Polymers: Lipids, Carbohydrates, Nucleic Acids, and Proteins

Fall 2004, AMS-691 Section 2
Topics in Applied Mathematics

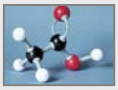
Introduction to Computational Structural Biology and Drug Design

Meeting 04, 09/13/04, Topics 1 and 2

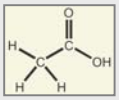
Robert C. Rizzo

*Adopted from http://web.uccs.edu/chemistry and other online sources

Graphical representations



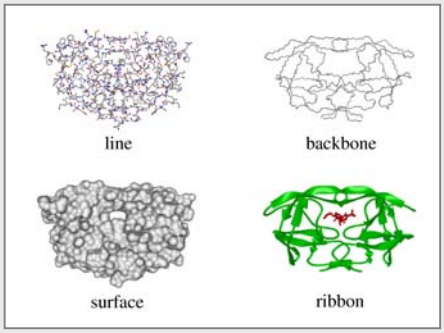
ball & stick



line

space filling

Graphical representations



H, O, C and N make up 99+% of atoms in the human body

ELEMENT	PERCENTAGE
Oxygen	63.0
Hydrogen	25.2
Carbon	9.5
Nitrogen	1.4

Their ability to form covalent bonds by sharing electron-pair unites H, O, C and N

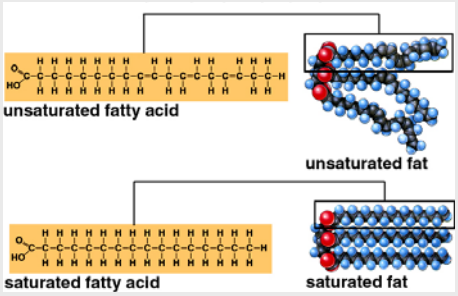
Four Major Groups of Biochemicals

- Lipids (fats)
- Carbohydrates (sugars)
- Nucleic Acids (DNA, RNA)
- Proteins (polymers of amino acids)

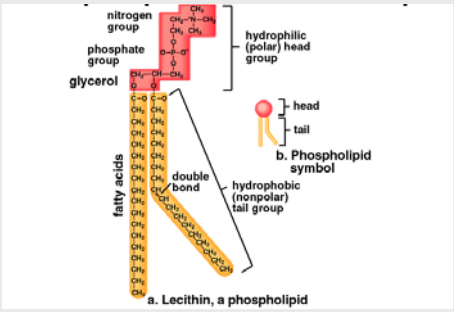
Lipids: fats, oils, phospholipids

- insoluble in water
- largely nonpolar
- hydrophobic
- phospholipids
- phosphate group is polar
- bipolar
- cell membranes

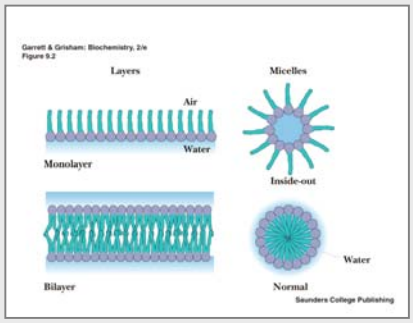
Lipids: Contain Long Hydrocarbon Chains (detailed view)



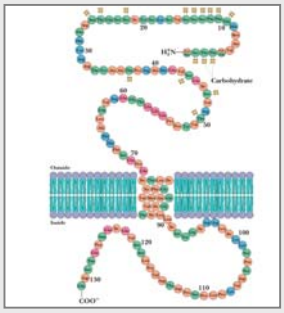
Phospholipids: Basic Structure (major constituent of cell membranes)



Phospholipids: Cell Membrane Bilayer arrangement (hydrophobic tails can associate)



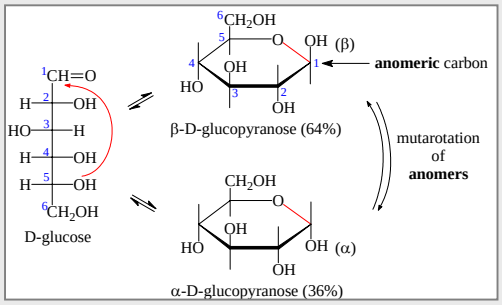
Phospholipids: The cell membrane lipid bilayer contains many proteins



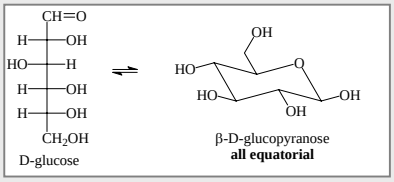
Carbohydrates (sugars)

- Carbohydrates are composed of H, C, O, and OH groups
- Monosaccharides (simple sugars)
- Oligosaccharides and Polysaccharides are polymers of the simple sugars

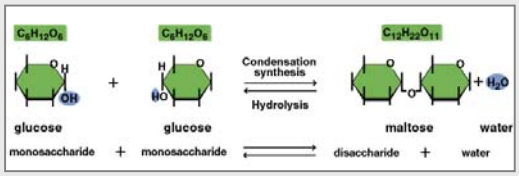
Carbohydrates: monosaccharides (cyclic structures: Haworth representation)



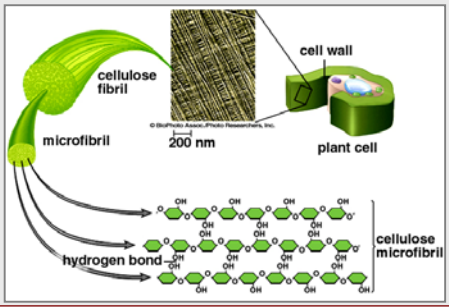
Carbohydrates: monosaccharides (cyclic structures: chair representation)



Carbohydrates: Can combine to form polymers (oligosaccharides)



Carbohydrates: Oligosaccharides make up cellulose fibrils (cell walls of plants)

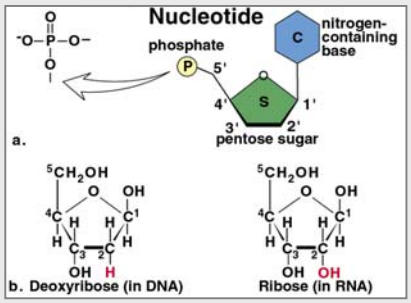


Nucleic Acids

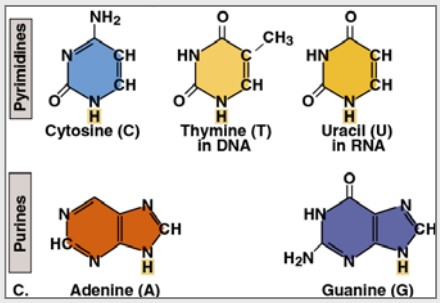
- Storage and blueprint of genetic information.
- DNA (deoxyribonucleic acid)
- RNA (ribonucleic acid)
- Polymers of nucleotides

- Each nucleotide consists of :
  - a **sugar** molecule (a carbohydrate)
  - a nitrogen containing **base**
  - a **phosphate** group

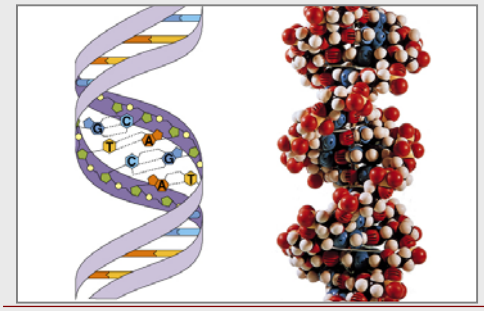
Nucleic acids: Are polymers of nucleotides: (base + sugar + phosphate)



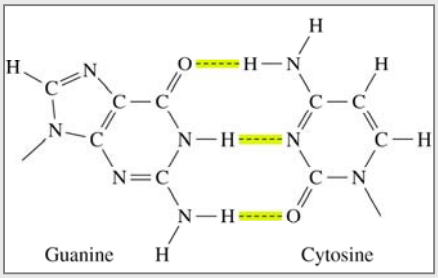
Nucleic acids: 5 types of bases A, C, T, G (DNA) and A, C, U, G (RNA)



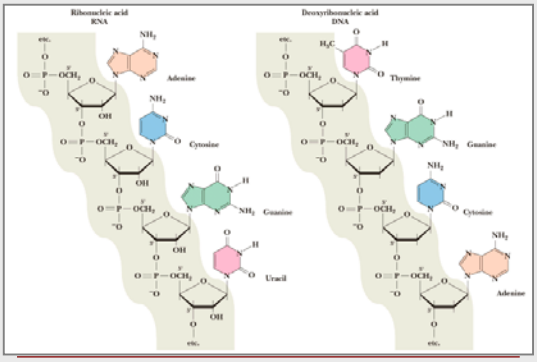
Nucleic acids: DNA double helix showing base-pairing (ribbon and space filling diagrams)



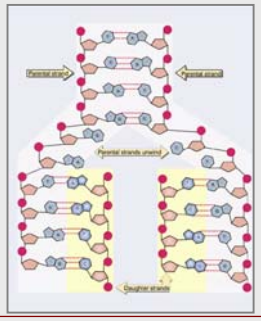
Nucleic acids: Hydrogen bonding between complementary bases



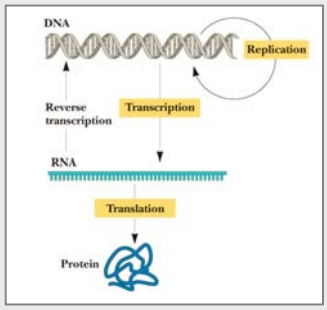
Nucleic Acids: RNA (left), DNA (right)



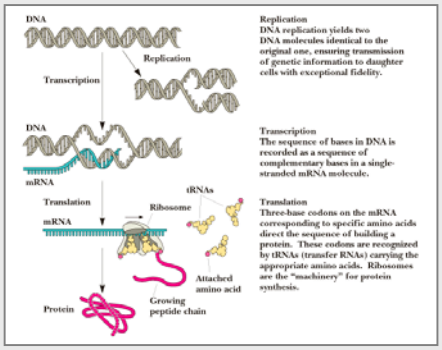
Nucleic acids: Base-pairing (H-bonding) provides a mechanism for replication



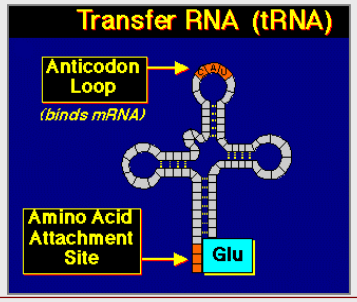
Nucleic acids: Contain blueprints (genes) which encode individual proteins



The Central Dogma of Molecular Biology



Transfer RNA (tRNA) recognize a complimentary messenger RNA (mRNA) and delivers the proper amino acid based on the genetic code



Nucleic acids: A sequence of 3 adjacent RNA bases (codon) encodes a particular amino acid (Genetic Code)

Genetic Code

1st base	2nd base U	2nd base C	2nd base A	2nd base G	3rd base
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met*	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

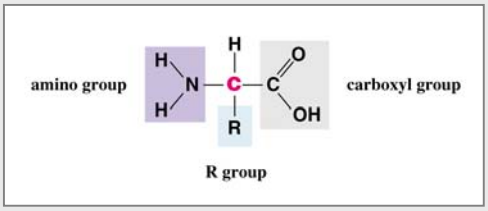
Proteins: Are the most structurally and functionally diverse of life's molecules

- contain C, H, O, N, S sometimes other elements
- polymers of amino acids
- 20 naturally occurring amino acids

Proteins: The primary worker molecule in the body

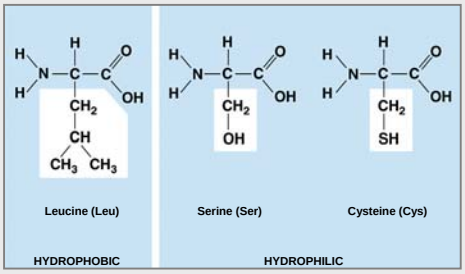
- Transport- hemoglobin in blood
- Storage- ferritin in liver
- Immune response- antibodies
- Receptors- sense stimuli, e.g. in neurons
- Channels- control cell contents
- Structure- collagen in skin
- Enzymes- catalyze biochemical reactions
- Cell functions- multi-protein machines

Proteins: Each amino acids (protein unit) has three components

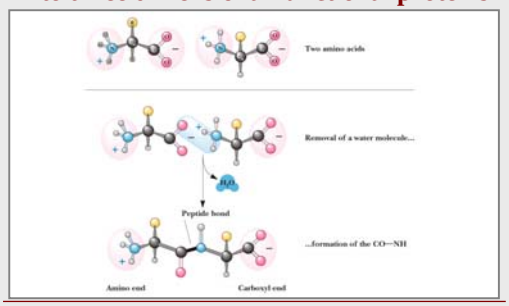


The R group determines the character of the 20 amino acids

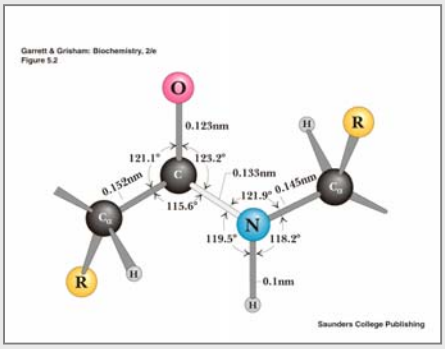
Proteins: R groups have specific properties (three examples)



Proteins: Amino acids combine to form linear polypeptides which later fold into three dimensional functional proteins



Proteins: Peptide Bond



Proteins: Amino acid R groups (hydrophobic)

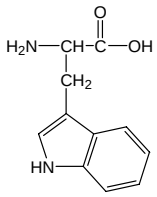
• Alanine – CH_3	Ala	A
• Valine – $CH(CH_3)_2$	Val	V
• Leucine – $CH_2CH(CH_3)_2$	Leu	L

Proteins: Amino acid R groups (hydrophobic)

• Isoleucine – $CH(CH_3)(CH_2CH_3)$	Ile	I
• Methionine – $CH_2SCH_2CH_3$	Met	M
• Phenylalanine – $CH_2(C_6H_5)$	Phe	F

Proteins: Amino acid R groups (hydrophobic)

• Tryptophan	Trp	W

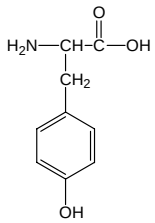


Proteins: Amino acid R groups (hydrophilic and uncharged)

• Serine – CH_2OH	Ser	S
• Threonine – $CH(OH)(CH_3)$	Thr	T

Proteins: Amino acid R groups (hydrophilic and uncharged)

- Tyrosine Tyr Y

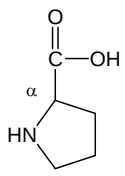


Proteins: Amino acid R groups (hydrophilic and uncharged)

- Glycine Gly G
  - H
- Asparagine Asn N
  - CH_2CONH_2
- Glutamine Gln Q
  - $CH_2CH_2CONH_2$
- Cysteine Cys C
  - CH_2SH

Proteins: Amino acid R groups (hydrophilic and uncharged)

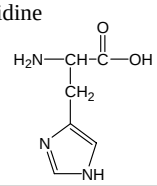
- Pro line Pro P



Rigid ring - can induce kinks in polypeptide chain

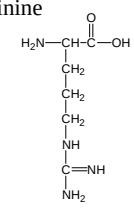
Proteins: Amino acid R groups (hydrophilic and charged)

- Aspartic acid — Asp — D (-1)
  - CH_2COOH
- Glutamic acid — Glu — E (-1)
  - CH_2CH_2COOH
- Histidine — His — H (0, +1)

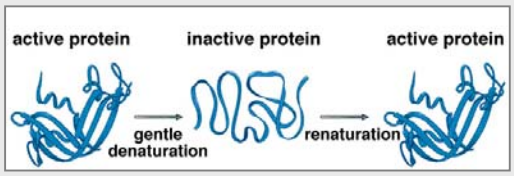


Proteins: Amino acid R groups (hydrophilic and charged)

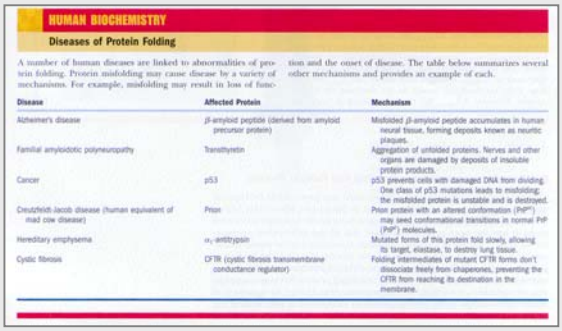
- Lysine — Lys — K (+1)
  - $CH_2CH_2CH_2CH_2NH_2$
- Arginine — Arg — R (+1)



Proteins: Protein function is connected to the 3D folded structure. Structure is determined by the amino-acid sequence



Proteins: Protein folding and disease



HUMAN BIOCHEMISTRY

Diseases of Protein Folding

A number of human diseases are linked to abnormalities of protein folding. Protein misfolding may cause disease by a variety of mechanisms. For example, misfolding may result in loss of function and the onset of disease. The table below summarizes several other mechanisms and provides an example of each.

Disease	Affected Protein	Mechanism
Alzheimer's disease	β-amyloid peptide (derived from amyloid precursor protein)	Misfolded β-amyloid peptide accumulates in human neural tissue, forming deposits known as neuritic plaques.
Familial amyloidotic polyneuropathy	Transthyretin	Aggregation of unfolded proteins. Nerves and other organs are damaged by deposits of insoluble protein products.
Cancer	p53	p53 prevents cells with damaged DNA from dividing. One class of p53 mutations leads to misfolding; the misfolded protein is unstable and is destroyed.
Creutzfeldt-Jacob disease (human equivalent of mad cow disease)	Prion	Prion protein with an altered conformation (PrP^{Sc}) may seed conformational transitions in normal PrP (PrP^{C}) molecules.
Hereditary emphysema	α_1-antitrypsin	Mutated forms of this protein fold slowly, allowing its target, elastase, to destroy lung tissue.
Cystic fibrosis	CFTR (cystic fibrosis transmembrane conductance regulator)	Folding intermediates of mutant CFTR forms don't dissociate freely from chaperones, preventing the CFTR from reaching its destination in the membrane.

Proteins: Sequence homology example

Cyt c amino acid residues from several species

Species	Sequence (20-29)
Human	VEKGGKHKTG
Horse	VEKGGKHKTG
Fruit fly	VEAGGKHKVG
Potato	VDKGAGHKEG

Proteins: Sequence determines structure

All of the information necessary for folding the peptide chain into its native structure is contained in the primary amino acid structure of the poly peptide.

Certain positions along the chain may act as nucleation points

Protein chain must avoid local energy minima

Chaperones (a type of helper protein) may assist folding

Proteins: Folding funnel model

Unfolded structures lie around the top. As the protein folds, it falls down the wall of the energy funnel to more stable conformations.

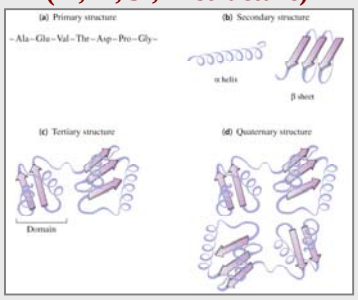
The native, folded structure is at the bottom.

Ken Dill, Nature Structural Biol. **4**, 10-19 (1997).

Proteins: 4 Elements of Protein Structure

- Primary
  - –Order of amino acids in linear chain
    - N - terminus (start, by convention) C-terminus (end)
- Secondary
  - –Portions of linear chain fold into regular conformations
    - α-helix β-sheet, reverse turns, etc
- Tertiary structure
  - –Overall shape
- Quaternary structure
  - –Protein consisting of several peptide chains

Protein fold: compact, globular folding arrangement of the polypeptide chain (1°, 2°, 3°, 4° structure)

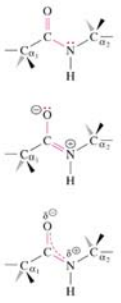


Proteins: The peptide bond has resonance

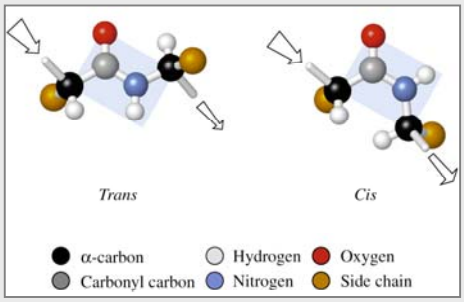
(a) Peptide bond shown as a C-N single bond

(b) Peptide bond shown as a double bond

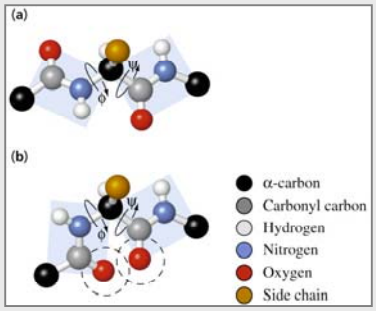
(c) Actual structure is a hybrid of the two resonance forms. Electrons are delocalized over three atoms: O, C, N



Proteins: Peptide bonds are usually trans (restricted rotation due to resonance)



Proteins: Rotation around the N-Cα (Phi) and Cα-C bonds (Psi) that link peptide groups



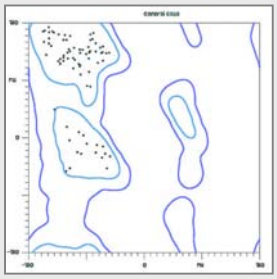
Proteins: Polypeptides adopt conformations

- Backbone conformation has 2 degrees of freedom per residue
  - These are two "torsion" angles (bond rotations) Phi: C-N-C^{α}-C and Psi: N-C^{α}-C-O
- Conformation can be illustrated on 2D plot of Phi versus Psi (Ramachandran plot)
- Steric interference limits possible values of Phi and Psi
  - Gly less restricted.
  - Pro more restricted

Proteins: Allowed combinations of N-Cα (Phi) and Cα-C bonds (Psi) (ramachandran plot)

- G. N. Ramachandran: plot phi vs psi & indicate regions of no steric conflict
- The sterically favorable combinations are the basis for preferred secondary structures

psi

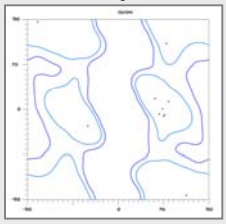


phi

Proteins: Ramachandran plot examples

Gly

psi

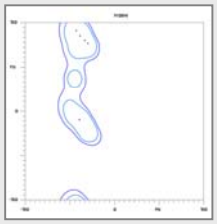


phi

Less restricted –
side chain is only H

Pro

psi



phi

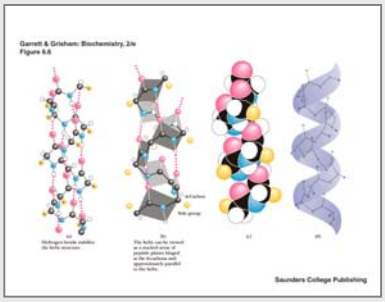
More restricted –
side chain of amino acid forms a ring

Proteins: Regular repeating structures

- Repetition of phi/psi combination
- Low energy structures – maximum formation of hydrogen bonds
- Proposed on the basis of models combined with experimental data:
  - **α-helix** from the X-ray pattern from hair
  - **β-sheet** from the X-ray pattern from silk

Proteins: Structural motif, alpha helix (α-helix)

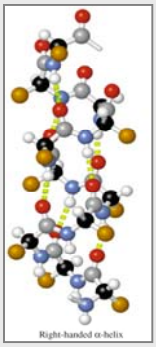
- First proposed by Linus Pauling and Robert Corey in 1951
- Identified in keratin by Max Perutz
- A common component of proteins
- Stabilised by H-bonds in backbone NH…O=C



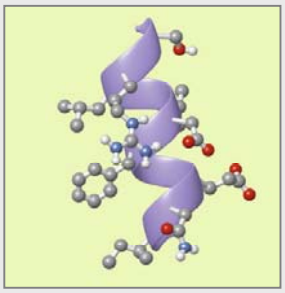
Proteins: Structural motif, alpha helix (α-helix)

- Each C=O (residue *n*) forms a hydrogen bond with the amide (H on N) hydrogen of residue *n*+4
- Helix is stabilized by many hydrogen bonds (which are nearly parallel to long axis of the helix)
- All C=O groups point toward the C-terminus (entire helix is a dipole with:

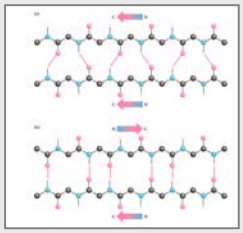
  (+) N terminus, (–) C terminus)

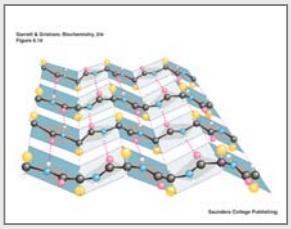


Proteins: Amino acid side chains (R groups) project outward from the alpha helix axis



Proteins: Structural motif, beta sheet (β-sheet





- Also first postulated by Pauling and Corey, 1951
- Strands may be parallel or antiparallel
- Nearly fully extended
- Sheets pleat to maintain correct H-bond stereochemistry
- Side chains point alternatively on opposite sides of the sheet

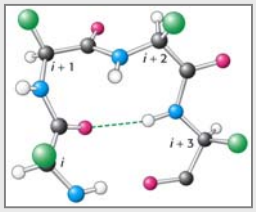
Proteins: β-sheets

- C=O, N-H groups form H-bonds between neighbouring strands
- Sheets pleat to maintain correct H-bond stereochemistry
- H-bonds are straighter & stronger in anti-parallel sheets
- Side chains point alternatively on opposite sides of the sheet
- Strands may be:

  parallel Phi = -119° Psi = 113°

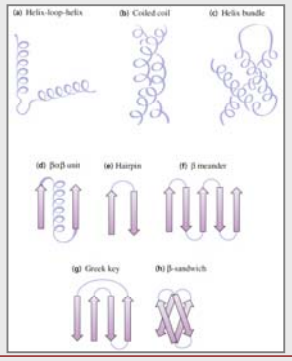
  antiparallel Phi = -139° Psi = 135°

Proteins: Structural motif, a reverse turn (note peptide bonds are trans, and H-bond)

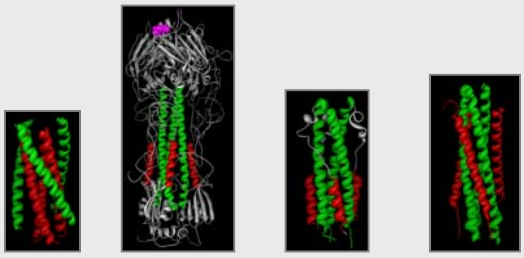


Beta turns allows the polypeptide chain to change direction quickly

Proteins: Other structural motifs

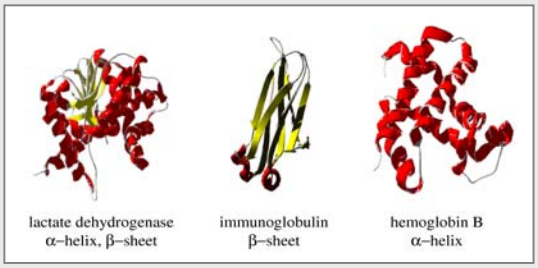


Coiled-coil motif common among membrane fusion proteins (N-helices in red, C-helices in green)

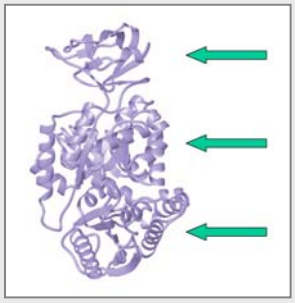


HIV (AIDS) **Influenza (flu)** **Ebola (hemorrhagic fever)** **HRSV (pneumonia)**

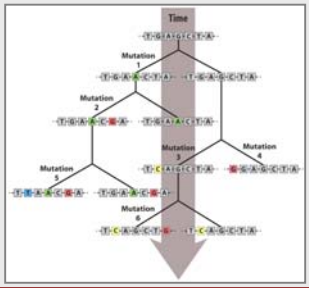
Proteins: Examples of proteins with different secondary structure elements



Proteins: Functional proteins are often composed of distinct domains (folded polypeptides)



Proteins: Mutations over time results in amino acid changes yet protein functional domains are often structurally conserved



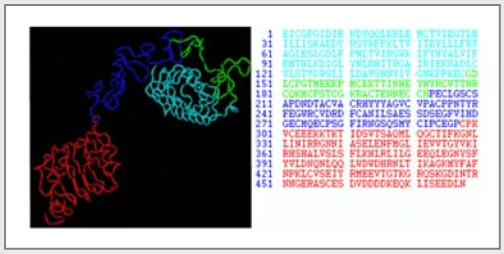
Proteins: Domains

A single domain may have a particular function (e.g. binding small molecules, catalyzing a single reaction)

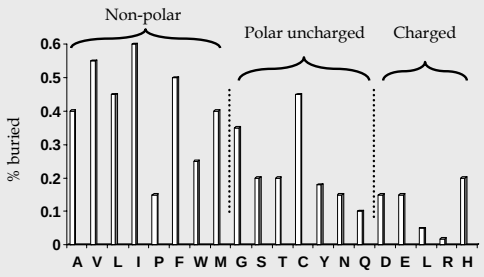
Interfaces between two separate domains provide grooves and pockets on the surface of a protein for binding sites

In multifunctional enzymes, each catalytic activity can be in a different domains

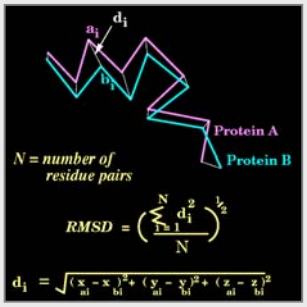
Proteins: Domains often correspond with the linear amino-acid primary sequence (single amino-acid codes shown)



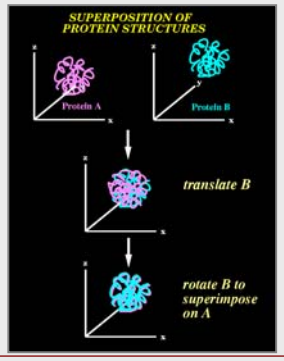
Proteins: Amino acids in tertiary structure



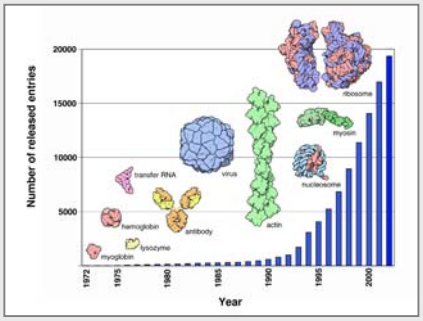
Proteins: Fold similarity (geometric similarity) using root-mean-square deviation (RMSD)



Proteins: Rigid body superposition



Experimental protein structures are stored in the protein data bank (PDB)



Summary: Four Major Groups of Biochemicals

- Lipids (hydrophobic)
- Carbohydrates (sugars)
- Nucleic Acids (codes protein primary sequence)
- Proteins are polymers of amino acids (20 types)

Summary: 20 amino-acids

AMINO ACID			SIDE CHAIN	AMINO ACID			SIDE CHAIN
Aspartic acid	Asp	D	negative	Alanine	Ala	A	nonpolar
Glutamic acid	Glu	E	negative	Glycine	Gly	G	nonpolar
Arginine	Arg	R	positive	Valine	Val	V	nonpolar
Lysine	Lys	K	positive	Leucine	Leu	L	nonpolar
Histidine	His	H	positive	Isoleucine	Ile	I	nonpolar
Asparagine	Asn	N	uncharged polar	Proline	Pro	P	nonpolar
Glutamine	Gln	Q	uncharged polar	Phenylalanine	Phe	F	nonpolar
Serine	Ser	S	uncharged polar	Methionine	Met	M	nonpolar
Threonine	Thr	T	uncharged polar	Tryptophan	Trp	W	nonpolar
Tyrosine	Tyr	Y	uncharged polar	Cysteine	Cys	C	nonpolar
POLAR AMINO ACIDS				NONPOLAR AMINO ACIDS			

Summary: Proteins fold into globular compact arrangements

- Polypeptide chain folds to optimize packing of the hydrophobic residues in the interior core of the protein
- Backbone conformation has 2 degrees of freedom per residue: Phi versus Psi Ramachandran plot
- Primary, secondary, tertiary, quaternary structures
- Conserved Domains (sequence/structure/function)

How many proteins drug targets?

	Organism	Number of genes in the genome
Bacteria	*Myscoplasma genitalium*	517
Yeast	*Saccharomyces cerevisiae*	6,275
	Arabidopsis thaliana	~ 20,000
	Caenorhabditis elegans	19,099
	Haemophilus influenzae	1,743
Fruit Fly	*Drosophila melanogaster*	13,601
	Neisseria meningitdis	2,158
Human	*Homo sapiens*	~ 30,000

HUMAN TARGETS
~30,00 genes
~200, 00 proteins
~10 % interesting:
20K human genome targets

PATHOGENS
Viruses, bacteria, fungi, parasites have smaller genomes but larger fraction could be targets

Say 100 organisims*1000 genes:
100K infectious disease targets