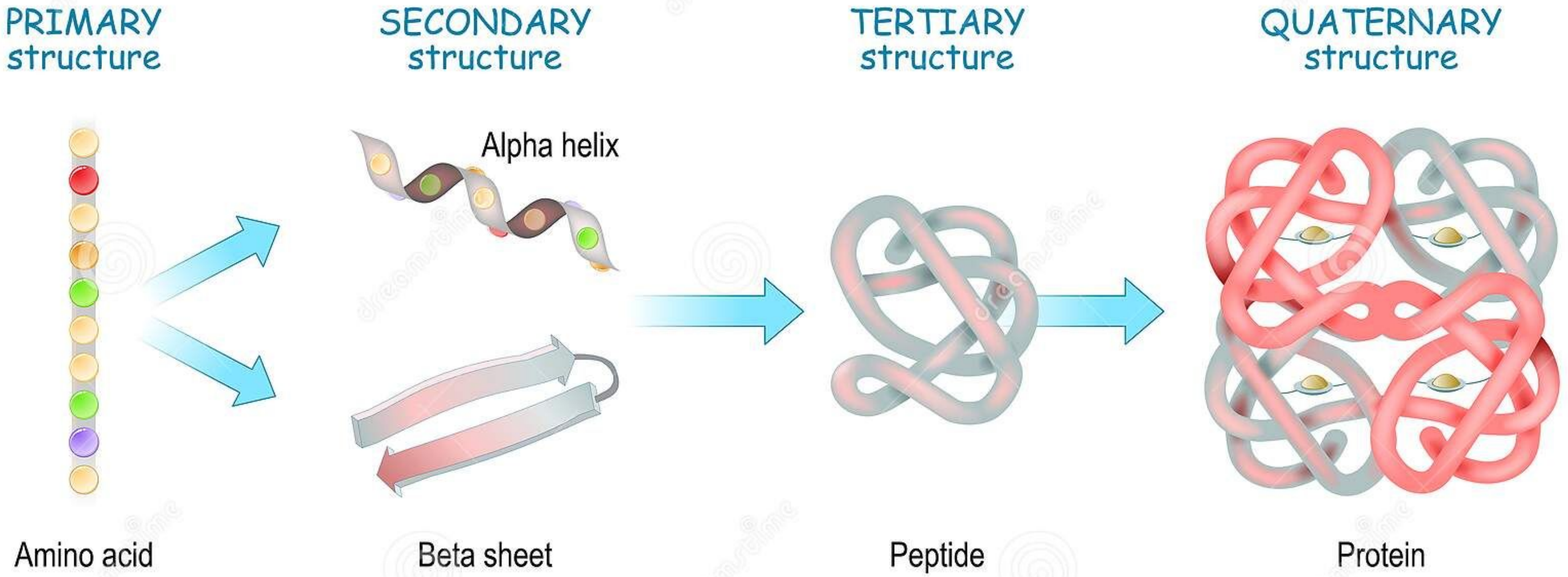
Algorithms for Protein Structure Prediction

Manu Madhavan

Outline

- Protein structures
- Alpha helix and Beta sheets
- Structure prediction algorithms

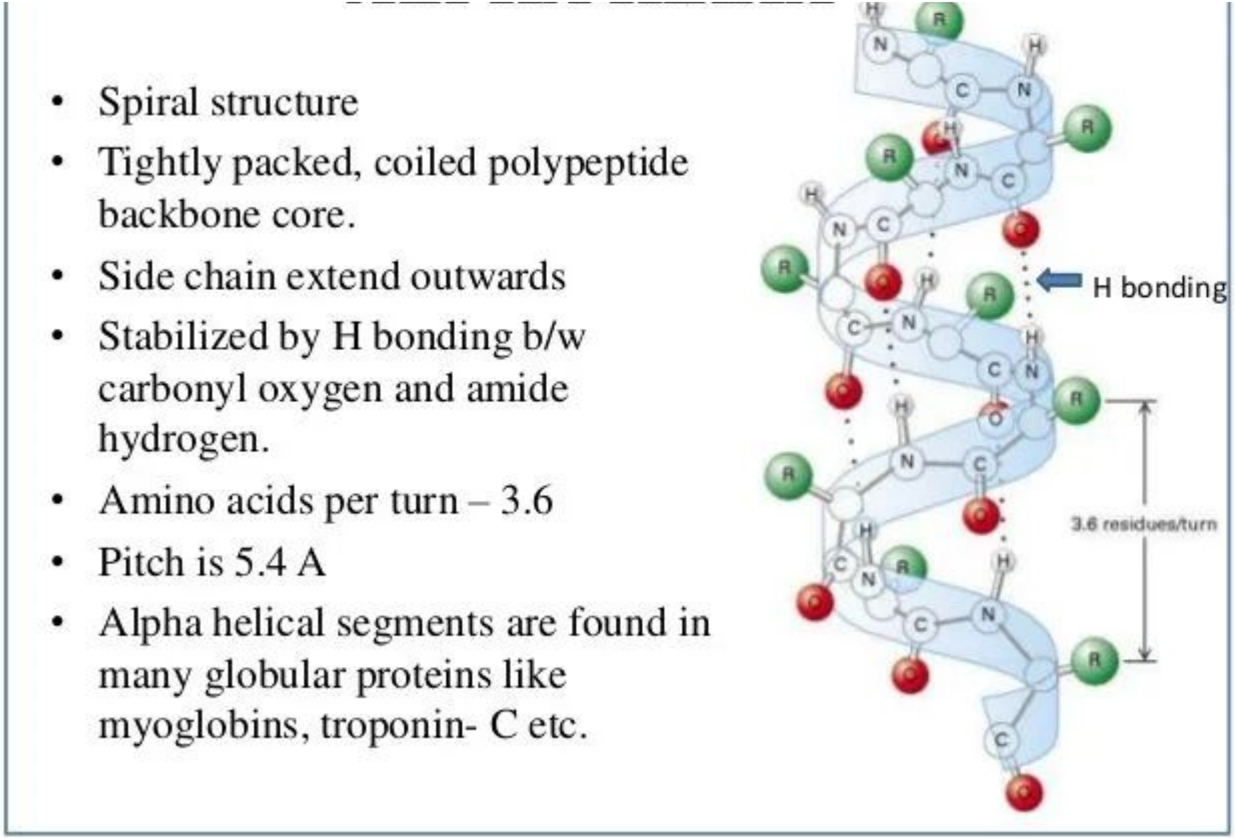
Protein structure



Protein Secondary Structures

- Regular, recurrent arrangement in space of adjacent amino acid residues in a polypeptide chain
- Maintained by hydrogen bonds between amide hydrogen and carbonyl oxygen of peptide backbone
- Commonly occurring secondary structures:
    - Alpha helices
    - Beta strands
    - Turns (bends)
    - Coil (irregular)

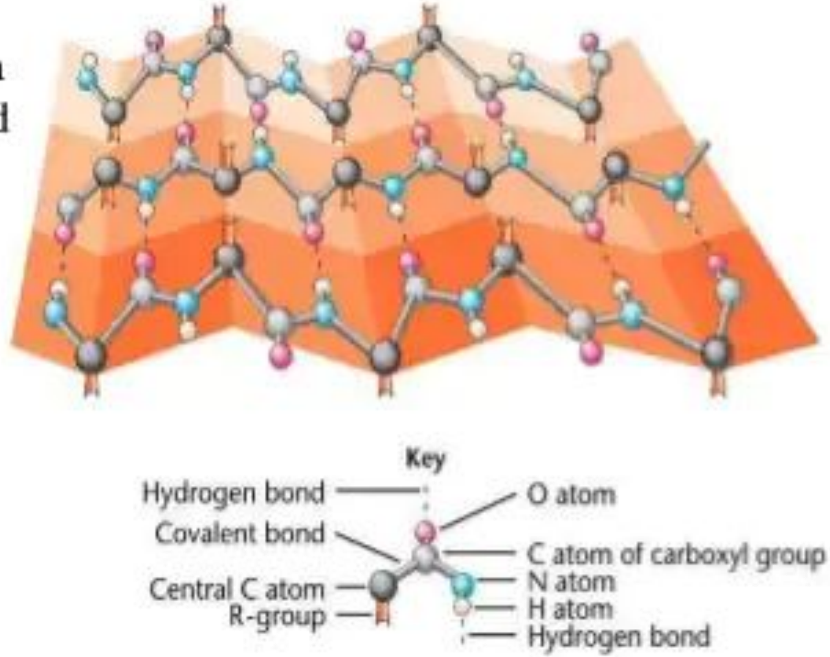
Alpha-Helix

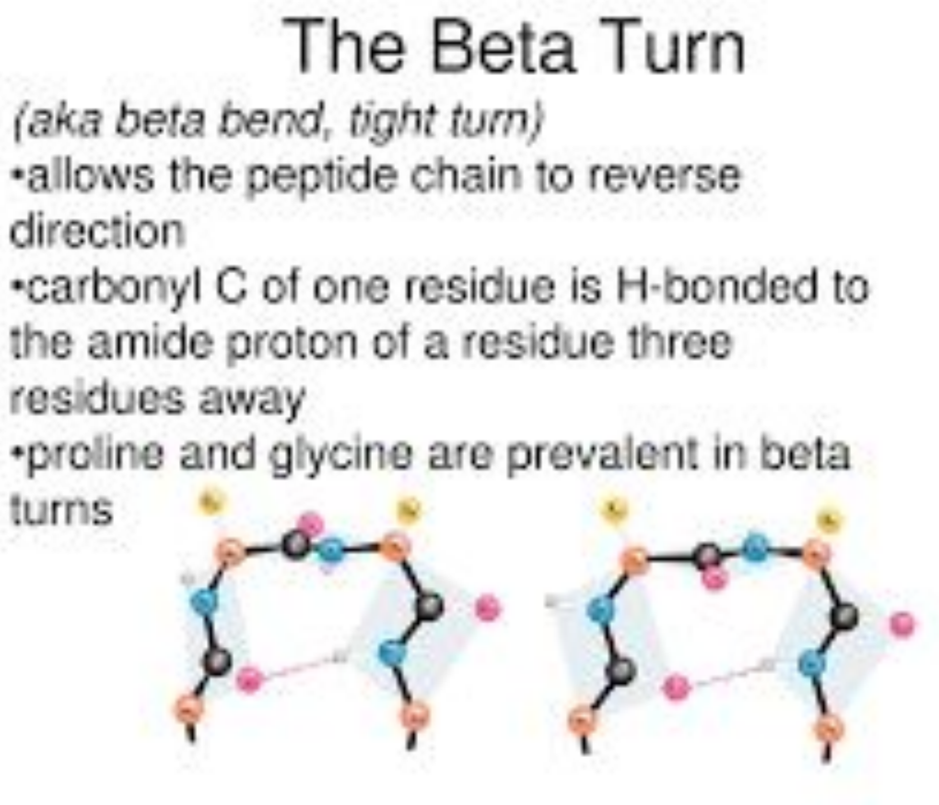


- Spiral structure
- Tightly packed, coiled polypeptide backbone core.
- Side chain extend outwards
- Stabilized by H bonding b/w carbonyl oxygen and amide hydrogen.
- Amino acids per turn – 3.6
- Pitch is 5.4 A
- Alpha helical segments are found in many globular proteins like myoglobins, troponin- C etc.

β – pleated sheet

- It is a secondary level of protein organization in which the backbone of the peptide chain is extended into a zigzag arrangement resembling a series of pleats, with the peptide bonds organized in planes of alternating slopes (alternating ascending and descending direction).
- The Beta pleated sheet can be formed between two peptide chains or between different segments of the same peptide chain.
- The large aromatic residues (tryptophan, tyrosine and phenylalanine) and Cβ-branched amino acids (isoleucine, valine, and threonine) prefer to adopt β-sheets conformations.





The Beta Turn

(aka beta bend, tight turn)

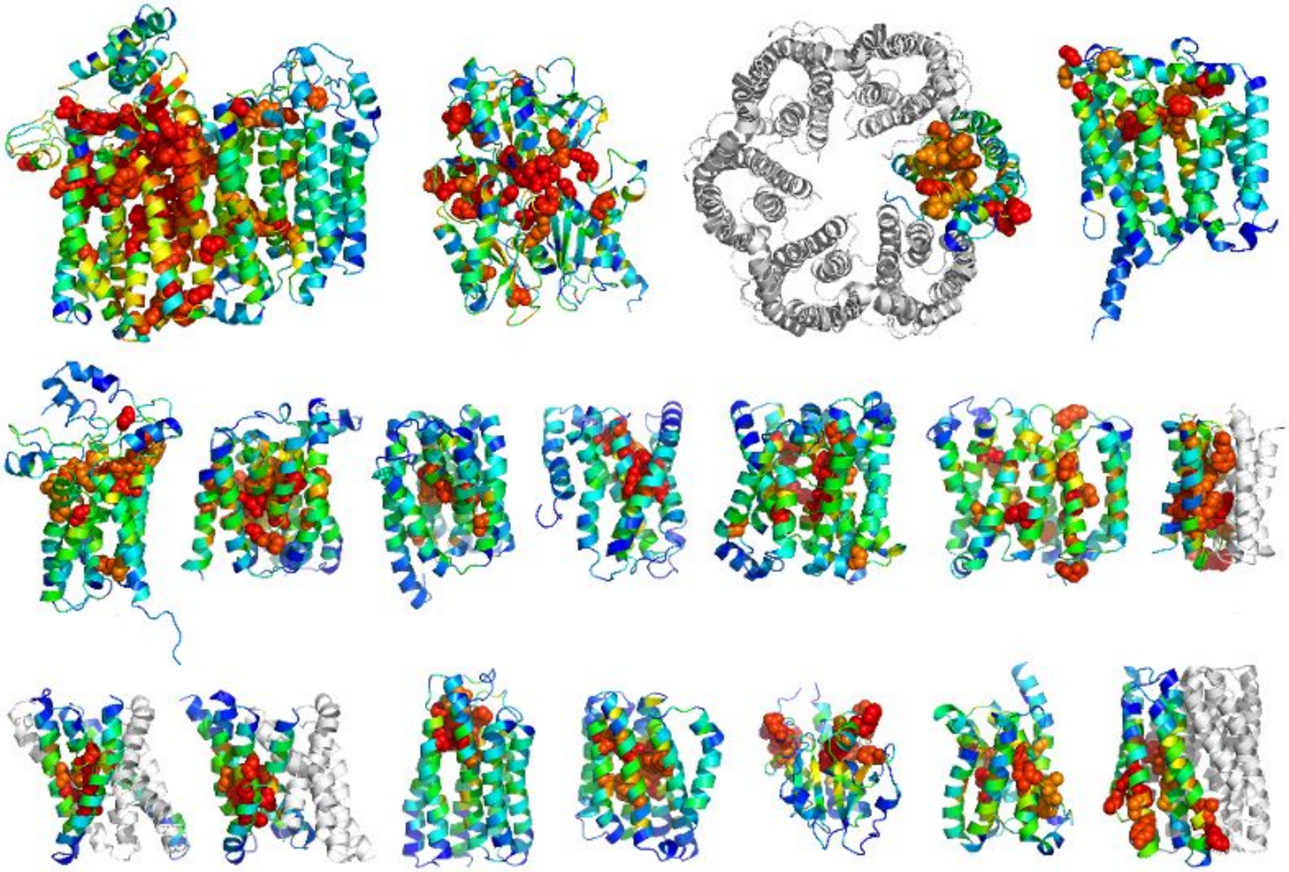
•allows the peptide chain to reverse direction

•carbonyl C of one residue is H-bonded to the amide proton of a residue three residues away

•proline and glycine are prevalent in beta turns

Protein Tertiary Structure

- Relative arrangement of secondary structure elements in 3D space
- a complete description of the precise relative position of every atom in the protein molecule constitutes the tertiary structure
- In the case of proteins whose functional form consists of only one polypeptide chain, the tertiary structure is the fully folded, final, functional structure of the protein.
- In the case of proteins made of more than one polypeptide chain, i.e. dimers, trimers, or in general multimeric proteins, there exists **quaternary structure.**



- While the fact that the primary structure specifies the secondary and the tertiary structure of a protein is a well-established one, the principles that govern this relationship are not yet fully known (exceptions are there)
- **Protein Structure Prediction:** refers to concepts and methods that address the use of computational methods to arrive at the three dimensional structure of a protein **from knowledge of its sequence alone**.
- **Protein Folding:** The study of the way in which ‘real’ proteins fold in ‘real’ time, in other words protein folding, is an important branch of experimental, theoretical and computational biophysics

Protein Secondary Structure Prediction

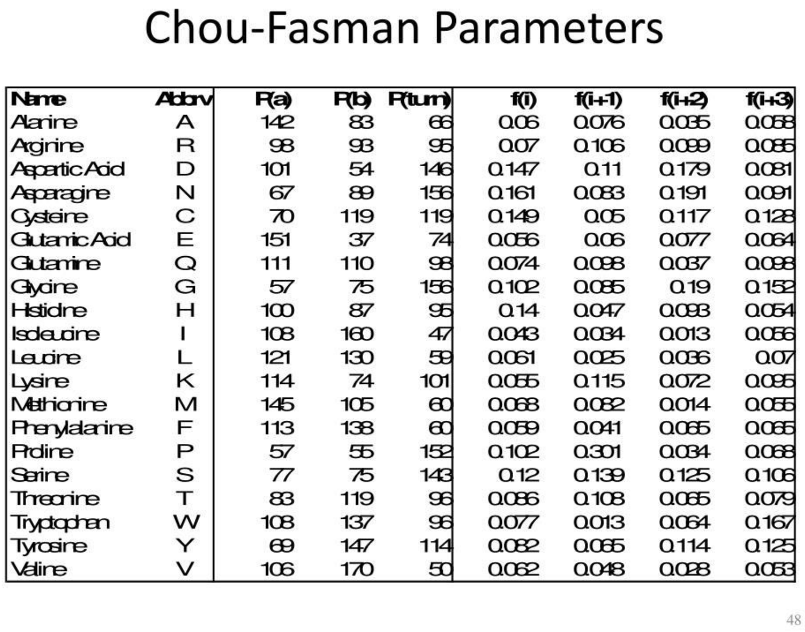
- For the purposes of prediction, every residue in a protein chain is always considered to exist in one of three (or four) secondary structural states. These are: helix, usually represented as H; beta strand,represented as B or E (for ‘extended’); and random coil, signified as C.
- The output of most secondary structure prediction algorithms and programs is the sequence of the protein along with one of the above symbols for each residue.
- secondary structure prediction is a residue-by¬residue process and not an estimate of the overall secondary structure content of the protein.

Chou–Fasman method

- Algorithm for assigning secondary structure
- The method is based on analyses of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography.
- Each amino acid is assigned several conformational parameters, **P(a), P(b), and P(turn)**- **representing the propensity** of each amino acid to participate in alpha helices, beta sheets, and beta turns, respectively, were determined based on observed frequencies in a set of sample

Chou-Fasman Algorithm

- Each amino acid is assigned four turn parameters, f(i), f(i+1), f(i+2), and f(i+3), corresponding to the frequency with which the amino acid was observed in the first, second, third, or fourth position of a hairpin turn.



Chou-Fasman Parameters

Name	Abbrv	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	A	142	83	66	0.06	0.076	0.035	0.058
Arginine	R	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	D	101	54	146	0.147	0.11	0.179	0.081
Asparagine	N	67	89	156	0.161	0.083	0.191	0.091
Cysteine	C	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	E	151	37	74	0.056	0.06	0.077	0.064
Glutamine	Q	111	110	98	0.074	0.098	0.037	0.098
Glycine	G	57	75	156	0.102	0.085	0.19	0.152
Histidine	H	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	I	108	160	47	0.043	0.034	0.013	0.056
Leucine	L	121	130	59	0.061	0.025	0.036	0.07
Lysine	K	114	74	101	0.055	0.115	0.072	0.095
Methionine	M	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	F	113	138	60	0.059	0.041	0.065	0.065
Proline	P	57	55	152	0.102	0.301	0.034	0.068
Serine	S	77	75	143	0.12	0.139	0.125	0.106
Threonine	T	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	W	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	Y	69	147	114	0.082	0.065	0.114	0.125
Valine	V	106	170	50	0.062	0.048	0.028	0.053

48

1. Identify alpha helices as follows:
   a. Find all regions where four of six contiguous amino acid residues have $P(\mathrm{a}) > 100$.
   b. For each region identified in part (a), extend the region in both directions until a set of four contiguous residues with $P(\mathrm{a}) < 100$ is encountered.
   c. For each region extended in part (b), compute $\Sigma P(\mathrm{a})$, the sum of $P(\mathrm{a})$ values for each residue in the region, and $\Sigma P(\mathrm{b})$. If the region is >5 residues in length, and $\Sigma P(\mathrm{a}) > \Sigma P(\mathrm{b})$, then the region is predicted to be an alpha helix.
2. Identify beta sheets using the same algorithm as in step 1, but search for regions where four of six residues have $P(\mathrm{b}) > 100$. Once the regions are extended (part b), a region is declared a beta strand if the average $P(\mathrm{b})$ over all residues in the region is greater than 100, and $\Sigma P(\mathrm{b}) > \Sigma P(\mathrm{a})$.

3. If any of the helices assigned in step 1 overlap a beta strand assigned in step 2, then the overlapping region is predicted to be a helix if $\Sigma P(\text{a}) > \Sigma P(\text{b})$, and a strand if $\Sigma P(\text{b}) > \Sigma P(\text{a})$.

4. Finally, identify beta turns as follows:

    a. For each residue, *i*, calculate the turn propensity, *P*(t), as follows: *P*(t) = the $f(i)$ of the residue i + the $f(i{+}1)$ value of the following residue + the $f(i{+}2)$ value of the subsequent residue (position $i + 2$) + the $f(i{+}3)$ of the residue at position $i + 3$.

    b. Predict a hairpin turn starting at each position, *i*, that meets the following criteria:

        i. $P(\text{t}) > 0.000075$

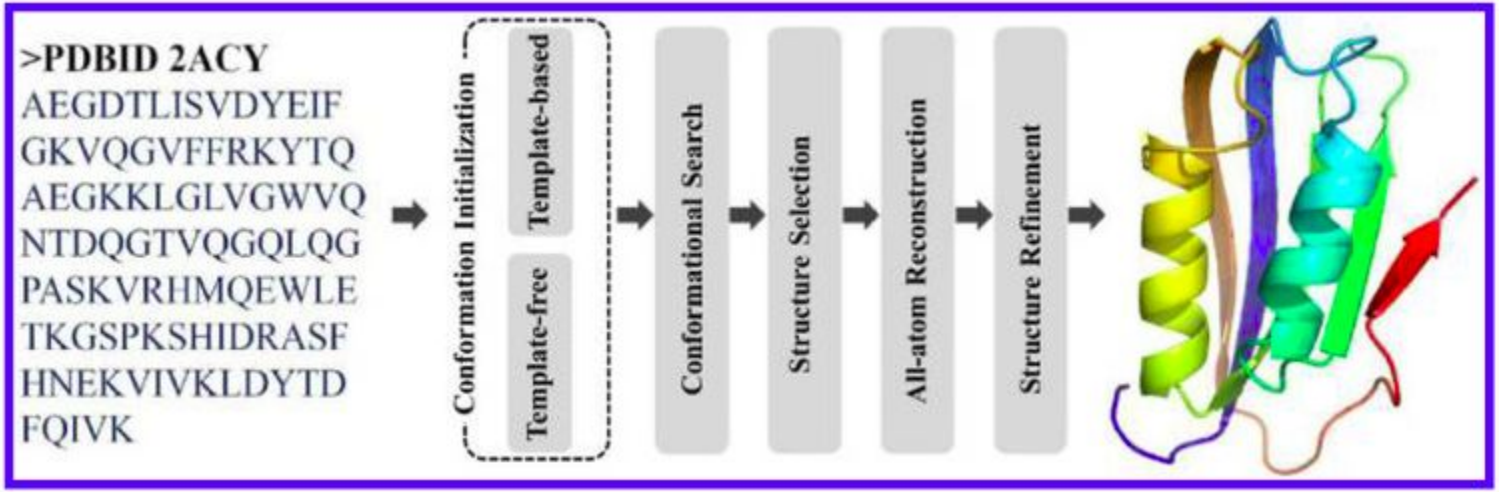
        ii. The average *P*(turn) value for the four residues at positions *i* through $i + 3 > 100$

        iii. $\Sigma P(\text{a}) < \Sigma P(\text{turn}) > \Sigma P(\text{b})$ over the four residues in positions *i* through $i + 3$.

Protein Folding Prediction

- Predicting the possible confirmation (folding) from Amino acid sequence and secondary structure
- Analysing homology with PDB templates
- Ab intio method use Physical/chemical properties to predict folding - computationally complex- Need HPC/GPU computations

The general flowchart of protein structure prediction.



>PDBID 2ACY
AEGDTLISVDYEIF
GKVQGVFFRKYTQ
AEGKKLGLVGWVQ
NTDQGTVQGQLQG
PASKVRHMQEWLE
TKGSPKSHIDRASF
HNEKVIVKLDYTD
FQIVK

Conformation Initialization

Template-based

Template-free

Conformational Search

Structure Selection

All-atom Reconstruction

Structure Refinement