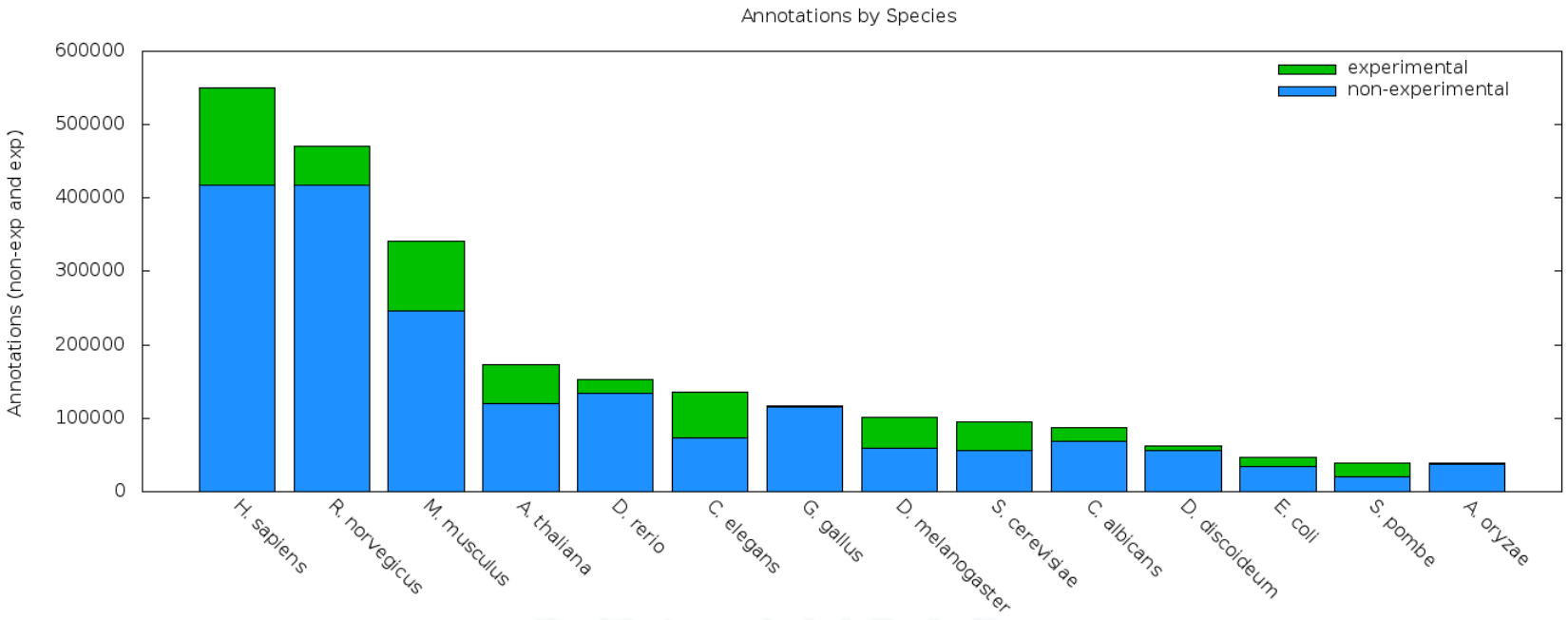
Gene Ontology (GO)

- Gene ontology, or GO, is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. More specifically, the project aims to:
  - Maintain and develop its controlled vocabulary of gene and gene product attributes;
  - Annotate genes and gene products, and assimilate and disseminate annotation data;
  - Provide tools for easy access to all aspects of the data provided by the project, and to enable functional interpretation of experimental data using the GO, for example via enrichment analysis.

http://geneontology.org/page/documentation

GO

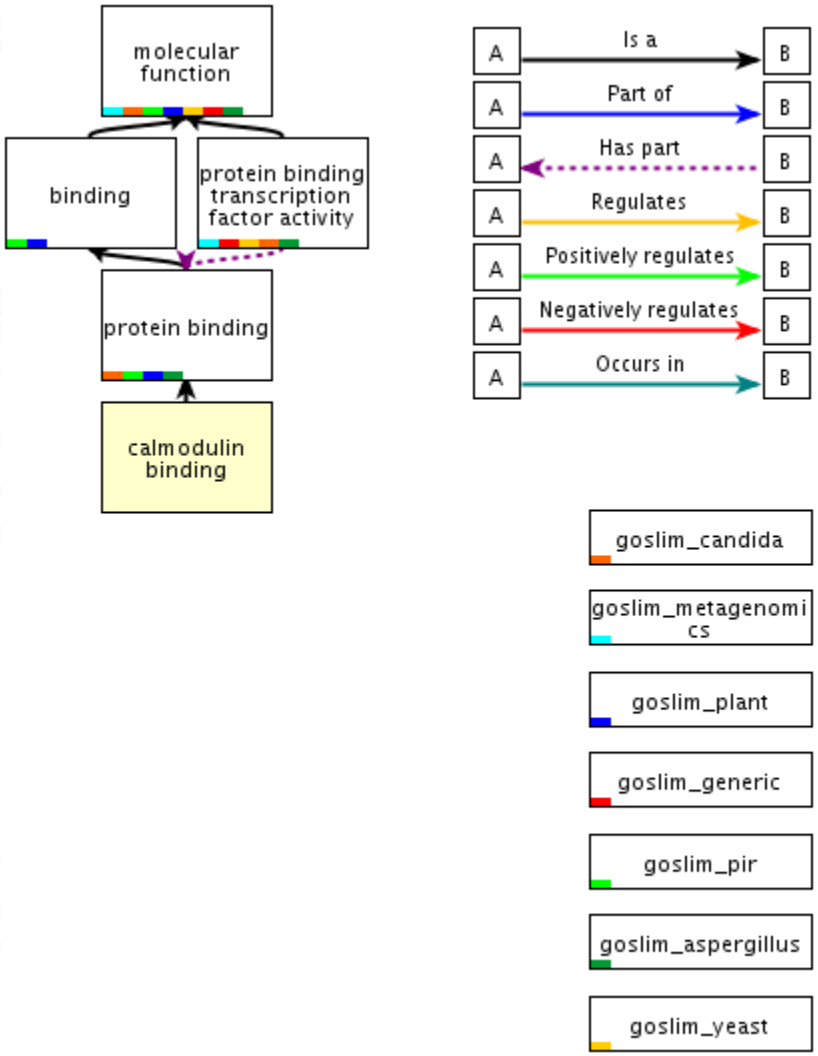
Current GO Statistics

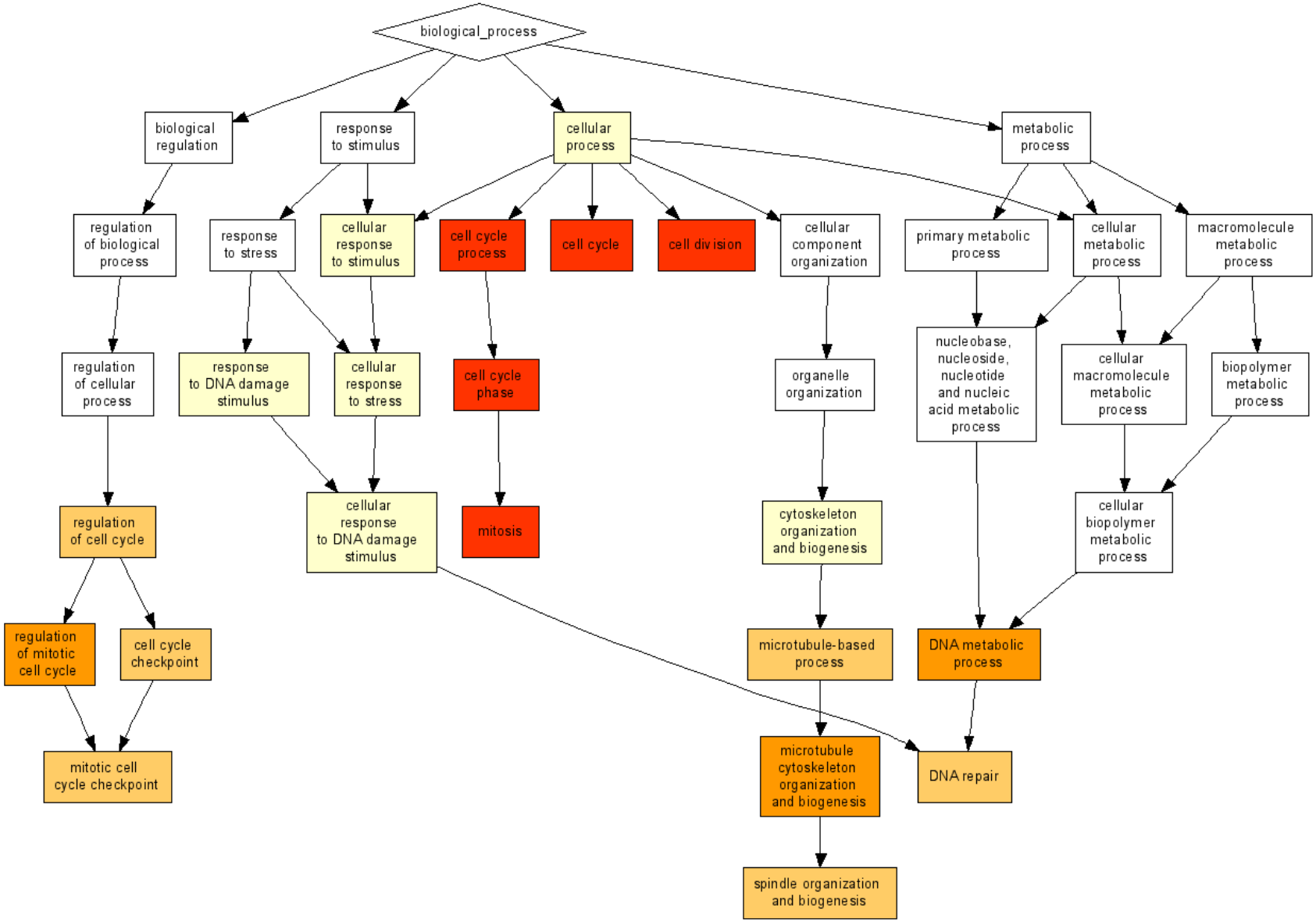


http://geneontology.org/page/current-go-statistics

Protein databases: Function annotations

- Gene Ontology Database
  - Hierarchically organized
  - Ontologies
    - Subcellular localization
    - Molecular Function
    - Biological Process
  - Example: “Calmdoulin Binding”
    - http://amigo.geneontology.org/amigo/term/GO:0005516




biological_process
biological regulation
response to stimulus
cellular process
metabolic process
regulation of biological process
response to stress
cellular response to stimulus
cell cycle process
cell cycle
cell division
cellular component organization
primary metabolic process
cellular metabolic process
macromolecule metabolic process
regulation of cellular process
response to DNA damage stimulus
cellular response to stress
cell cycle phase
organelle organization
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
cellular macromolecule metabolic process
biopolymer metabolic process
regulation of cell cycle
cellular response to DNA damage stimulus
mitosis
cytoskeleton organization and biogenesis
cellular biopolymer metabolic process
regulation of mitotic cell cycle
cell cycle checkpoint
microtubule-based process
DNA metabolic process
mitotic cell cycle checkpoint
microtubule cytoskeleton organization and biogenesis
DNA repair
spindle organization and biogenesis

End of Lecture