

**Bioinformatics**  
**Prof. Michael Gromiha**  
**Department of Biotechnology**  
**Indian Institute of Science Education and Research**

**Lecture – 31b**

**Overview II**

Then also we can also derive some features, some properties using amino acids. So, what are the various properties we obtained from amino acid sequences?

(Refer Slide Time: 00:26)

**Protein Sequence Analysis**

- Amino acid occurrence
- Composition
- Molecular weight
- Residue pair preference

(4) AIALALALST  
 $Comp(A) = 4/10 = 0.4$   
 $\sum_{i=1}^n n(i)w(i) - 18(n-1)$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Student: (Refer Time: 00:27).

Amino acid occurrence; that means, each time how many time a amino acid occur for example, if this is a sequence. So, with the occurrence of alanine.

Student: Four.

Four right this 4. So, what is the composition of alanine?

Student: 4 by 10.

This is equal normalized by chain length this is equal to 0.4. So, then again get the molecular weight how to get the molecular weight?

Student: Some of the.

For each residue we have molecular weight, add up right. So,  $n$  of  $i$  into  $w$  of  $i$ .

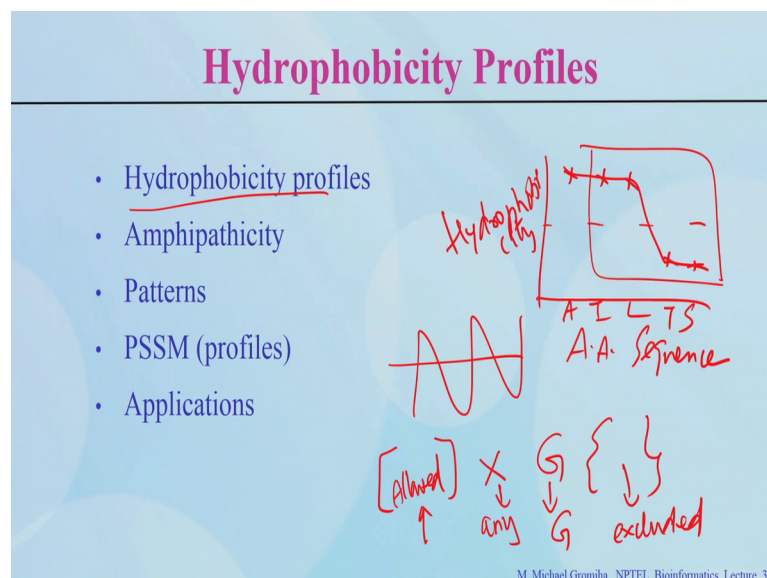
Student: Minus.

Minus.

Student: 18 into.

18 into  $n$  minus 1 right you can get the molecular weight, then you can write the residue pair preference how for each 2 residues right dipeptides, they come close to each other for example, here A is with the L, this is A with T, this is A with C and so on.

(Refer Slide Time: 01:22)



You can get the residue pair preference, then we did the hydrophobicity profile. What is hydrophobicity profile?

Student: Sequence.

Right you can see the amino acid sequence.

Student: Hydrophobicity.

So, here you can see the hydrophobicity right. So, for each amino acid for example, this is AITSLTS right. So, you can have the values right you can get the values and getting a

profile, this is hydrophobicity profile what is a hydrophobic characteristics? So, what is the hydrophobic characteristics of alpha helices and beta strands?

Student: pattern of hydrophobic

Pattern right, for example, the helices we have the we can take the 4 residues approximately because of the 3.6 residues. So, 2 from one side, 2 from other side. So, 2 are hydrophobic and 2 are the hydrophilic regions right that we will see the hydrophobicity right. For example, if we put like this, these 4 residues right they form this.

Then for the beta strand you can see alternate behavior right you can see pattern like this right fine. So, then we discussed about patterns. So, what is the pattern, how to define a pattern? We can use a square bracket, we can use X right we can use any specific residue and you can use bracket what is the meaning of this.

Student: Allowed residue.

Allowed, these are the residues are allowed. This means.

Student: Any residue.

Any residue here.

Student: Particular residue.

Particular residue is conserved, because we cannot change here.

Student: Not allowed.

Not allowed this is excluded right you can use the patterns. So, what is the PSSM profiles.

Student: Profile.

Is process which is scoring matrices right we can see how for each residue is occupied with that same residues or different residues. When you align the multiple sequence alignment form that you can derive this position weight matrix; how many amino acids among the how many amino acids for same position, now how many times each residue

or nucleotide occurs right. Then normalize with this frequency then we can convert this into the weight matrix right we can, but the applications of PSSM? See you have wide range of applications because its based on this scoring matrices. So, we can predict the secondary structures binding sides right agregating peptides and so on right.

(Refer Slide Time: 03:40)

**Construction Of Non-redundant Datasets**

- Large scale data analysis
- Development of non-redundant datasets
- Parameters
- Software

Handwritten notes in red: CD-HIT, Blastclust, PISCES, 30%

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

When we derive any parameters right for analysis right if you deal about the large scale data, it is very important to derive the datasets. So, in this case we have to do with non-redundant dataset.

What is non-redundant dataset?

Student: less redundancy.

Now, less redundancy because they if you add the same sequence several times, this introduce a bias. So, in order to avoid the bias right we need to take the unique sets right. So, we can make any cut off for example, 30 percent then if you see with 2 sequences right which are sharing the identity of more than 30 percent what you have to do?

Student: (Refer Time: 04:18).



Right keep one and discard the other right. So, in this case we need to derive the non-redundant dataset right for reducing redundancy. So, what are the various programs we discussed? CD-HIT

Student: Blastclust.

Blastclust.

Student: Pisces.

And Pisces right. So, what is the program you use for the algorithm, what is the principle used CD-HIT?

Student: k-means clustering.

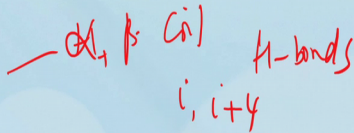
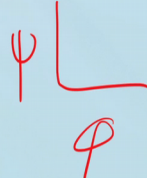
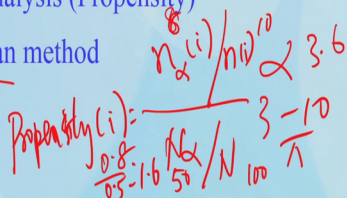
Clustering, k-means clustering. So, you used the different clusters then you pick up the different data from each classes right likewise you can this work like the clustering techniques. So, then we moved on to the secondary structures. So, because all this aspects we can do from primary sequences right. The earlier days the residues mainly focused on primary sequence right after that the applications of the primary sequence data was less because this the reliability become less, then move to secondary structures and tertiary structures .

Now, current scenario again they came back to the primary structure because now we have the due to the next generation sequencing data. So, we have plenty of data available for the sequences, not for the structures right. The structure data is still limited. So, we have more number of sequence data now they are deriving various parameters and algorithms to align the sequences right getting more attention right again the sequence analysis right.

From the sequence then when the attention to mainly structure then went back again to the sequencing because of these next generation sequencing techniques.

(Refer Slide Time: 05:44)

# Protein Secondary Structures

- Protein secondary structures 
- Ramachandran plot
- Dictionary of secondary structures of proteins
- Statistical analysis (Propensity) 
- Chou-Fasman method 

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

Now, secondary structures what are the various secondary structures?

Student: Helix.

Helix.

Student: Strand.

Strand.

Student: Coil.

And the coil right and coil right. So, what are the major interactions which are for the secondary structures?

Student: Hydrogen bond.

Hydrogen bonds right. So, hydrogen bonds between main chain atoms and side chain atoms.

Student: Main chain

It is mainly main chain atoms right you main chain atoms. how the helices are formed?

Student: i and i plus 4.

i and i plus 4 between the NH and CO groups right then we will discuss Ramachandran plot. What is Ramachandran plot ?

Student: Phi and psi plot.

Phi and psi because if you have a dipeptide or any longer peptides. So, the rotational available, is possible in the dihedral angles right phi and psi angles, that is rotational angle of N - C $\alpha$  and C $\alpha$  - C. So, then if you plot phi and psi, so in this case only some specific positions are allowed to have these specific conformation for example, alpha helices or the beta strands. You can see the beta strands is very wide and the alpha helices is in narrow range right. So, you can see these are the allowed positions for residues to be in alpha helices and beta strands. Then what is DSSP?

Student: Dictionary of secondary structure.

Dictionary of secondary structure of proteins and how they derive these secondary structure?

Student: Based on hydrogen bonds.

Based on hydrogen bonding patterns right because the hydrogen bonding patterns are known, for the alpha helices or the beta strand or turns right, because this is hydrogen bonding patterns right, Kabsch and Sander derived 8 classes. Different types of helices what are different types of helices?

Student: 3 10 helix.

Alpha helices.

Student: 3 10.

3 10 helix.

Student: pi helix.

And pi helix right this is the how many turns in alpha helix how many residues per turn.

Student: 3 point six.

3 point.

Student: 6 residues.

6 residues right. So, likewise you can the 5 helix and 3 time helix.

Student: 3 and 5.

3 and 5 right. So, this is the variation between different helices; likewise 3 strands beta bridge and the beta strand and coil, turn and bend right fine. So, now, from this assignment, we discussed about the propensity right. So, how to get the propensity of residues in alpha helix and beta strand? For example, for alpha helices what is the propensity.

Student: (Refer Time: 08:05).

Propensity of i equal to.

Student: Number of residues of i.

Number of residues of i, if it is alpha helix you put alpha, right n of i.

Student: By total.

By N alpha divided by N right. You have 100 residues right and the 50 residues are helix right and the alanine if there are 10 alanine and 7 are, 8 are in helix right what is the propensity?

Student: 0.8 by 0.5.

0.8 by.

Student: 1.6.

0.5, this is equal to 1.6, this is preferred or nor preferred one.

Student: Preferred.

You preferred one because it is more than 1. So, you can see this is a preferred 1 likewise you can do it for the all the residues as well as the different secondary structure right. See

Chou Fasman derived this method, this is named after their names as Chou Fasman method right.

(Refer Slide Time: 09:01)

**Protein Secondary Structure Prediction**

- ❖ Protein secondary structure prediction methods
- ❖ Statistical analysis
- ❖ Information theory *GOR AIA TLSTV*
- ❖ Hydrophobicity profiles *[Handwritten graph]*
- ❖ Multiple sequence alignment
- ❖ Machine learning techniques

*Consensus*

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then we discussed about secondary structure predictions and what are the various secondary structure prediction we discussed? The statistical analysis. So, for statistical analysis, how you get the secondary structure based on statistical analysis?

Student: Based on chou-fasman

Right if you have the 20 residues like classified into different groups, helix former, strong helix former, right, weak helix former, then helix breaker, helix strong breaker and then the helix weak breaker. So, based on that we assign values 1 or 0 or 0.5 or minus 1, and if it is 4 then we considered as nucleating helix. The main aim is there should be at least 4 helix farmers, should not more than 2 1 helix breaker right. Then we can use the actual values because the propensity values we have, based on that your extend the length if the value is more than 4 you add the residues right if it is less than 4 you cut and then a select the segment.

Right then the what is the information theory? They use information theory right, to get the secondary structures, based on the information regarding to central residue as well as 8 residues on both the sides right. GOR developed this theory right. So, there is the information theory, then hydrophobicity profiles right.

If you have the profile right. So, you can see any patterns right. For example, what the helix you can this type of pattern and the strand this type of pattern or you can have this type of pattern right mainly the buried strand right and you can see some turn like this. Based on the profiles we can see which secondary structures right in any particular segment. Then we also used the weighted average you see the window length average of 8 residues or 4 residues, and see the peak and they probably nucleate helix or strand right you can see various methods to get these profiles.

Then multiple sequence alignments, they how to get the secondary structure using multiple sequence alignments.

Student: (Refer Time: 11:11).

So, you have the aligned sequences, 2 ways we can do either in the aligned sequences use GOR method and then you use it or you can use the experimental data right whether, it belongs to helix or strand right the majority of what you can get to this residue belongs to helix or strand. Then there are various machine learning algorithms right neural networks as the support vector machines right they map the input data mainly the residue neighboring residues right then to predict this secondary structures. Then also we have the consensus method what is the meaning of consensus?

Student: (Refer Time: 11:41).

Now, either you see the meta or the ensemble based method. For example, if you have 10 methods, get the values for 10 methods wherever, more than 5, then you can say this could be probably helix or strand. Or we can get the output from the different servers put it in the new server, right train and then see the output right you can also do in this 2 ways to use this consensus method.

(Refer Slide Time: 12:04)

The slide is titled "Protein Tertiary Structure" in a pink font. Below the title is a list of topics in blue font, each preceded by a square bullet point. Handwritten in red ink are the coordinates "X, Y, Z" next to "Protein tertiary structure", the text "(PDB)" next to "Protein Data Bank", and a red underline under "Protein Data Bank".

- Protein tertiary structure X, Y, Z
- Protein Data Bank (PDB)
- Contents
- Visualization tools
- Pymol

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then go with tertiary structure. So, what is the tertiary structure which information we required to get from tertiary structure?

Student: atom coordinate.

Coordinates right if we get the xyz coordinates, mainly we get the xyz coordinates then we get the occupancy and the B-factor right. So, where we get the data for this 3d structures?

Student: Protein data bank.

Protein data bank it is widely used database right you can get the structures. So, what are the various contents in the PDB?

Student: header information.

Right we have the header information, name of the protein and the resolution and the secondary structures, sequence and mainly the xyz coordinates right along with the occupancy, and the temperature factors, and hetero atoms right water molecules and so on. So, there are various tools to visualize the protein structures right; what are the various tools to visualize the protein structures?

Student: Pymol.

Pymol is one of the widely used ones because it can provide give the high quality figures right also we can do various options like you can calculate the distance, and the bond angle, tertiary angle right you can use one different models for example, the space-fill model and the carton model right you can make. Like the other molecular, other programs Jmol and then different other programs are available right rasmol right kings various software tools are available, but pymol is a widely used right tool for the visualization.

(Refer Slide Time: 13:24)

**Protein Structure Analysis**

- ❖ Structural classes of proteins
- ❖ SCOP and CATH
- ❖ Contact maps
- ❖ Solvent accessibility

Handwritten notes: all- $\alpha$ , all- $\beta$ ,  $\alpha + \beta$ ,  $\alpha/\beta$ , A-helix, A-h-residue.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then if you have these proteins right then you have different structure class of proteins what is the various structure class of proteins?

Student: All alpha.

All alpha, all beta, alpha plus beta, alpha beta. What is the all alpha.

Student: mainly alpha.

We mainly alpha, all beta is mainly beta, alpha plus beta then have alpha helices and beta sheet but they are segregated. alpha by beta

Student: (Refer Time: 13:48).



Intermixes with each other or alternate helix and strand. So, 2 databases called the SCOP what is SCOP? Structure structural classification of proteins right and the CATH. What is CATH class?

Student: Class.

Architecture.

Student: Topology.

Topology.

Student: Homo.

And homologous superfamily right you can that. So, you can do this 2 databases this will give you the information regarding classes, fold, certain families, super families other things then once you have this data you can derive various features. So, the most important once and the easiest once they get the contact maps right what is the meaning of contact map how to construct contact map?

Student: distance between residues

Yeah this is the plot connecting these amino acid residues right if these 2 residues for example, the different residue numbers or in contacts then you put a dot right likewise we can construct a map right. If it they are in contact then you can put a dot we can consider a contact. So, in this case we require some information is to make this contact first we need to a cut off which atoms we need to consider, which atom we usually consider?

Student: Either Calpha.

Either Calpha, you can see Cbeta.

Student: Or heavy atoms.

Or any heavy atoms or center of any residues right that you can use. Then the distance cut off as for the based on the atom you use, you can use you can use different distance cut-off like if it all atoms you can use less distance cut off 4.5 to 5 angstrom, and if we take the c alpha atoms or c beta atoms it takes 6 to 10 angstrom right you can use it. So,

based on these thresholds you can derive these contact maps right then till then this is the 3D representation converted to.

Student: 2 D graphs.

2 D graphs you can easily see right which residues are close to each other in the 3D structures based on this graph; then what is solvent accessibility?

Student: How much.

How much is. each residue right in a protein molecule is accessible to the solvent, by rolling water molecular of radius 1.4 angstrom, you can see how much is accessible to the particular solvent molecule right then we class we classified into different groups.

(Refer Slide Time: 15:58)

**Protein Structure Analysis**

- Solvent accessibility
- Buriedness
- Contact order
- Long-range order
- Hydrophobic/disulfide/cation- $\pi$  Interactions
- Superposition of protein structures

Handwritten notes on the slide include:

- $\frac{\sum \Delta S_{ij}}{N \cdot L}$
- $\sum n_{ij}/N$
- $n_{ij} = 1 \text{ if } |i-j| > 12$
- RMSD
- ASAView
- Buried
- Exposed

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

We can see its buried what is a meaning of buried?

Student: low solvent accessibility

So, for example, if you if you have a protein right some of them are interior core right this is called the buried and you can see this is the exposed. So, what is the program used to identify these residues which are displayed, pictorially represent these buried and exposed residues?

Student: ASAView

ASAvue right ASAvue can tell you the location of these residues, whether they are in the buried or in the exposed right also the type of these residues and the size right gives you the accessibility. Then based on solvent accessible surface area, we did another parameters like reduction solvent accessibility, buriedness and that. Then based on the contact we discussed some of the parameters like contact order right, what is contact order?

Student: (Refer Time: 16:51).

This will give you how for the residue which are given close space right and how for they are close in the sequence right. They are mainly it is  $\Delta S_{ij}$  right normalized by  $N$  into  $L$ , one is the the length and this is the number of contacts right we can quantify the contacts based on contact order, what is long range order right you can see  $n_{ij}$  by  $N$  right. So, if  $n_{ij}$  equal to 1 if  $i$  minus  $j$  is less than.

Student: (Refer Time: 17:31).

Greater than equal to 12 right. So, this will occurred only for the long range contacts right; then also we have the hydrophobic contacts disulfide interactions, cation pi interactions based on the contact between some specific residuals right whether the carbon atoms or the positive charge and aromatic residues right or the cysteine so on right.

Then we discussed about superposition protein structures what is superposition of protein structures?

Student: Aligning

Aligning; that means, the how much the difference between 2 structures. Which is measured in terms of.

Student: RMSD.

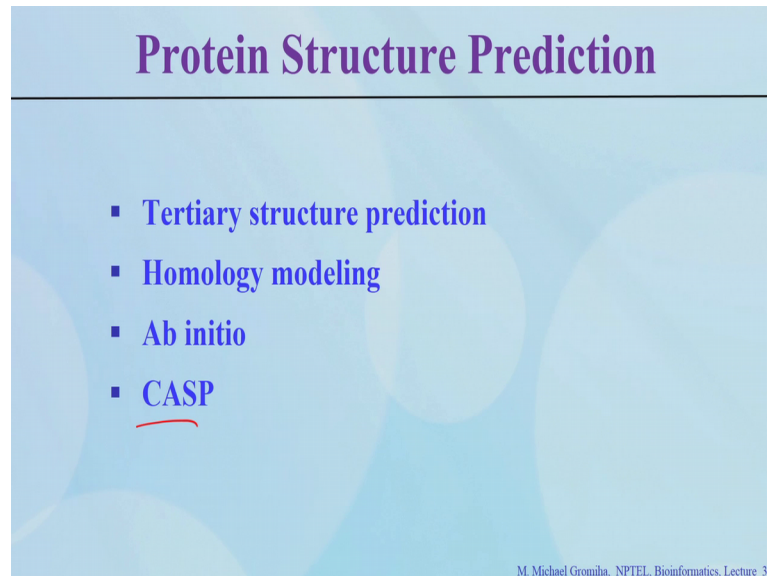
RMSD if it is less; that means, they are close right if it is more than the structures are not similar right. So, when we discussed about the program that is the PDBparam right to see to calculate all these parameters right. Then we can discuss with the applications, mainly

structure prediction, you can do the homology modeling right what is the principle used in homology modeling?

Student: homologous sequences

Right if the two sequences are similar. So, the structures are similar right.

(Refer Slide Time: 18:31)



So, there are various steps to the homology modeling first you get the template then do the alignment. Alignment correction and loop modeling right you can see the after the alignment correction, you do the main chain, you can construct main chain, then loop modeling, sidechain construction then optimization and validation right you have to do all the steps to do homology modeling. Likewise you have to do ab initio modeling and fold recognition. What is CASP?

Student: Critical assessment.

Critical Assessment for the structure of proteins. So, it is a blind test. So, collect the data from the different communities right and then see when they publish the data. So, then can compare right how far the methods they work it is a reliable or not then you go with the stabilities.

What is the protein stability this is the free energy change between the folded and unfolded state.

(Refer Slide Time: 19:24)

**Protein Stability**

- ❖ Protein stability
- ❖ Experimental techniques
- ❖ Energetic contribution ←  $\Delta G$
- ❖ Prediction of protein stability

5-20 kcal/mol

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Generally it is marginal, that is about 5 to 20 kilo cal per mol right. So, there are various ways to get the stability with the experiments right the differential scanning calorimetry, circular dichroism right. So, also you can get from thermal denaturation or denaturant denaturation right. When you have these structures you can get the energetic contributions hydrophobic, electrostatic, van der Waals, hydrogen bonds, disulfide bonds, right covalent bond and non-bonded interactions.

And see how far we can relate this is experimental stability right obtained from the thermodynamic experiments plus the energetics right this is the folded minus unfolded free energy say you found that the hydrophobicity is the major factor whereas, the hydrogen bonds and the other interaction right give the shape and then maintain this stability.

(Refer Slide Time: 20:09)

The slide has a light blue background with a darker blue header. The header contains the title 'Stabilizing Residues in Protein Structures' in a bold, pink font. Below the header, there is a bulleted list in blue text. The third item in the list, 'Thermodynamic database', is underlined in red. To the right of this item, the word 'ProTherm' is handwritten in red. At the bottom right of the slide, there is a small line of text: 'M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31'.

## Stabilizing Residues in Protein Structures

- Stabilizing residues in protein structures
- Comparison with experiments
- Thermodynamic database ProTherm
- Features and utilities

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then we discussed about the stabilizing residues in protein structures, we see the some residues right which are stabilizing if you mutate these residues that will alter the stability and function. Mainly its based on hydrophobic interactions, long-range interactions and plus the conservation score. Then we obtain the data then compare with the experiments mainly the data from thermodynamic database, what is database for the thermodynamics?

Student: Protherm database.

Protherm database, it has the data for the wildtype proteins plus the mutants.

Student: Mutants.

(Refer Slide Time: 20:40)

## Stability of Proteins Upon Mutations

- ❖ Protein stability upon mutation
- ❖ Physicochemical properties and protein stability
- ❖ Prediction of protein stability upon mutation
- ❖ Structure
- ❖ Sequence

$\Delta P = P_{mut} - P_{wild}$

$\Delta \Delta G$

HP

ATAKALL

Inverse HP Hydrophobic Effect

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

So, if you have the mutations then we can see the important parameters, what are factors which influences the stability of this proteins right for example, this is buried mutation we showed that the hydrophobicity right which have direct relationship, with the stability right we can we will get this direct relationship.

In the case of buried mutations, in the exposed mutation we get the inverse relationship what is the inverse hydrophobic effect? So, increase in hydrophobicity will decrease the stability right you get this type of patterns right. So, the mainly the exposed mutation, in the coil region. So, if you increase the hydrophobicity that decreases the stability right this is the inverse hydrophobic effect.

So, this is delta G and this is hydrophobicity right. Then we discussed about the structure information we can add the residues which are neighboring occurring with the limit of 8 angstrom to account for the structural information, and also if we have the sequence right if the central residue we can get the neighboring residue information to include the sequence information right. Now in this case you can add this sequence information right if there are same mutations that are different locations right.

Otherwise we will have same results, if you use the properties right because we use the delta P for the different property, this is equal to P mutation minus P wild for example, hydrophobicity right difference values. Likewise you have delta G for the mutation value right then you can combine these 2, relate using correlations right correlation coefficient.

(Refer Slide Time: 22:15)

**Protein Folding Rates**

- Protein folding rates
- Structure based parameters and folding rates
- Prediction of protein rates

CO  
LRD  
MCI

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then we discussed the folding rate. What is the folding rate? This will tell you the how protein right takes time whether it is slow or fast to fold from the amino acid sequence to the folded 3D structures right.

So, depends for example, if you see the all alpha proteins, which usually fold right faster than the all beta proteins then we account this folding rate based on the structure based parameters mainly the contact order, long range order, multiple contact index or different parameters. So, you can get this numbers and directly relate with the folding rates right say direct relationship or inverse relationship? Is inverse relationship right.

So, these parameters inversely related with this folding rate because the more number of contacts then the slows down the folding process right then we predict the folding rates based on the sequence right or from the structures right.



(Refer Slide Time: 23:05)

## Protein Interactions

- Protein interactions
- Identification of binding sites
- Binding propensity
- Important interactions for binding
- Database for binding affinity

*Distance*  
*ASA*  
*Energy*  
*Proximate*     *hotspot > 2 kcal/mol*

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

We can predict this folding rates. Then we discussed about the folding interactions right mainly protein-protein interactions, protein-DNA interactions, protein-RNA interactions right protein-ligand interactions right.

So, if we have the complexes we can identify the binding sites right. What are the various ways to identify the binding sites?

Student: Distance.

So, distance based approach in this case we see the distance between one protein to the another protein or protein and DNA and make any cut off if these atoms are within a specific cut off we can say they are in the interface. Then we can accessible surface area based approach in this case we take the accessible surface area where the whole complex and cut into pieces and take this unbound one right and then see the difference if there is a difference in ASA then means that they are in the interface right.

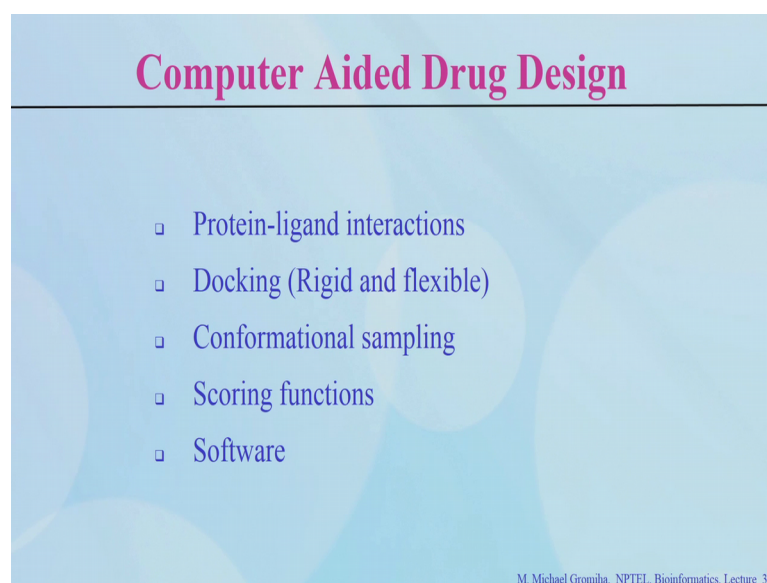
Then using the then we did the energy based approach. So, we calculate interaction energy based on interaction energy you can make a cut-off say a one kilo cal per mole then if you see it less than that we can see this is in the interface. Then using this information we identified the binding sites, then from that binding sites we calculate the binding propensity, what is the binding propensity? This is the tendency of each residues

to be at the interface right then you can see for 10 alanine, all the 8 are in the interface, you can see that they prefer to be at the interface.

Then we identify the residues and residue pairs, from that we can see what are the probable interactions. Now protein residue interactions we discussed importance of few interactions like electrostatic interactions, cation- $\pi$  interactions, aromatic interactions right then we can get the importance of specific residues using the binding affinity upon mutation right, if you can see the free energy change of more than 2 kilo cal per mole right.

Then we term as hotspot residues right if you have at least 2 kilo cal per mole, more than 2 kilo cal per mole right. So, where shall we get the information, what is the database proximate right we can get this  $\Delta G$ . So, to get the binding affinity upon mutation, you can use this for the understanding the affinity upon mutation.

(Refer Slide Time: 25:07)



Then the final part the applications we discussed about a computer aided drug design.

So, what is the protein-ligand interactions? So proteins interact with small molecules right. So, small molecules they trigger this proteins change the conformation right in this case they can be used as a drug like molecule like a lead compounds. So, what are the different types of docking? Its rigid docking and a flexible docking.

In the rigid docking. So, protein is fixed, ligand also we can fix and make the interactions, ligand we can make various configurations right with different conformation we can dock and get the score. For the flexible you can give the conformation changes right for the various protein is flexible right the ligand is flexible you can see the probable pose and the type of interactions, where we can find between the protein as well as the ligand, but this is time consuming.

What are the 2 different aspects we have to consider in docking? Conformational sampling under the scoring function right. So, conformational sampling, 2 types of a conformational sampling we discussed, one is the stochastic sampling one is the systematic sampling. Systematic means we have to do systematically, but takes time right, but we can get 1 low energy value. For the stochastic sampling, we used various aspects right.

What are the various techniques we discussed simulated annealing and genetic algorithm let us say the sample right the data right different conformation and pick the best.

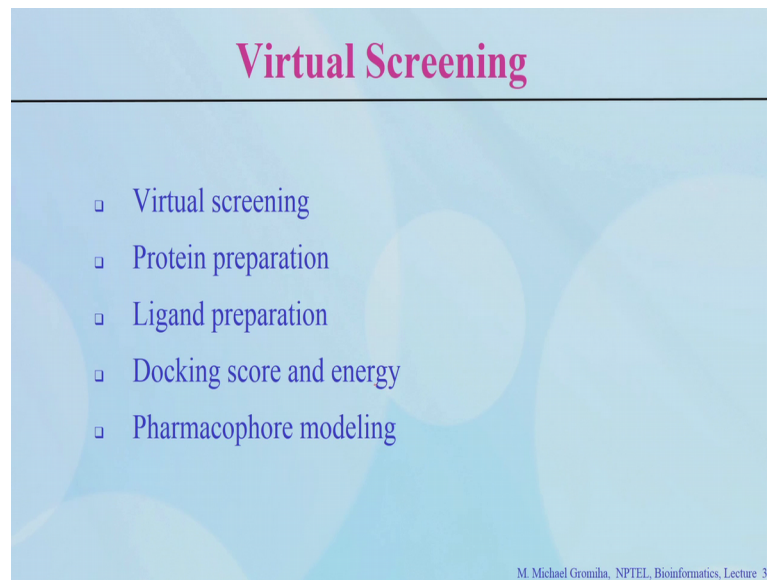
What is scoring function, what are the various types of scoring functions we discussed?

Student: shape (Refer Time: 26:42).

Shape complementarity directory and chemical complementarity, empirical functions, energy contributions and propensity values, knowledge based approach and the consensus. There are the various ways to get the scoring functions right and we discuss various software like Glide or Autodock right. Its available in the literature we can use to get the scoring functions as well as to understand the probable pose.

Then the virtual screening, first we need to get the protein right.

(Refer Slide Time: 27:17)



## Virtual Screening

- Virtual screening
- Protein preparation
- Ligand preparation
- Docking score and energy
- Pharmacophore modeling

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

We have to model the protein, homology modeling or the abinitio modeling, then ligand we have to prepare, then we need to do the docking. So, we get the docking score as well as the energy. Based on that then we can identify the probable ligands to be a drug-like molecules.

Then we can also do pharmacophore modeling, surrounding the protein pocket. So, what are the possibility of the different types of interactions right, based on the hydrogen bond donors, hydrogen bond acceptors right or the polar surface area right or the molecular weight you can see how we can model this particular ligand.

(Refer Slide Time: 27:42)

## Quantitative Structure Activity Relationship

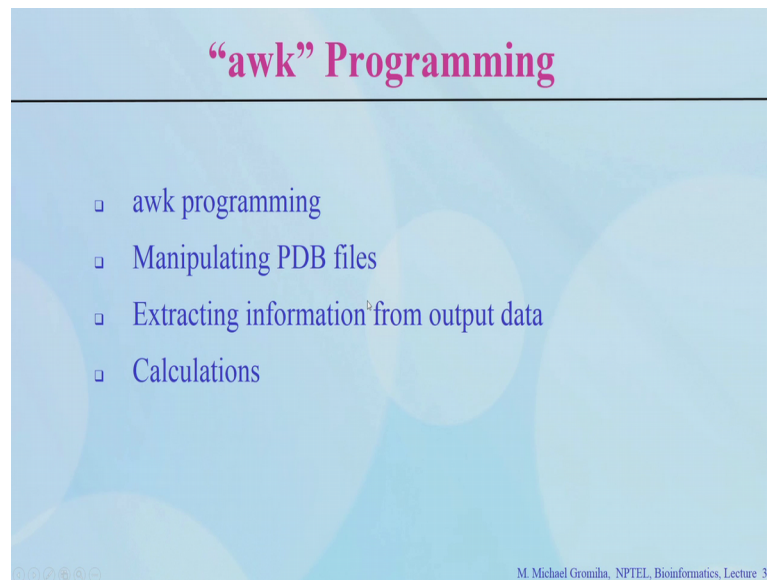
- Quantitative structure activity relationship (QSAR)
- Molecular descriptors
- Structure-activity
- Measures of performance
- Case study

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then what is QSAR? It is the quantitative structure activity relationship, the principle behind is QSAR is the activity can be represented as a function of molecular descriptors like the physical properties or chemical properties or 2 dimensional connectivities or the structural information right we can explain the activity in terms of the descriptors right physical and electronic properties of this ligands.

So, we did get the descriptors and finally, we relate the structure with the activity, and we can measure the performance right we can use the correlation coefficient, cross-validation techniques right to see whether this will fit the model and then use the model for identifying the new compounds right by changing the chemical groups .

(Refer Slide Time: 28:28)



A presentation slide with a light blue background and a darker blue header. The header contains the title “awk” Programming in a bold, pink font. Below the header, there is a list of four topics in a blue font, each preceded by a small square icon. At the bottom left, there are navigation icons, and at the bottom right, there is a footer text.

## “awk” Programming

- awk programming
- Manipulating PDB files
- Extracting information from output data
- Calculations

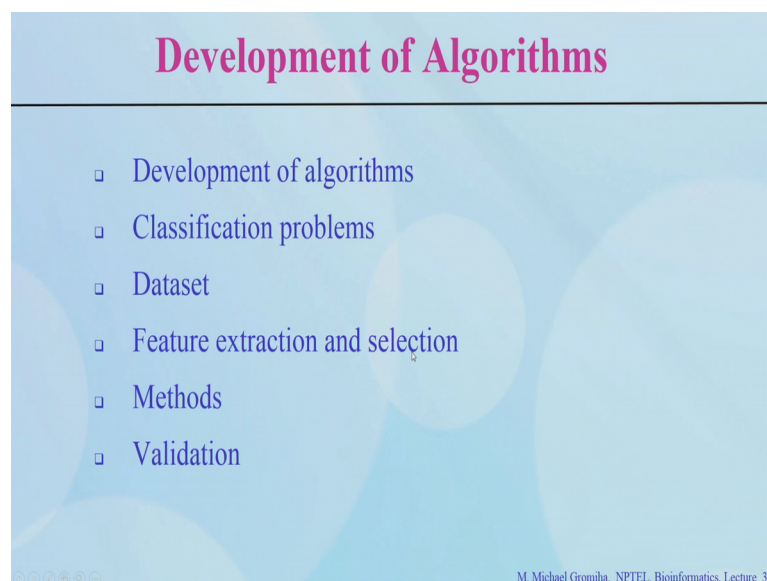
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then we discussed the programming, what is awk programming?

Student: pattern

It is a scripting language, it is the pattern driven language right, data driven language. So, you can use it systematically and effectively for handling a large dataset right mainly in the different databases, or you can get the output from the different programs right you can use it. Specifically on manipulating PDB files right and the extracting data from the many output data.

(Refer Slide Time: 28:51)



A presentation slide with a light blue background and a darker blue header. The header contains the title Development of Algorithms in a bold, pink font. Below the header, there is a list of six topics in a blue font, each preceded by a small square icon. At the bottom left, there are navigation icons, and at the bottom right, there is a footer text.

## Development of Algorithms

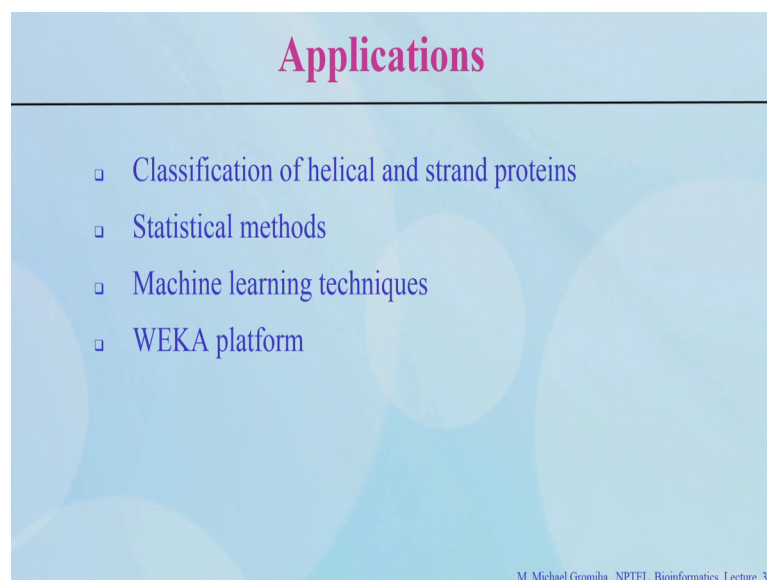
- Development of algorithms
- Classification problems
- Dataset
- Feature extraction and selection
- Methods
- Validation

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then we discussed about the different types of algorithms right, importance of the bioinformatics in classification problems. So, what are the different classification problems we discussed? Whether DNA binding or not, or the binding proteins are not, transmembrane region or not and the driver mutation, passenger mutation or the malignant or the this benign tumors and so on. So, various steps we have to do, first we have to do take the dataset and then feature selection and develop the method and assess the performance and validate.

So, these are the various aspects we need to consider when we are developing the algorithm.

(Refer Slide Time: 29:26)



Then we discussed about the applications like classifying, the helical and strand proteins right. Mainly we discussed about statistical methods and machine learning techniques and specifically and the applications of WEKA right it contains various machine learning techniques right to classify in the different types of proteins and so on.

Essentially the this course we mainly focused on the algorithms, techniques and the fundamentals and databases right and some of the applications and protein folding stability and as well as interactions. So, we cover the basic aspects as well as the on some of the applications, I hope you enjoyed these lectures and you understood the aspects of the basics as well as the for carrying the different projects, how to approach right, how to choose a problem, how to approach a specific problem, and what are the ways different

ways you need to proceed right for the successful completion of your project right. So, our best wishes for your bright career and I hope the program will be useful to you.

Thank you very much.