

Protein structure prediction



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Dr. Jaspreet Kaur Dhanjal

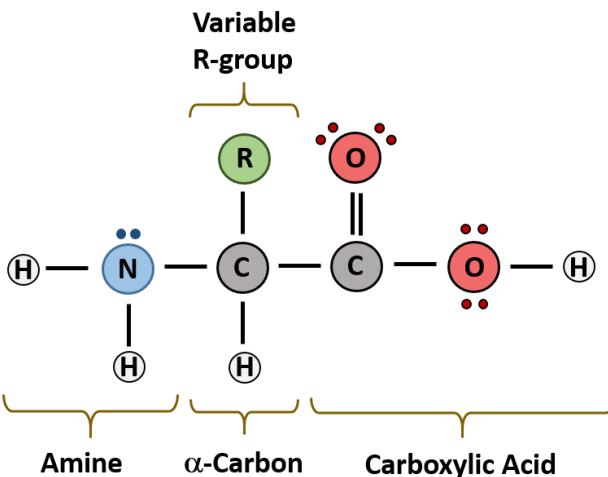
Assistant Professor, Department of Computational Biology

Email ID: jaspreet@iiitd.ac.in

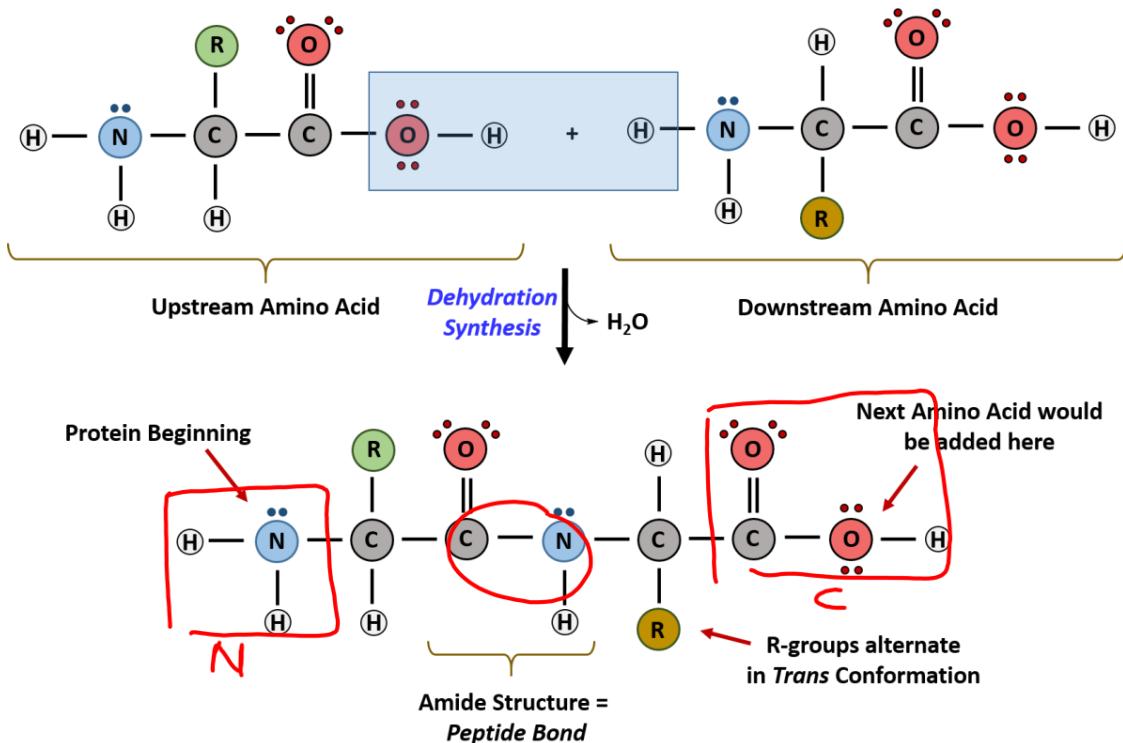
October 03, 2025

Amino acids, the building blocks of protein

Basic structure of an amino acid

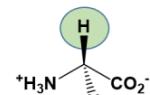


Formation of peptide bond

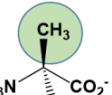


Different types of Amino acids

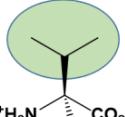
Nonpolar (Hydrophobic) Amino Acids



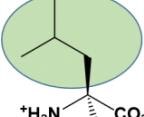
Glycine
Gly, G



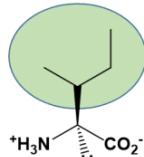
Alanine
Ala, A



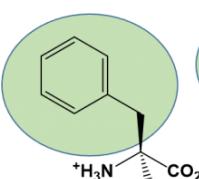
Valine
Val, V



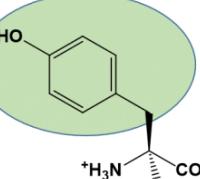
Leucine
Leu, L



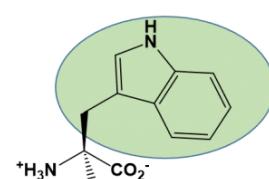
Isoleucine
Ile, I



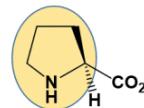
Phenylalanine
Phe, F



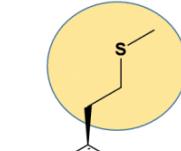
Tyrosine
Tyr, Y



Tryptophan
Trp, W

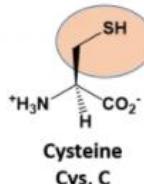


Proline
Pro, P

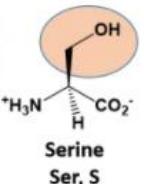


Methionine
Met, M

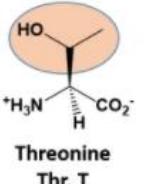
Polar (Hydrophilic) Amino Acids



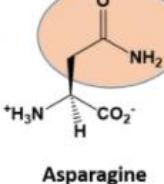
Cysteine
Cys, C



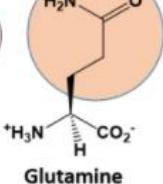
Serine
Ser, S



Threonine
Thr, T

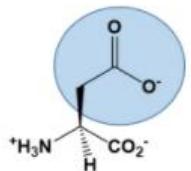


Asparagine
Asn, N

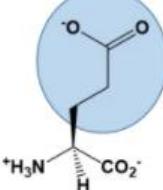


Glutamine
Gln, Q

Acidic Amino Acids

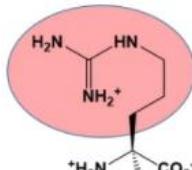


Aspartic Acid
Asp, D

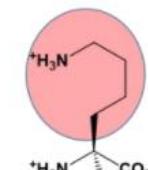


Glutamic Acid
Glu, E

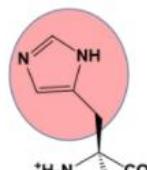
Basic Amino Acids



Arginine
Arg, R



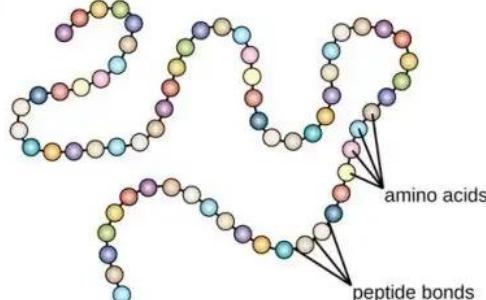
Lysine
Lys, K



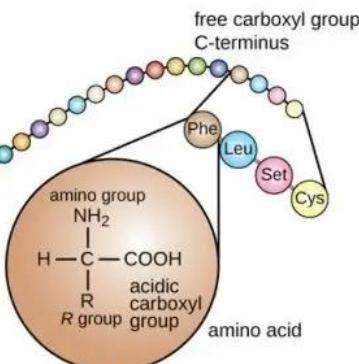
Histidine
His, H

Structure of protein

free amino group,
N-terminus



The primary protein structure
is the chain of amino acids
that makes up the protein.

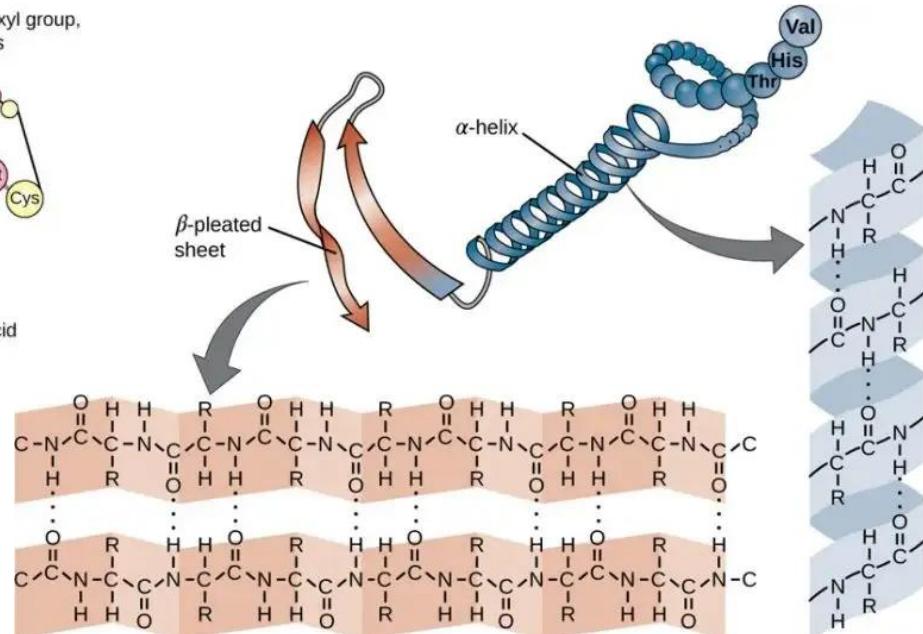


1. Primary structure of protein

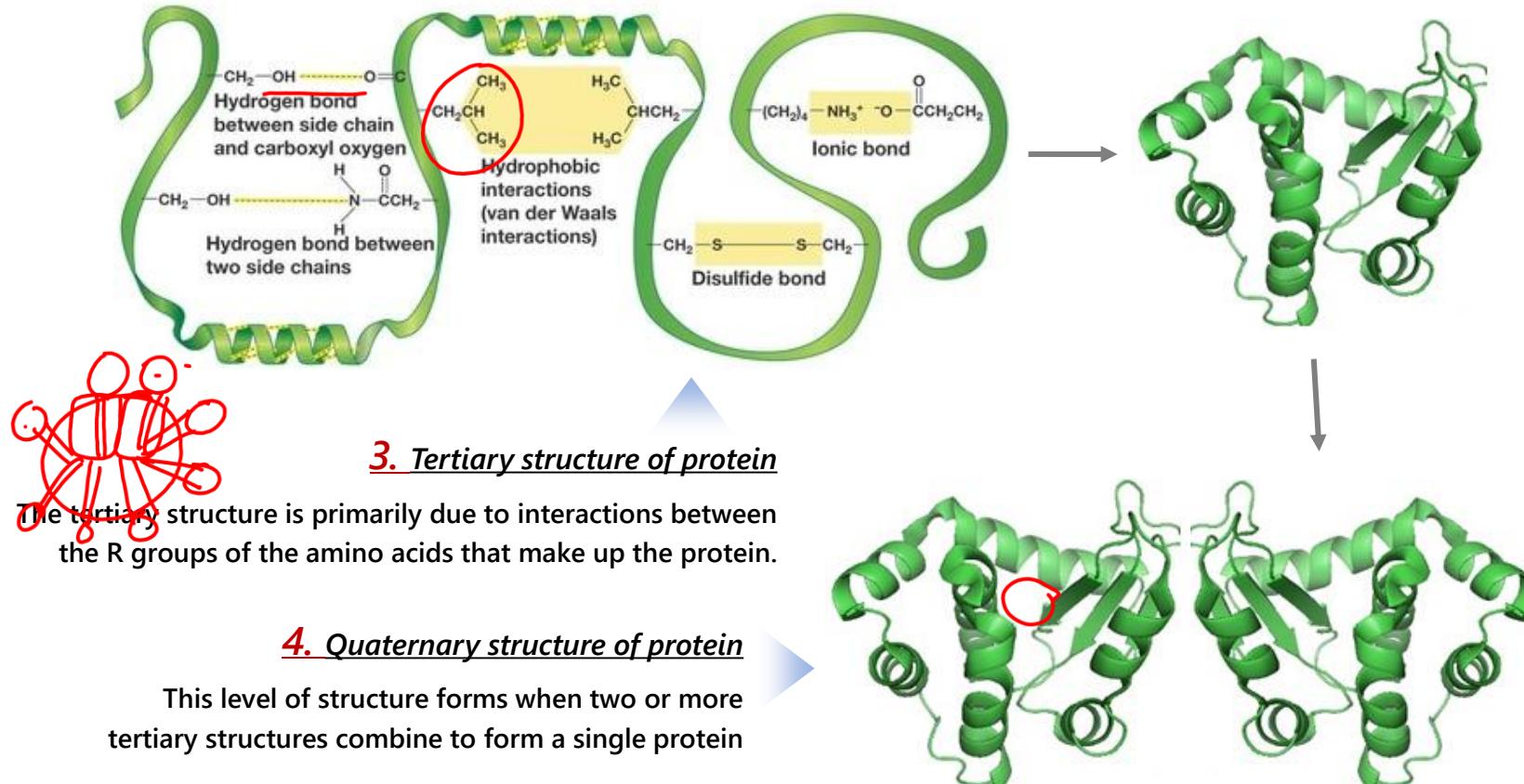
This level of structure is determined by the sequence
of amino acids that join to form a polypeptide.

2. Secondary structure of protein

Hydrogen bonding between amino acids cause the
polypeptide to form an alpha helix or a pleated sheet.



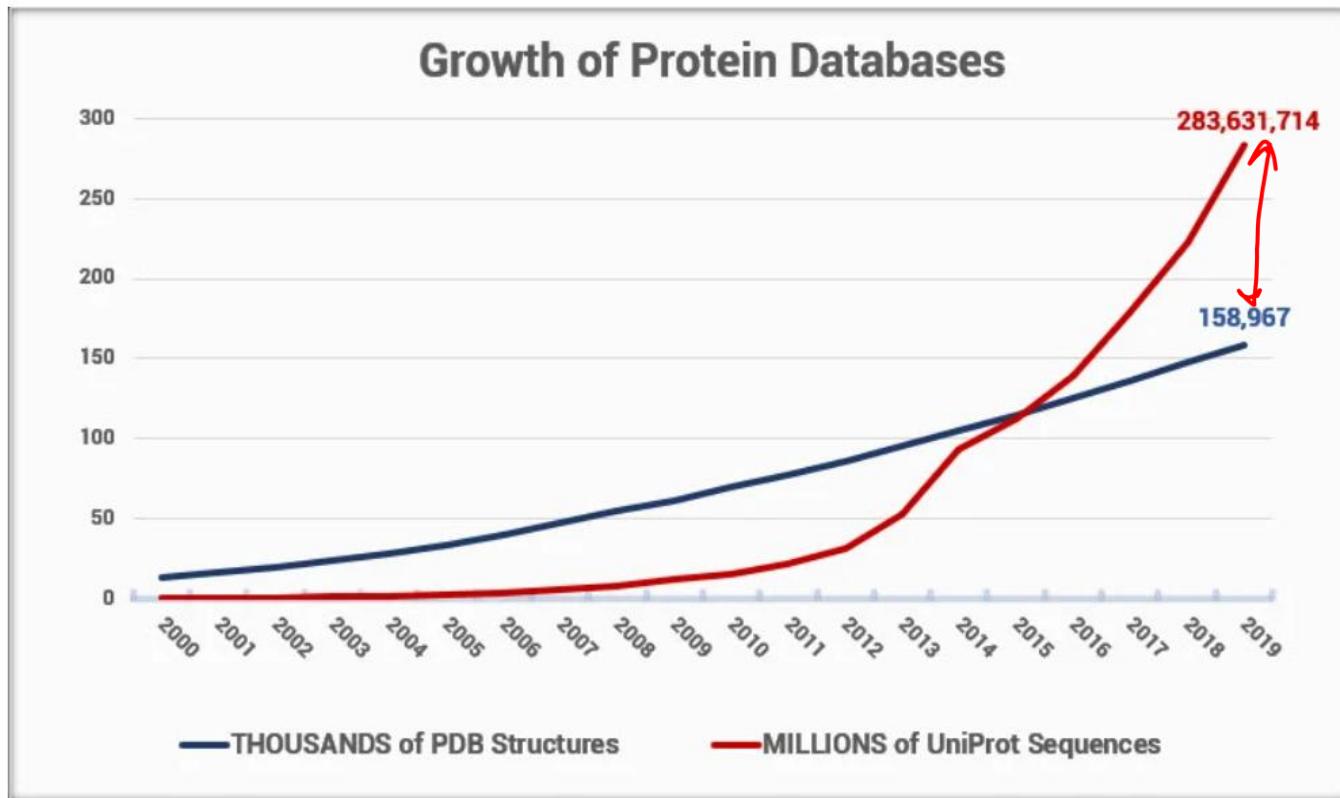
Structure of protein



Importance of protein structure prediction

- A protein's biological function is dictated by the arrangement of the atoms in the three-dimensional structure.
- This could be the arrangement of catalytic residues in an active site or how a protein interacts with other proteins for structural or other regulatory purposes.
- Having a protein structure provides a greater level of understanding of how a protein works, which can allow us to create hypotheses about how to affect it, control it, or modify it.
- For example, knowing a protein's structure could allow to design site-directed mutations with the intent of changing function.
- Or you could predict molecules that bind to a protein for developing its inhibitors.

Gap between known proteins and structures solved



2025/10/03

Swiss-Prot
(573,661)
TrEMBL
(253,061,697)
PDB
(242,874)

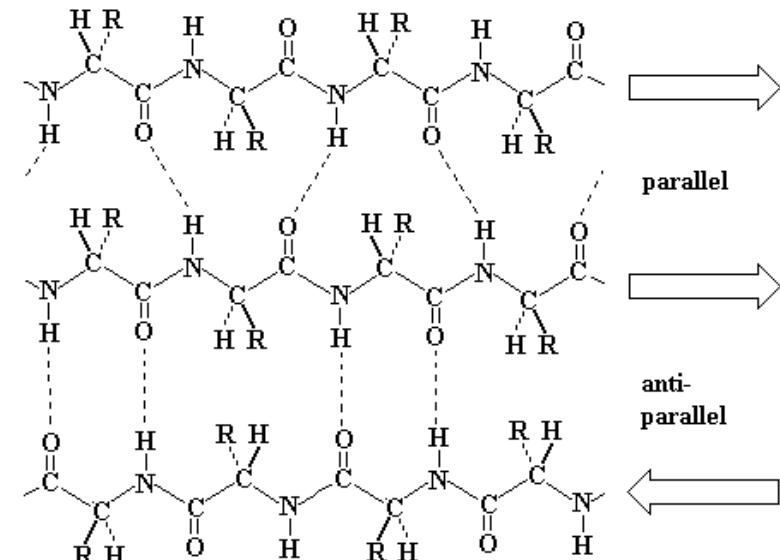
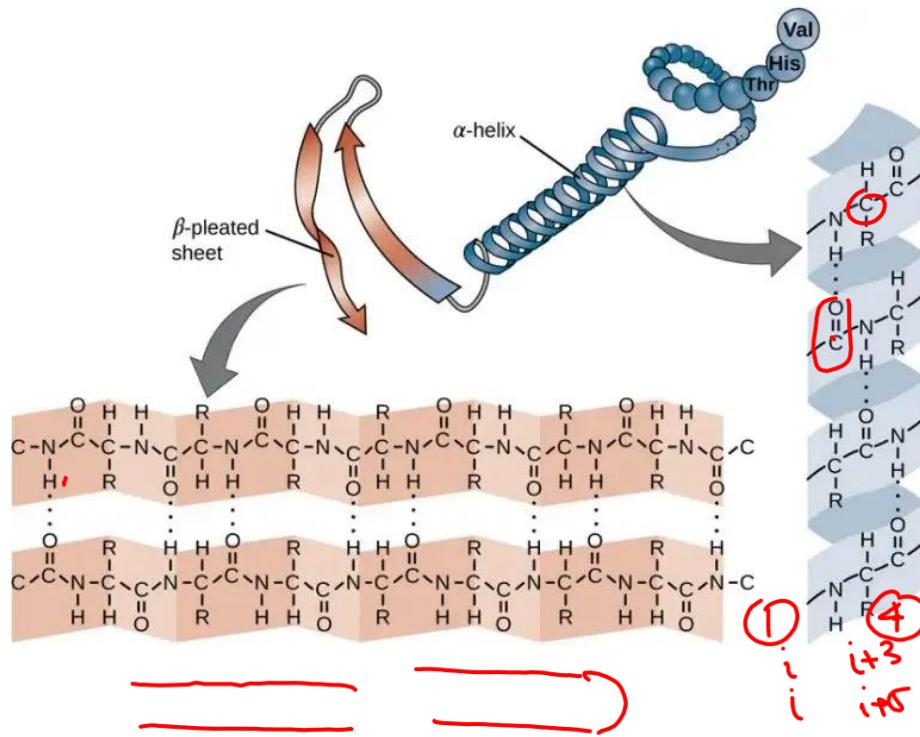
Figure 1. Growth of protein sequence and structure databases over time

Prediction of protein structure

The structure of protein is predicted at two different levels:

1. Secondary structure prediction
2. Tertiary structure prediction

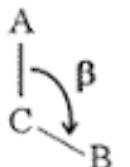
Hydrogen bonds for secondary structure assignment



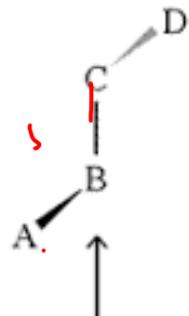
Hydrogen bonding pattern

Dihedral bonds for protein backbone conformation

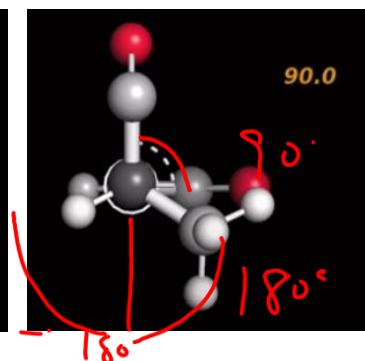
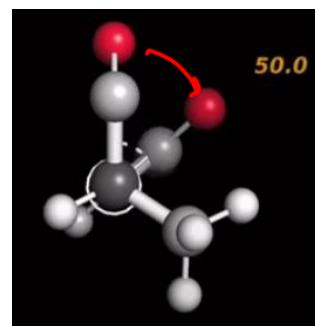
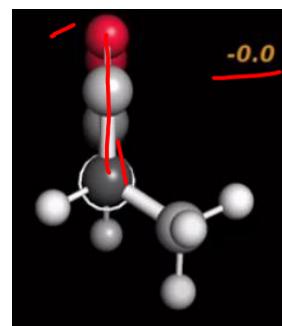
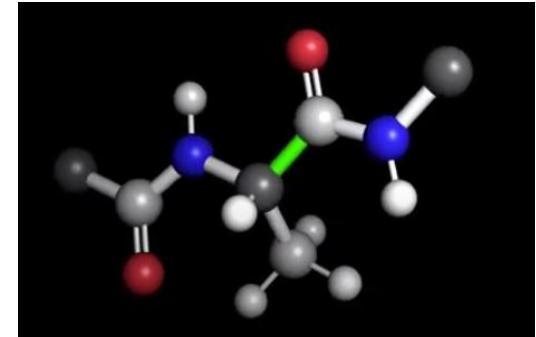
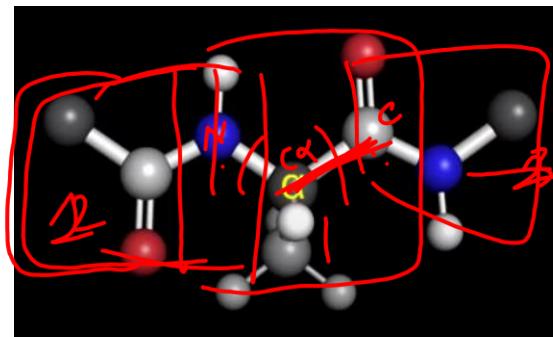
Dihedral/torsional angles



bond angle

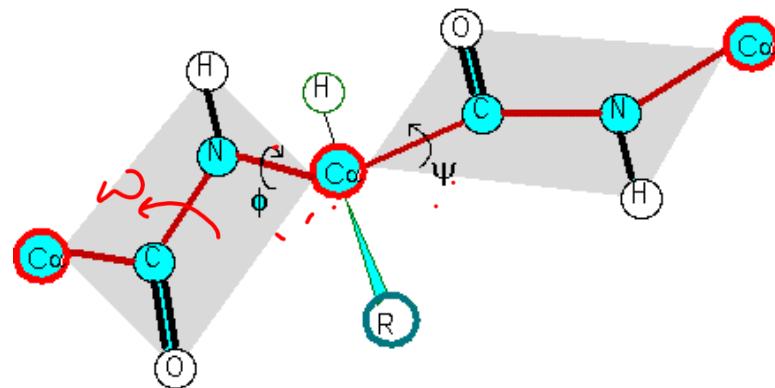


torsion angle

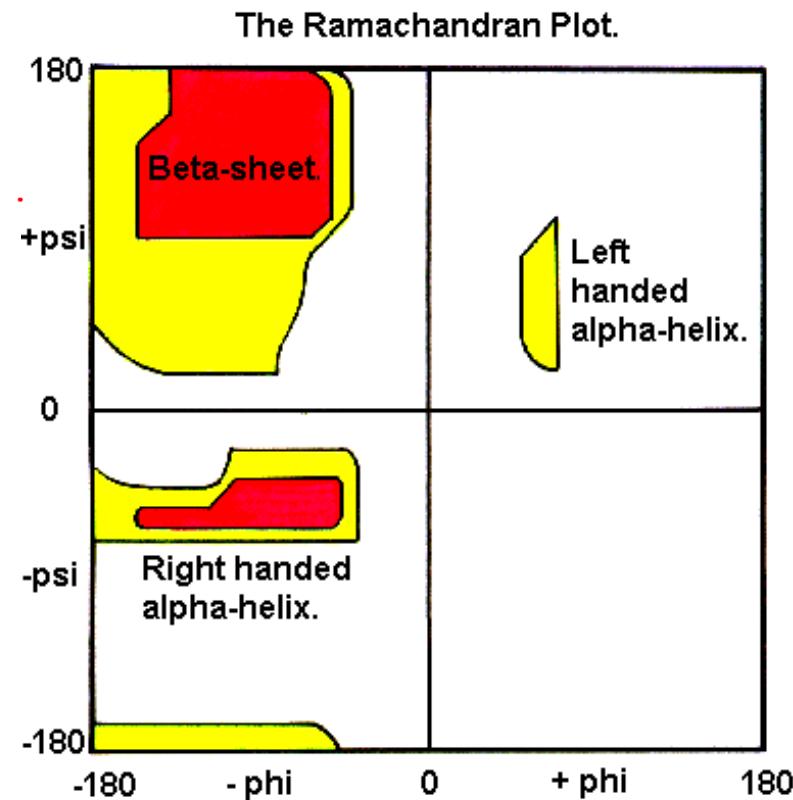


<https://www.youtube.com/watch?v=JyUMLSsbeI>

Dihedral bonds for protein backbone conformation



Dihedral angels



Secondary structure assignment

DSSP (Dictionary of Protein Secondary Structure) - The DSSP program works by calculating the most likely secondary structure assignment given the 3D structure of a protein.

There are eight types of secondary structure that DSSP defines:

G = 3-turn helix (3_{10} helix). Min length 3 residues.

H = 4-turn helix (α helix). Minimum length 4 residues.

I = 5-turn helix (π helix). Minimum length 5 residues.

T = hydrogen bonded turn (3, 4 or 5 turn)

E = extended strand in parallel and/or anti-parallel β -sheet conformation. Min length 2 residues.

B = residue in isolated β -bridge (single pair β -sheet hydrogen bond formation)

S = bend (the only non-hydrogen-bond based assignment).

C = coil (residues which are not in any of the above conformations).

Secondary structure assignment

DSSP (Dictionary of Protein Secondary Structure) - Computer Program for secondary structure assignment

Other examples: Stride and Pcurve

Secondary structure prediction

Methods:

1. Statistical analysis

(Preference of residues, by Chou and Fasman in 1974)

2. Information theory (GOR method, by Garnier, Osguthorpe, and Robson in 1978)

3. Hydrophobicity Profile

4. Multiple sequence alignment

5. Machine learning techniques

(Neural networks, support vector machines, etc.)

6. Consensus (Joint)

Statistical analysis: Propensity

The propensity of an amino acid residue i in any conformation (helix or strand or turn or coil) has been defined as the percentage of residue i in that conformation to the percentage of all residues in the same conformation.

$$\text{propensity}_{\alpha}(i) = \% \text{ of residue } i \text{ in } \alpha\text{-helix} / \\ \% \text{ of all residues in } \alpha\text{-helix.}$$

$$\% \text{ of residue } i \text{ in } \alpha\text{-helix} = n_{\alpha}(i)/N(i)$$

$n_{\alpha}(i)$ = number of residues of type i in α -helix

$N(i)$ = number of residues of type i in the whole dataset

$$\% \text{ of all residues in } \alpha\text{-helix} = n_{\alpha}/N$$

n_{α} = total number of residues in α -helix

N = total number of residues in the whole dataset

Propensity

VLS**E**GEWQLV LH**WAKV**EAD **VAGHGQDIL**I RLFKSH**PETL** EKFDRFKHLK
HHHHHHHH HHHHHHHGGG HHHHHHHHHHH HHHHH HHHH HT GGGTT

T**E****A****M**KASED LKKHGVTVL**T** ALGA**LL**KKKG HH**E****A****E**L**K**PLA QSHAT**K**HKIP
SHHHHHHH HH HHHHHHHHHHH HHHHHHTTTT HHHHHHHHH HHHHHTS

I**K****Y****L****E****F****I****S****E****A** **I****I****H****V****L****H****S****R****HP** GDFG**ADA****Q**GA MN**K****A****L****E****F****R****K** DIAAKY**K**ELG
HHHHHHHHHHHH HHHHHHHHH G GGS HHHHHH HHHHHHHHHHH HHHHHHHHHHT

YQG

E.g. **Ala**: % of Ala in α -helix = $N_a(\text{Ala})/N(\text{Ala})$
= 15/16 = 0.94

% of all residues in α -helix = $N_a/N = 115/153 = 0.75$

Propensity of Ala = $0.94/0.75 = \underline{1.25} > 1$

Propensity of Gly: $0.5/0.75 = 0.66$

Algorithm

1. Compute the occurrence of 20 residues in helix
2. Compute the occurrence of 20 residues in whole protein
3. Compute the ratio
4. Compute total number of residues in helix
5. Compute the ratio: number of residues in helix/ total number of residues in the protein
6. Divide 3 by 5 to get the propensity of all the 20 amino acid residues in helix

Chou-Fasman method

TABLE 5.2 Chou-Fasman parameters

Residue	P _α	Residue	P _β	Residue	P _t
Glu	H _α 1.53	H _β Met	1.67	Asn	1.68
Ala	1.45	Val	1.65	Gly	1.68
Leu	1.34	Ile	1.60	Ser	1.56
His	h _α 1.24	h _β Cys	1.30	Pro	1.54
Met	1.20	Tyr	1.29	Asp	1.26
Gln	1.17	Phe	1.28	Tyr	1.25
Trp	1.14	Gln	1.23	Cys	1.17
Val	1.14	Leu	1.22	Trp	1.11
Phe	1.12	Thr	1.20	Lys	1.01
Lys	I _α 1.07	Trp	1.19	Arg	1.00
Ile	1.00	I _β Ala	0.97	Thr	1.00
Asp	i _α 0.98	i _β Arg	0.90	Phe	0.71
Thr	0.82	Gly	0.81	His	0.69
Ser	0.79	Asp	0.80	Met	0.67
Arg	0.79	b _β Lys	0.74	Ile	0.58
Cys	0.77	Ser	0.72	Ala	0.57
Asn	b _α 0.73	His	0.71	Gln	0.56
Tyr	0.61	Asn	0.65	Leu	0.53
Pro	B _α 0.59	Pro	0.62	Glu	0.44
Gly	0.53	B _β Glu	0.26	Val	0.30

H_α: Strong helix former

h_α: Helix former

I_α: Weak helix former

i_α: Weak helix breaker

b_α: Helix breaker

B_α: Strong helix breaker

Rules for identifying Helix

Helix:

- Values of the six parameters are $H_\alpha = h_\alpha = 1$; $I_\alpha = 0.5$; $i_\alpha = 0$; $B_\alpha = b_\alpha = -1$;
- Scan for window of 6 residues, where score ≥ 4 , i.e. at least four helix formers and not more than one helix breaker;
- Extend the length in both directions until the score is less than 4;

- Continue the search and locate all helical regions in the sequence.
- Refinement: Pro, Asp, Glu: N-terminal; His, Lys, Arg: C-terminal; Pro: Not in inner helix or C-terminal

TABLE 5.2 Chou-Fasman parameters

Residue	P_α
Glu	H_α 1.53
Ala	1.45
Leu	1.34
His	h_α 1.24
Met	1.20
Gln	1.17
Trp	1.14
Val	1.14
Phe	1.12
Lys	I_α 1.07
Ile	1.00
Asp	i_α 0.98
Thr	0.82
Ser	0.79
Arg	0.79
Cys	0.77
Asn	b_α 0.73
Tyr	0.61
Pro	B_α 0.59
Gly	0.53

Rules for identifying Helix

KVFGRCELAAAMKRH GLDNYRGYSLGNWVCAAKFESNFNT
QATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDV
QAWIRGCRL

$$\text{KVFGRC: } 0.5+1+1-1+0+0 = 1.5$$

$$\text{VFGRCE: } 1+1-1+0+0+1 = 2$$

$$\text{FGRCEL: } 1-1+0+0+1+1 = 2$$

$$\text{GRCELA: } -1+0+0+1+1+1 = 2$$

$$\text{RCELA: } 0+0+1+1+1+1 = 4$$

Score

$$\text{MKRH: } 1.20+1.07+0.79+1.24 = 4.3$$

$$\text{KRHG: } 1.07+0.79+1.24+0.53 = 3.63$$

TABLE 5.2 Chou-Fasman parameters

Residue	P_{α}
Glu	$H\alpha$ 1.53
Ala	1.45
Leu	1.34
His	$H\alpha$ 1.24
Met	1.20
Gln	1.17
Trp	1.14
Val	1.14
Phe	1.12
Lys	$I\alpha$ 1.07
Ile	1.00
Asp	$i\alpha$ 0.98
Thr	0.82
Ser	0.79
Arg	0.79
Cys	0.77
Asn	$b\alpha$ 0.73
Tyr	0.61
Pro	$B\alpha$ 0.59
Gly	0.53

Rules for identifying beta sheet

- The values of the six parameters are $H_{\beta} = h_{\beta} = 1$; $I_{\beta} = 0.5$; $i_{\beta} = 0$; $B_{\beta} = b_{\beta} = -1$;
- Scan for window of 5 residues, where score > 3, i.e. at least three strand formers and not more than one strand breaker;
- Extend the length in both directions until the segment has the average propensity < 1;
- Continue the search and locate all strand regions in the sequence.

Sequence and secondary structure for 4LYZ chain A

1	KVFGRCELAA AMKRHGLDNY RGYSLGNNVVC AAKFESNFNT QATMRNTDGS B HHHHHHH HHHHTT TTB TTB HHHHHHH HHHHHHHHTBS S EEE SSS
51	TDYGILQINS RWWCNDGRTP GSRNLCNIPC SALLSSSDITA SVNCAKKIVS EEETTTTEET TTT B SS T T SS SBG GGGGSS HH HHHHHHHHHHTT
101	DGNGMNAWVA WRNRCKGTDV QAWirGCRL TSSGGGGSHH HHHHTTTS G GGGSTT

Rules for identifying Beta sheet

Conflicting situation:

A region containing overlapping helical and strand assignments is considered as a helix (or strand) if average propensity of alpha-helix (beta-strand) is greater than that of beta-strand (alpha-helix).

GOR method (Garnier–Osguthorpe–Robson)

- Information theory-based method for the prediction of secondary structures in proteins.
- Assumes amino acids up to 8 residues on each side influence the ss of the central residue.
- Frequency of amino acids at the central position in the window, and at -1, ..., -8 and +1, ..., +8 is determined for alpha helices, beta strands and turns (later other or coils) to give three 17×20 scoring matrices.
- Calculate the score that the central residue is one type of SS and not another.
- Correctly predicts ~64%.

- **Information (i) for each residue**

Central residue, 8 neighbors on each side (window length of 17 residues); 4 states (helix, strand, turn and coil)

GOR method (Garnier–Osguthorpe–Robson)

Information content

$$I(SS_i=X:\sim X;aa) = \ln(P(SS_i=X|aa) / P(SS_i=\sim X|aa)) - \ln(P(S_i=X) / P(S_i=\sim X)),$$

$SS_i \rightarrow$ secondary structure at position i in the sequence

$X \rightarrow$ any secondary structure: helix (H), strand (E), turns (T) and
coil (C)

$aa \rightarrow$ any amino acid residue

GOR method

	Helix	~Helix	Total		
Alanine (aa= A)	210	90	300	$P(SS=H aa=A) = 210/300 = 0.70$	
All residues	810	990	1800	$P(SS=\sim H aa=A) = 90/300 = 0.30$	$P(SS=H) = 810/1800 = 0.45$

$$\begin{aligned} I(SS=H:\sim H;aa=A) &= \ln(0.70/0.30) - \ln(0.45/0.55) \\ &= 0.847 - (-0.20) = 1.047 \end{aligned}$$

GOR method

Directional information measure for the α -helical conformation†

Amino acid residue	Residue position‡ (centinats)												
	<i>j</i> - 8	<i>j</i> - 6	<i>j</i> - 4	<i>j</i> - 2	<i>j</i>	<i>j</i> + 2	<i>j</i> + 4	<i>j</i> + 6	<i>j</i> + 8				
Gly	-5	-10	-15	-20	-30	-40	-50	-60	-86	-60	-50	-40	-30
Ala	5	10	15	20	30	40	50	60	65	60	50	40	30
Val	0	0	0	0	0	5	10	14	10	5	0	0	0
Leu	0	5	10	15	20	25	28	30	32	30	28	25	20
Ile	5	10	15	20	25	20	15	10	6	0	-10	-15	-20
Ser	0	-5	-10	-15	-20	-25	-30	-35	-39	-35	-30	-25	-20
Thr	0	0	0	-5	-10	-15	-20	-25	-26	-25	-20	-15	-10
Asp	0	-5	-10	-15	-20	-15	-10	0	5	10	15	20	20
Glu	0	0	0	0	10	20	60	70	78	78	78	78	70
Asn	0	0	0	0	-10	-20	-30	-40	-51	-40	-30	-20	-10
Gln	0	0	0	0	5	10	20	20	10	-10	-20	-20	-10
Lys	20	40	50	55	60	60	50	30	23	10	5	0	0
His	10	20	30	40	50	50	50	30	12	-20	-10	0	0
Arg	0	0	0	0	0	0	0	0	-9	-15	-20	-30	-40
Phe	0	0	0	0	0	5	10	15	16	15	10	5	0
Tyr	-5	-10	-15	-20	-25	-30	-35	-40	-45	-40	-35	-30	-25
Trp	-10	-20	-40	-50	-50	-10	0	10	12	10	0	-10	-50
Cys	0	0	0	0	0	0	-5	-10	-13	-10	-5	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35
Pro	-10	-20	-40	-60	-80	-100	-120	-140	-77	-60	-30	-20	-10

GOR method

87654321012345678

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYF

10+0+10-15-80+40-10+30-**26**-40+5+0++30+20+40-10+0

$$I(H_9; \text{MVLSPADKTNVKA} \text{AWGK}) = 4$$

Similarly calculate for other secondary structure states.