- Unit 1.2

- Primary Structure:
  - Building blocks of proteins are 20 naturally occurring amino acids
  - Each molecules has NH_2, COOH and C_α which is attached to a hydrogen and a side chain group (R)
  - Amino acids can group into several categories based on chemical and physical properties of the side chains
  - Side chain groups can be divided into:
  - Small
    - Non polar
      - Glycine
      - Alanine
      - Proline
    - Polar
      - Cysteine
      - Serine
      - Theronine
  - Large
    - Polar
    - Glutamine
    - Asparagine
  - Polar (Hydrophillic Positively charged basic)
    - Arginine
    - Lysine
    - Histidine

- ➔ Polar (Hydrophillic Negatively charged acidic)
  - ➔ Glutamate
  - ➔ Aspartate
- ➔ Nonpolar (Hydrophobic Aliphatic)
  - ➔ Isoleucine
  - ➔ Luecine
  - ➔ Methionine
  - ➔ Valine
  - ➔ Nonpolar (Hydrophobic Aromatic)

- ➔ Peptide formation
  - ➔ Two amino acids covalently joined together between the carboxyl group of one and amino group of another is called peptide formation
  - ➔ A linear polymer of more than 50 amino acids is called a polypeptide
  - ➔ The end of the peptide containing the amino groups is known as the N-terminus, the end with the carboxly group is known as the C-terminus
  - ➔ Atoms that form the peptide bond are called backbone atoms
- ➔ Dihedral Angles
  - ➔ Peptide bond is a partial double bond because of the shared electron between O=C-N atoms
  - ➔ Rigid double bond forces the atoms associated with the peptide bond to lie in the same plane
  - ➔ Due to this there is a restriction on the rotational freedom of the bonded pairs of atoms
  - ➔ The angle of rotation of this bond is known as the dihedral angle
  - ➔ The dihedral angle along the N-C bond is called phi
  - ➔ Dihedral angle along C-C bond is called psi
  - ➔ The combinations of these bonds allow the protein to fold in many different ways

- Ramachandran Plot
  - Rotation of phi and psi is not free, hence there is only a limited range of peptide conformation
  - When there angles are plotted against each other the resulting diagram is Ramachandran plot
  - This plot maps the entire conformational space of the peptide and shows sterically allowed and disallowed regions
  - Useful to evaluate quality of proteins
- Heirarchy of protein structures
  - Protein strutures are organized into four levels
    - Primary
      - Simplest form
      - Linear amino acid sequence
      - Held together through peptide bonds
    - Secondary
      - Local conformation of peptide chain
      - Highly regular and repeated arrangment of amino acid residues
      - Stabilized by hydrogen bonds
      - Hydrogen bonds are present between main chain atoms of the C=O group and NH group
    - Tertiary
      - Three dimensional assembly of all amino acids of a single polypeptide chain
    - Quaternary
      - This refers to the association of several polypeptide chains intro a protein complex
    - Super Secondary
      - Intermediate between secondary and tertiary

- ➔ Defined as two or three secondary structural elements forming a unique functional domain

➔ Stabilizing forces

- ➔ Secondary to quaternary structures are maintained by non-covalent forces
- ➔ Noncovalent forces:
  - ➔ Electrostatic interactions
    - ➔ Significant stabilizing force
    - ➔ Occur when excess negative charge in one region is neutralizied by positive charge in another region
    - ➔ Result in formation of salt bridges
    - ➔ Can function within relatively long range
    - ➔ Forms withing range of 10 angstorm
  - ➔ van der Waals forces
    - ➔ These are instantaneous interactions between atoms when they become transient dipoles
    - ➔ Transient dipole can induce another transient dipole nearby
    - ➔ The dipoles moment ocillates resulting in an attractive force
    - ➔ Weaker than electrostatic and hydrogen bonds
  - ➔ Hydrogen bonding
    - ➔ Similar to dipole-dipole interactions involving hydrogen from one residue and oxygen from another
    - ➔ Can occur between main chain atoms and side chain atoms
    - ➔ Hydrogen donor from N-H is slightly positive and oxygen from C=O is slightly negative
    - ➔ Dominant factor in detemining different types of protein secondary structure
    - ➔ Forms withing range of 3 angstrom
  - ➔ In addition there are Disulfide bridges

- ➔ They are bonds between sulfur atoms and cysteine residue

➔ Secondary structures

- ➔ Local structures with regular conformations
- ➔ Stabilized by hydrogen bonds
- ➔ Chief elements are
  - ➔ Alpha Helix
    - ➔ Main chain backbone conformation
    - ➔ Are all right handed helixes
    - ➔ 3.6 amino acids per helical turn
    - ➔ Stabilized by hydrogen bonds
    - ➔ Average phi angle is 60 and psi angle is 45
    - ➔ Visible in narrowly defined region lower left region of Ramachandran plot
    - ➔ Hydrophobic residues face inside, hydrophillic residures face outside
    - ➔ Every third residue is hydrophobic
  - ➔ Beta Sheet
    - ➔ Fully extended configuration built from several spatially adjacent regions
    - ➔ Each regions is formed by beta strands
    - ➔ Beta strands
      - ➔ It is pleated, main chain backbone is zigzagginhg and side chains are alternativly on opposite sides
      - ➔ They are stabilized by hydrogen bonds
      - ➔ Beta strands near the surface are alternatively hydrophobic and hydrophillix
      - ➔ Beta strands at the core are all hydrophobic
      - ➔ Running in same direction = parallel sheet
      - ➔ Running in opposite diretion = antiparallel sheet
    - ➔ Psi and phi bonds are distributed in the upper left region in Ramachandram

plot

- ➔ It is more difficult to predict beta sheet than alpha helix

➔ Coils and loops

- ➔ They are local structures that do not belong to regular secondary structure
- ➔ Loops
  - ➔ Sharp turns
  - ➔ Hairpin structures
  - ➔ Irregular connecting regions = random coils
- ➔ Residues in loop or coil regions are charged and polar and location on the surface

➔ Coiled coils

- ➔ They are a special type of supersecondary structure
- ➔ Have a bundle of two or more alpha helices wrapping around each other
- ➔ Hydrophibicity repeats every seven residues (5 hydrophobic and 2 hydrophillic)

Potential energy maps

- ➔ Potential energy related to structural arrangement of atoms or molecules
- ➔ Potential energy surface
  - ➔ This describes the energy of a system, especially a collection of atoms, generally the positions of the atoms
  - ➔ It can be used to explore properties of structures composed of atoms for eg:
    - ➔ Finding the minimum enerft shape of a molecule
    - ➔ Computing the rate of a chemical reaction
- ➔ They illustrate the charge distribution of molecules three dimensionally
- ➔ Allow use to visualize variably charged regions of a molecule

- ➔ Electrionstatic potential maps are invaluable to predicting the behaviour of complex molecules

Coordinate systems:

- ➔ It is a system that uses one or more numbers to uniquely determine the position of the points or other geometric element
- ➔ In 3D space catesian coordinate system is used:
  - ➔ This has three mutually perpendicular axes:
    - ➔ X axis
    - ➔ Y axis
    - ➔ Z axis

Anfinsen Experiment

- ➔ Created by Christian Anfinsen
- ➔ Proposed that the 3D structure of a protein is a direct consequence of its primary structure

IM DONE MAKING NOTES. ABHI READ KAREGA

Protein secondary structure prediction

- ➔ Plays a key role in its function, if protein loses its shape it will no longer be functional
- ➔ 50% fold intro alpha helix and other intno beta sheet
  - ➔ look up for alpha helix and beta sheet im not writing again
- ➔ Secondary structure prediction refers to prediciton of the conformational state of each amino acid residue of a protein sequence based on ther three possible states
  - ➔ Helix
  - ➔ Strand
  - ➔ Coil
- ➔ Prediction is based on the fact that secondary structures have regular arrangements of amino acids
- ➔ This structural regularity serves as fundation of prediction algorithms
- ➔ There are three generations of seconday structure prediction
  - ➔ Each new generation has 10% higher accuracy then methods from previous generation

  - ➔ First generation
    - ➔ Primary prediciton methods are based on two approached
      - ➔ Physico and stereo-chemical analyses
        - ➔ A set of rules were deribed from this approach by using helices and strands
      - ➔ Statistical method of prediction
      - ➔ Prediction rules were based on principles that govern secondary structure prediction
      - ➔ These methods achieved 56% accuracy
      - ➔ They were based on single residue statistics
      - ➔ Chou Fasman

- ➔ Frequencies from the alpha helices, beta sheets and beta turns were computed from 14 amino acids
- ➔ This information was used to predict statistically the secondary structures in other proteins with 50% accuracy

➔ GOR

- ➔ The prediction is by assigning each residue in the primary sequence 4 states
  - ➔ Alpha helix
  - ➔ Beta sheet
  - ➔ Beta turn
  - ➔ Loop
- ➔ Completely determined statistically by the residues with same primary sequence
- ➔ This had 57% accuracy

➔ Second Generation

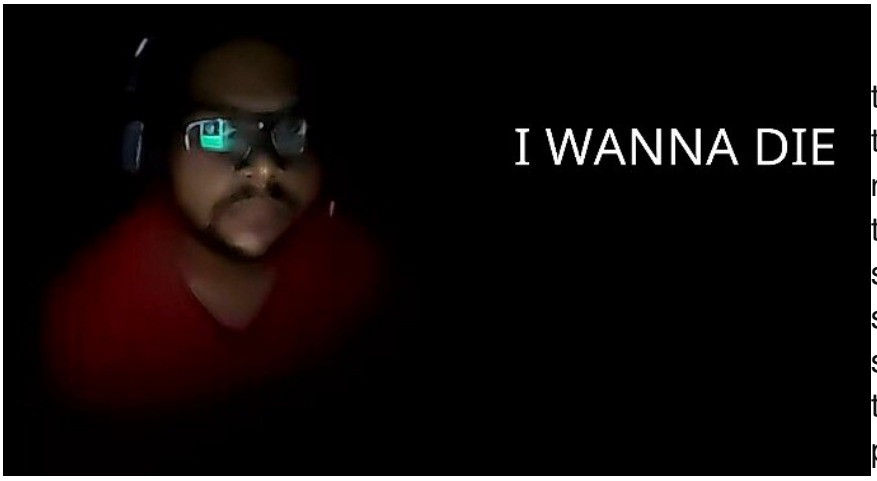
- ➔ Main improvement from second generation was by increase in the size of the database as well as statistics based on segments
- ➔ Accuracy levels were 60%
- ➔ Algorithms used were
  - ➔ Physico-chemical properties
  - ➔ Sequence patterns
  - ➔ 3 layered neural networks
  - ➔ graph theory
  - ➔ Multivariate statistics
  - ➔ expert rules
  - ➔ nearest neighbour algorithms
- ➔ New version of GOR had 63% accuracy

- ➔ Third Generation
  - ➔ Methods are based on sequence homology
  - ➔ In homologous prediction, database is scanned to detect homology between segments of query and template sequences
  - ➔ Prediciton is based on nearest-neighbor algorithm
  - ➔ This method has an averafe of 60% accuracy
- ➔ Application of protein structure prediction
  - ➔ Prediction models are already been used for durg screening
  - ➔ Used as a screening process for proteins that are no freely available for experimentation
  - ➔ Also can we used for protein engineering to reach a level of accuracy where the predicted strucutre can be used to find binding partners

Tertiary structure prediction:

- ➔ This aims to predict 3D shape of protein molecules by describing spatial disposition of each atom
- ➔ There are methods to resolve molecular structure with high precision but they are time and resource consuming
- ➔ Computation based software techniques can predict tertiary structure of protein with acceptable precision with high efficiency
- ➔ Solving for protein takes 1 to 3 years
- ➔ some proteins cant be resolved by Xray and NMR and sequence data for many important proteins is available by the 3D structure is unknown
- ➔ It is necessary to obtain approximate protein structure through computer modelling to gain full understanding of their biological roles
- ➔ There are three computational approaches to protein 3D structural prediction
  - ➔ Homology Modeling
    - ➔ Predicts protein structures based on sequence homology with known structures

- Called comparative modeling
- It consists of six steps
  - Template selection
    - Template selection involves searching PDB for homologous proteins
    - Search can be performed by using a heuristic alignment program such as BLAST or FASTA
    - Template should have at least 30% sequence identity to the query sequence
    - If multiple database structures with significant similarity can be found as a result of the search
    - The structures with the highest identity percentage and highest resolution should be chosen as template
  - Sequence Alignment
    - Template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment
    - The realignment is the most important step
    - Errors made in this step cant be corrected later
    - Best possible alignment algorithms such as Praline and T-Coffee should be used for this purpose
  - Backbone model building



    - Now the residues in the aligned regions of the target can ssume a similar structure as the template proteins
    - In backbone modeling it is simplest to use one template structure, best

quality and highest resolution is chosen

- ➔ This structure carries the fewest errors

- ➔ Loop modeling
  - ➔ Loop modeling is mini protein modeling by itself
  - ➔ Database method
    - ➔ finds spare parts from known protein structures in a database that fit in the gap regions
    - ➔ Usually many different alternative segments that fit are available. The best loop is selected based on sequence similarity as well as minimal steric clashes
  - ➔ Ab initio method
    - ➔ Generates many random loops and searches for the one that does not clash with nearby side chains
  - ➔ If the loop is short (three to five residues), correct models can be built
  - ➔ If loops are longers it is difficult to create a reliable model
- ➔ Side chain refinement

- ➔ Evaluation

- ➔ Threading
- ➔ Ab initia prediction

Modeller

- ➔ Computer program for comparative protein structure modelling
- ➔ Input is alignment sequences to be modelled
- ➔ Automatically calculates model containing all non hydrogen atoms
- ➔ Apart from model building it can perform auxiliary tasks
  - ➔ Fold assignment
  - ➔ Alignment of two protein sequences
  - ➔ Multiple alignment
  - ➔ Calculation of phylogenetic trees
  - ➔ De novo modelling