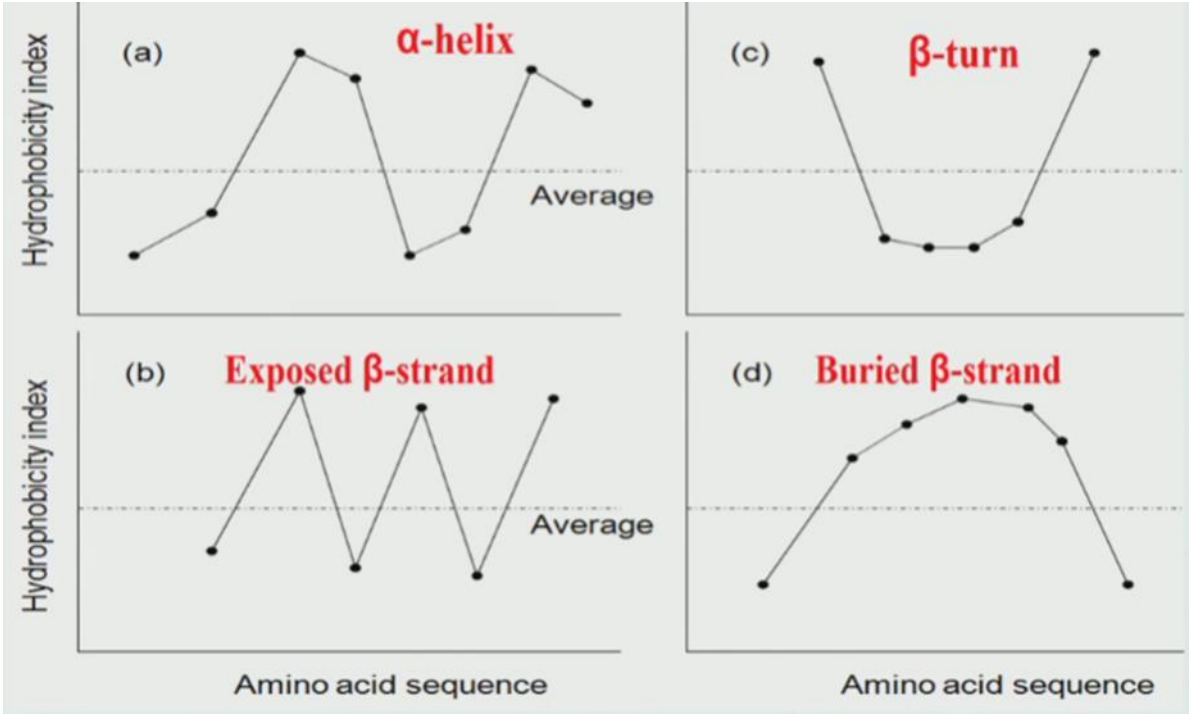
# Protein structure prediction

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

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

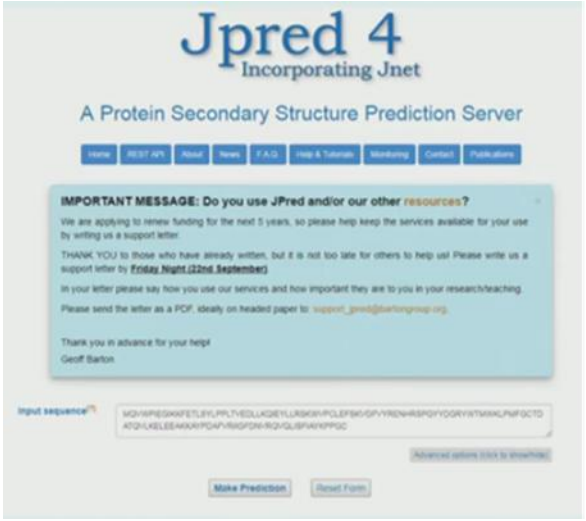
***October 07, 2025***

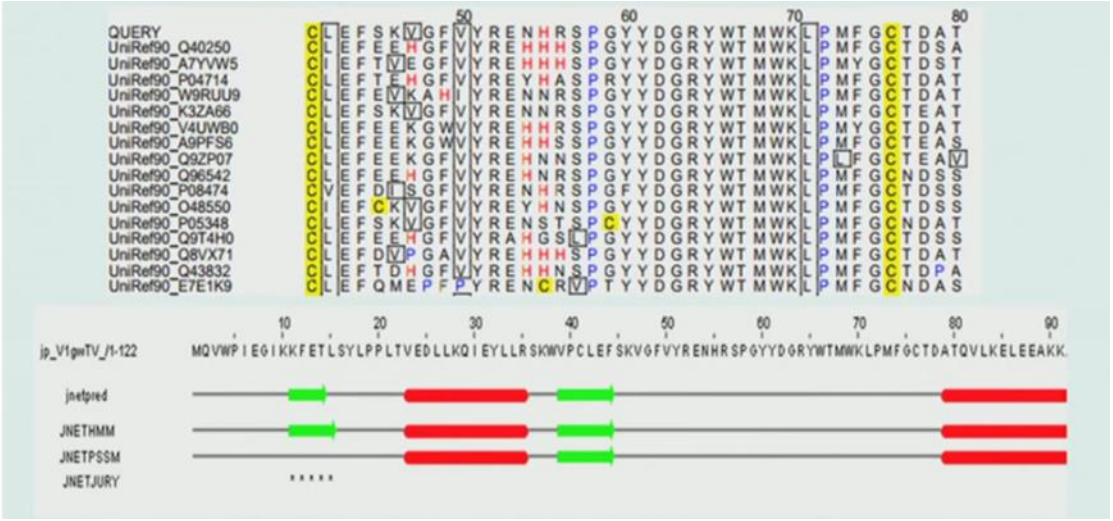
# Hydrophobicity Profile



# Multiple sequence alignment

- Average of helix, strand, coil and turn parameters (GOR) for all the aligned residues at each position
- Insertion is given a value of zero
- Confidence of the prediction is related to the conservation score (0-1) at the position
- Zpred – First SS predictor using MSA

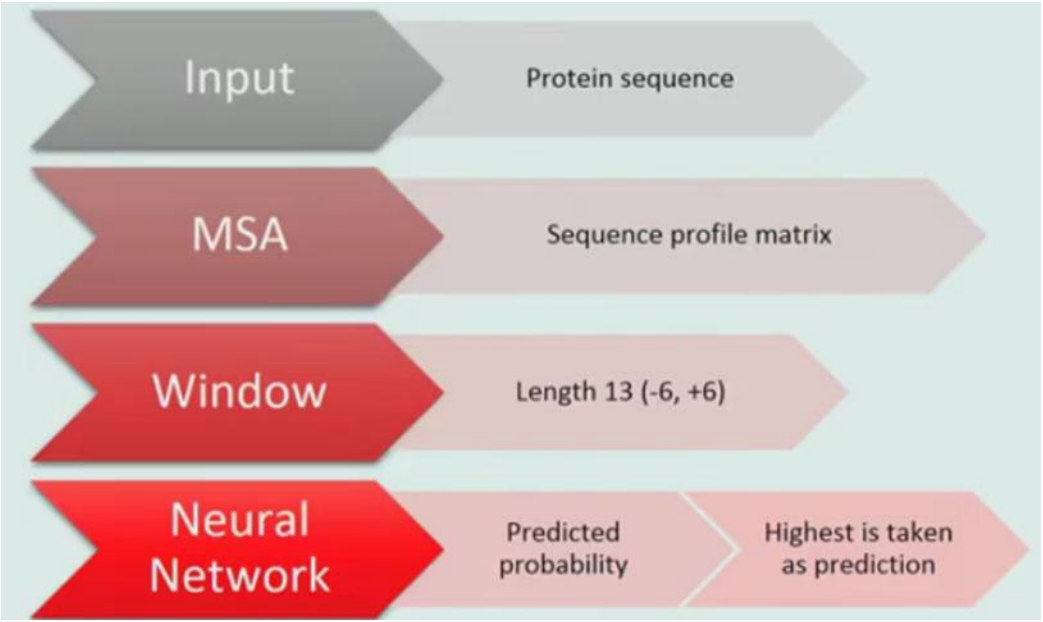


# Machine learning techniques

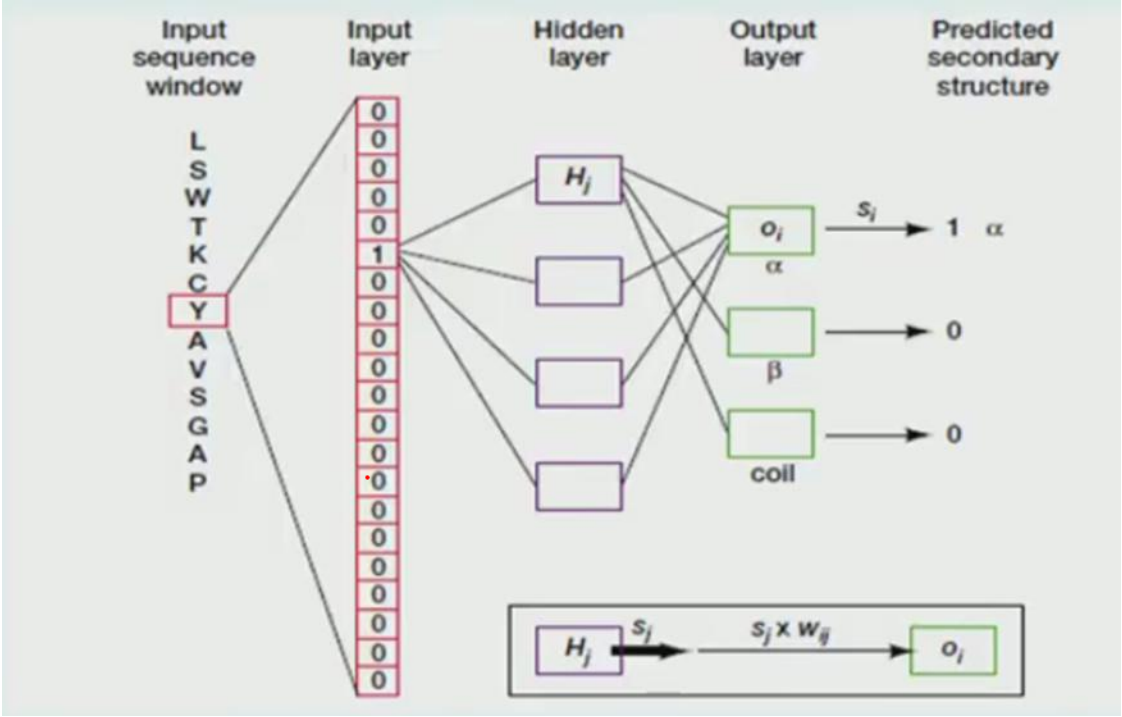
## Profile neural network systems from HeiDelberg (PHD)

- Aspects:
  - Multiple sequence alignment → Evolutionary information
  - Neural networks → Learning secondary structure – amino acid residue pattern/information
  - Window approximation → Captures effect of neighboring residues



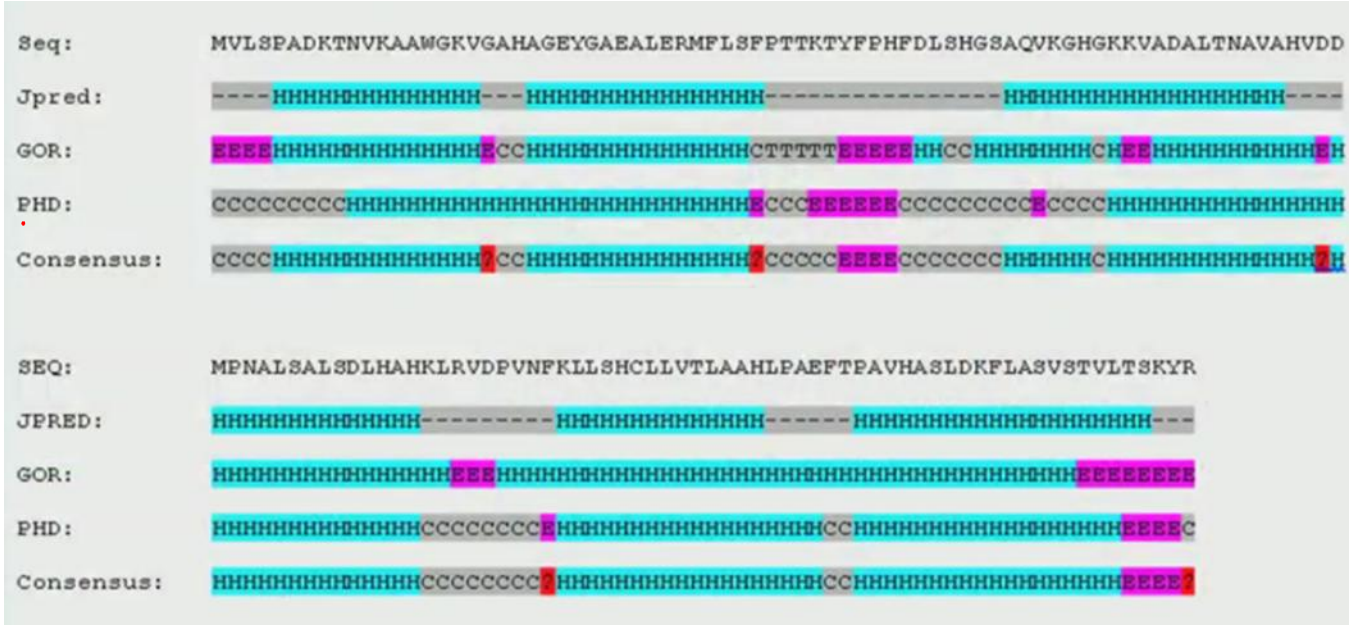
# Machine learning techniques

## Profile neural network systems from HeiDelberg (PHD)



# Consensus (joint) prediction methods

1. Consensus
   - Input: Prediction results from other methods
   - Strategy: Majority of voting
   - Output: Based on high vote



# Consensus (joint) prediction methods

1. Meta prediction
   - Input: Prediction results from other methods
   - Strategy: Train on machine learning methods
   - Output: Improved outcome

# Summary of secondary structure prediction methods

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)

# Knowledge based prediction of protein tertiary structure

Homology Modeling

- Need homologues of known protein structure
- Backbone modelling
- Side chain modelling
- Fails in absence of similarity with other proteins

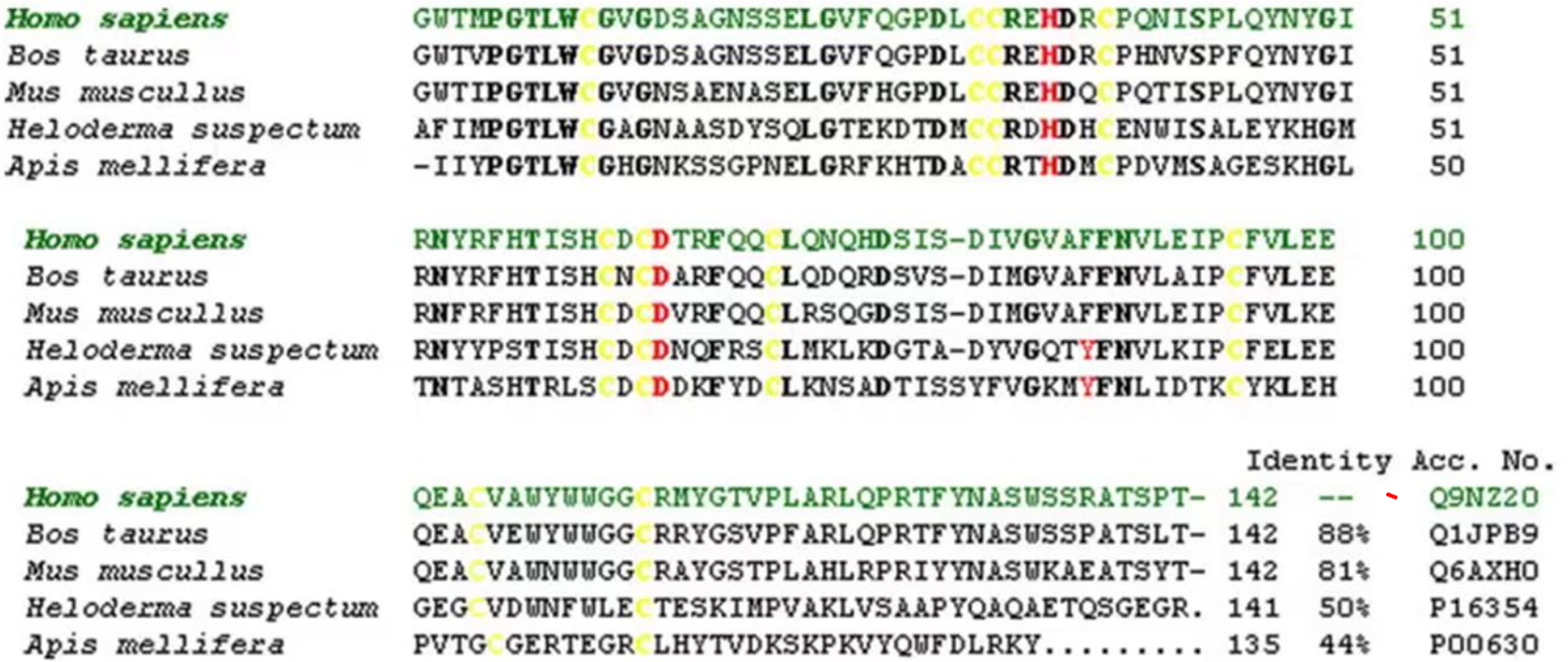
Threading Based Methods

- New way of fold recognition
- Sequence is tried to fit in known structures
- Motif recognition
- Loop & Side chain modelling
- Fails in absence of known example

# Homology modeling of proteins

- Structure of a protein is uniquely determined by its amino acid sequence. Therefore, knowing the sequence should, at least in theory, suffice to obtain the structure.
- During evolution, a structure is more stable and changes at a slower rate as compared to its sequence, so that similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures.
- If the percentage of similarity or homology is 40%, two sequences are practically guaranteed to adopt a similar structure.

# Step 1- Finding the best template



# Step 1- Finding the best template

## human plasmakallikrein protease domain in complex with active site directed inhibitor [Homo sapiens]

**Sequence ID:** 6T7P_A **Length:** 242 **Number of Matches:** 1

**Range 1: 1 to 237** GenPept Graphics ▼ Next Match ▲ Previous Match

| Score | Expect | Identities | Positives | Gaps |
|---|---|---|---|---|
| 213 bits(543) | 7e-66 | 103/240(43%) | 149/240(62%) | 8/240(3%) |

```
Query  25
Sbjct  1
Query  31
Sbjct  59
Query  37
Sbjct  11
Query  43
Sbjct  17
```

## Crystal structure of the extracellular region of the transmembrane serine protease hepsin with covalently bound preferred substrate. [Homo sapiens]

**Sequence ID:** 1Z8G_A **Length:** 372 **Number of Matches:** 1

**Range 1: 18 to 360** GenPept Graphics ▼ Next Match ▲ Previous Match

| Score | Expect | Identities | Positives | Gaps |
|---|---|---|---|---|
| 211 bits(537) | 2e-63 | 118/346(34%) | 174/346(50%) | 17/346(4%) |

```
Query  158  LQVYSSQRKSWHPVCQDDWNENYGRAACRDMGYKNNFYSSQGIVDDSGSTSFMKLN-TSA  216
            L V+     +W  +C    N      +C +MG+      S+  V  +G+
Sbjct  18   LMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRALTHSELDVRTAGAAGTSGFFCVDE  77

Query  217  GNVDIYKKLYHSDA---CSSKAVVSLRCIACGVNLNSSRQSRIVGGESALPGAWPWQVSL  273
            G +   ++L    +   C     ++  C  CG         RIVGG     G WPWQVSL
Sbjct  78   GRLPHTQRLLEVISVCDCPRGRFLAAICQDCG--RRKLPVDRIVGGRDTSLGRWPWQVSL  135

Query  274  HVQNVHVCGGSIITPEWIVTAAHCVEKPLNNPWHWTAFAGILRQSFMFYGAGYQVEKVIS  333
                 H+CGGS+++ +W++TAAHC  +       W  FAG + Q+   +G    V+ V+
Sbjct  136  RYDGAHLCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAVAQASP-HGLQLGVQAVVY  194

Query  334  HPNY------DSKTKNNDIALMKLQKPLTFNDLVKPVCLPNPGMMLQPEQLCWISGWGAT  387
            H  Y      +S+  +NDIAL+ L  PL   + ++PVCLP  G  L   ++C ++GWG T
Sbjct  195  HGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALVDGKICTVTGWGNT  254

Query  388  EEKGKTSEVLNAAKVLLIETQRCNSRYVYDNLITPAMICAGFLQGNVDSCQGDSGGPLV-  446
            +  G+ + VL  A+V +I    CN    Y N I P M CAG+ +G +D+CQGDSGGP V
Sbjct  255  QYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGGIDACQGDSGGPFVC  314

Query  447  ---TSKNNIWWLIGDTSWGSGCAKAYRPGVYGNVMVFTDWIYRQMR  489
                S+   W L G  SWG+GCA A +PGVY  V  F +WI++ ++
Sbjct  315  EDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIFQAIK  360
```

**Related Information**

Structure - 3D structure displays

**Crystal structure of the extracellular region of the transmembrane serine protease hepsin with covalently bound preferred substrate. [Homo sapiens]**

Sequence ID: 1Z8G_A Length: 372 Number of Matches: 1

Range 1: 18 to 360 GenPept Graphics ▼ Next Match ▲ Previous Match

| Score | Expect | Identities | Positives | Gaps |
|---|---|---|---|---|
| 211 bits(537) | 2e-63 | 118/346(34%) | 174/346(50%) | 17/346(4%) |

```
Query  158  LQVYSSQRKSWHPVCQDDWNENYGRAACRDMGYKNNFYSSQGIVDDSGSTSFMKLN-TSA  216
            L V+     +W  +C    N      +C +MG+      S+  V  +G+
Sbjct  18   LMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRALTHSELDVRTAGAAGTSGFFCVDE  77

Query  217  GNVDIYKKLYHSDA---CSSKAVVSLRCIACGVNLNSSRQSRIVGGESALPGAWPWQVSL  273
            G +   ++L    +   C     ++  C  CG          RIVGG     G WPWQVSL
Sbjct  78   GRLPHTQRLLEVISVCDCPRGRFLAAICQDCG--RRKLPVDRIVGGRDTSLGRWPWQVSL  135

Query  274  HVQNVHVCGGSIITPEWIVTAAHCVEKPLNNPWHWTAFAGILRQSFMFYGAGYQVEKVIS  333
                 H+CGGS+++ +W++TAAHC  +       W  FAG + Q+   +G    V+ V+
Sbjct  136  RYDGAHLCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAVAQASP-HGLQLGVQAVVY  194

Query  334  HPNY------DSKTKNNDIALMKLQKPLTFNDLVKPVCLPNPGMMLQPEQLCWISGWGAT  387
            H  Y      +S+  +NDIAL+ L   PL    + ++PVCLP  G  L   ++C ++GWG T
Sbjct  195  HGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALVDGKICTVTGWGNT  254

Query  388  EEKGKTSEVLNAAKVLLIETQRCNSRYVYDNLITPAMICAGFLQGNVDSCQGDSGGPLV-  446
            +  G+ + VL  A+V +I    CN    Y N I P M CAG+ +G +D+CQGDSGGP V
Sbjct  255  QYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGGIDACQGDSGGPFVC  314

Query  447  ---TSKNNIWWLIGDTSWGSGCAKAYRPGVYGNVMVFTDWIYRQMR  489
               S+   W L G  SWG+GCA A +PGVY  V  F +WI++ ++
Sbjct  315  EDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIFQAIK  360
```

**Related Information**

Structure - 3D structure displays

**Human hepsin protease in complex with the Fab fragment of an inhibitory antibody [Homo sapiens]**

Sequence ID: 3T2N_A Length: 372 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 18 to 360 GenPept Graphics ▼ Next Match ▲ Previous Match

| Score | Expect | Identities | Positives | Gaps |
|---|---|---|---|---|
| 211 bits(537) | 2e-63 | 118/346(34%) | 174/346(50%) | 17/346(4%) |

```
Query  158  LQVYSSQRKSWHPVCQDDWNENYGRAACRDMGYKNNFYSSQGIVDDSGSTSFMKLN-TSA  216
            L V+     +W  +C    N      +C +MG+      S+  V  +G+
Sbjct  18   LMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRALTHSELDVRTAGANGTSGFFCVDE  77

Query  217  GNVDIYKKLYHSDA---CSSKAVVSLRCIACGVNLNSSRQSRIVGGESALPGAWPWQVSL  273
            G +   ++L    +   C     ++  C  CG          RIVGG     G WPWQVSL
Sbjct  78   GRLPHTQRLLEVISVCDCPRGRFLAAICQDCG--RRKLPVDRIVGGRDTSLGRWPWQVSL  135

Query  274  HVQNVHVCGGSIITPEWIVTAAHCVEKPLNNPWHWTAFAGILRQSFMFYGAGYQVEKVIS  333
                 H+CGGS+++ +W++TAAHC  +       W  FAG + Q+   +G    V+ V+
Sbjct  136  RYDGAHLCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAVAQASP-HGLQLGVQAVVY  194

Query  334  HPNY------DSKTKNNDIALMKLQKPLTFNDLVKPVCLPNPGMMLQPEQLCWISGWGAT  387
            H  Y      +S+  +NDIAL+ L   PL    + ++PVCLP  G  L   ++C ++GWG T
Sbjct  195  HGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALVDGKICTVTGWGNT  254

Query  388  EEKGKTSEVLNAAKVLLIETQRCNSRYVYDNLITPAMICAGFLQGNVDSCQGDSGGPLV-  446
            +  G+ + VL  A+V +I    CN    Y N I P M CAG+ +G +D+CQGDSGGP V
Sbjct  255  QYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGGIDACQGDSGGPFVC  314

Query  447  ---TSKNNIWWLIGDTSWGSGCAKAYRPGVYGNVMVFTDWIYRQMR  489
               S+   W L G  SWG+GCA A +PGVY  V  F +WI++ ++
Sbjct  315  EDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIFQAIK  360
```

**Related Information**

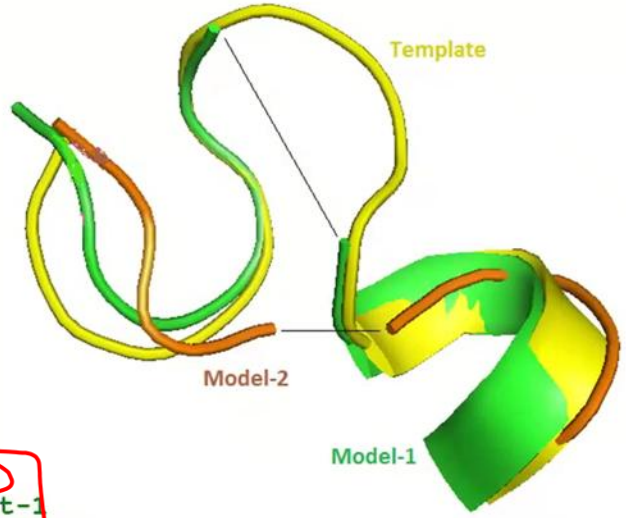
Structure - 3D structure displays

Identical Proteins - Identical proteins to 3T2N_A

# Step2- Correct sequence alignment

- Sometimes it may be difficult to align two sequences in a region where the percentage sequence identity is very low.

- A simple sequence alignment gives the higher score based on matching/pairing of residues on the two sequences

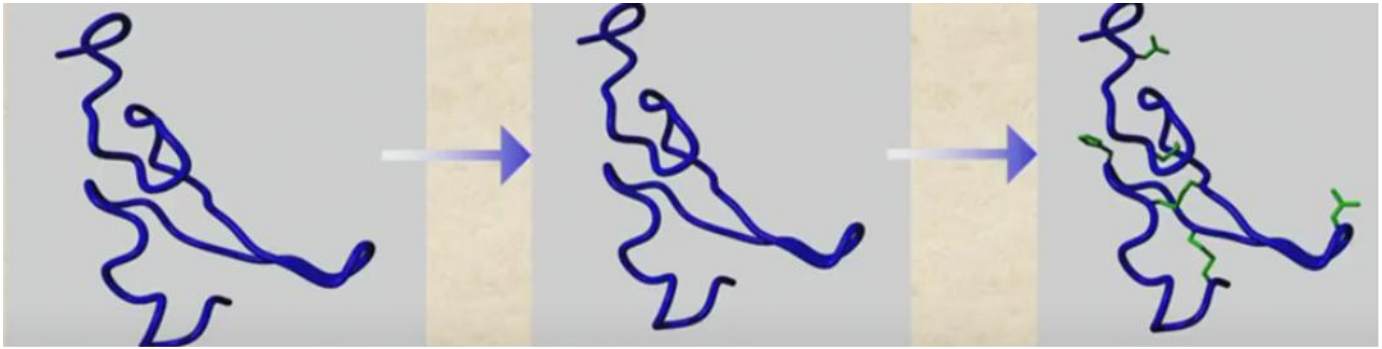



```
 1   2   3   4   5   6   7   8   9  10  11  12  13  14
ALA SER ILE CYS ARG LEU PRO GLY SER ALA GLU GLY VAL CYS Template
ALA ASN VAL CYS ARG THR PRO --- --- --- GLU GLY ILE CYS Alignment-1
ALA ASN VAL CYS ARG --- --- --- THR PRO GLU GLY ILE CYS Alignment-2
```

- Model-2 for alignment-2 is correct, because it leads to a small gap and can be corrected by small backbone shifts.

# Step3- Generating the backbone

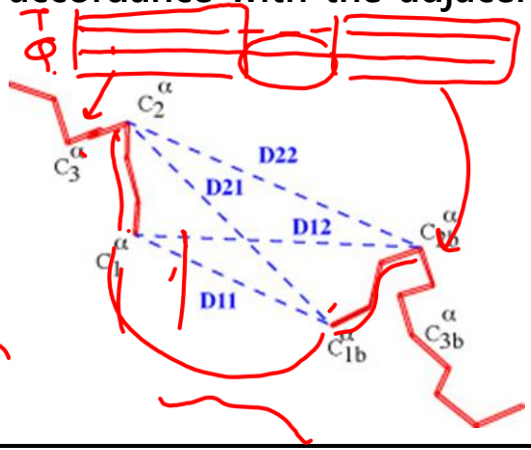
- Generate backbone coordinates from the template for the aligned regions.
- If most of aligned residues are identical, back bone too will be identical
- Template that generates a back bone with least errors is chosen.
- Back bone of secondary structural elements like helices and beta sheets are usually maintained, while variations may be seen in the loops and turns,



# Step4- Loop modeling

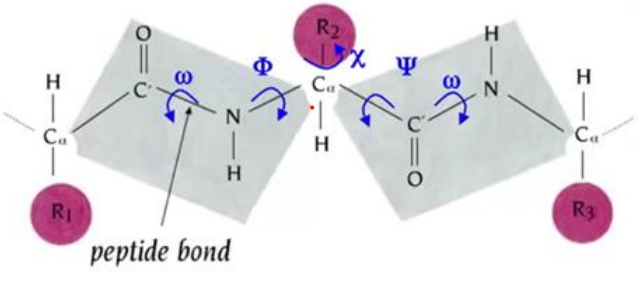
- Region of query sequence that is not aligned to template is modeled as a loop.
- Loop modeling is done by knowledge-based method where PDB is searched for known loops, or by Energy based method where long chains are built by sampling Ramachandran conformations randomly.
- For deletions, residues from the template is omitted, creating a void in the model that is filled by bringing together the adjacent residues' coordinates.
- For insertions, the coordinates for the missing residues is given a loop conformation in most cases and simulated for it to take up a secondary structure in accordance with the adjacent structures in the template.

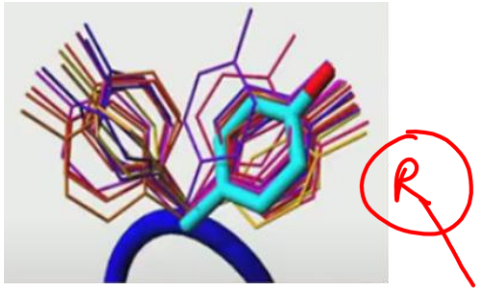
*Selecting protein fragments for loop modelling: In the example shown here, the gap in the framework is characterized by 4 distances (D11, D12, D21 and D22) between the 2 Cα preceding the gap and the 2 Cα following the gaps. The library of protein structure fragments of appropriate size (i.e. Ngap + 4, where Ngap is the number of residues in the gap and 4 corresponds to the 2 anchoring residues on each side of the gap) is scanned and the fragments with the proper geometry (i.e. such that the distances between the 2 first and last residues match the 4 characteristic distances of the gap) are retained has potential loops to fill in the gap.*



# Step5- Side chain modeling

- Involves placing side chains coordinates either obtained from a template structure or generated from *ab initio* modeling simulations or both.
- Side chain remains more flexible on the surface and tend to adopt multiple conformations as compared to the core where residues have restricted environment and limited conformational liberty.
- At high levels of sequence identity χ1-angles adopt positions from template structure and for sequence identity <35% combinatorial approach using rotamer libraries are used.
- Wrong backbone will affect the side chain building process.
- Prediction accuracy is highly dependent on scoring functions.



TYR

# Step6- Model optimization and structure refinement

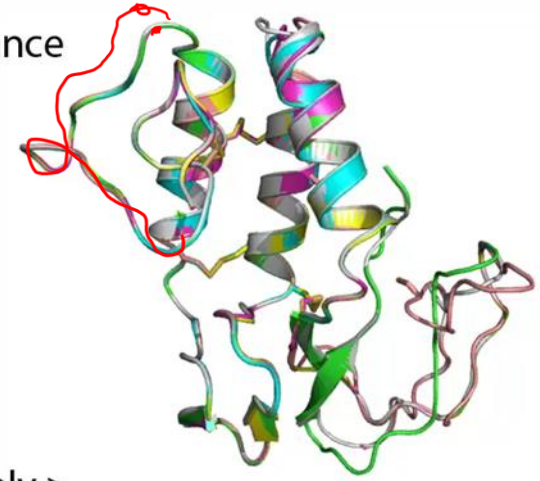
- Model optimization tries to bring it closer to real in-vivo conformation.
- Commonly achieved by Molecular Dynamics Simulation.
- Simulation follows the motions of the protein on a femto-second timescale and mimics true folding process.
- This helps to get a true structural model having minimum energy configuration

# Step6- Validation of the developed model

## Root mean square deviation (RMSD)

- Root mean square deviation is the measure of the average distance between the atoms of two superimposed proteins.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$



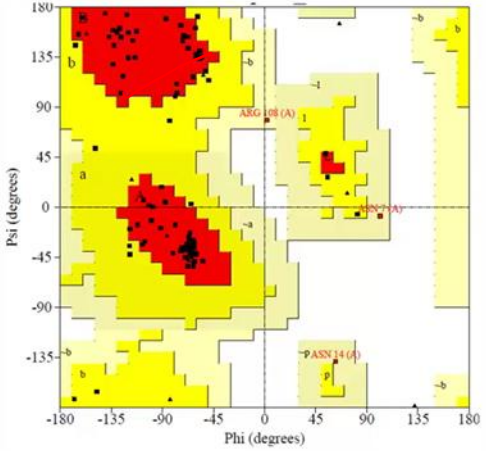
- δ is the distance between number of pairs of equivalent atoms

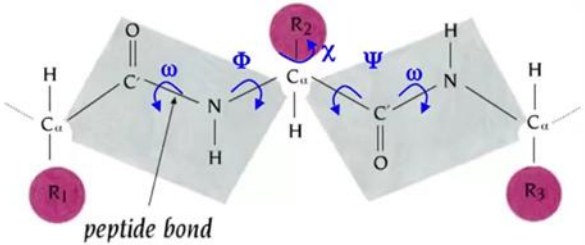
- 0.0-0.5Å : Essentially Identical; <1.5 Å: Very good fit; < 5.0 Å: Moderately > 7.0 Å: Dubious; > 12.0 Å: Completely unrelated

# Step6- Validation of the developed model

## Ramachandran Plot

- Atoms are treated as hard spheres with dimensions corresponding to their van der Walls radii
- Residues with phi and psi angles which cause atoms to collide are considered to be in sterically disallowed conformations
- Most allowed region occupied by residues with no steric clashes
- Dis-allowed region with residues whose atoms come closer than their van der Wall radii.
- Additionally allowed region with residues have atoms with inter-atomic distance slightly shorter than van der Walls radii.



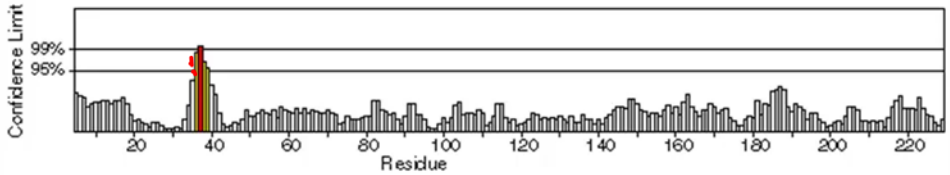



# Step6- Validation of the developed model

## ERRAT Plot

- Of three heavy atoms carbon, nitrogen, and oxygen, there exist six different combinations of pairwise non-covalently bonded interactions (CC, CN, CO, NN, NO, and OO).

- In a database of 96 reliable protein structures, these set of pairwise interactions are characterized; ERRAT finds regions of model structures that have incorrect pattern of non-bonded interactions that result in incorrectly folded regions

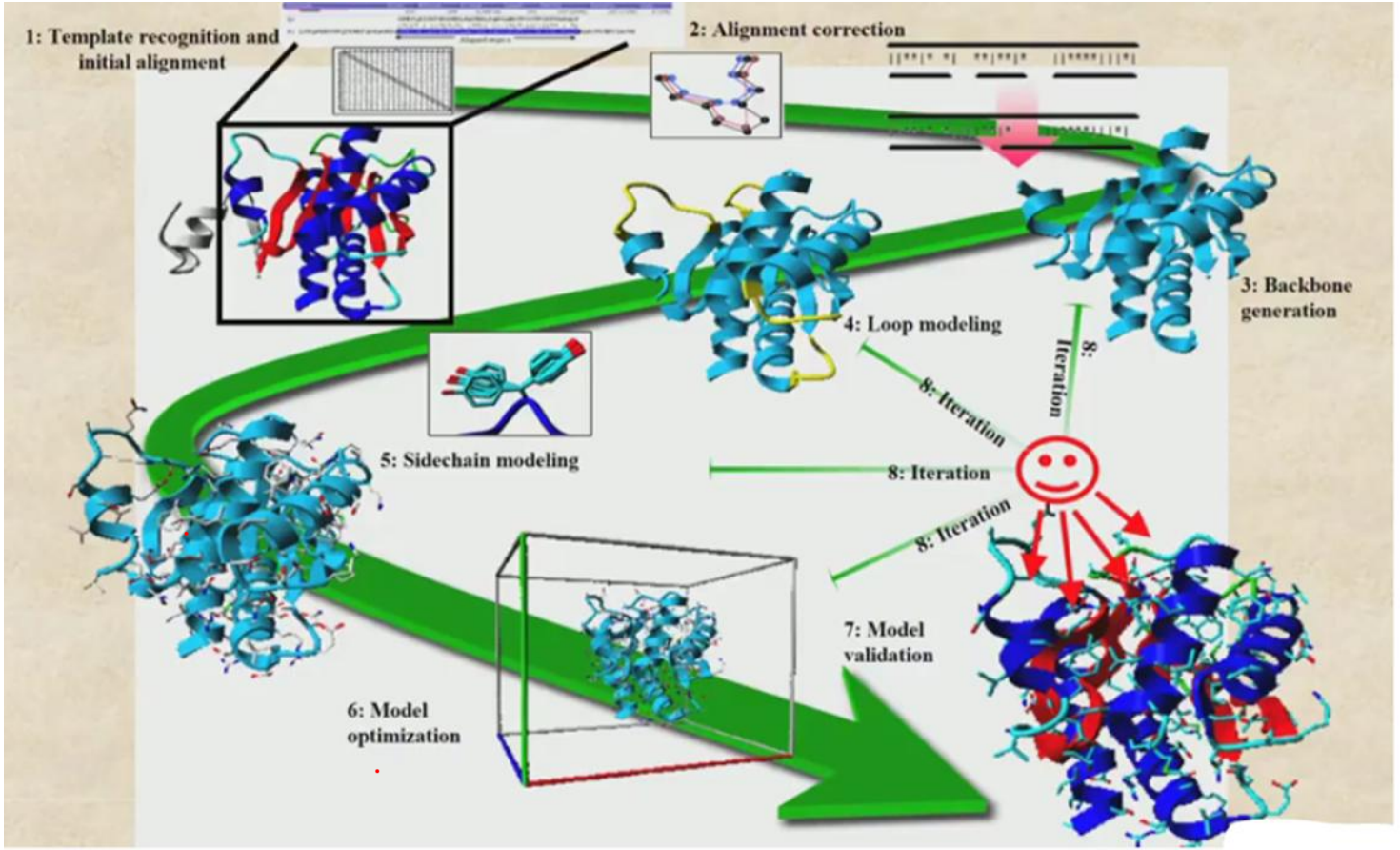



Regions of the structure that can be rejected at the 95% confidence level are shown in yellow and regions that can be rejected at the 99% level are shown in red.

# Step6- Validation of the developed model

- WHAT IF http://www.cmbi.kun.nl/gv/servers/WIWWWI/
- SOV http://predictioncenter.llnl.gov/local/sov/sov.html
- PROVE http://www.ucmb.ulb.ac.be/UCMB/PROVE/
- ANOLEA http://www.fundp.ac.be/pub/ANOLEA.html
- ERRAT http://www.doe-mbi.ucla.edu/Services/ERRATv2/
- VERIFY3D http://shannon.mbi.ucla.edu/DOE/Services/Verify_3D/
- BIOTECH http://biotech.embl-ebi.ac.uk:8400/
- ProsaII http://www.came.sbg.ac.at
- WHATCHECK http://www.sander.embl-heidelberg.de/whatcheck/

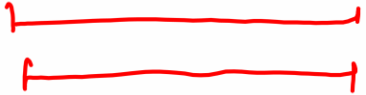
# Revisiting the protocol for homology modeling



# Advantages and disadvantages of homology modeling

## Advantages

- Requires amino acid sequence and a 3D structure of a similar protein
- Does not involve any bench work or consumables
- Less time consuming and less costly
- Most of the softwares are freely available online
- Most promising method among the three computational methods



## Disadvantages

- Loop regions are difficult to model
- Experimentally derived structure for homologous protein is essential
- Does not explain aspects of protein folding.

# Programs available for homology modeling of proteins

- SWISS Model : http://www.expasy.org/swissmod/SWISS-MODEL.html
- WHAT IF : http://www.cmbi.kun.nl/swift/servers/
- The CPHModels Server : http://www.cbs.dtu.dk/services/CPHmodels/
- 3D Jigsaw : http://www.bmm.icnet.uk/~3djigsaw/
- SDSC1 : http://cl.sdsc.edu/hm.html
- EsyPred3D :http://www.fundp.ac.be/urbm/bioinfo/esypred/

- COMPOSER http://www.tripos.com/sciTech/inSilicoDisc/bioInformatics/matchmaker.html
- MODELER http://salilab.org/modeler
- InsightII http://www.msi.com/
- SYBYL http://www.tripos.com/