Protein structure prediction

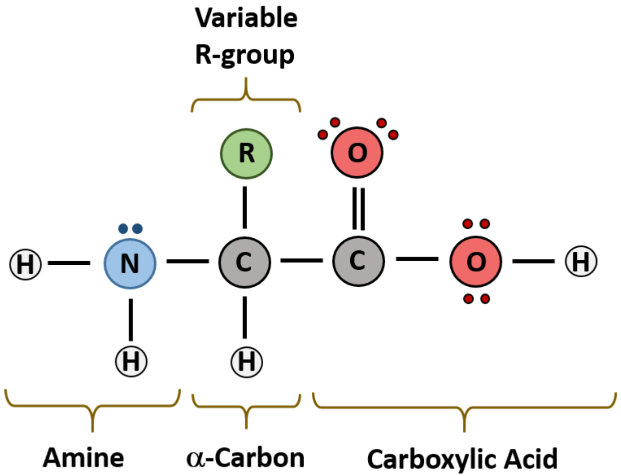
INDRAPRASTHA INSTITUTE of
INFORMATION TECHNOLOGY DELHI

Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Department of Computational Biology
Email ID: jaspreet@iiitd.ac.in

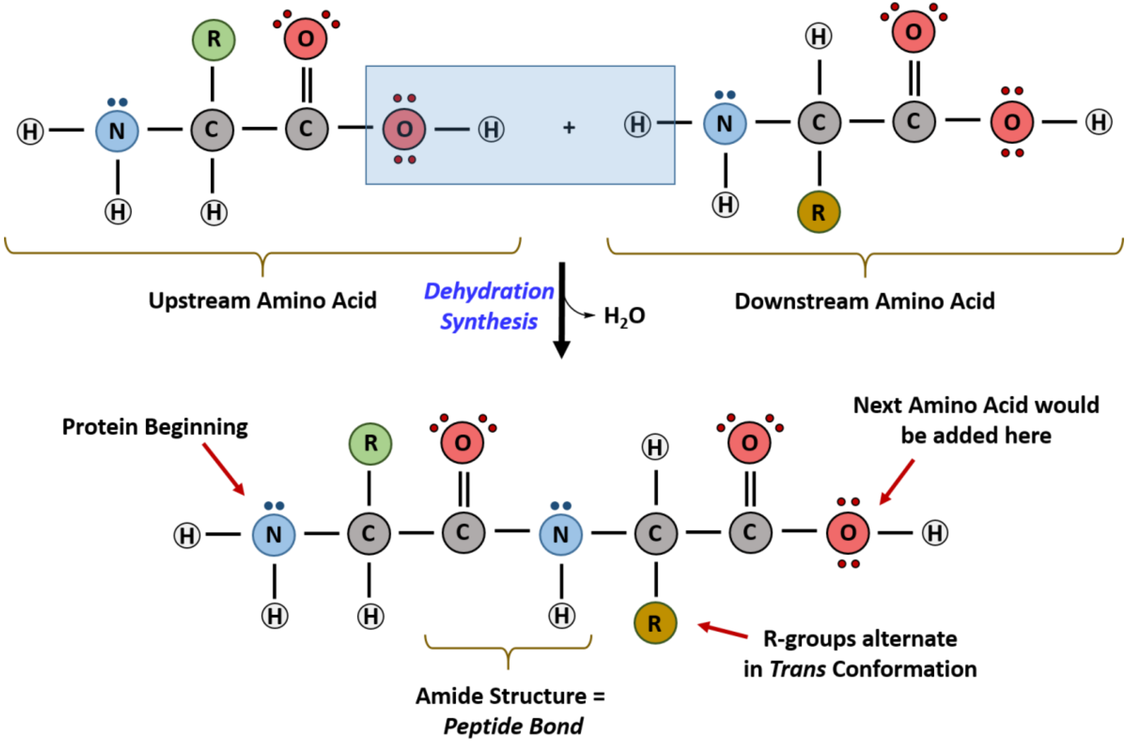
October 03, 2025

Amino acids, the building blocks of protein

Basic structure of an amino acid

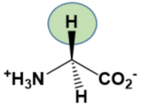


Formation of peptide bond

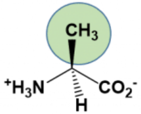


Different types of Amino acids

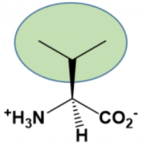
Nonpolar (Hydrophobic) Amino Acids



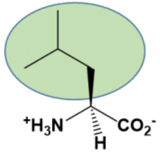
Glycine
Gly, G



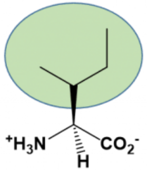
Alanine
Ala, A



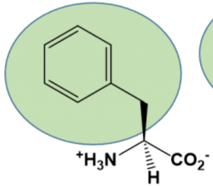
Valine
Val, V



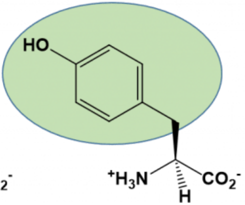
Leucine
Leu, L



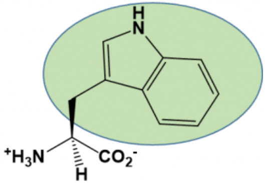
Isoleucine
Ile, I



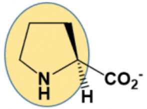
Phenylalanine
Phe, F



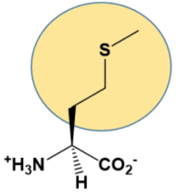
Tyrosine
Tyr, Y



Tryptophan
Trp, W

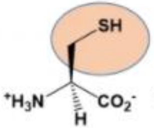


Proline
Pro, P

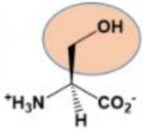


Methionine
Met, M

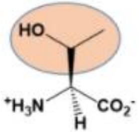
Polar (Hydrophilic) Amino Acids



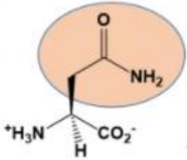
Cysteine
Cys, C



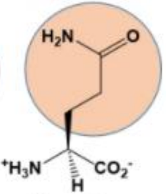
Serine
Ser, S



Threonine
Thr, T

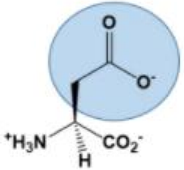


Asparagine
Asn, N

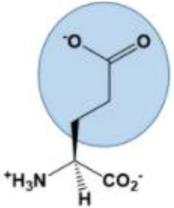


Glutamine
Gln, Q

Acidic Amino Acids

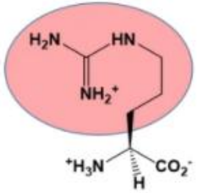


Aspartic Acid
Asp, D

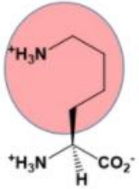


Glutamic Acid
Glu, E

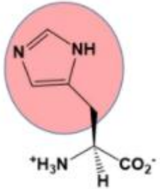
Basic Amino Acids



Arginine
Arg, R

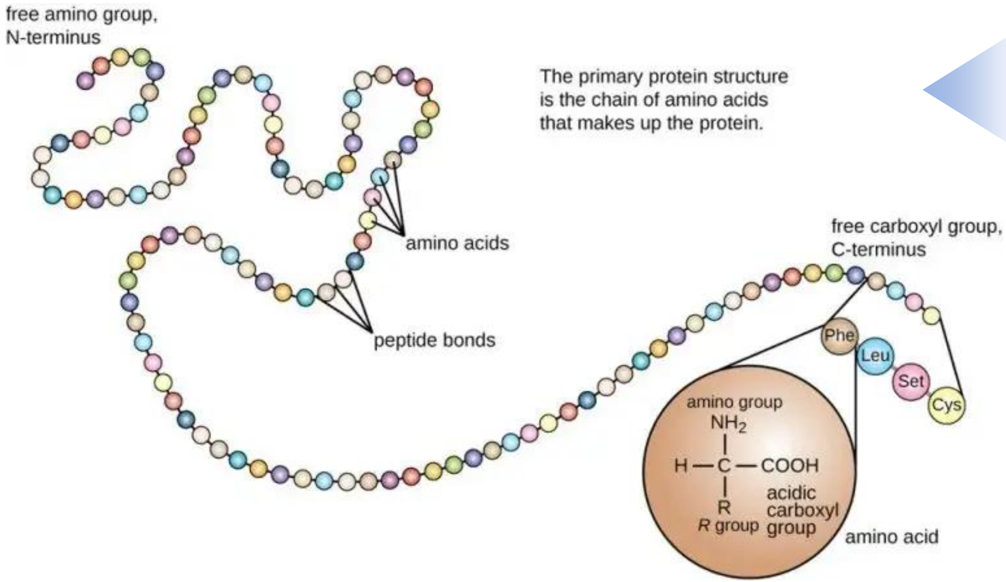


Lysine
Lys, K



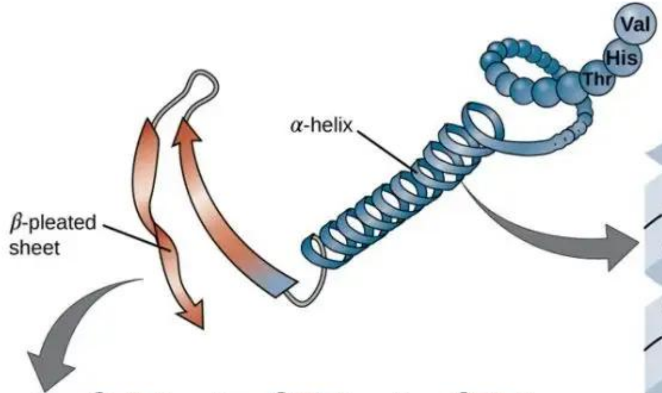
Histidine
His, H

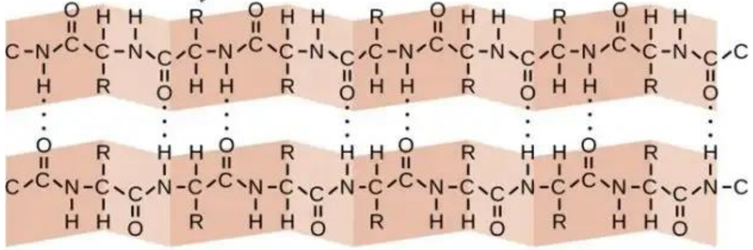
Structure of protein

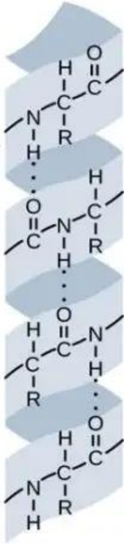

free amino group, N-terminus
The primary protein structure is the chain of amino acids that makes up the protein.
amino acids
peptide bonds
free carboxyl group, C-terminus
Phe
Leu
Set
Cys
amino group
NH_2
H — C — COOH
R
R group
acidic carboxyl group
amino acid

1. Primary structure of protein

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



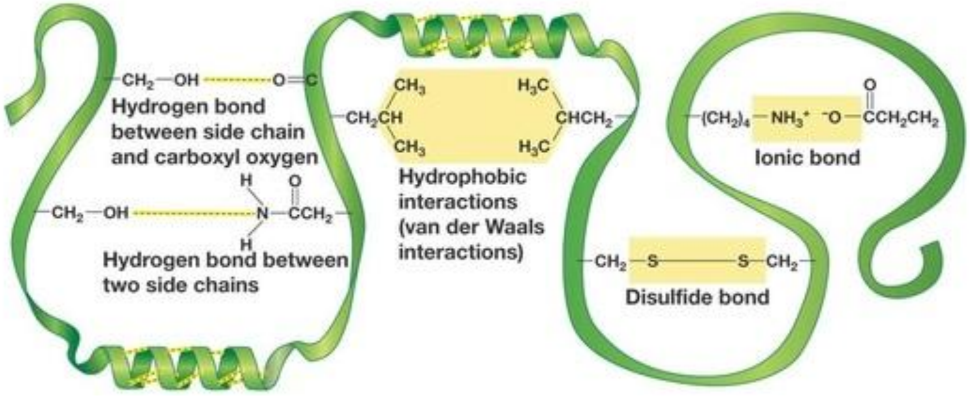




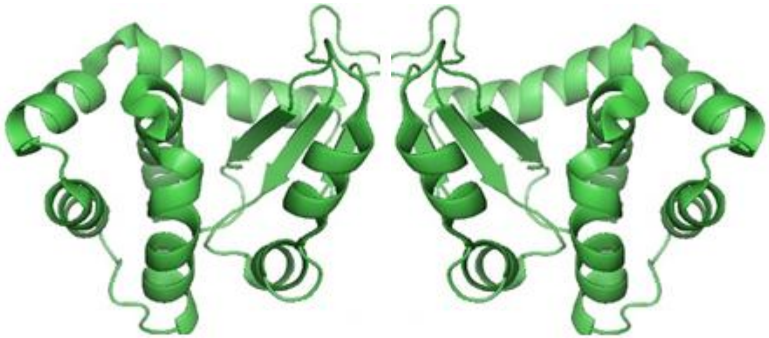
2. Secondary structure of protein

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

Structure of protein



CH_2-OH ········ $O=C$
Hydrogen bond between side chain and carboxyl oxygen
CH_2-OH ········ $N-\overset{O}{C}CH_2$
Hydrogen bond between two side chains
CH_3 H_3C
CH_2CH $CHCH_2$
CH_3 H_3C
Hydrophobic interactions (van der Waals interactions)
$(CH_2)_4-NH_3^+$ $^-O-\overset{O}{C}CH_2CH_2$
Ionic bond
$CH_2-S-S-CH_2$
Disulfide bond



3. *Tertiary structure of protein*

The tertiary structure is primarily due to interactions between the R groups of the amino acids that make up the protein.

4. *Quaternary structure of protein*

This level of structure forms when two or more tertiary structures combine to form a single 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

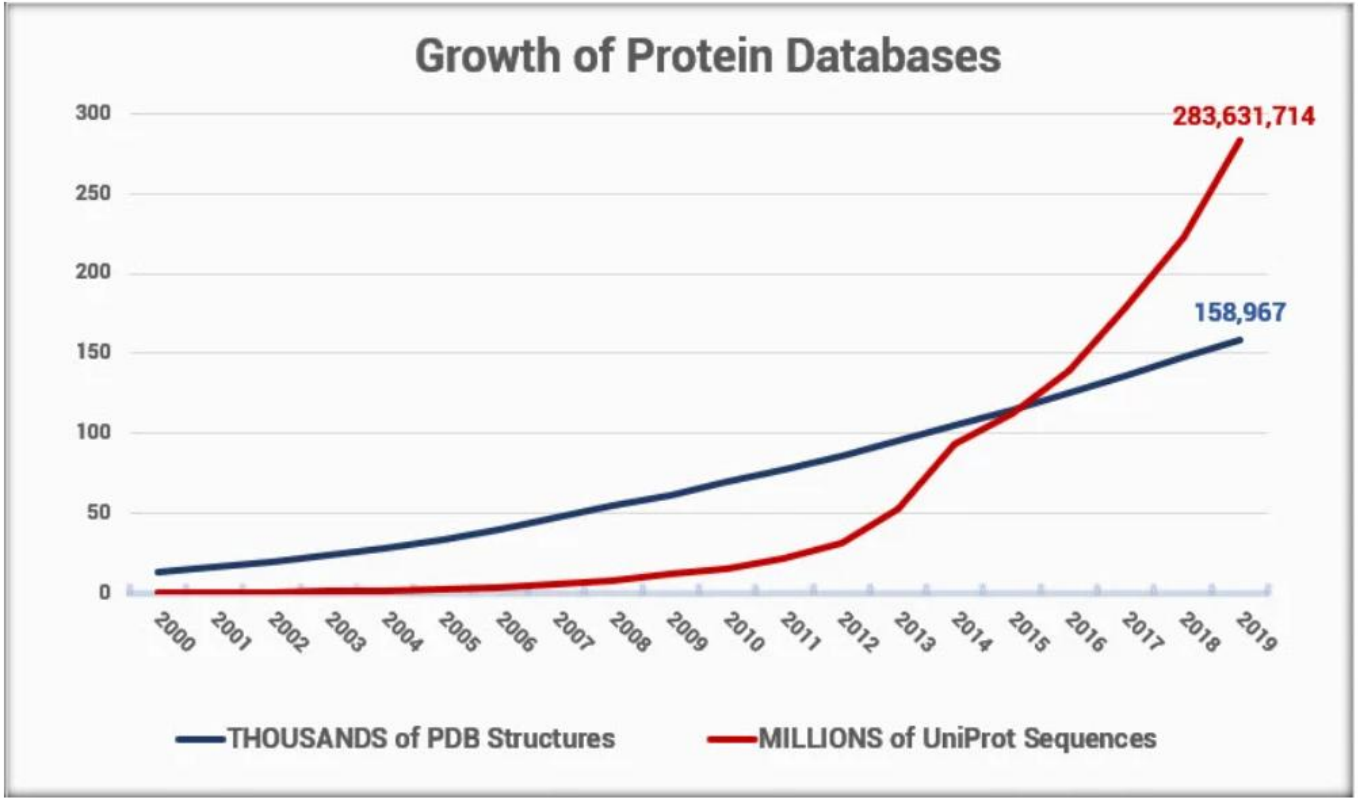


Figure 1. Growth of protein sequence and structure databases over time

2025/10/03

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

PDB
(242,874)

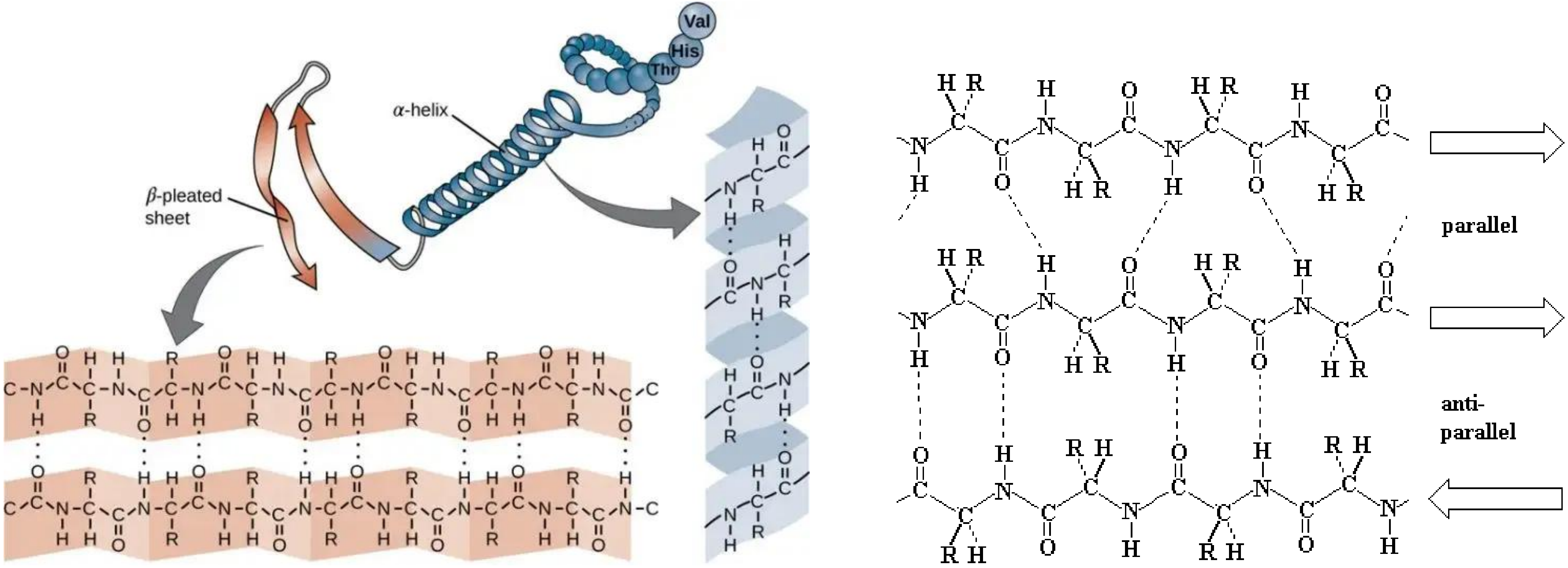
Prediction of protein structure

The structure of protein is predicted at two different levels:

1. Secondary structure prediction

2. Tertiary structure prediction

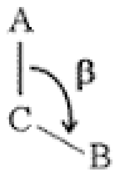
Hydrogen bonds for secondary structure assignment



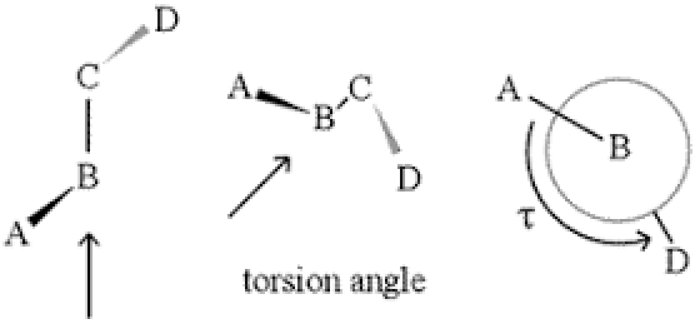
Hydrogen bonding pattern

Dihedral bonds for protein backbone conformation

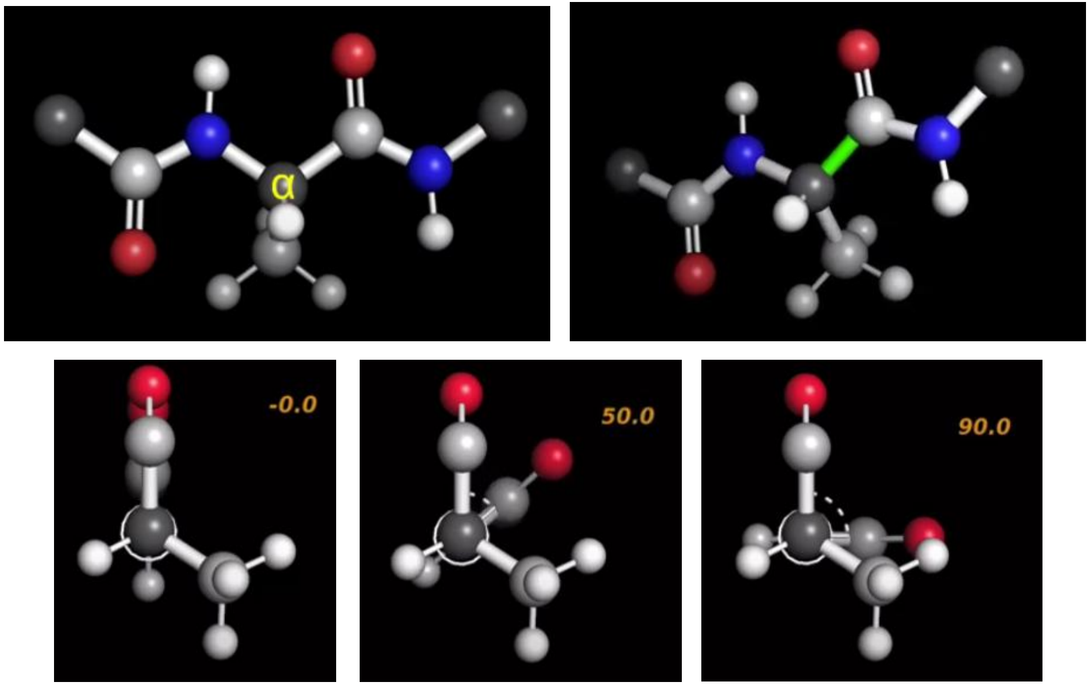
Dihedral/torsional angels



bond angle

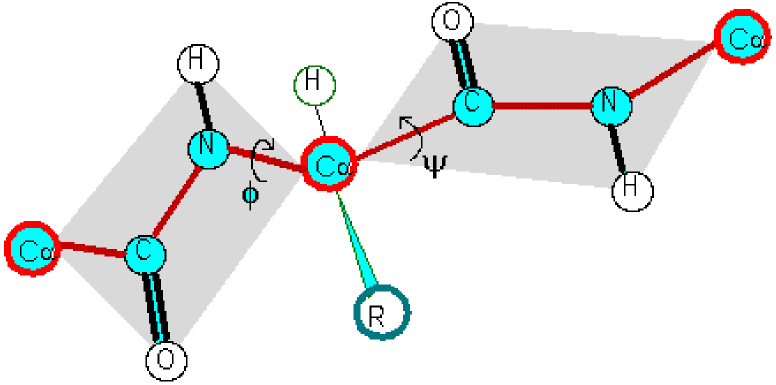


torsion angle

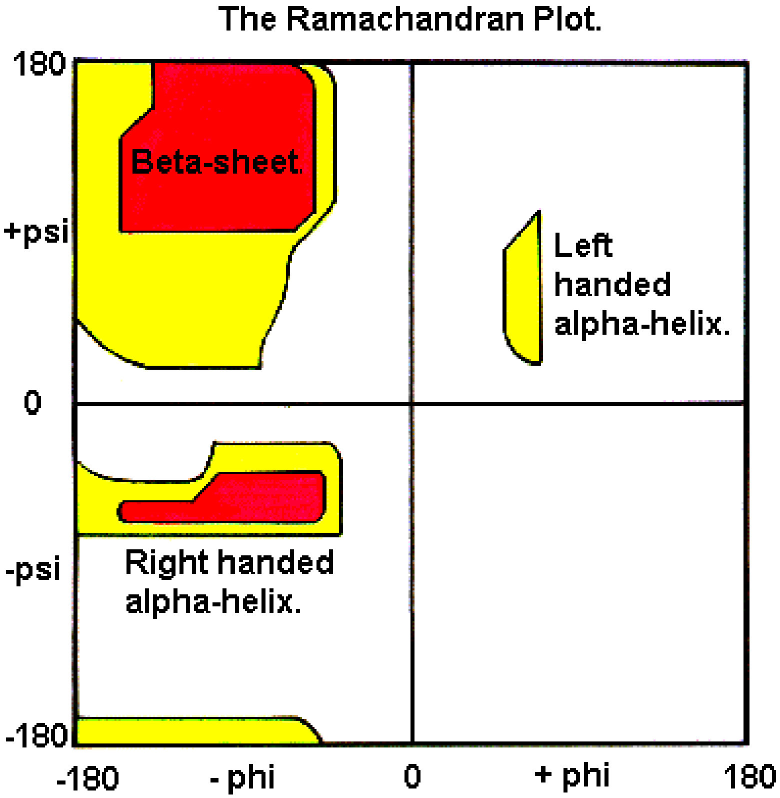


https://www.youtube.com/watch?v=JyUMLSsbecI

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.

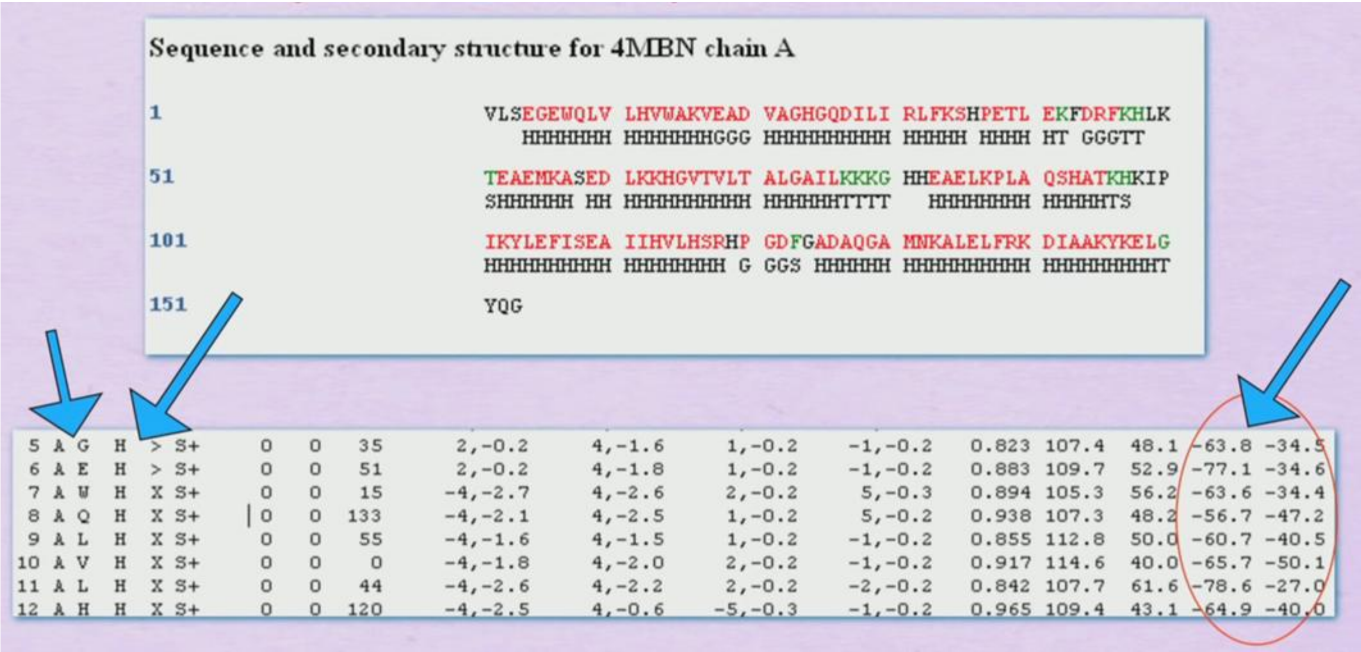
B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation)

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



Sequence and secondary structure for 4MBN chain A

```
1      VLSEGEWQLV LHVWAKVEAD VAGHGQDILI RLFKSHPETL EKFDRFKHLK
          HHHHHHH HHHHHHHGGG HHHHHHHHHH HHHHH HHHH HT GGGTT

51     TEAEMKASED LKKHGVTVLT ALGAILKKKG HHEAELKPLA QSHATKHKIP
       SHHHHHH HH HHHHHHHHHH HHHHHHTTTT   HHHHHHHH HHHHHTS

101    IKYLEFISEA IIHVLHSRHP GDFGADAQGA MNKALELFRK DIAAKYKELG
       HHHHHHHHHH HHHHHHHH G GGS HHHHHH HHHHHHHHHH HHHHHHHHHT

151    YQG
```

```
 5 A G  H  > S+     0   0   35     2,-0.2     4,-1.6     1,-0.2    -1,-0.2   0.823 107.4  48.1 -63.8 -34.5
 6 A E  H  > S+     0   0   51     2,-0.2     4,-1.8     1,-0.2    -1,-0.2   0.883 109.7  52.9 -77.1 -34.6
 7 A W  H  X S+     0   0   15    -4,-2.7     4,-2.6     2,-0.2     5,-0.3   0.894 105.3  56.2 -63.6 -34.4
 8 A Q  H  X S+    |0   0  133    -4,-2.1     4,-2.5     1,-0.2     5,-0.2   0.938 107.3  48.2 -56.7 -47.2
 9 A L  H  X S+     0   0   55    -4,-1.6     4,-1.5     1,-0.2    -1,-0.2   0.855 112.8  50.0 -60.7 -40.5
10 A V  H  X S+     0   0    0    -4,-1.8     4,-2.0     2,-0.2    -1,-0.2   0.917 114.6  40.0 -65.7 -50.1
11 A L  H  X S+     0   0   44    -4,-2.6     4,-2.2     2,-0.2    -2,-0.2   0.842 107.7  61.6 -78.6 -27.0
12 A H  H  X S+     0   0  120    -4,-2.5     4,-0.6    -5,-0.3    -1,-0.2   0.965 109.4  43.1 -64.9 -40.0
```

Other examples: Stride and Pcurve

Secondary structure prediction

Methods:

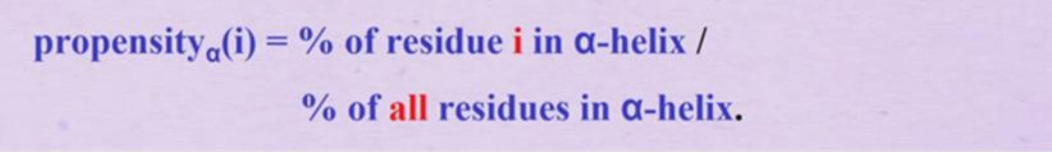
1. Statistical analysis

   (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)

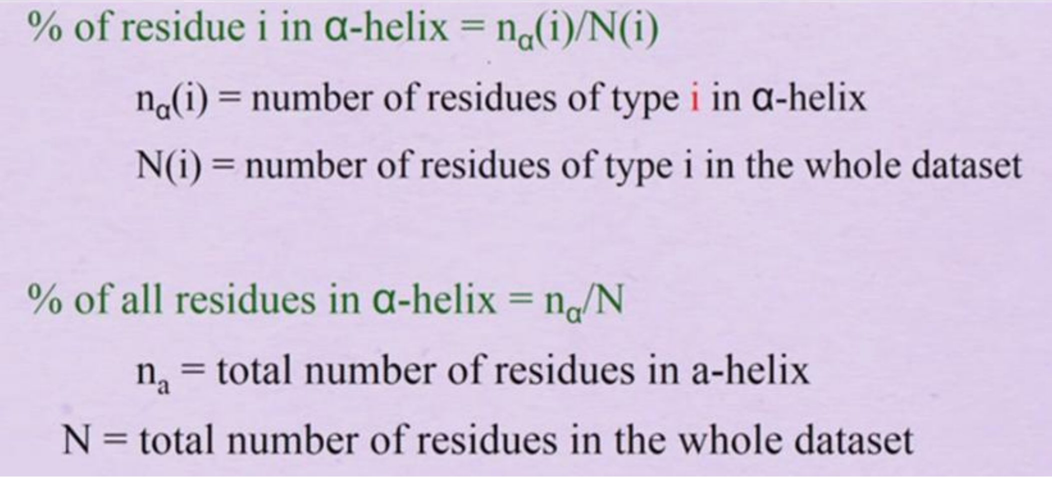
Statistical analysis: Propensity

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.



propensity$_\alpha$(i) = % of residue i in α-helix /

% of all residues in α-helix.



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

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

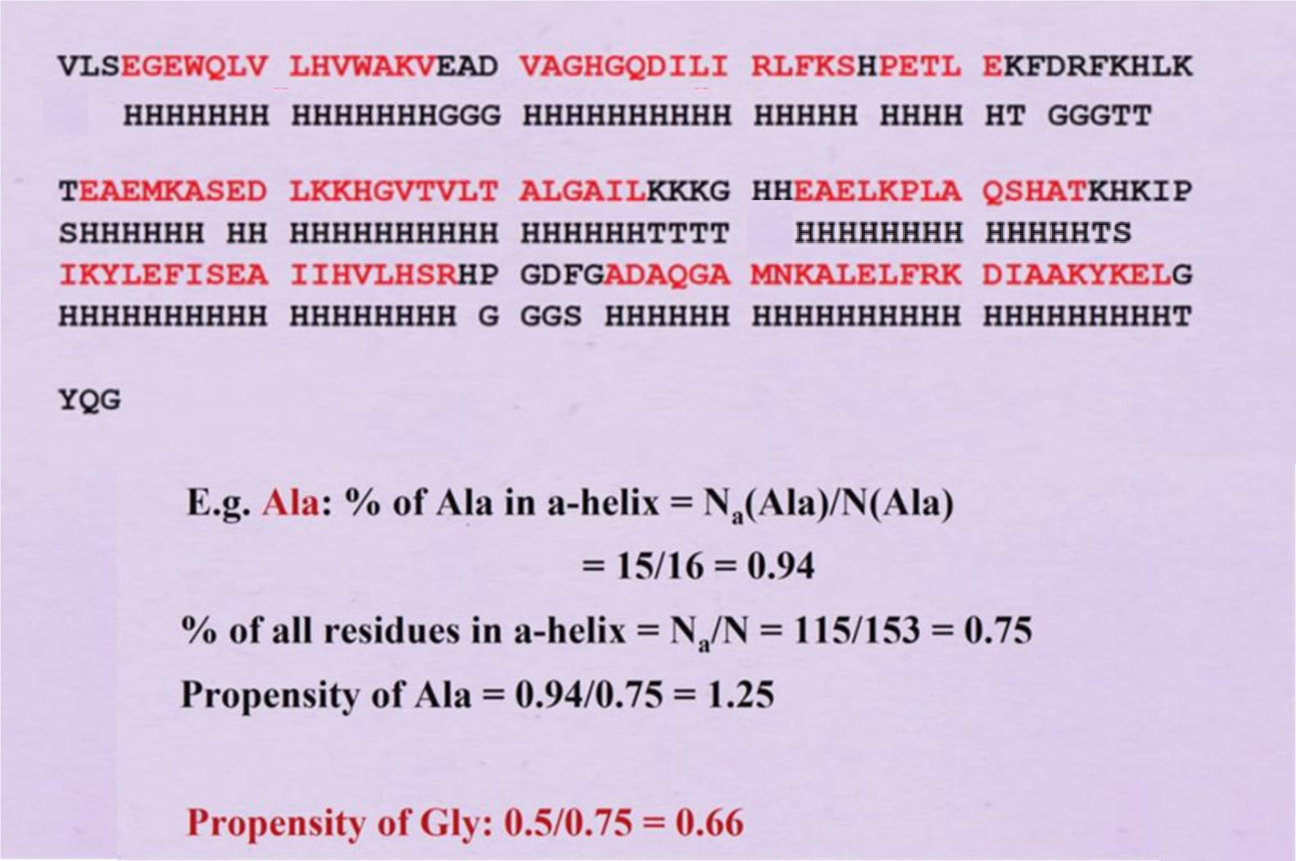
$N(i)$ = number of residues of type i in the whole dataset

% of all residues in α-helix = n_α/N

n_a = total number of residues in a-helix

N = total number of residues in the whole dataset

Propensity



```
VLSEGEWQLV LHVWAKVEAD VAGHGQDILI RLFKSHPETL EKFDRFKHLK
   HHHHHHH HHHHHHHGGG HHHHHHHHHH HHHHH HHHH HT GGGTT

TEAEMKASED LKKHGVTVLT ALGAILKKKG HHEAELKPLA QSHATKHKIP
SHHHHHH HH HHHHHHHHHH HHHHHHTTTT   HHHHHHHH HHHHHTS
IKYLEFISEA IIHVLHSRHP GDFGADAQGA MNKALELFRK DIAAKYKELG
HHHHHHHHHH HHHHHHHH G GGS HHHHHH HHHHHHHHHH HHHHHHHHHT

YQG
```

E.g. Ala: % of Ala in a-helix = N_a(Ala)/N(Ala)

= 15/16 = 0.94

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

Propensity of Ala = 0.94/0.75 = 1.25

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

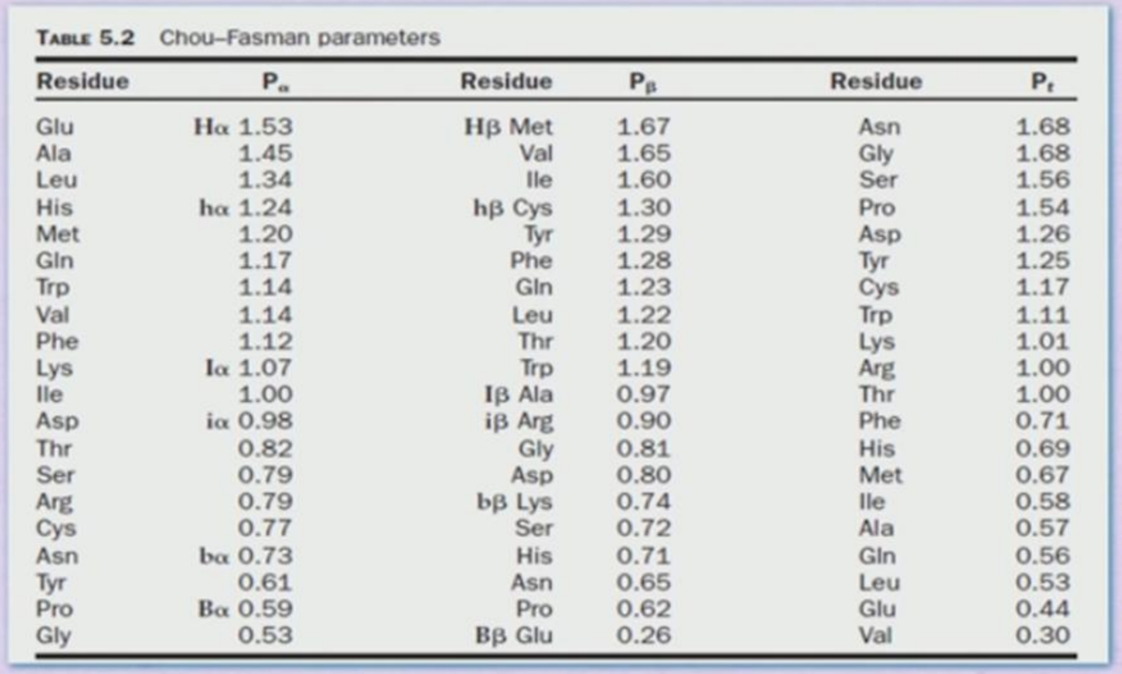


TABLE 5.2 Chou–Fasman parameters

Residue	P_α	Residue	P_β	Residue	P_t
Glu	Hα 1.53	Hβ 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	hα 1.24	hβ 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	Iα 1.07	Trp	1.19	Arg	1.00
Ile	1.00	Iβ Ala	0.97	Thr	1.00
Asp	iα 0.98	iβ 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	bβ Lys	0.74	Ile	0.58
Cys	0.77	Ser	0.72	Ala	0.57
Asn	bα 0.73	His	0.71	Gln	0.56
Tyr	0.61	Asn	0.65	Leu	0.53
Pro	Bα 0.59	Pro	0.62	Glu	0.44
Gly	0.53	Bβ Glu	0.26	Val	0.30

Hα: Strong helix former

hα: Helix former

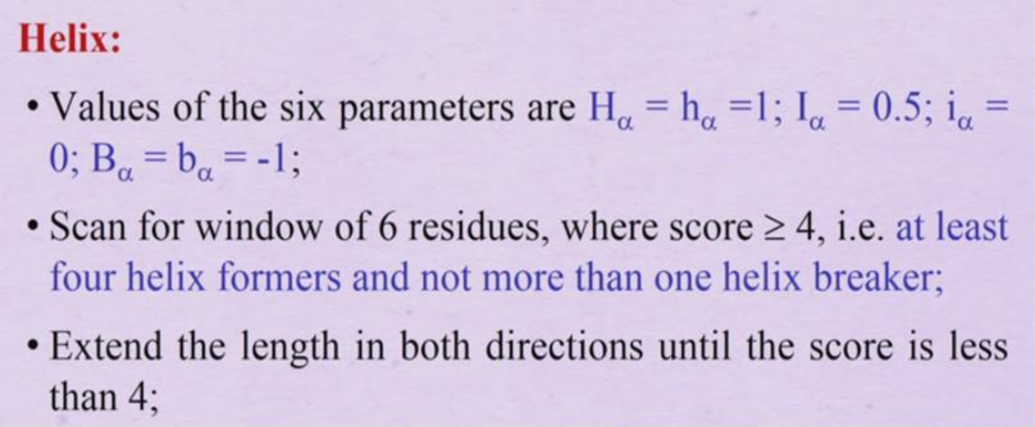
Iα: Weak helix former

iα: Weak helix breaker

bα: Helix breaker

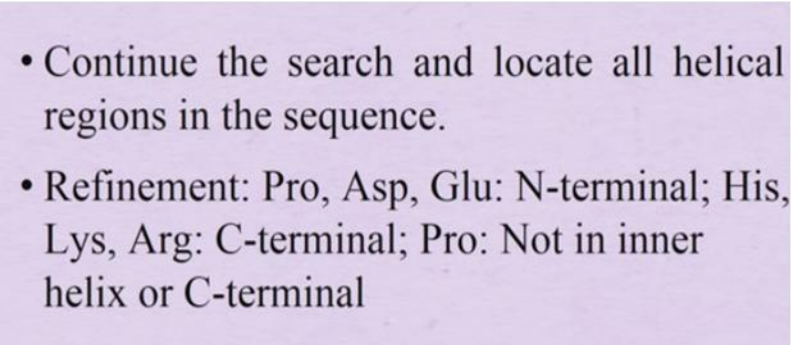
Bα: Strong helix breaker

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

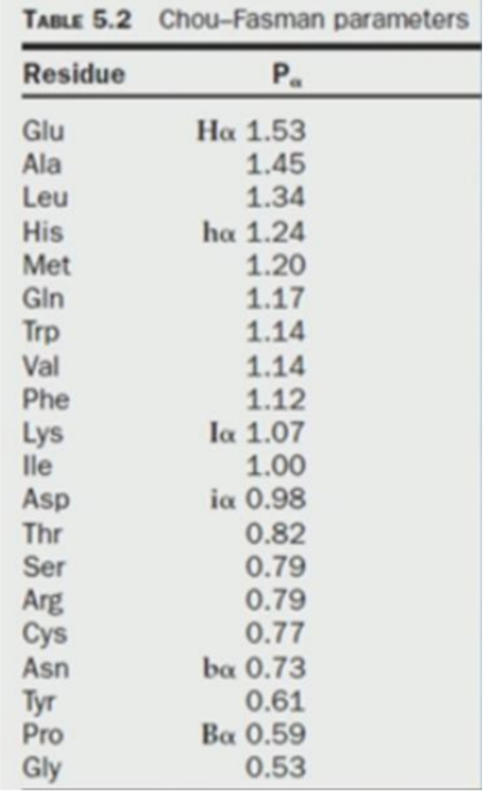
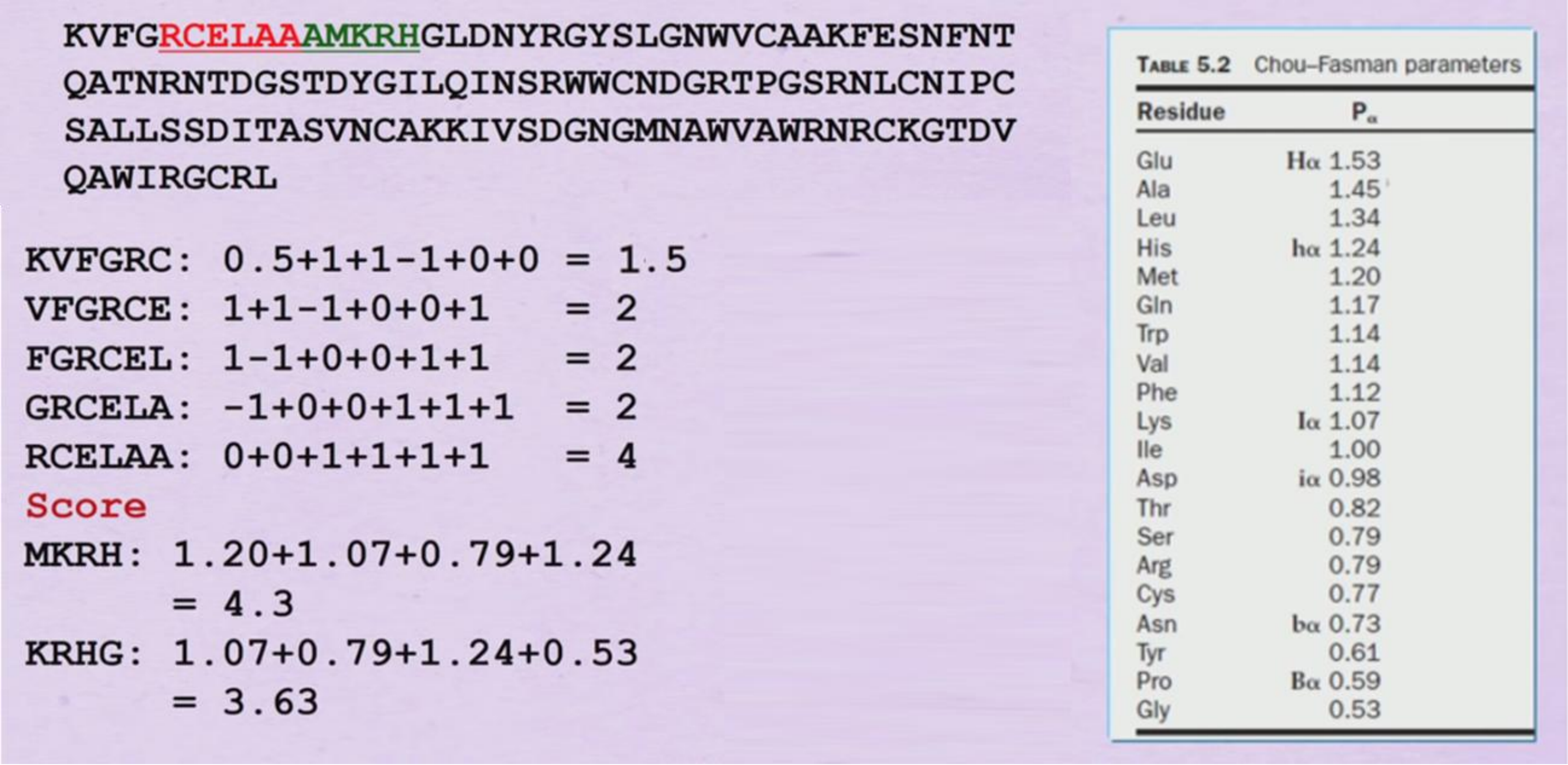


TABLE 5.2 Chou–Fasman parameters

Residue	P_α
Glu	Hα 1.53
Ala	1.45
Leu	1.34
His	hα 1.24
Met	1.20
Gln	1.17
Trp	1.14
Val	1.14
Phe	1.12
Lys	Iα 1.07
Ile	1.00
Asp	iα 0.98
Thr	0.82
Ser	0.79
Arg	0.79
Cys	0.77
Asn	bα 0.73
Tyr	0.61
Pro	Bα 0.59
Gly	0.53

Rules for identifying Helix



KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNT
QATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDV
QAWIRGCRL

```
KVFGRC:  0.5+1+1-1+0+0 = 1.5
VFGRCE:  1+1-1+0+0+1   = 2
FGRCEL:  1-1+0+0+1+1   = 2
GRCELA:  -1+0+0+1+1+1  = 2
RCELAA:  0+0+1+1+1+1   = 4
```

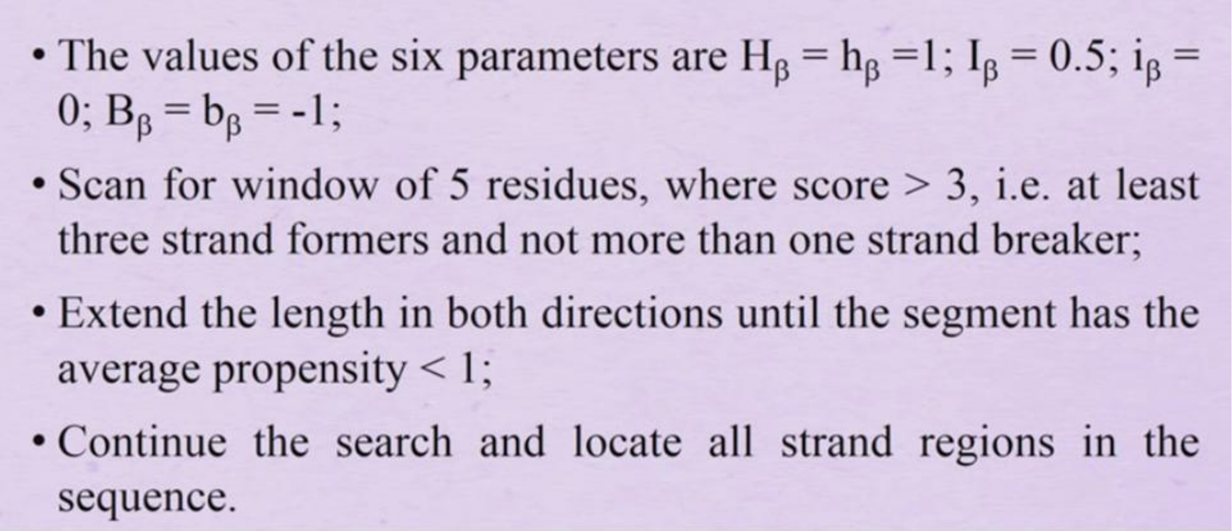
Score

```
MKRH:  1.20+1.07+0.79+1.24
       = 4.3
KRHG:  1.07+0.79+1.24+0.53
       = 3.63
```

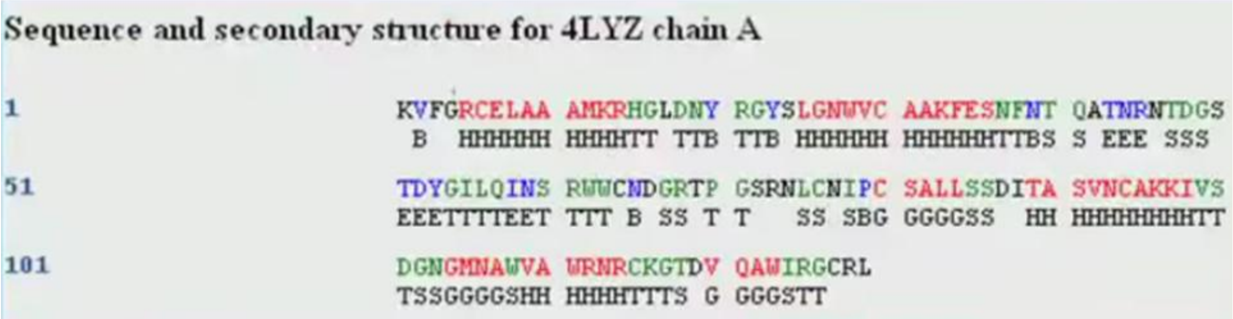
TABLE 5.2 Chou–Fasman parameters

Residue	P_α
Glu	Hα 1.53
Ala	1.45
Leu	1.34
His	hα 1.24
Met	1.20
Gln	1.17
Trp	1.14
Val	1.14
Phe	1.12
Lys	Iα 1.07
Ile	1.00
Asp	iα 0.98
Thr	0.82
Ser	0.79
Arg	0.79
Cys	0.77
Asn	bα 0.73
Tyr	0.61
Pro	Bα 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 RGYSLGNWVC AAKFESNFNT QATNRNTDGS
        B  HHHHHH HHHHTT TTB TTB HHHHHH HHHHHHTTBS S EEE SSS

51     TDYGILQINS RWWCNDGRTP GSRNLCNIPC SALLSSDITA SVNCAKKIVS
       EEETTTTEET TTT B SS T T   SS SBG GGGGSS  HH HHHHHHHHTT

101    DGNGMNAWVA WRNRCKGTDV QAWIRGCRL
       TSSGGGGSHH HHHHTTTS G GGGSTT
```

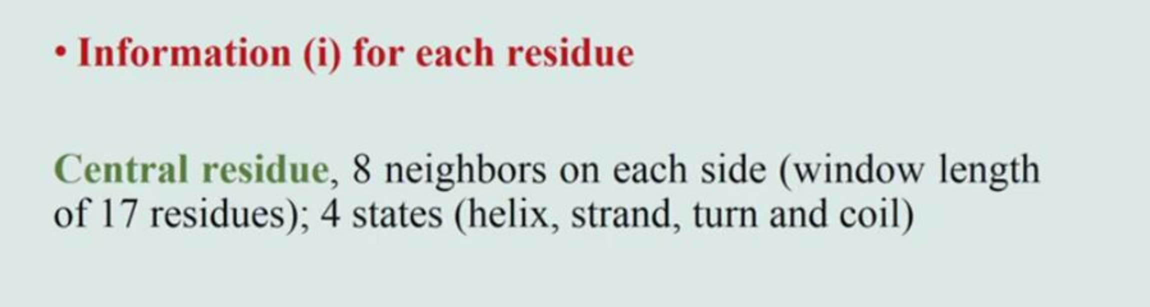
Rules for identifying Beta sheet

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 bets-strand (alpha-helix).

GOR method (Garnier–Osguthorpe–Robson)

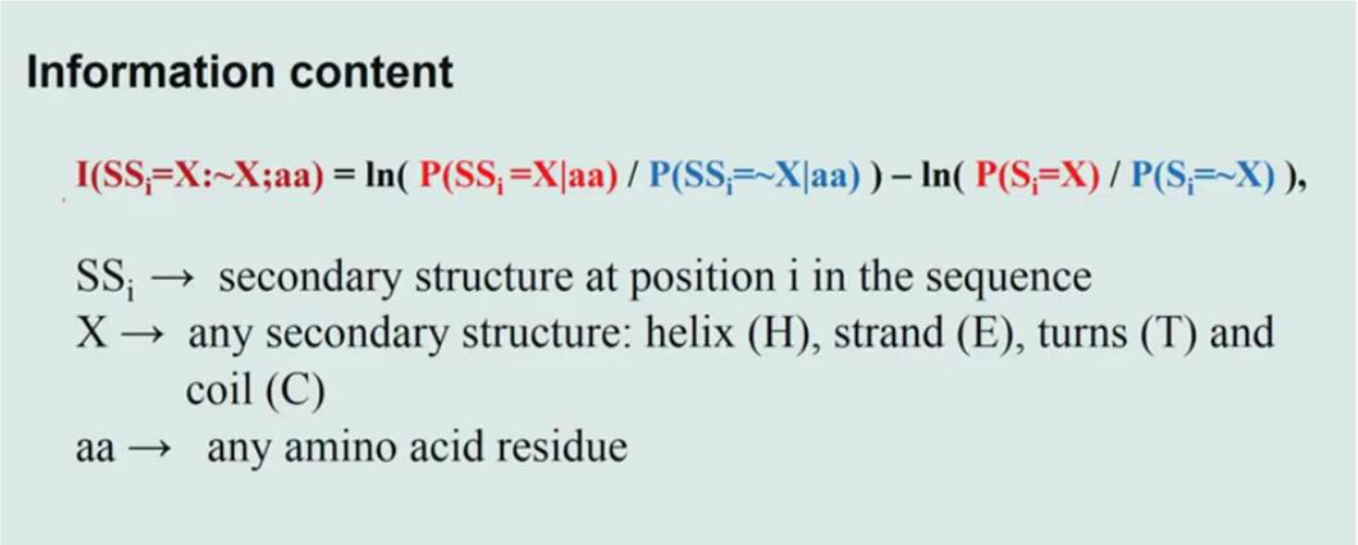
- Information theory-based method for the prediction of secondary structures in proteins.
- Assumes amino acids up to 8 residues on each side influence the ss of the central residue.
- Frequency of amino acids at the central position in the window, and at -1, ... -8 and +1,....+8 is determined for alpha helices, beta strands and turns (later other or coils) to give three 17 x 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

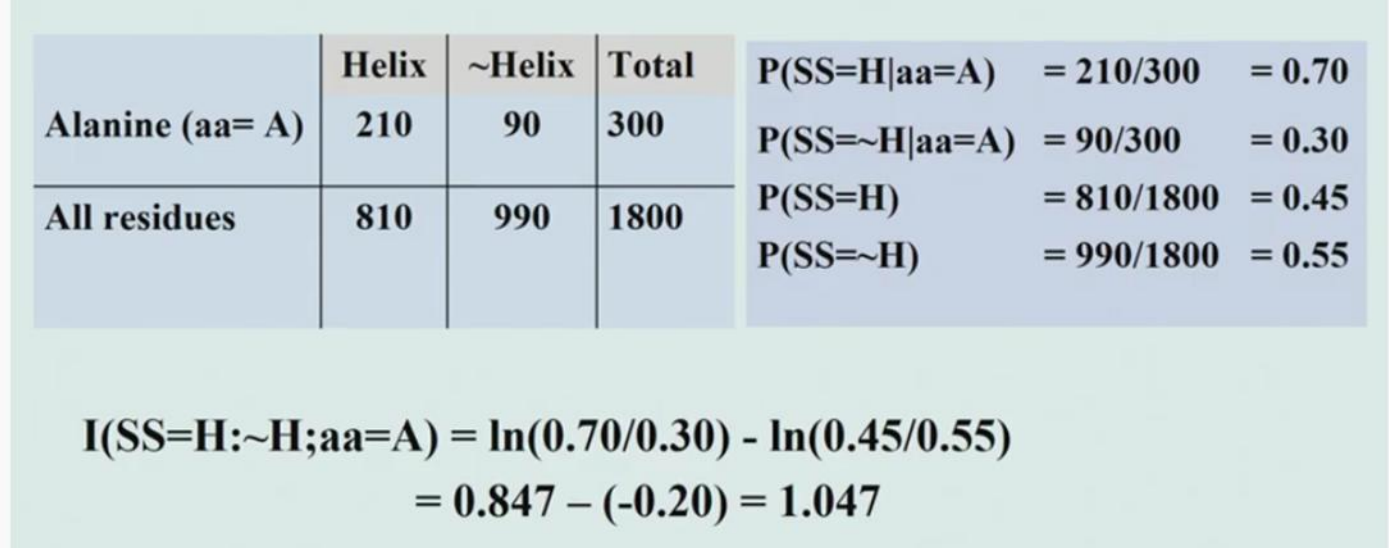
$$I(SS_i{=}X{:}{\sim}X;aa) = \ln\left(P(SS_i{=}X|aa) \,/\, P(SS_i{=}{\sim}X|aa) \right) - \ln\left(P(S_i{=}X) \,/\, P(S_i{=}{\sim}X) \right),$$

SS_i → secondary structure at position i in the sequence

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

aa → any amino acid residue

GOR method

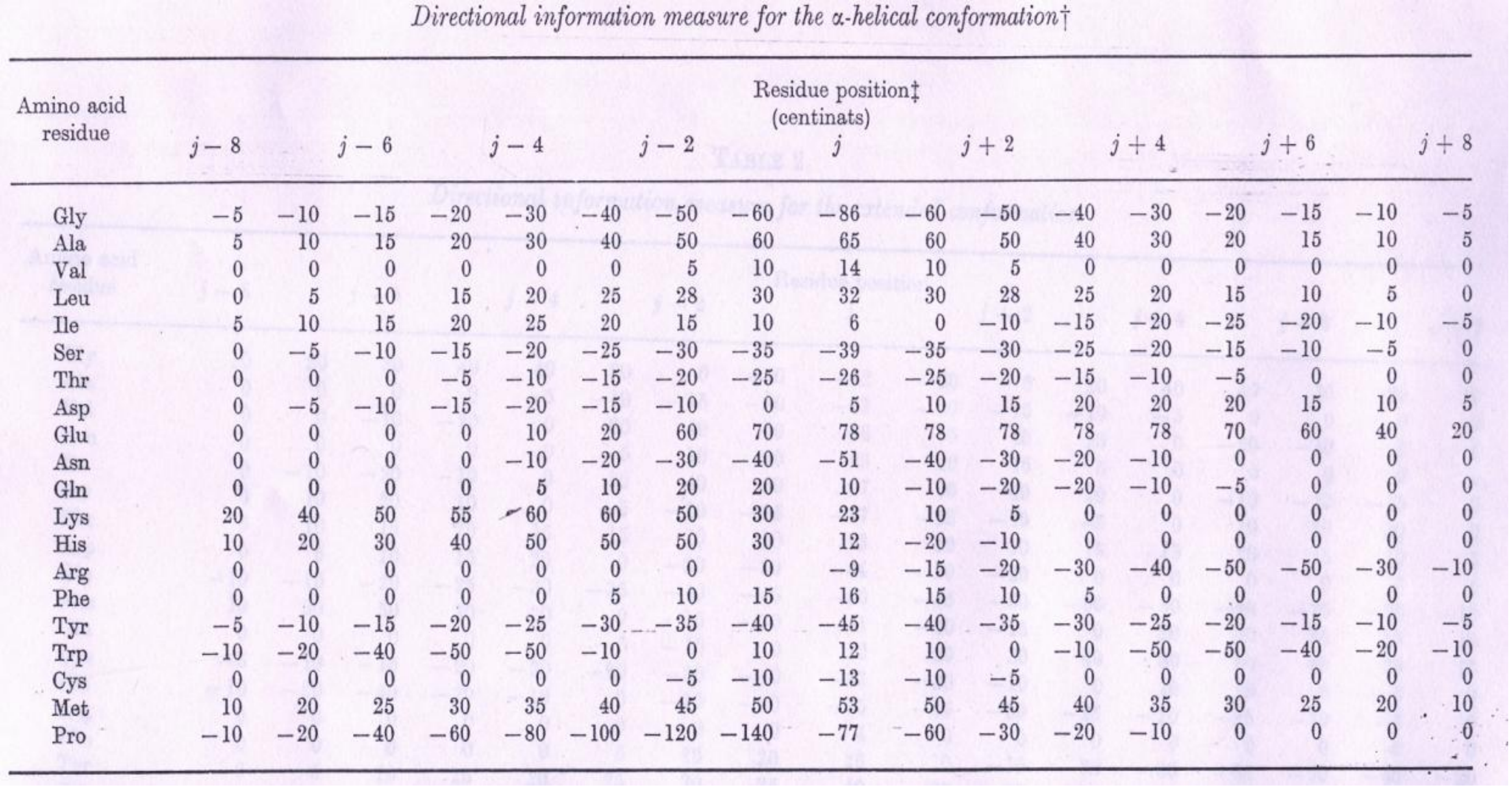


	Helix	~Helix	Total
Alanine (aa= A)	210	90	300
All residues	810	990	1800

P(SS=H|aa=A)	= 210/300	= 0.70
P(SS=~H|aa=A)	= 90/300	= 0.30
P(SS=H)	= 810/1800	= 0.45
P(SS=~H)	= 990/1800	= 0.55

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

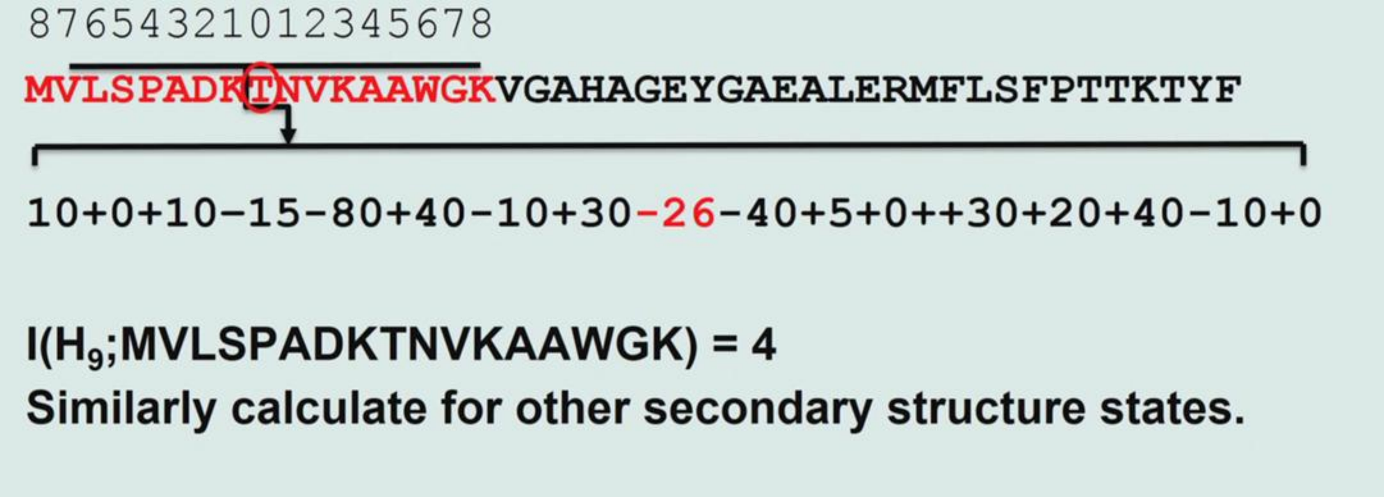
GOR method



Directional information measure for the α-helical conformation†

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

GOR method



87654321012345678

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYF

10+0+10−15−80+40−10+30−26−40+5+0++30+20+40−10+0

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

Similarly calculate for other secondary structure states.