01/2024

Bioinformatics

↳ biological information → used to extract useful inferences

Biological Data/Information → use of comp. sci concepts

DBMS, Programming

- 1D Data → genome sequences (A, T, G, C)
- Structural Data / 3D Data → information about structure (coordinates of macromolecules) → PDB Data

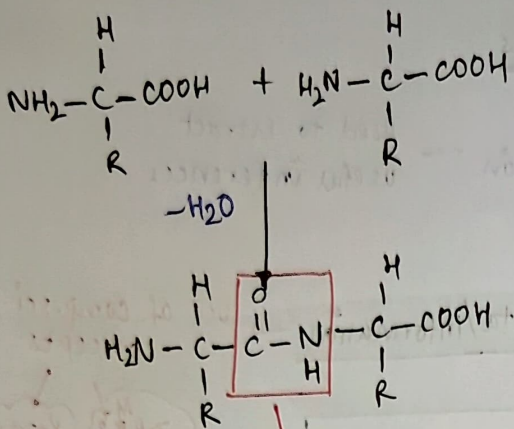
* 3D Data: generated from some experiments

Proteins
- structural → scaffolds that make up organisms
- functional → enzymes

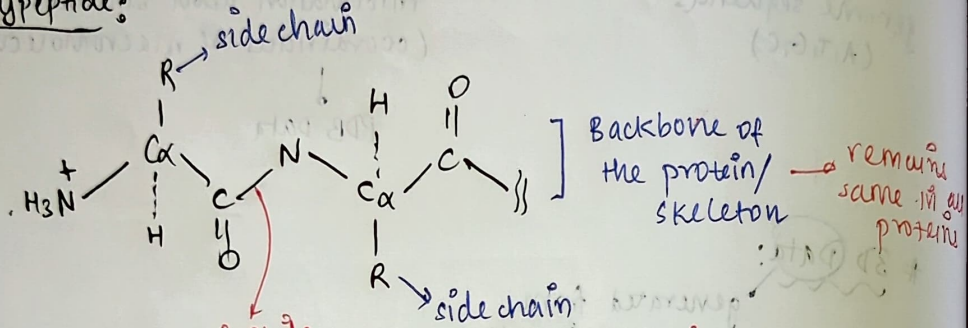
polymer of amino acid residues → should be folded into a 3D structure → should perform/exhibit some biological function

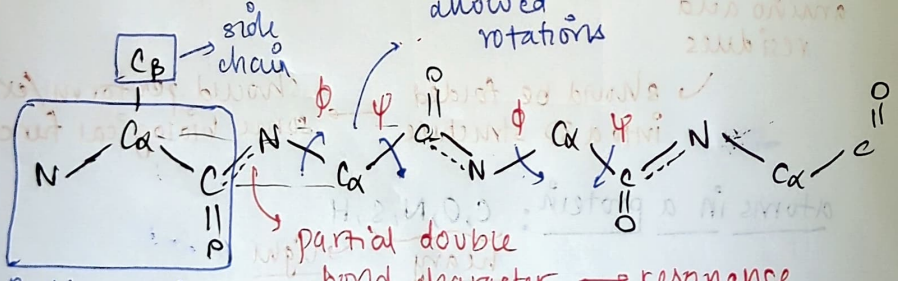
atoms in a protein: C, O, N, S, H
- heavy atoms → coordinates in PDB
- light atom (H) → coordinates not in PDB

structures solved by X-ray Crystallography
- H does not diffract X-ray → as only one e⁻ we cannot solve the Phase problem
- molecule in sol^{n} state → NMR → we get H coordinates



Polypeptide:





more energy than a single bond
↓
restricted rotation

trans conformation

R group & carbonyl oxygen on opp. sides.

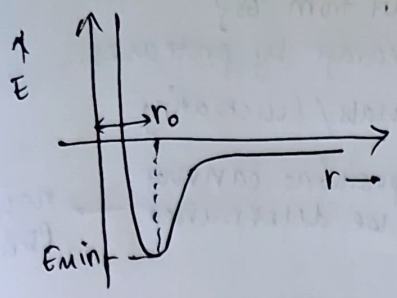
preferred!

found in protein structure

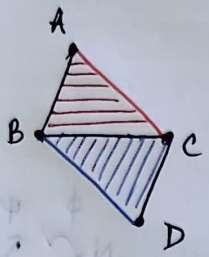
cis conformation

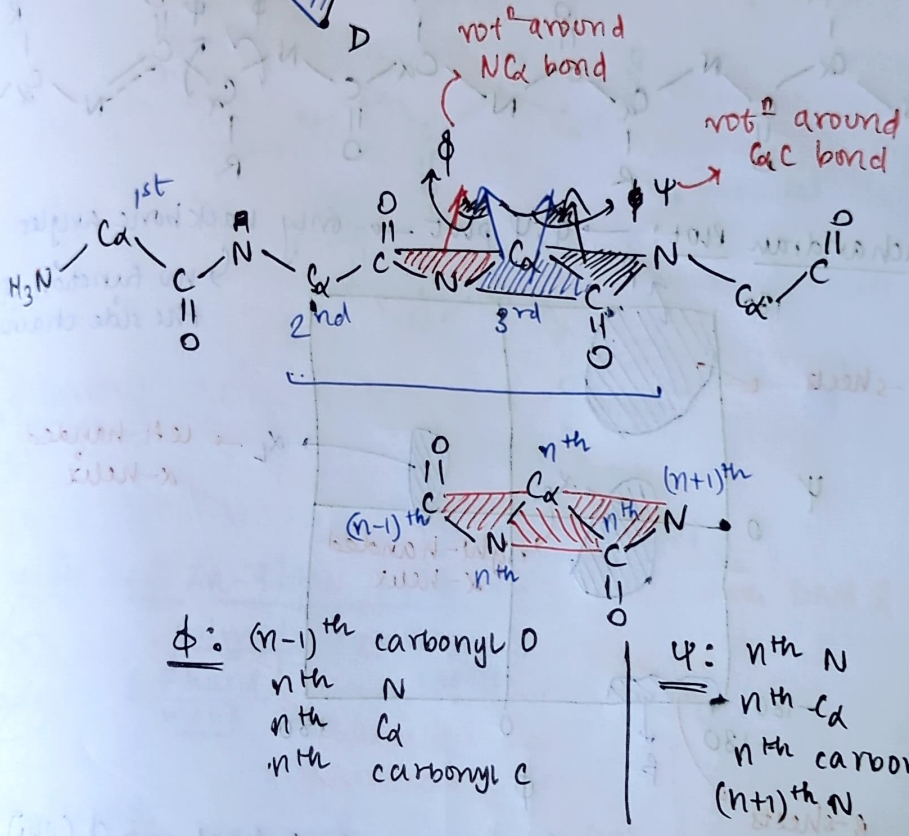
R group & carbonyl oxygen on the same side

interference of vdw radii ← steric hindrance



Lennard-Jones Pot^n





ϕ: $(n-1)^{th}$ carbonyl O
n^{th} N
n^{th} Cα
n^{th} carbonyl C

ψ: n^{th} N
n^{th} Cα
n^{th} carbonyl C
$(n+1)^{th}$ N

molecular visualization software
- PyMOL → most famous
- VMB
- RasMOL

protein crystallization → water molecules get trapped in the crystal → can see in a visualization software → HETATM

H.W.

Q. Read PDB Manual

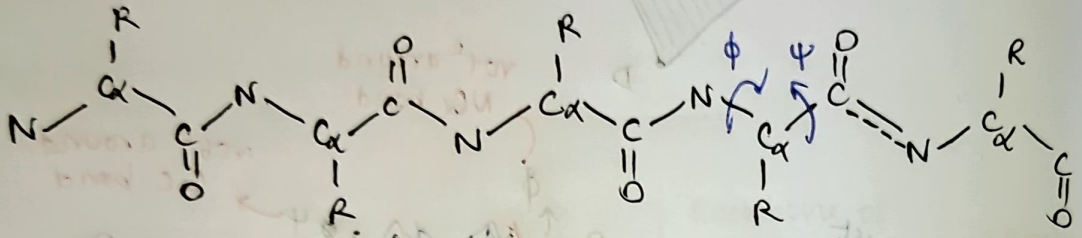
Q. Why does residue no. start from 68?

(1.) N- or C-terminal cleavage by proteases.

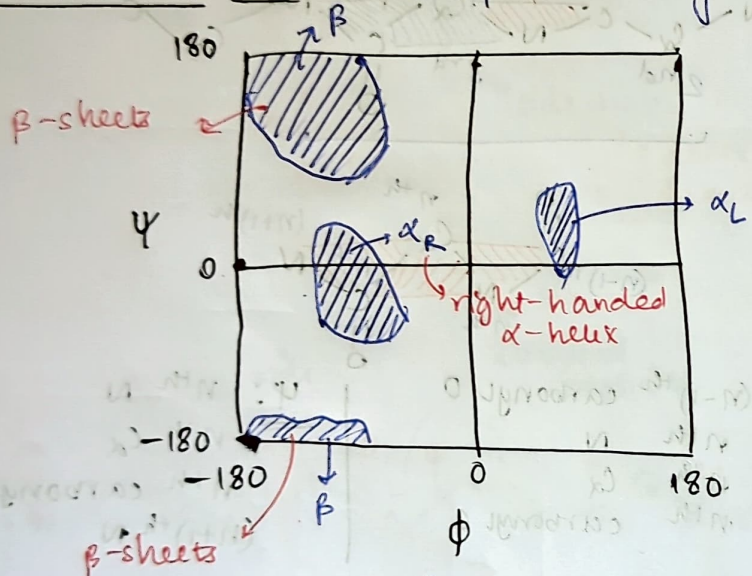
(2.) disordered terminals/fluctuating
↳ positions cannot be determined → not in the PDB file.

24/1/23

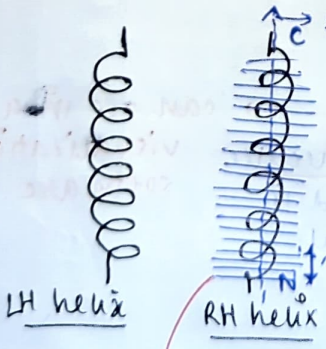
Trans-conformation



Ramachandran Plot: → ϕ;ψ plot → only backbone angles
↳ no function of the side chains.



rotation around N-Cα bond → ϕ (phi)
rotation around Cα-C bond → ψ (psi)



α-helix

per turn = 3.6 residues
rise per residue = 1.5 Å
rise per turn = 5.4 Å → pitch of the helix

Helices found in protein structures

- α-helices
- π-helices
- 3_{10}-helices

can be used to calculate length of helix

3_{10} helix: → if we compress α-helix

per turn = 3 residues

π-helix: → if we extend the α-helix.

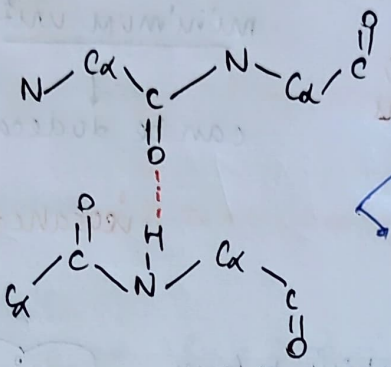
per turn = 5 residues

α-helix, β-sheet → 2° structures → stabilized by H-bonds → supports that hold helical structure

Regular 2° structure

α-helix H-bond pattern: i^{th} N–H··· / $(i+4)^{th}$ C=O··· (or) i^{th} C=O··· / $(i+4)^{th}$ N–H···

β sheet.



→ parallel

→ anti-parallel

Irregular structure → crowns, loops

Motifs (Super 2° structure) — not fully formed/folded but a consortium of 2° structures

eg: Zn-finger motifs → interact with DNA & RNA

Calmodulin → Ca^{2+}-binding motif → bind Ca^{2+}

EF hand motif → E helix & F-helix connected by a loop

E & H helices → Hb where porphyrin ring is present

can be used to perform bioinformatic analyses

eg: presence of EF hand motif can predict Ca^{2+} binding ability of protein at hand.

RNA-binding proteins → Arg-rich motifs.

Gamma-crystalin → detect photons from EM waves in our eyes & convert it into nerve signals.

Tertiary structure → single polypeptide chain fold to form func^{n} protein

eg: Myoglobin (muscles)
Lysozyme (tears)

Quarternary structure → assembly of more than one polypeptide chain & perform functions

eg: Hemoglobin → dimer of dimer

↓ assembled by non-covalent interactions → mostly hydrophobic & vdw interactions, electrostatic

↓ some can be covalently linked ↓ by disulphide bonds

eg: insulin, Antibodies, heat-shock proteins

minimum unit → dimer

↓ can be dodecamers (football-like) → viral capsids

icosahedral viruses

Propensity → ability to form α-helix → helical residues → more likely to form helices

proline → helix capping residue / helix breaker

↳ if we want to terminate the helix.

same protein

↓ 300 residues

↳ 50 leucine in total

LRR structure (leucine-rich)

helix 1

L
I
R
H
K
L
H
R
A
G

helix 2

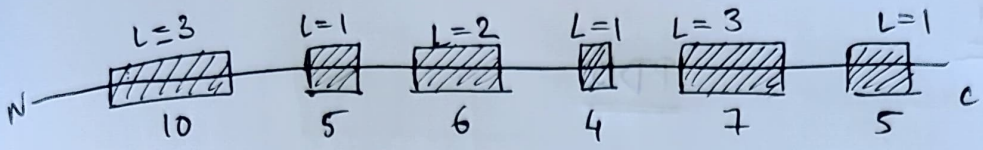
(say) Leucine (L)

freq. = 2/10

freq. (L) = 3/19 } 5 leucines in 19 helical residues

$$\text{Propensity} = \frac{5/19}{50/300}$$

} if we don't divide by this, we cannot take into account the fact that proteins have



$$P = \frac{\frac{3+1+2+1+3+1}{10+5+6+4+7+5}}{\frac{70}{300}}$$

$\frac{3+1+2+1+3+1}{10+5+6+4+7+5}$ → $f(h)$ sec. struct. (helical)

$\frac{70}{300}$ → $f_p(h)$ (total)

P = propensity

▨ helical region
300 total a.a.
70 total h in seq.

1DFU

PyMOL:

S - show ← cartoon, ribbon — only backbone, sticks

PyMOL > remove solvent ... # to remove water molecules.

viewing mode.

Edit mode → select 2 atoms → distance
" 3 atoms → angles
" 4 atoms → dihedral angle

H.W.

PyMOL manual — 18 pages

PyMOL > select polymer. protein // to select protein only
PyMOL > select polymer. nucleic // to select nucleic acid only.
PyMOL > select chain M // to select chain M of nucleic acid
PyMOL > color red, chain M // coloring chain M to red
PyMOL > color blue, resi 91 // coloring residue 91
PyMOL > color orange, chain P and ALA // coloring all alanines of chain P.

31/1/24

PDB

Resolution

2FBD

Asymmetric Unit

some proteins → monomeric
some proteins → oligomers
} → the functional unit of protein

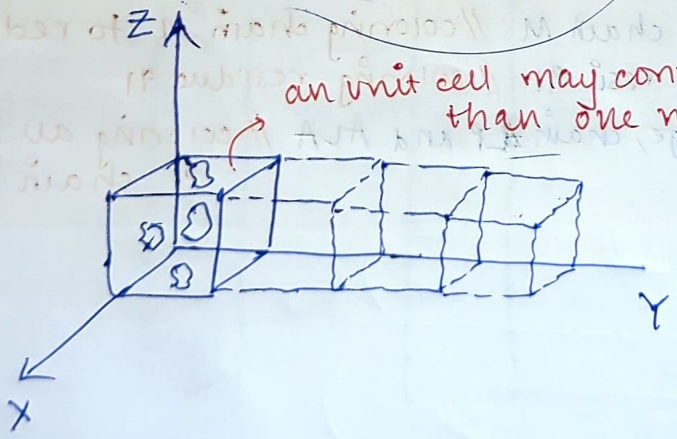
Biological Assembly

homodimer:
two identical polypeptide chains
eg: tRNA synthetase (Asp)

eg: $\alpha_2\beta_2$ of Hb (all 4 chains are required for the function)
↓
dimer of dimers

heterodimer:
two different polypeptide chains
eg: light & heavy chains in Ab

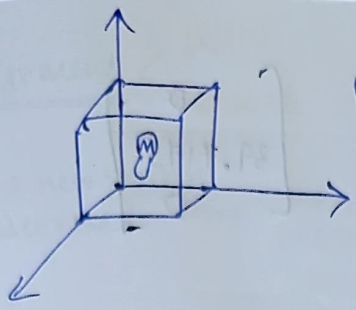
Assymetric Unit
↳ property of the crystal



an unit cell may contain more than one molecule

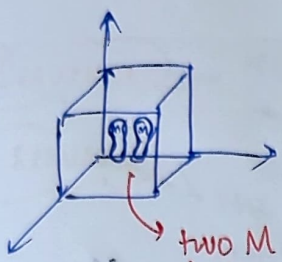
(M) Myoglobin (say)

crystallization is not a deterministic process.
↳ we cannot determine which lattice/space group we are going to obtain



Myoglobin

bio. assembly = monomeric
asymm. unit = monomeric

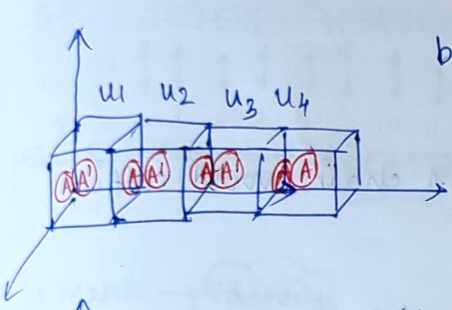


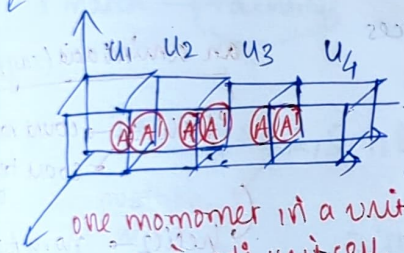
bio. assembly = monomer.
asymm. unit = dimer

two M come together due to forces during crystallization

Asp tRNA synthetase

biological assembly = (A)(A') = dimer
asym. unit = monomer





Biological assembly

one monomer in a unit cell & in adj. unit cell, we have the other monomer

Unit cell parameters in a PDB file:

CRYST1 36.523 79.435 45.203 90.0 102.97 90.00 P 1 21 1

a b c α β γ

2 fold screw axis in y-direction

transformation matrix

$$T' = \underset{3\times3}{[R]} \overset{X}{\cancel{\ }} \underset{3\times1}{[T]}$$

Symmetry Operator

X	Y	Z
-X	Y+1/2	-Z

↳ half unit cell translation along y axis

$\frac{79.435}{2} = 39.7175$

- ✓ Point group symmetry
- ✓ Space group symmetry

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 39.7175 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} x_1' & y_1' & z_1' \\ x_2' & y_2' & z_2' \\ \vdots & \vdots & \vdots \\ x_n' & y_n' & z_n' \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{bmatrix}$$

aa

Assignment:

↳ propensity of only A chain would suffice.

check residue stretch of helices

eg: HELIX 1 RESIDUES 4-15

Read the residues 4-15 & count

an amino acid (say Asp)

↳ helix 1 → count no. of A
count total no of aa

↳ helix 2 → no. of Asp
count # of aa

↳ helix 8 → no of Asp
# of aas

→ total Asp
→ total aa

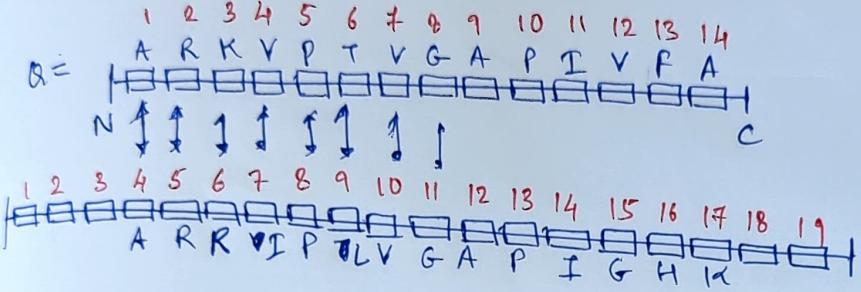
$$Ala_p = \frac{\left(\frac{\#\ Ala\ in\ all\ (8)\ helices}{\#\ all\ aa\ in\ 8\ helices}\right)}{\left(\frac{total\ \#\ Ala\ present\ in\ protein\ chain\ A}{total\ \#\ aa\ in\ protein\ chain\ A}\right)}$$

Sequence Alignment < Local / Global

when a new protein is discovered → checking whether this protein is similar to some existing protein → we can then draw info. about evolution

Query Sequence → the protein we are trying to study

Target Sequence → the proteins with which we are trying to align our Query sequence



a match → identity

both R & K are +ve charged

Identity: 1, 2, 5, 7, 8 (position w.r.t. Q)

Similar: 3, 4

Different: 6

✓ if the aligned position has same amino acid → Identity

✓ if the aligned position has same type of amino acid residue (eg: hydrophobic, aromatic) → Similarity

eg: A R & K
D & E
T, P, F, W
N & Q.

% identity = (no. of identical positions / total no. of aligned positions) X100

% similarity = (no. of similar positions / total no. of aligned positions) X100

similarity in nts:

A & A → identical

A & G → similar (both purines).

7/2/24

Homology Modelling of Proteins

Prediction of 3D structure of a target protein from the amino acid seq (1° structure) of a homologous (template) protein for which an X-ray or NMR structure is available.

N-term ——— a.a. seq ——— C-term.

prediction ↓

Model structure → may or may not be correct.

Methods:

1. homology modelling / comparative modelling
2. Threading.

1. Comparative / homology modelling:

N-term ——— a.a. seq. ——— C-term } Target sequence: seq. for which we want to build the model.

PDB → 3D structures of proteins
Prosite database → 1D sequence of proteins

↓ similarity search → [BLAST] → blastp (align only at seq. level)

search against the database → nrPDB → non-redundant protein database → only 1 structure for a protein (not like PDB where 1 protein may have 1000s of structures)

↓ sequence of proteins which have 3D structures

↓

N ——— Target seq. ——— C

→ possible similar seq.

their 3D structures are in nrPDB ←

Output:

% identity
% similarity
Query Coverage → how many aa. are covered → poor query coverage makes homology modelling difficult

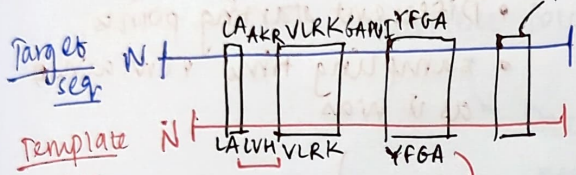
selection of one structure that would act as template structure → used as a reference for homology modelling.

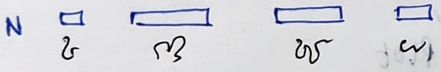
Target ↕ matching
Template

Conditions:

1. Sequence identity b/w tg and tm > 30% — 20-30% Twilight zone
2. Query coverage → how much % of tg is aligned with tm > 95%

identical regions / aligned regions



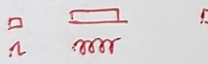


→ now we know the structures of these fragments

the unstructured parts → fragmented parts of homology modeling

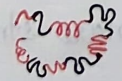
take each fragment and do the BLAST search again on nrPDB

eg: AKR → we will get similarity for such a small fragment

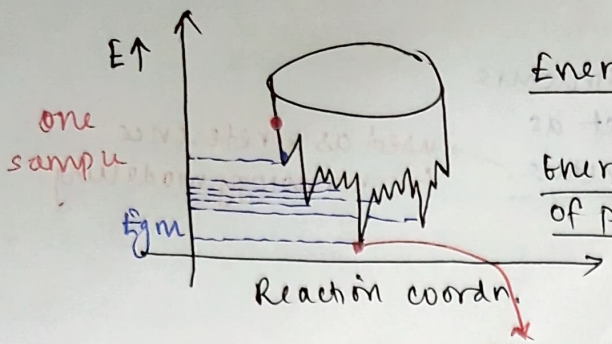


coordinates for each & every fragment

next step is to connect/ stitch the fragments together



→ Energy minimization of the entire structure



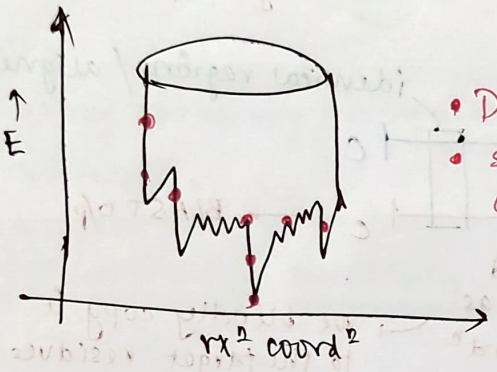
Energy funnel Diagram / Energy landscape of protein folding

Global minima

more sampling → to obtain the global minima. → sample might get stuck to any local minima if it is deep enough

solution : 10 different starting structures

Monte Carlo



multiple samples

- Different starting points
- sampling time remains as it was

↓

chances of reaching global minum ↑.

once model is selected, check Ramachandran Plot

Homology Modelling Softwares

1. SWISS-MODEL
2. Modeller

CASP Critical assessment of structure prediction

John Hopkins University

✓ However, for homology modelling, template structure is a must.

(2) Threading :

→ performed when template structure is not available

3D structure space

what/how many different types of fold are possible in protein universe?

C- Class
A- Architecture
T- Topology
H- Homologous Superfamily

CATH Database

unit → Fold space

1200 folds are possible in protein universe

↳ any protein we take from any species, their architecture will be 1 of these 1200 unit folds

↓

take target sequence & thread with the 1200 possible folds

↳ Energy minimizⁿ → Model structure