Genomics and Bioinformatics: Hot New Areas of Biotechnology

- **Genomics** – cloning, sequencing, and analyzing entire genomes
  - Shotgun sequencing or shotgun cloning
    - The entire genome is cloned and sequenced
    - Produces thousands of fragments to be sequenced
    - Individual genes are sorted out later through **bioinformatics**
      - Computer programs are used to align the sequenced fragments based on overlapping sequence pieces

- Bioinformatics
  - An interdisciplinary field that applies computer science and information technology to promote an understanding of biological processes
- Application of Bioinformatics
  - Databases to store, share, and obtain the maximum amount of information from protein and DNA sequences
  - GenBank

- The Human Genome Project
  - Started in 1990 by the U.S. Department of Energy
  - International collaborative effort to identify all human genes and to sequence all the base pairs of the 24 human chromosomes
  - 20 centers in 6 countries: China, France, Germany, Great Britain, Japan, and the United States

- The Human Genome Project
  - April 14, 2003, map of the human genome was completed
  - Consists of 20,000 to 25,000 protein-coding genes

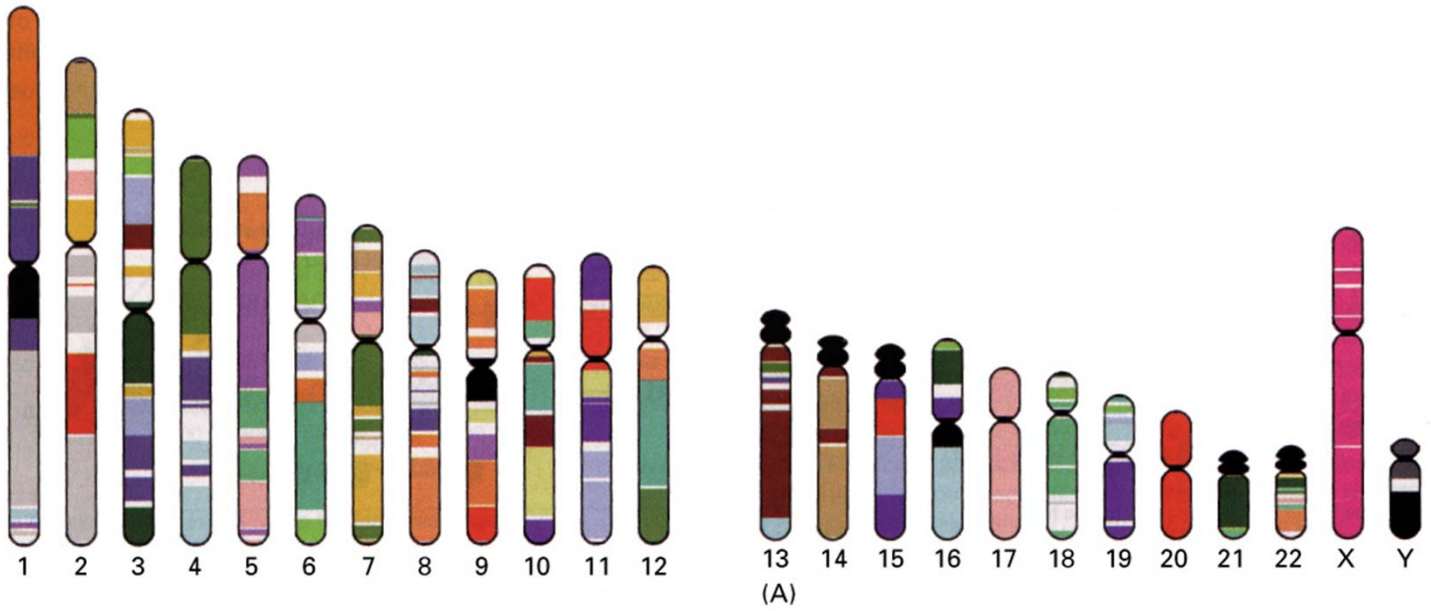


Figure 4–18 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

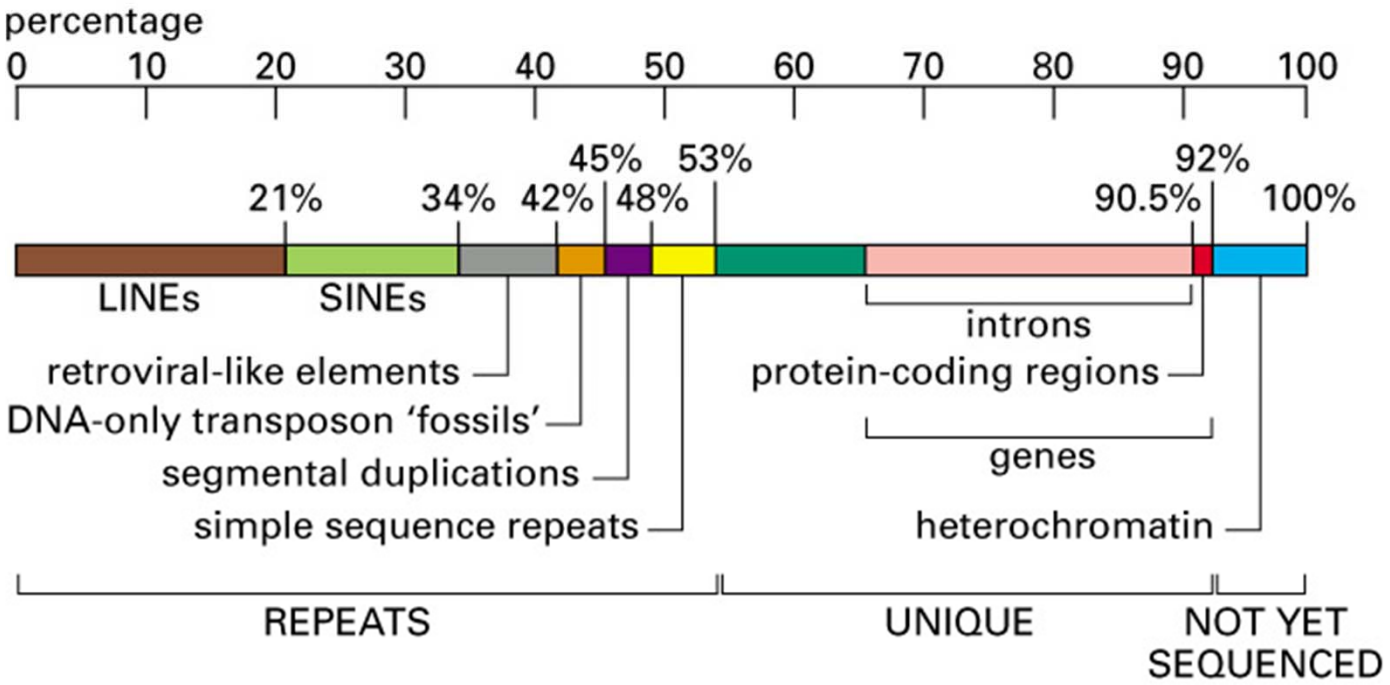


Figure 4–17. Molecular Biology of the Cell, 4th Edition.

Genomics and Bioinformatics: Hot New Areas of Biotechnology

- The Human Genome Project
  - Started an “omics” revolution
    - Proteomics
      - study of the proteome (the entire set of proteins expressed by a genome, cell, tissue or organism)
    - Metabolomics
      - Study of the meabolome (the complete set of small-molecule metabolites (such as metabolic intermediates, hormones and other signalling molecules, and secondary metabolites) to be found within a biological sample, such as a single organism.

- Glycomics
  - the study of glycomes (the entire complement of sugars, whether free or present in more complex molecules, of an organism),
- Interactomics
  - a discipline at the intersection of bioinformatics and biology that deals with studying both the interactions and the consequences of those interactions between and among proteins, and other molecules within a cell
- Transcriptomics
  - The study of transcriptome (the set of all RNA molecules, including mRNA, rRNA, tRNA, and non-coding RNA produced in one or a population of cells)

- Nutrigenomics
  - the study of the effects of foods and food constituents on gene expression.
  - It is about how our DNA is transcribed into mRNA and then to proteins and provides a basis for understanding the biological activity of food components

Genomics and Bioinformatics: Hot New Areas of Biotechnology

- Comparative Genomics
  - Mapping and sequencing genomes from a number of model organisms
  - Allows researchers to study gene structure and function in these organisms in ways designed to understand gene structure and function in other species including humans

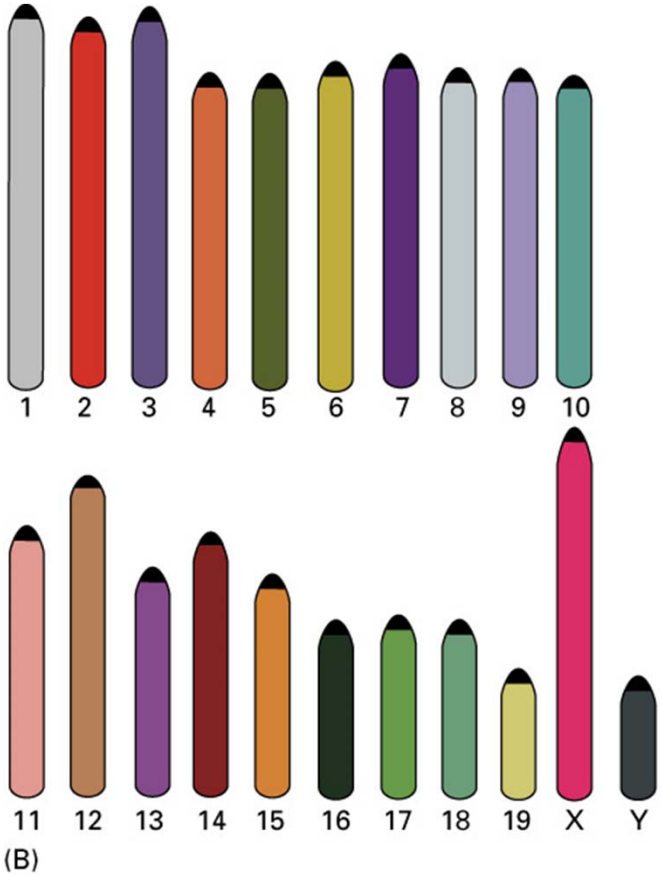


Figure 4–18 part 2 of 2. Molecular Biology of the Cell, 4th Edition.

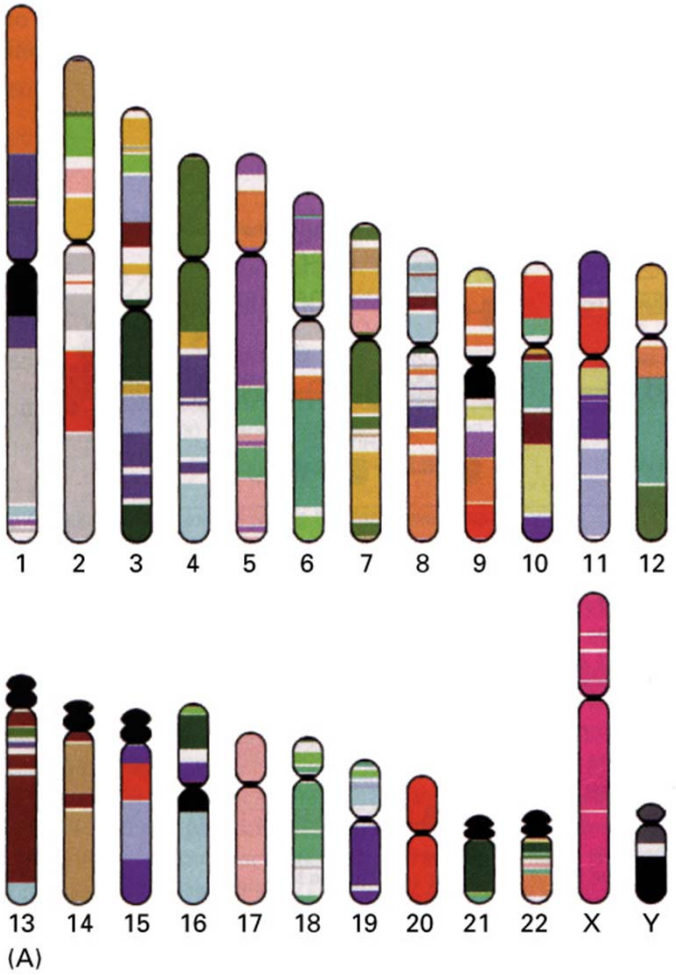


Figure 4–18 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

Genomics and Bioinformatics: Hot New Areas of Biotechnology

- Stone Age Genomics (paleogenomics)
  - Analyzing “ancient” DNA

What is Bioinformatics

Interdisciplinary subject involving computer and biological sciences

Useful Server Websites

USA

- NCBI http://www.ncbi.nlm.nih.gov/

Europe

- EBI http://www.ebi.ac.uk/

Google search

www.google.com

Tools

- Bioinformatic Software
  - $$$ buy from company
    - Vector NTI
    - Mac Vector
    - Sequencer
    - Bio edit (free software)
  - Free web-based software
    - DNA sequence analysis
    - Primer design
    - DNA translation tools
    - Structure prediction
    - Restriction site analysis
    - Sequence alignments
    - etc

Biological Databases

- Primary Database
- Secondary Database
- Specialized Database

Primary Database

- Raw nucleic acid sequence
  - Genbank
  - EMBL
  - DDBJ
    - Use different format to present data
    - These databases are closely connected and exchanged data daily
- 3D structure : PDB
    - Protein and nucleic acid structure
    - Atomic coordinates from X-ray and NMR

Secondary Database

- Computational processed information
- Provide sequence annotation
- SWISS-PROT
- TrEMBL (translated nucleic acid sequence from EMBL)
- Other : UniPort, Pfam, Blocks, DALI

Specialized Database

- Focus on particular organisms
  - Flybase
  - Wormbase
  - AceDB, TAIR
- Focus on functional analysis
  - Genbank EST
  - Microarray gene expression database

Important

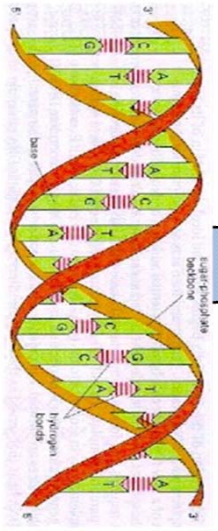
- Many database are connected
  - NCBI are most integrated
- Reliability !
- There are many errors in the database
  - Sequencing error (especially before 1990s)
  - Redundancy
    - Non-redundant database
    - UniGene (coalesce EST)

Information Retrieval

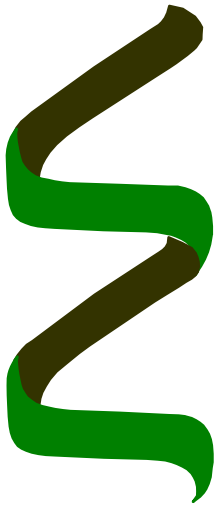
- Use Boolean operation = join a series of keywords
- Text-based search
- Provide access to multiple database for retrieval of integrated search result
- Entrez (NCBI)
- SRS (Seq retrival system from EBI)

What Are proteins? How are they made?

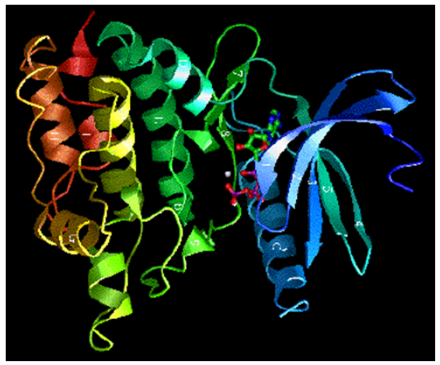
Central Dogma



Transcription



Translation



Gene (DNA) mRNA Protein

Hair (Keratins)

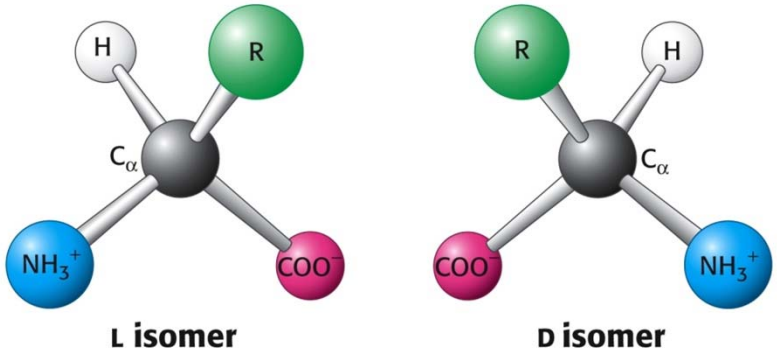
Silk (fibroin)

Enzymes

Proteins

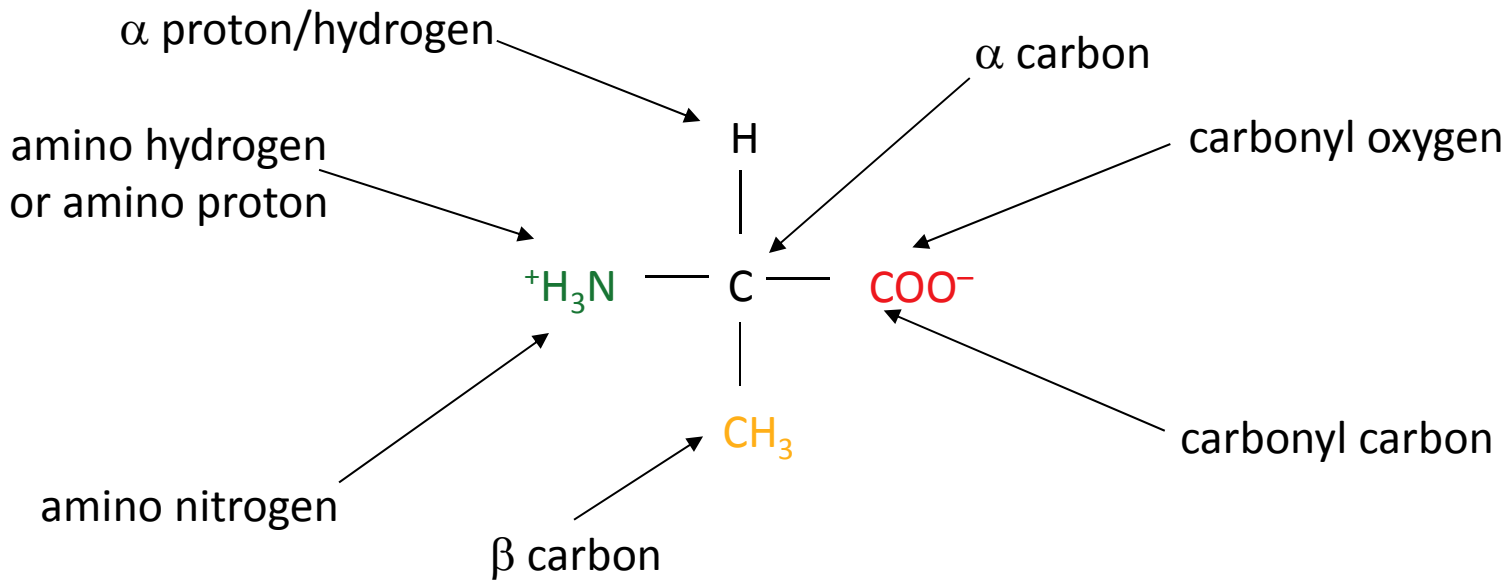
Amino Acids

Amino acids: Basic constituent of proteins



- Proteins are composed of amino acids
- All 20 amino acids have central carbon atom (Cα) to which are attached a hydrogen atom, an amino group (NH_2) group and a carboxyl group (COOH).The L amino acids are so designated if one looks downward from hydrogen atom to Cα then **CO,R and N** substituents are in clockwise direction.
- L amino acids are found in most of the naturally occurring proteins.

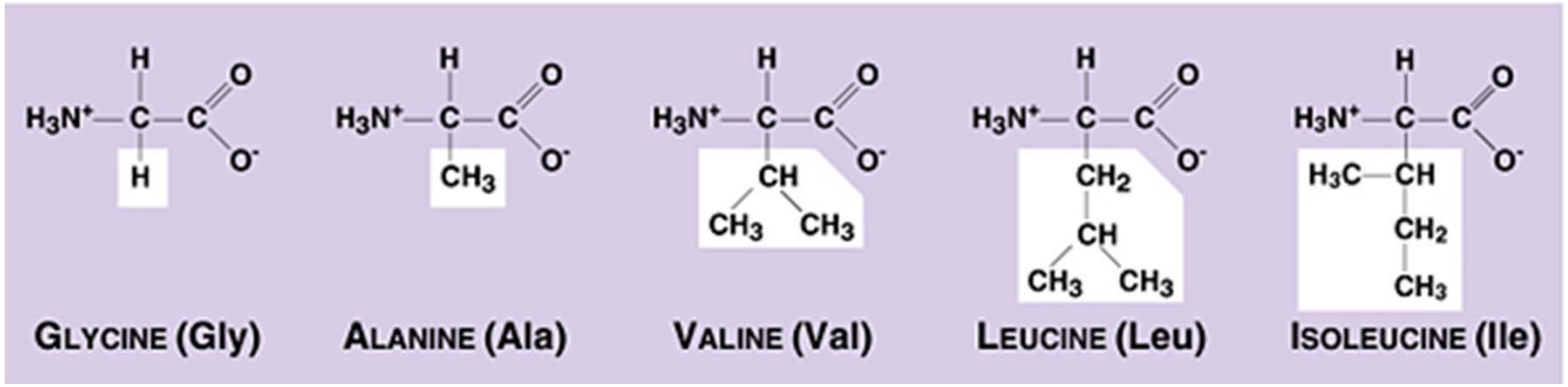
Naming amino acids



alanine
(a polar amino acid)
R = CH_3
(a methyl functional group)

Non-polar amino acids

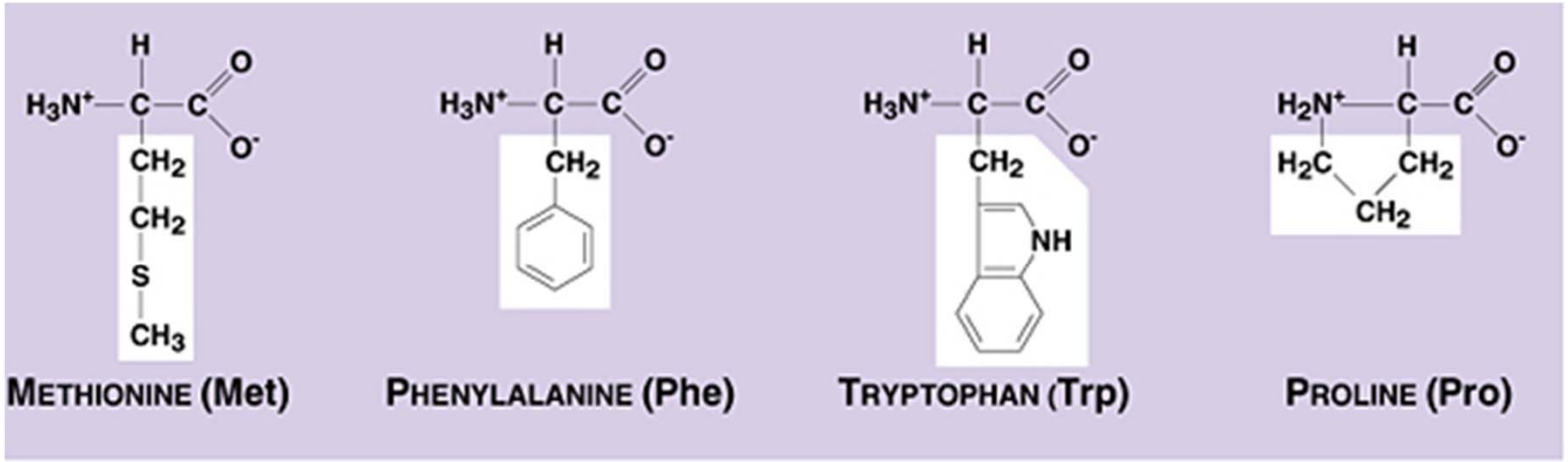
- R group consists of carbon chains



Leucine and isoleucine are structural isomers

Nonpolar amino acids

- R group consists of carbon chains



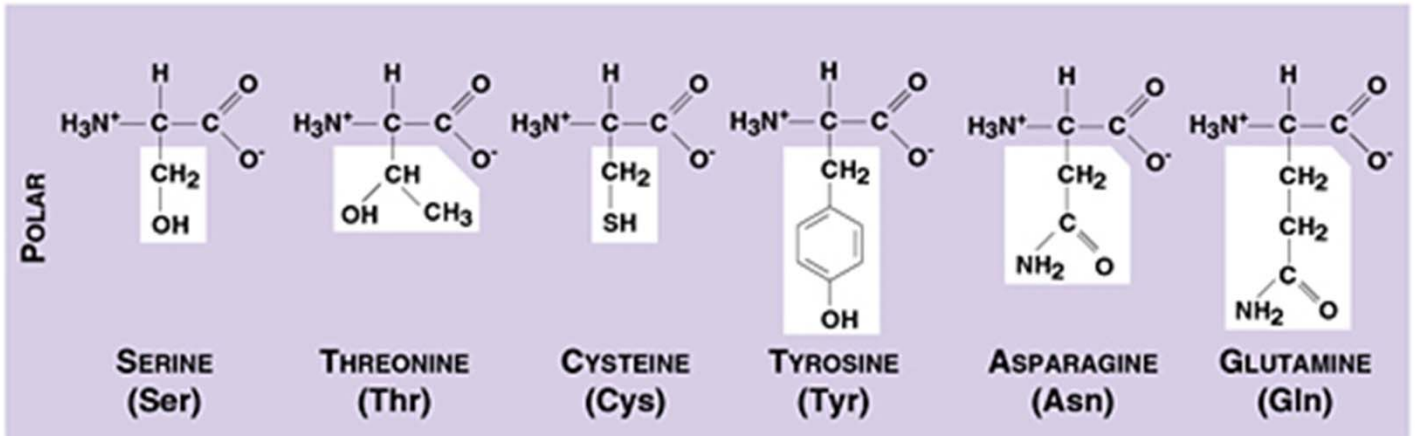
Methionine has a sulphur atom in its sidechain

Phenylalanine and tryptophan have aromatic rings

Proline has its R group bound to the amino nitrogen to form a ring network

Polar amino acids

- R group consists of carbon, oxygen and nitrogen atoms together they make the sidechain more hydrophilic

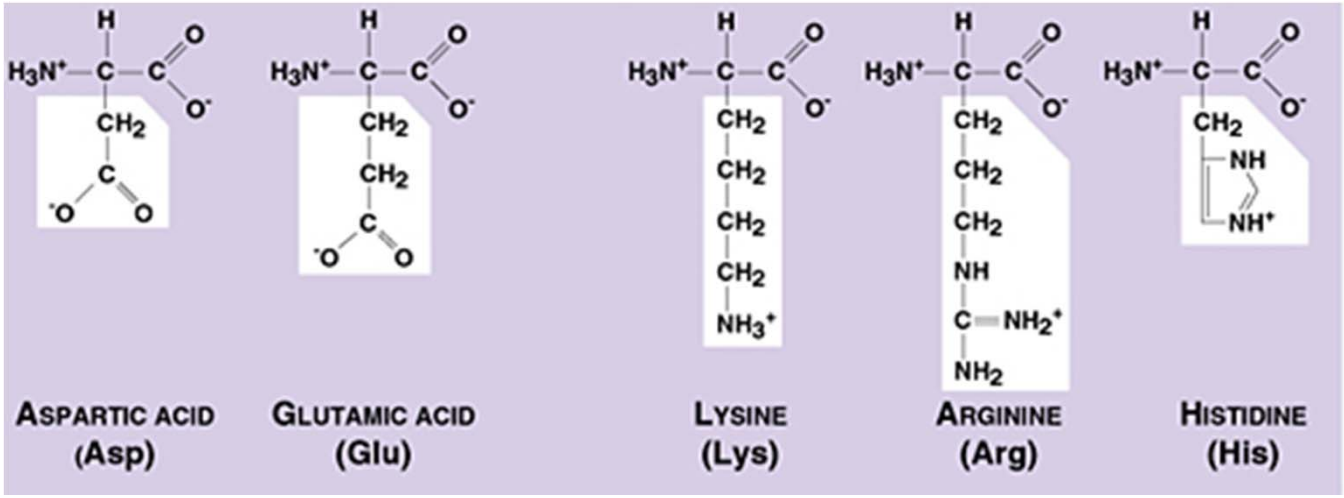


Ser and thr are a mix of carbon chains and hydroxyl functional groups (-OH). Cysteine has a thiol group (-SH) which is otherwise structurally similar to serine but not chemically similar

Asn and gln have an amide functional group

Charged amino acids

- R group has a charge at physiological pH (7.4). pK of the charged groups vary



carboxyl group | carboxyl group | amino group | guanidinio group | imidazole group

Describing amino acids

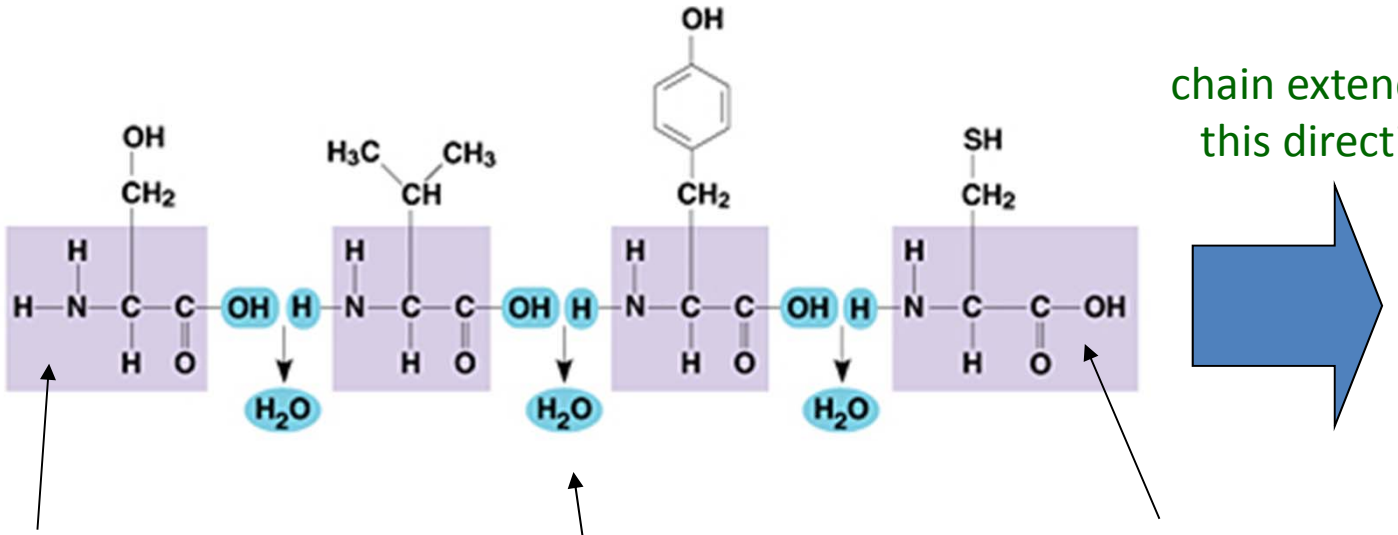
• amino acids have a full name (glycine), a short three-letter name (gly) and an even shorter one-letter name (G)

- nonpolar
- polar
- acidic (negative charge)
- basic (positive charge)

A	ala	alanine
C	cys	cysteine
D	asp	asparatic acid
E	glu	glutamic acid
F	phe	phenylalanine
G	gly	glycine
H	his	histidine
I	ile	isoleucine
K	lys	lysine
L	leu	leucine
M	met	methionine
N	asn	asparagine
P	pro	proline
Q	gln	glutamine
R	arg	arginine
S	ser	serine
T	thr	threonine
V	val	valine
W	trp	tryptophan
Y	tyr	tyrosine

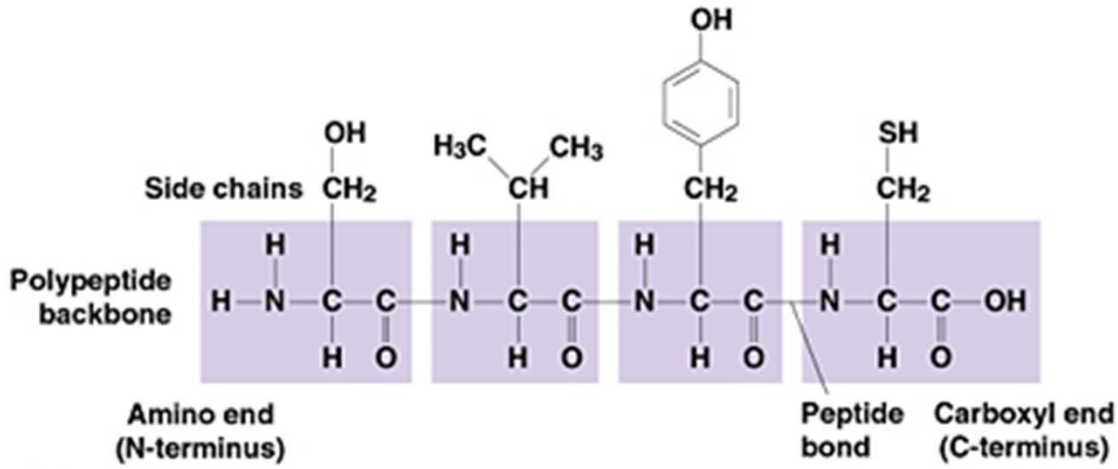
Joining amino acids together

• In a cell, a complex set of proteins called a ribosome catalyze a dehydration reaction (loss of water) to join amino acids together



Joining amino acids together

• In a cell, a complex set of proteins called a ribosome catalyze a dehydration reaction (loss of water) to join amino acids together



• a peptide bond (like an amide bond C-O-N) joins each amino acid

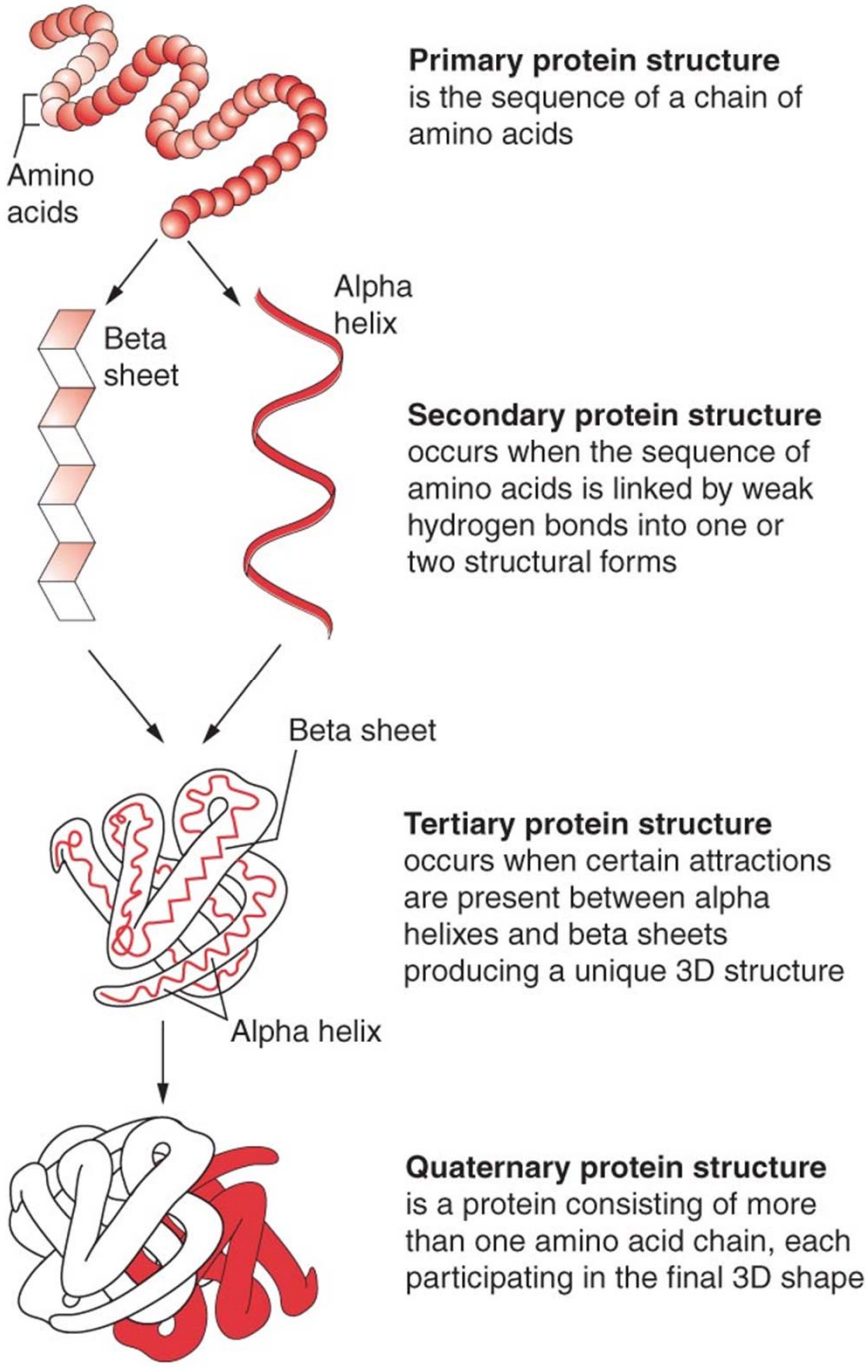
• the invariant purple part of the polypeptide is generally called the backbone

• it's the sidechains that give a protein its unique chemical and structural character

Protein Structures

- Structural Arrangement – four levels
  - **Primary structure** is the sequence in which amino acids are linked together
  - **Secondary structure** occurs when chains of amino acids fold or twist at specific points
    - Alpha helices and beta sheets
  - **Tertiary structures** are formed when secondary structures combine and are bound together
  - **Quaternary structures** are unique, globular, three-dimensional complexes built of several polypeptides

Protein Structures

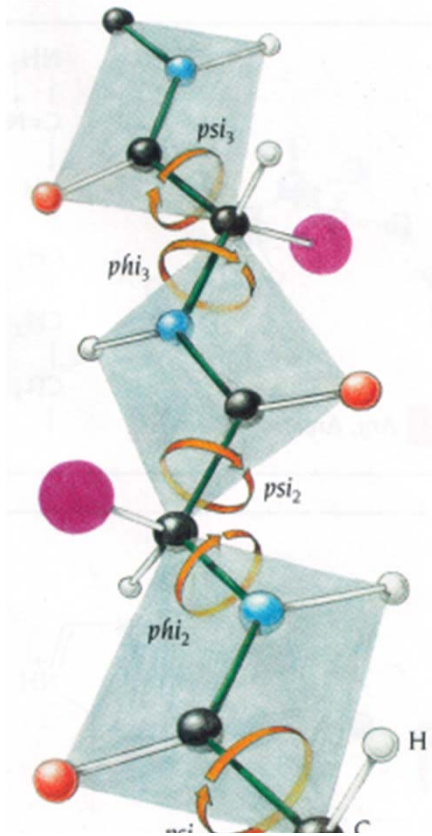


Peptides as building blocks of protein structure

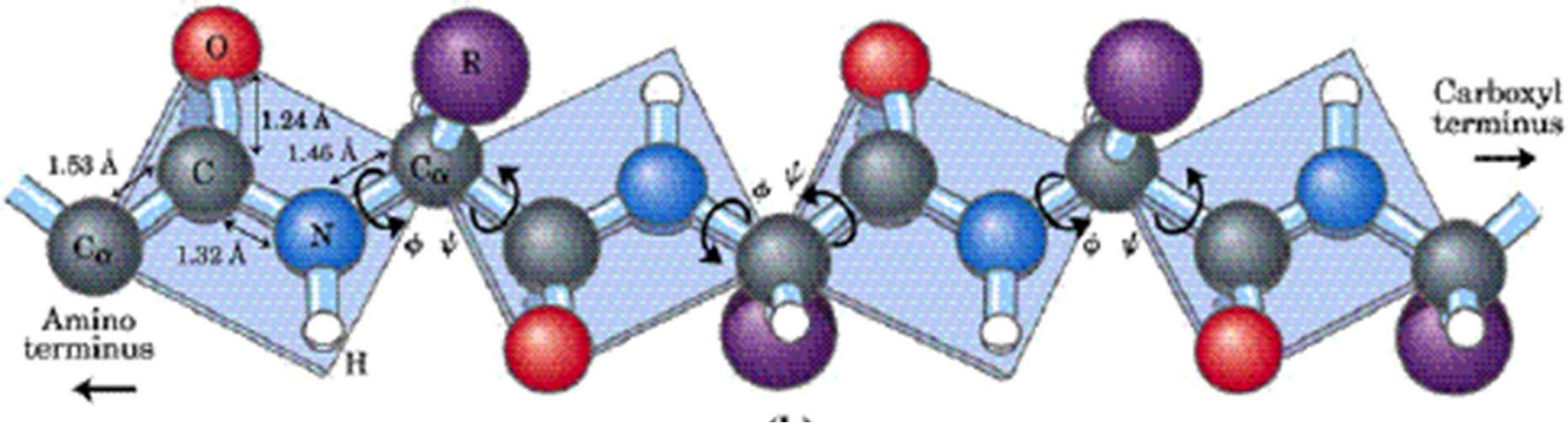
- The polypeptide chain can be divided into peptide units going from one Cα atom to the next Cα atom.

- The peptide units are rigid groups linked to the chain by covalent bonds at the Cα atom, the only degree of freedom is rotation around these bonds.

.



Peptides as building blocks of protein structure

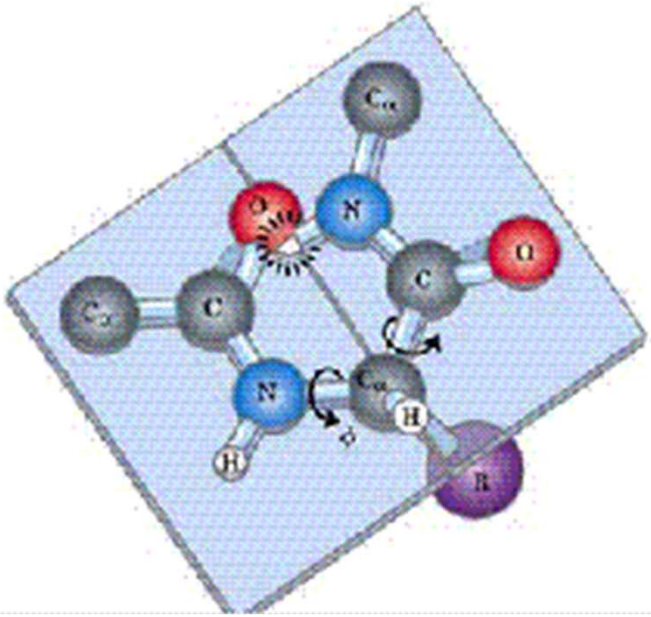


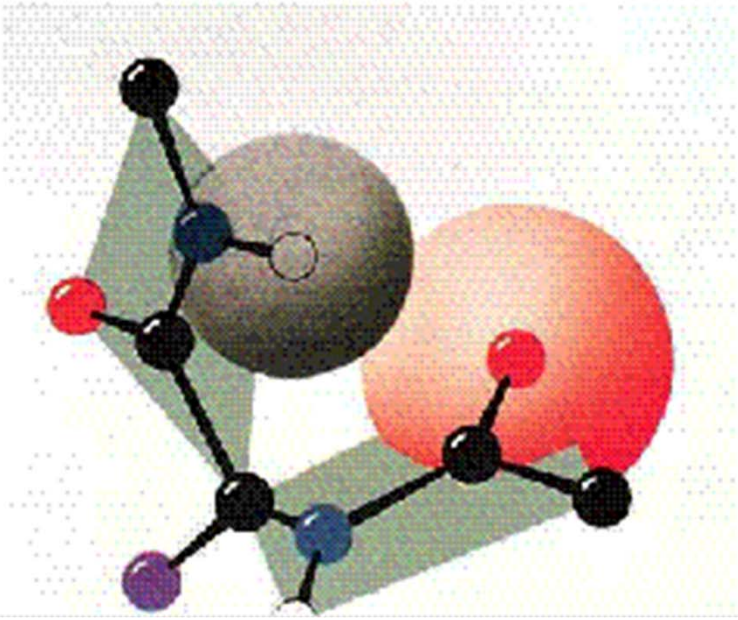
- These bonds are designated as psi (φ) for rotational angle between Cα and N of the amino group and phi (ϕ) for rotational angle between Cα and C of the carboxyl group.

- Each amino acid residue is associated with two conformational or dihedral angles and besides this is there slight contribution by side chain groups also. If all the dihedral angles for each amino acid residue is defined with high accuracy then the conformation of the main chain is completely determined

Steric hindrance

Steric hindrance (van der Waals volume overlap) of back bone and side chain atoms limits back bone flexibility



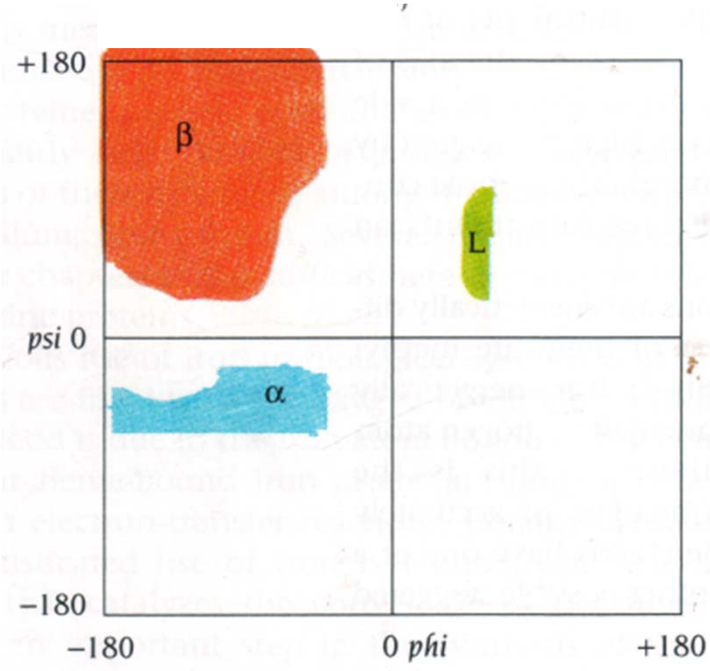


Ramachandran Plot

- Most of the dihedral angles φ and θ are not allowed for an amino acid because of the steric collisions between side chain and main chain.

- Based on the Van der waals radius of the amino acids and side chains the allowed combination of dihedral angles for amino acids can be calculated.

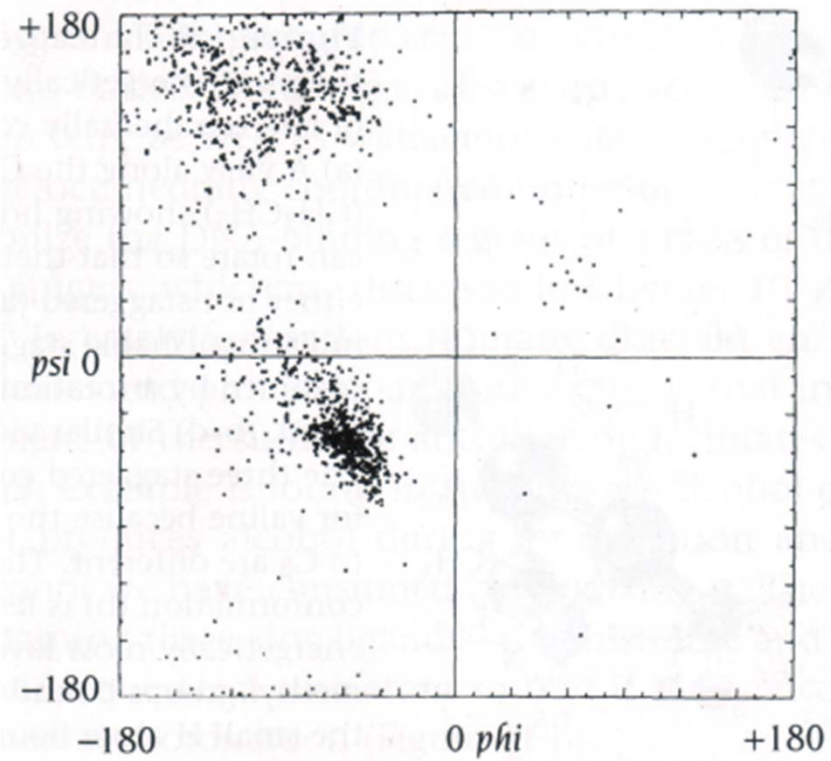
- Ramachandran plot depicting sterically allowed dihedral angles as colored areas. the areas denoted as α, β and L correspond to conformational angle found for the usual right-handed α helices, β-strands and left-handed α helices

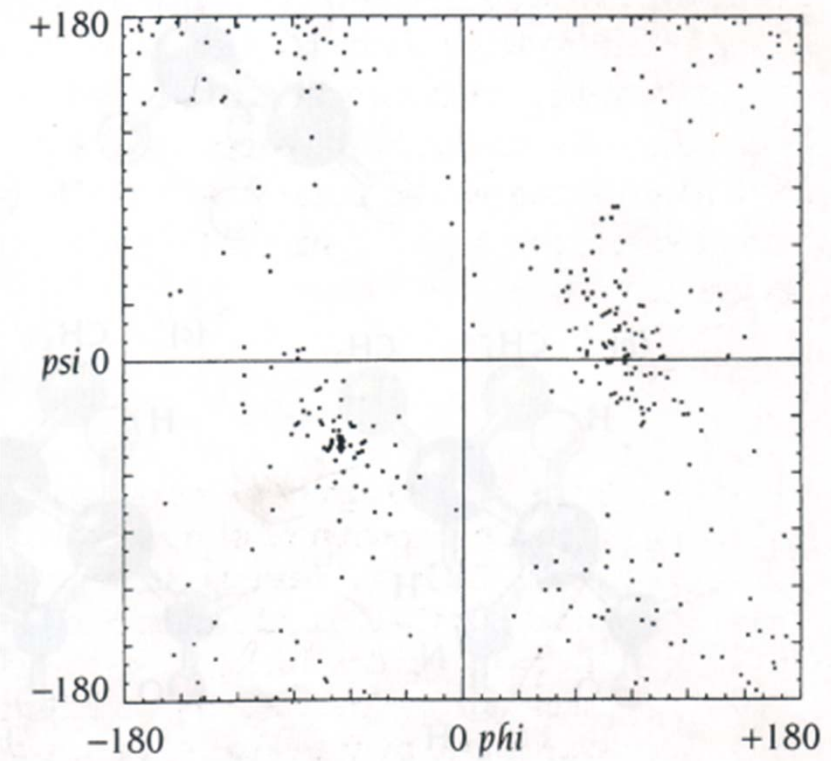


Ramachandran Plot

Observed values of dihedral angles for all residues except glycine in high resolution X-ray structure

Observed values of dihedral angles for glycine





Each point represents φ and θ value for an amino acid residue

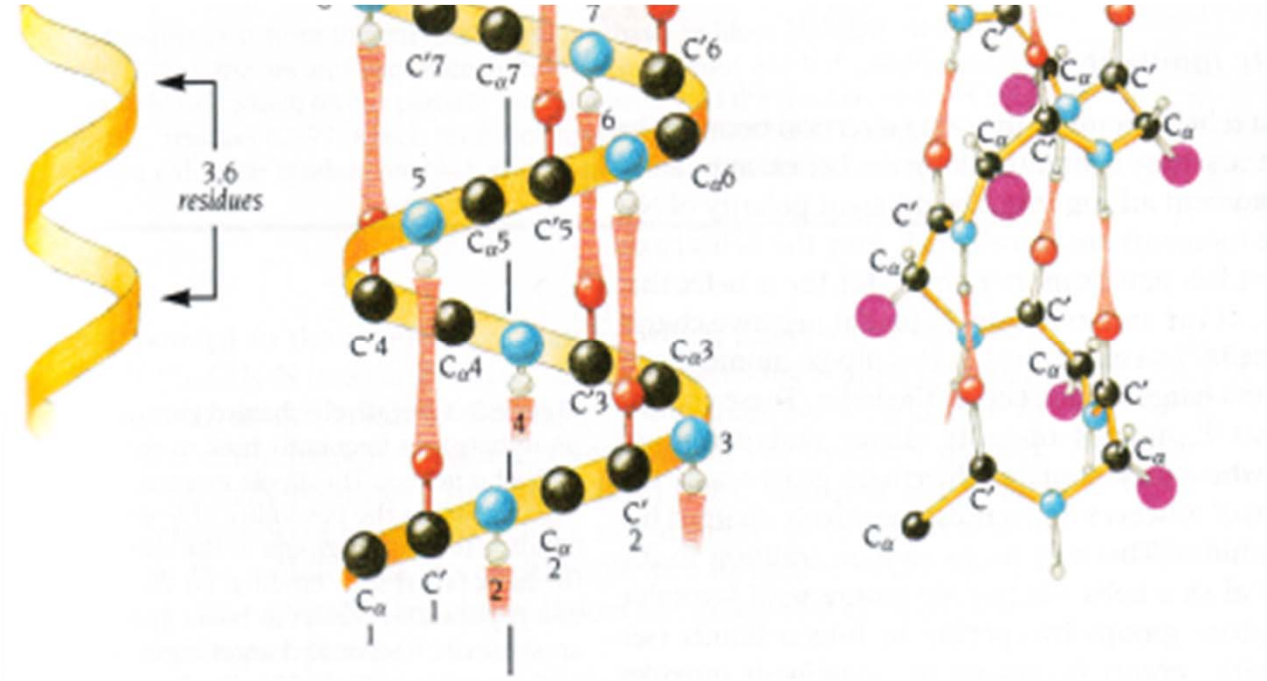
Secondary Structure of Proteins

- Regular arrangements of amino acids
- Allow hydrogen bonding propensities of backbone N-H and C=O groups to be satisfied in the absence of water
- Held together by hydrogen bonds

  Two major classes:

  - Alpha helices
  - Beta sheets

α-Helix



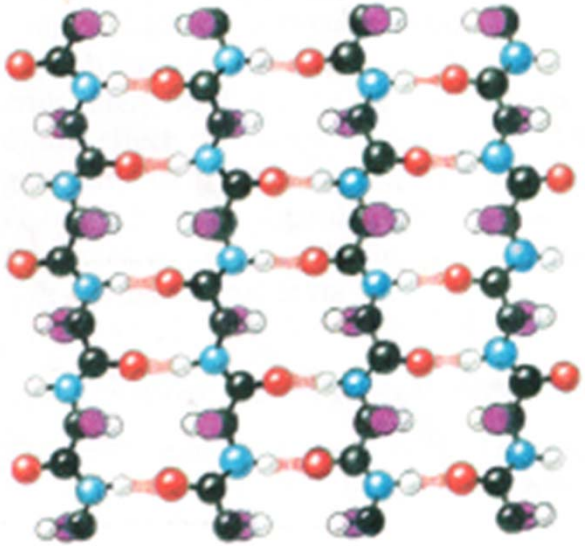
• α-Helix conformation was first predicted by Linus Pauling to be energetically favorable and stable in protein.

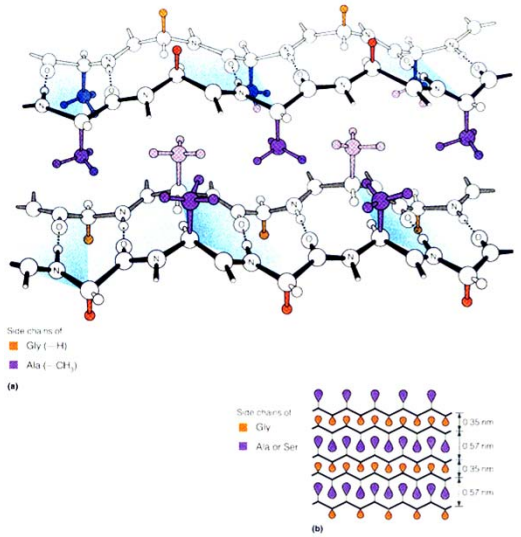
• α-Helices are found in proteins when a stretch of consecutive aa residues all have ϕ. φ angle pair approximately -60° and -50° (upper portion of bottom left quadrant in Ramachandran plot).

α-Helix

- An α-helix has 3.6 residues per turn with hydrogen bonding between C=O of residue n and NH of residue n+4
- In an α-helix, hydrogen bonding potential of all groups is satisfied except for first NH and last CO group.This makes the ends of α-helix polar and consequently are always found on the surface of proteins.
- α-Helix varies from more than 40 residues to 5 residues though average of 10 residues (three turns) are seen.The rise per residue of an α-helix is 1.5 angstrom

β-sheet

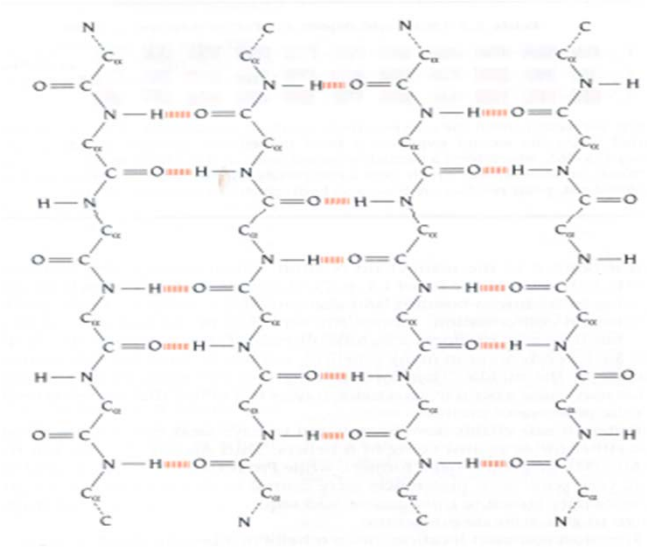




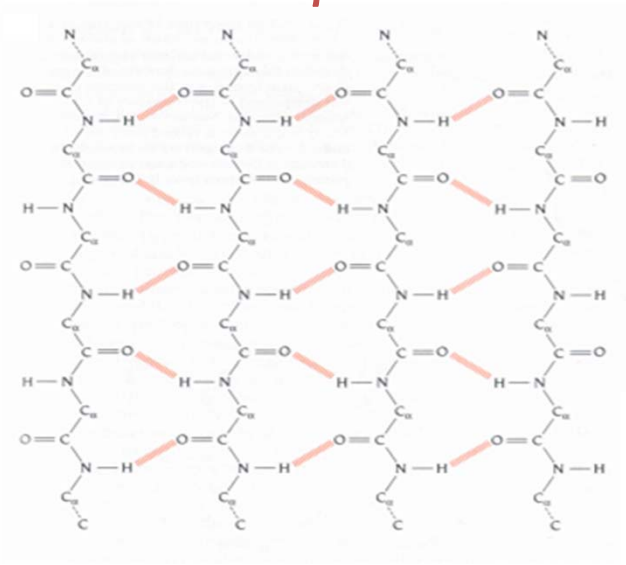
- β-sheet is composed of stacks of β-strands which are usually from 5 to 10 residues long and are aligned adjacent to each other allowing hydrogen bonding between C'=O group of one β-strand and NH group on an adjacent β-strand and vice versa.
- The β-sheet is said to be pleated because Cα successively are little below and little above the plane of β-sheet.
- The side chains follow the pattern of Cα atoms and therefore they point alternately above and below the β-sheet.

β-sheet

Anti-parallel β-sheet



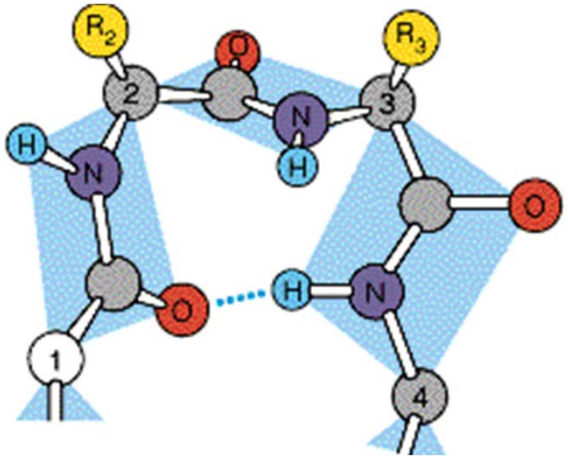
Parallel β-sheet



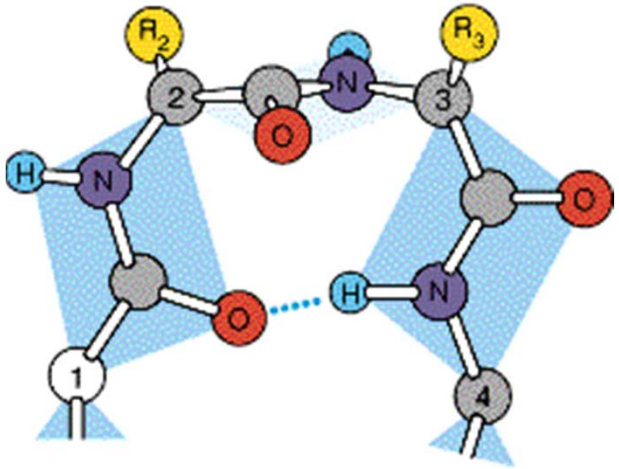
- β-Strands interacts with other β-strands running in same direction or opposite direction to form parallel or anti- parallel β-sheet respectively.
- The anti-parallel β-sheet has a narrowly spaced hydrogen bonding alternating with widely spaced hydrogen bonding.
- Parallel β-sheet have evenly spaced hydrogen bonding but at an angle.

Loop regions connectors of secondary structures

Hairpin loop Type I



Hairpin loop Type 2



- Loop regions are often seen as connectors of secondary structures
- The loop regions that connect the two anti parallel adjacent β-strands are known as hairpin loops or reverse turns.
- Loop regions are seen at the surface of the molecule therefore are rich in hydrophilic amino acids.