

**Bioinformatics**  
**Prof. M. Michael Gromiha**  
**Department of Biotechnology**  
**Indian Institute of Technology, Madras**

**Lecture - 18a**  
**Protein Structure Analysis III**

In this lecture, I will continue on the development of structure based parameters using protein 3D structures. Previously we discussed on various aspects specifically on the structural classes of proteins. So, what are the various structural classes of proteins?

Student: All alpha

All alpha proteins.

Student: All beta.

All beta proteins.

Student: Alpha plus beta.

Alpha plus beta proteins and alpha by beta proteins right all alpha helices are dominated by alpha helices and all beta proteins right are induced with the beta strands.

You can see the percentage of helices and strands in the alpha plus beta and alpha by beta proteins, in alpha plus beta proteins these aggregate right and then in the alpha by beta proteins they mixed with each other.

Then we discussed about the databases which contained information on different classes fold families, and super families etcetera right what are major databases which contain this information.

Student: SCOP.

SCOP and CATH right and then we tried to develop various parameters, the simplest one is the contact maps right what is the contact map?

Student: It will give you the information regarding in the closeness of two residues.

Right. So, it is giving the 3D information in the two dimensional form. It gives you the information regarding the contact between amino acid residues, at any cutoff distance right what are two parameters required for constructing contact maps?

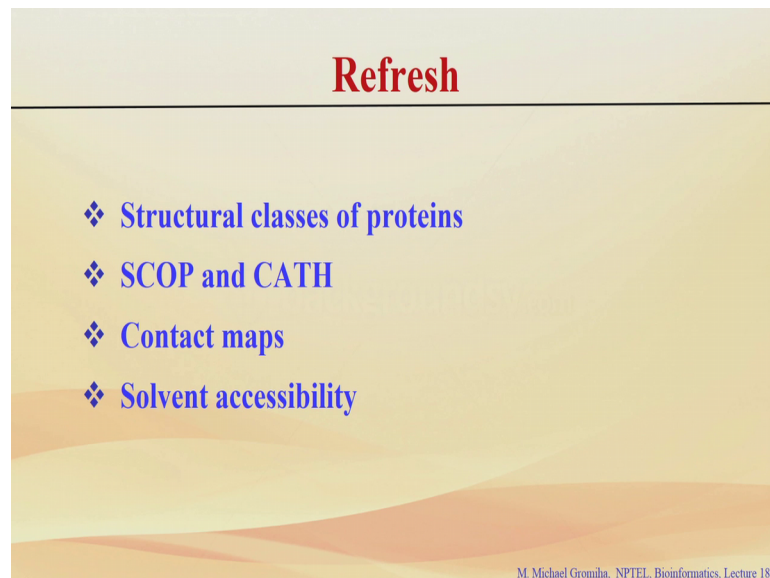
Student: Atom choice.

Atom choice.

Student: Distance cutoff.

Another distance right you can either change this  $C\alpha$   $C\beta$  or any heavy atoms or likewise you can see the cutoff distance 4 Å or 6 Å and 8 Å so on.

(Refer Slide Time: 01:53)



So, when you have the contact maps, we can see the distribution of the residues which are in contact with each other right, which are close in space and how far they will be distributed in the sequence. Based on the distance we can classify the contacts into the short range contacts, medium range contacts, long range contacts right and long range contacts are mainly influenced with the residues which are far away in the sequence.

So, now the discussed some of the examples right you can see the different types of globular proteins right, all beta proteins are dominated with long range interactions.

Student: Long range.

Whereas all alpha proteins are dominated with?

Student: Short range medium range.

Medium range interactions, then we discussed about solvent accessibility right what is solvent accessibility?

Student: So, how far the residues are accessible to the solvent.

Right they will tell you the for the probe radius of water molecule right with the different atoms or residues in a protein, you can tell you the how far these residues are accessible to the solvent right to the van der Waals distance radius of each atom.

So, there are various programs are developed to get the values right what are the various programs we discussed?

Student: ACCESS.

ACCESS NACCESS

Student: DSSP.

DSSP.

Student: GETAREA.

GETAREA right and ASA and so on. So, the advantage of using DSSP?

Student: It provides both

It provide both secondary structure and solvent accessibility, what are the limitations?

Student: It cannot give atom wise

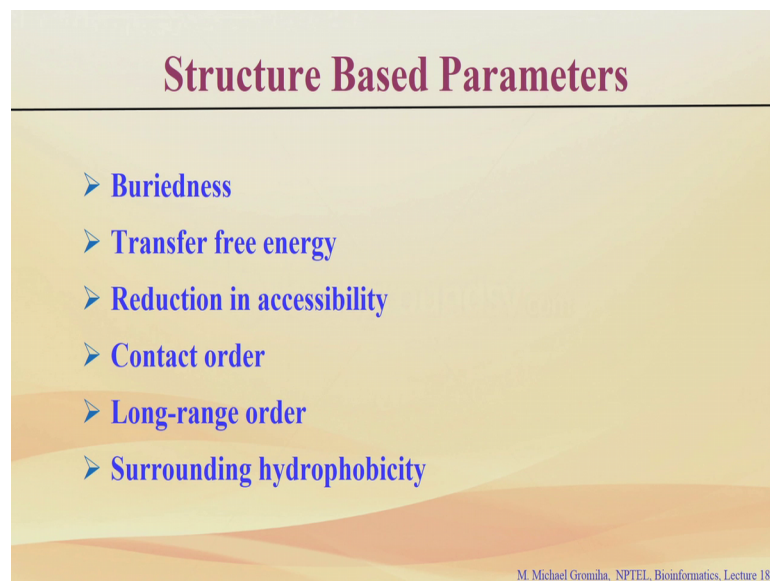
It is only for the all residues not for all atoms right fine.

So, then we discussed about the pictorial representation of residues right. So, what is the program we used to get the pictorial representation?

Student: ASAView.

ASAView, it will give you spiral plot as well as the line plot right we will give you how far these atoms residues, they are at the interior of the protein or at the surface of the particular protein. So, today based on these solvent accessibility, we can derive some other parameters as well as the known 3D structures are used to derive different other types of parameters right we will discuss some of the important properties right in this structure.

(Refer Slide Time: 03:53)

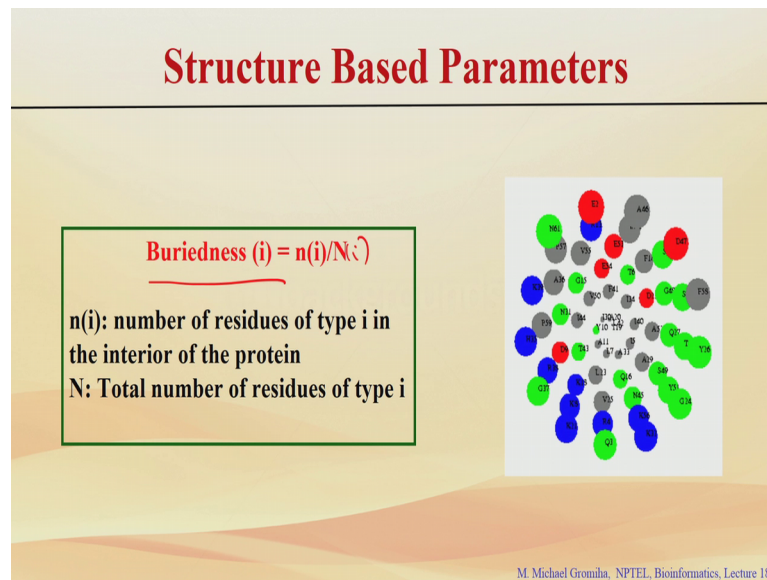


Some of the parameters which I describe today is buriedness, transfer free energy, how the residues the reduce the accessibility right from the unfolded state the folded state; and how far they are in contact with each other any parameters you can quantify the contacts in a full protein right as well as how far these residues determine the hydrophobicity of any residues right we can that is all called surrounding hydrophobicity.

So, buriedness will tell you the tendency of each amino acid residue, to be located in the interior of the protein.



(Refer Slide Time: 04:30)



See if we see this figure right. So, I this is the ASAView of a particular protein right. So, the globular protein.

So, you can see some residues which are preferred to be an interior of the protein you can see here interior of the protein, and some residues which are at the surface. So, you have a set of proteins and see whether any preference of residues, which are located at the interior of the protein. So, that will give you the buriedness index of any residues, here i varies for 1 to 20. So, for the 20 different amino acid residues.

See enough I gives you the number of residues of type i and the interior of the protein, and you needs the total number of residues of type i for example, you can see n of i right for example, there are 10 alanines and 9 are the interior then what is the buriedness index for the alanine?

Student: 90 percent.

90 percent or 0.9 and likewise you take the set of proteins calculate this solvent accessibility and for each residues see how many residues in all the proteins and how many residues are in the interior of the protein. So, this ratio will give you the buriedness of any particular residue.

So, if you relate this with hydrophobicity earlier we will discussed with the hydrophobicity, what could be the relationship between hydrophobicity and buriedness?

Student: Highly correlated.

Highly correlated right because the hydrophobic residues right they prefer to be at the interior of the protein. So, usually the hydrophobic residues valine alanine isoleucine right you can see the preference of these residues in the interior of the protein. There also you have high hydrophobicity. So, you can see a good correlation between buriedness as well as the hydrophobicity.

(Refer Slide Time: 06:15)

**Structure Based Parameters**

**Transfer free energy (i) =  $-RT \ln f(i)$**   
 **$f(i) = \frac{n_{int}(i)}{n_{ext}(i)}$**

$n_{int}(i)$ : number of residues of type i in the interior of the protein  
 $n_{ext}(i)$ : number of residues of type i in exterior of the protein

**Reduction in accessibility (i)**  
 **$= 1 - \frac{A^f(i)}{A^{unf}(i)}$**

$A^f(i)$ : Accessible surface area in folded state  
 $A^{unf}(i)$ : Accessible surface area in unfolded state

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 18

Then you can also calculate transfer free energy, in this case it will consider the preference of residues both the interior the exterior of the protein right. You can see the transfer free energy which can calculate using this formula  $-RT \ln f(i)$ , this  $f(i)$  you can calculate the number of residues any type i at the interior of the protein that is interior divided by the same type of residue i the exterior of the protein.

This will give you the relative occurrence of amino acid residues in the interior of the protein as well as at the exterior of the protein right. For example, if we take the hydrophobic residues what will happen?

Student: They will be more interior.

They prefer to be in the interior the protein. So, you can see this  $f(i)$  that is positive. So, you can get the transfer free energy right this  $-RT \ln$  to these are positive values then you will get the negative values right.

Likewise you can see for 20 different amino acid residues, the tendency of each residue right to be the interior as well as the exterior of the protein, this will give you the transfer free energy of that particular residue  $i$  in any particular protein.

So, then also you can also obtain the reduction accessibility, this will as I tell you how far this residue right which reduce the accessibility from unfolded state to the folded state. So, here if this is a ratio  $A^f(i)$  divided by  $A^{unf}(i)$ , here  $A^f(i)$  stands for the accessible surface area in the folded state, and  $A^{unf}(i)$  gives you the accessible surface area in the unfolded state this will give you the ratio, then how far this is reduced right. In this case you subtract from 1 right then 1 minus this quantity will give you the reduction accessibility. It is how far the specific residue which reduce from the un folded state to the unfold unfolded or unfolded state to folded state.

So, these are the various parameters we discussed based on solvent accessibility right if you have solvent accessibility, you can calculate accessible surface area you can represent in the form of pictures and we can derive the various parameters right.

Now, from the contacts point of view, you will discuss different types of contacts, short range contacts, medium range contacts and long range contacts, and you can quantify these contacts in terms of some parameters.

(Refer Slide Time: 08:25)

## Contact order

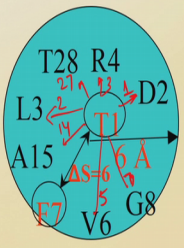
Relative importance of local and non-local contacts

$$CO = \frac{\sum \Delta S_{ij}}{L \cdot N}$$

$\Delta S_{ij}$ : sequence separation between contacting residues  $i$  and  $j$  (6.8 Å)

$L$ : total number of residues

$N$ : total number of contacts



$\sum \Delta S_{ij} = 65$

Plaxco et al. (1998) J. Mol. Biol. 277, 985-994.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 18

One of the well known parameters is contact order, here this will classify these contacts and it will include the importance of local and non local contacts. Local means you can see medium range or short range, non local means you can see long range contacts they give weightage to these contacts and they develop the parameter.

So, if you see the contact order it is given as  $\sum \Delta S_{ij}$  where  $\Delta S_{ij}$  you can see the sequence separation. For example, here T 1 first T 1 and F 7. So, what is sequence separation?

Student: 6.

6 residues right. So, likewise you add up for all the residues and normalized with the L and N where L is the total number of residues N is the total number of contacts. In the developing the contact order, they use all atoms and they are considered to be contact if the distance is 6 Å. This is actually 6 Å say they tried 6, 7 and 5; different distances and finally, the optimized right to have the distance of 6 Å right between the any residues in the space.

So, take the T 1 they construct a sphere of radius 6 Å. So, these are the residues right which are having contact or any atoms with the limit of 6 Å. If we take T 1 and F 7 delta is equal to six likewise you can calculate the total separation.

So, what is total separation for this residue? 1 and 2 this equal to 1 right. So, 1 and 4?

Student: 3.

This is 3, this case it is 2 this case 27.

Student: 14.

14 here 6 5 7 right. So, you can add up in this case,  $\sum \Delta S_{ij}$  equal to what is total value? 27, 30, 31, 33, 42, 47, 52, 58.

Student: 65.

65 right this is 65 this is for this particular residues. So, you can do it for all the residues in the protein, then add up everything. So, you will get a number total number of contacts, then normalize with the number of residues and number of contacts then get a number. So, this how we will quantify different types of contacts in protein structures

with the one number. So, in the 3D structures, when you make the contact maps will make in the 2D maps.

Now, here if you see this one you will get a number, this number will tell you how far the residues are distributed in 3D structures. Whether several long range contacts or several short range contacts, depending upon the length of this particular protein.

So, these contact order considers all the contacts the short range contacts, medium range contacts and long range contacts, but the effects they considered based on the  $\Sigma \Delta S_{ij}$  sequence separation between i and j. So, then later on this another parameter called long range parameter, it has been developed using the knowledge of only long range contacts; that means, several cases the rate limiting step in protein folding is how many residues which can make long range contacts. It is easy to make the short range contacts or the medium range contacts, but takes time right to make long range contacts.

In this case if your protein there are several residues which have tendency of making more number of long range contacts, that the automatically it needs more time to fold.

(Refer Slide Time: 12:12)

## Long-range order

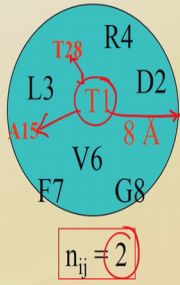
Obtained from the knowledge of long-range contacts (contacts between two residues that are close in space and far in sequence)

$$LRO = \frac{\sum n_{ij}}{N}; n_{ij} = 1 \text{ if } |i-j| > 12;$$

$= 0 \text{ otherwise.}$

i and j: two residues in which  $C_\alpha$  distance between them is  $\leq 8 \text{ \AA}$

N: the total number of residues in a protein.



$n_{ij} = 2$

Gromiha, M.M. and Selvaraj, S. (2001) J. Mol. Biol. 310, 27-32.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 18

So, based on that assumption. So, they developed the concept of long range order, this is from the knowledge of long range contacts what is long range contacts?

Student: Greater than 4.

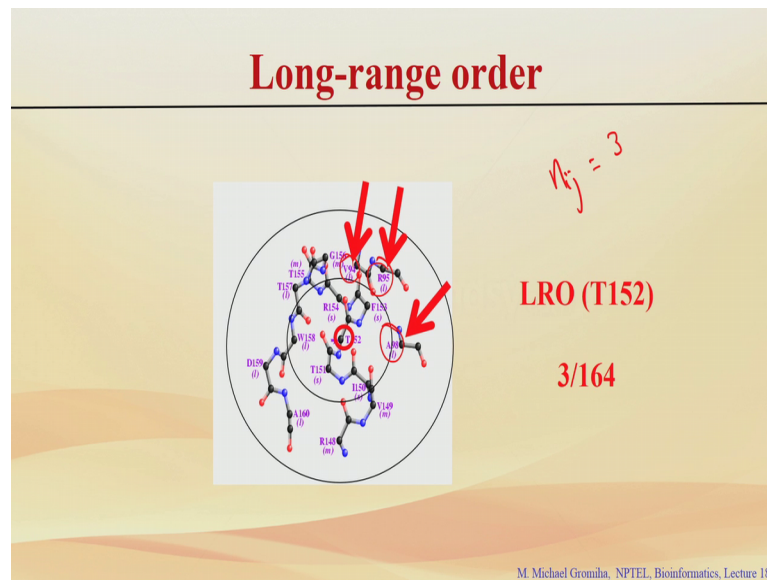
It is a contact between two residues right any contacts, but there are close in space and which are far in the sequence, but far in the sequence how far in the sequence, whether 4 residues 10 residues 20 residues right. So, that we have to fix.

So, in this the long range order you try to use various distance cut offs and finally, they identified the value of 12 residues far apart. In this case for each residue they look at the contacts, and they count only if the contact residues are separated by minimum of 12 residues otherwise it is 0.

In this case for example, if T is the central residue. So, they construct the 8 Å sphere, they construct the sphere of 8 Å and these are the residues which are located within the limit of 8 Å. Say aspartic acid 2, arginine 4, threonine 28, leucine 3, A 15, V 6 F 7 and G 8. Among all these residues how many residues which are separated by at least 12 residues.

For example this and this these are separated by 12 residues, and here they are separated by 12 residues and all others they are not having 12 residues. So, in this case  $n_{ij}$  equal to 2 and here you can see that  $\sum n_{ij}$  means you can do it for all the residues for example, D 2. So, you can construct a sphere around D 2 and identify the residues which are occurring within the limit of 8 Å, and see which residues are separated by distance of 12 residues you can add with a numbers. Similarly for the whole protein you can add up right and normalized with the n, n is the length of the protein right total number of amino acid residues right in your protein.

(Refer Slide Time: 14:10)



So, in this case you can calculate the long range order for example, this is the protein, I showed in the last class, this is the contact residues of threonine 152 in lysozyme. So, you can see they are occurring within the limit of 8 Å. So, what is the value of  $n_{ij}$  here? Because you have to see the contact residues, and how many residues which are separated by at least 12 residues.

So, here this is 94 and 152 this is 1, here this is another 1 2 and here 1 3. So,  $n_{ij}$  equal to 3 for example, if you want to have the LRO for the particular residue then normalized by the length here T for lysozyme is 164 residues, in this case LRO equal to 3 by 164 right they will give you the number.

So, if each residues you have these numbers, then based on that numbers you can tell that how many residues which are forming more number of long range contacts. Then for the whole protein if the numbers is very high, then you can see that there are several residues in that particular protein, which are forming long range contacts. So, here also you can tell you that the one number can explain the type of interactions right you can make the different types of the globular protein structures.

So, generally if you look into the all alpha proteins and all beta proteins how will the LRO vary?

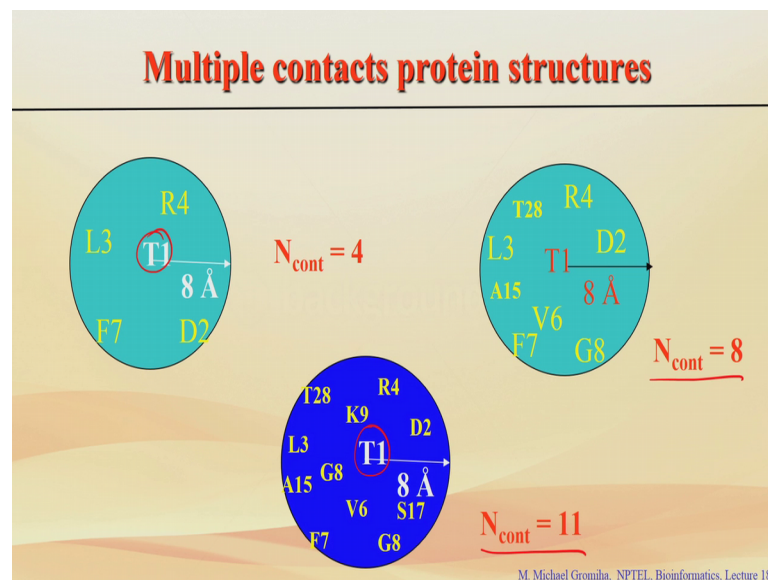
Student: higher for the all beta.



Right higher for all beta proteins because usually they make more number of long range contacts right and if it is low for the case of all alpha proteins because you mainly due to the short medium range interactions.

So, now first we discussed about the contact order, then we discussed about long range order that takes only the long range contacts, and among the long range contacts and some residues they try to form more number of contacts.

(Refer Slide Time: 16:08)



For example in this case T 1 how many contacts T 1 has?

Student: 4 4.

Four contacts 1 2 3 4 if you take second one right how many contacts in this case.

Student: 8.

Eight contacts right and the third example it is very crowded, but the space is same because the distance right in the 3D structure that is the same everything is 8 Å, but if you see this one how many contacts this residue has? 1 2 3.

Student: 11

4, 5, 6, 11, 8, 9, 10, 11 contacts right. See if you look into this a with 3 different examples, which this is more important in protein structure.



Student: Third one.

Third one right because it will make a many contacts, the contacts are influenced with different types of interactions right. Maybe you see the electrostatic interactions or the van der Waals interactions and hydrophobic interactions, and if you disturb this particular residue right and all the interactions are broken, in this case a this will destabilize the protein structure as well as the stability and lost the function.

So, then in this case you can define a parameter, based on the residues which are having more number of contacts.

(Refer Slide Time: 17:21)

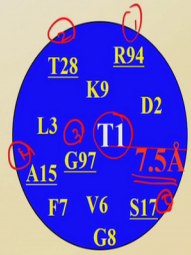
### Multiple contact index

**Parameters:**

- ✓1. Distance between residues
- ✓2. Sequence separation
- ✓3. Residues with multiple contacts

$$n_{ci} = \sum n_{ij}; n_{ij} = 1 \text{ if } r_{ij} < 7.5 \text{ \AA};$$
$$|i-j| > 12 \text{ residues};$$
$$= 0 \text{ otherwise};$$

$$MCI = \sum n_{mi} / N; n_{mi} = 1 \text{ if } n_{ci} \geq 4; 0 \text{ otherwise}$$



$n_{ci} = 5; n_{mi} = 1$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 18

So, this is called the multiple contact index, here we have 3 different parameters, first one is a distance between residues right this is used for all the parameters like contact order or long range order or the multiple contact index, and sequence separation. In contact order they utilize the sequence information in terms of.

Student:  $\Delta S_{ij}$ .

$\Delta S_{ij}$ . In the case of long range order they utilize the sequence separation right with the condition that to consider only if the two residues are separated by 12 residues.

So, in addition to that, in multiple contact index they consider how many residues they have multiple contacts. So, here you see these multiple contact index they define this is

the number of contacts residues with the multiple contacts normalized by  $n$ , this  $n$ ,  $n$  is a number of residues in a protein and  $n_{mi}$  equal to 1 if the contacts is more than equal to 4.

So, this  $n_{ci}$  we define the radius of 7.5 residues, 7.5 Å and the distance 12 residues then we define  $n_{ij}$  equal to 1. A two-step procedure, first we see the  $n_{ci}$  this is a number of contacts based on the same as long range order, here the distance is 7.5 Å and the 12 same into 12 residues and then if the same case they see how many of them having more number of contacts, at least four contacts there based on that they can calculate the multiple contact index.

For example next one example here this is the central residue T 1 with the distance 7.5 Å, how many long range contacts it can make right this is 1, 2, 3, 4, 5 right in this case the  $n_{ci}$  equal to 5, because that there is 5 contacts. If a if  $n_{ci}$  equal to 5 then its greater than four in this case  $n_{mi}$  equal to 1.

So, this will consider the residues which are closed in space which are separate with 12 residues as well as the residues which contain at least four contacts. The numbers are not very high. So, there are less number of residues are having these type of contacts and if you see how many residues they have this type of information, they are acting as a key residues right from that you can see the how many key residues there which are important for the stabilizing or the folding in a particular protein.

So, we discussed 3 different parameters based on contacts, what are 3 different parameters we discussed?

Student: Contact order.

Contact order.

Student: Long range order.

Long range order.

Student: multiple contact index.

And the multiple contact index right we discuss various parameters using solvent accessibility, and some parameters about the contact amino acid contacts, and here we discussed about hydrophobicity. See earlier we discussed about the hydrophobicity scales

experimentally you can obtain the hydrophobicity values for the 20 amino acid residues using the relative solubility of each amino acid residue in water and organic solvents like octanol or ethanol right organic solvents.

So, here we will the 20 values for the 20 different amino acid residues, but if we look into this protein environment the behavior of the amino acid residues are not the same; some residues they behave to be closely with the polar residues and some environment with the nonpolar residues.

So, if you look into a protein environment and checked a tendency of each residue, to be surrounded with which type of residues; this will give you the hydrophobicity in protein environment.

(Refer Slide Time: 21:11)

### Surrounding Hydrophobicity

It characterizes the hydrophobic behavior of the 20 amino acid residues in protein environment.

$$H_j = \sum n_{ij} h_i$$

A, C, G, M, Y: 1
F, I, L, V, W: 2
D, E, H, K, R: -2
N, P, Q, S, T: -1

$H_j$ : Surrounding hydrophobicity of the central residue j

$n_{ij}$ : Number of residues of type i around j

$h_i$ : Experimental hydrophobicity of residue i

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 18

That means each residue characterizes the hydrophobic behavior in protein environment right it is not just the one value for the same residues right. The threonine is here, the behavior of these threonine is different from this was for this is take this is 1, this is take 4. So, T 1 and T 4 they have different hydrophobicity values depending upon the location of this residue and the residues which are surrounded by any specific residues.

Based on this information, I calculate the  $H_j$  for any residue j in a protein, using this formula  $\sum n_{ij}$  number of residues of type i which is surrounded by the residue j multiplied

by the hydrophobicity index of these side chains these particular residues. This side chains you can take the experimental value right.

So, for example, the octanol experiments or the ethanol experiments right and then see the neighboring residues are surrounding residues from the 3D structures. So, here we combine the experimental data plus the location of each residues in the environment. For example, if we take the 8 Å and we located residues which are within the limit of 8 Å. So, this is the approximate values are used for different residues for simplicity, but actually we have different experimental values for all 20 residues. So, here if you see this is the nonpolar residues, I give the value of 2 and the charged residues I give -2, polar residues -1 and the hydrophobic residues 1.

If you have these numbers, what is the hydrophobicity value for the central threonine?

Student: -1

Now, what is D? D is -2 arginine.

Student: -2.

-2 this T

Student: -1.

-1; this one.

Student: 2.

2.

Student: 2.

2.

Student: 2.

2. So, the total value will be 2, 4, 6, 7.

Student: 2.

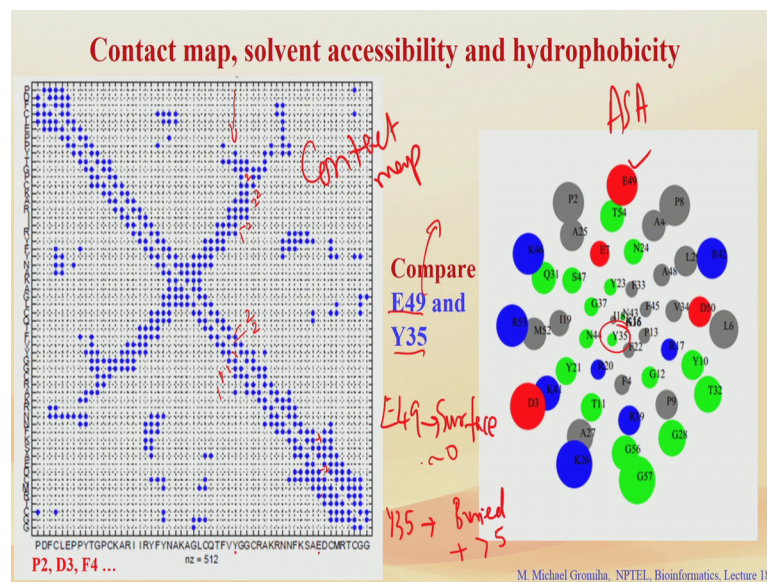
6, 4, 2 right. So, here the hydrophobicity value is 2, these 2 is determined by these surrounding residues if you got these threonine here T4. So, to get this hydrophobicity value of this T what you have to do?

Student: Get the surrounding residues.

We have to make this sphere right of 8 Å and then identify the residues which are occurring on this limit and substitute the values then you will get that. So, for getting the 20 different values same protein is different amino acid residues are same amino acid residues at different locations you have different numbers.

Now, you take average, and then you can get the average values then you will get 20 different values. If you look into these numbers and the experimental values some of them are you can see the correlation and some residue they have difference in behavior, depending upon the location of the residues in protein environment right. Likewise you can calculate this hydrophobicity and this is wide applications right in protein structure and function.

(Refer Slide Time: 24:18)



Now, I have one example here what is the left hand side this one.

Student: Contact map.

This is a contact map what is this one?

Student: ASA

This is ASA accessible surface area for different residues. So, you can see two residues I put for example, E49 and Y35, how these residues behave in terms of contacts accessible surface area and hydrophobicity.

So, if you look into the accessible surface area where E49 is located E49 is here, this is at the surface are buried.

Student: Surface.

Surface right E49 surface how about Y35?

Student: Buried.

Here right, it is relatively buried. Now let could the contacts right what is Y this is start from P2 D3 and so on; where is the Y35? Somewhere here or here

Student: Lot of contacts.

This one right then where is E49? This one E49 and E49, 1, 2, 3, 4, 5, 6, 7 yeah I think this is E 49.

So, if we look in to this Y right how many contacts this Y has? Many contacts right you can see here. Starting from this line right then what is a hydrophobicity value for this Y 35? This is this is alanine.

What is value for A; this is equal to one. So, this is the cysteine 2 and here this is a glycine 2, this is glycine 2, this one valine this is 2, and here this is also 2; am I right yeah. So, here Y this is equal to 1 and here arginine is -2, this is R RY is it this one right R I I I I this is 2, 2, 2, 2, this is C is 2 right.

So, if we see these numbers its highly hydrophobic and the numbers are positive right you can see the buried and this plus I think it is more than 5. And look in to this E49 these E49 these E49 you can see these are the residues right. So, serine is -1 and here the d is -2 right -2 and you can see this is around zero or this is -1.

So, if we look into these residues which are located at the buried or the surface you can see the opposite trend in the hydrophobicity, which are the buried they are having high

hydrophobicity this is more hydrophobic compared with the exposed ones. Likewise the contact also you can see here more number of contacts because the buried here it has more number of contacts compared with the residues which are at the exposed and for example, this is E49. So, you can relate different structure based parameters which you obtained from 3D structures right for example, the contact maps or the hydrophobicity or the accessibility and so on. So, they are and they are interrelated with each other.