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

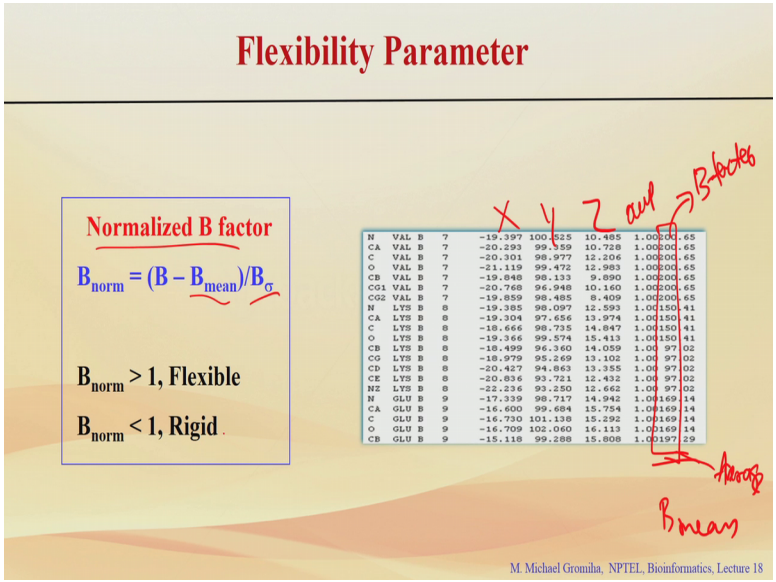
Lecture – 18b
Protein Structure Analysis IV

We will explain few more parameters, one is the flexibility parameters B-factor what is B factor?

Student: Thermal index.

Yeah, you can see thermal fluctuations right. So, for we can accommodate right within this electron density map.

(Refer Slide Time: 07:29)



So, this is the values. So, we can see this is a xyz coordinates is occupancy and here this is the B factor. So, we have different numbers in a protein, and if we want to compare this B-factor with different proteins, because it changes from protein to protein. So, in order to normalize that we can derive the normalized B-factor using the formula of B normalized equal to B minus B mean right what is B mean?

Student: Average of (Refer Time: 00:58).

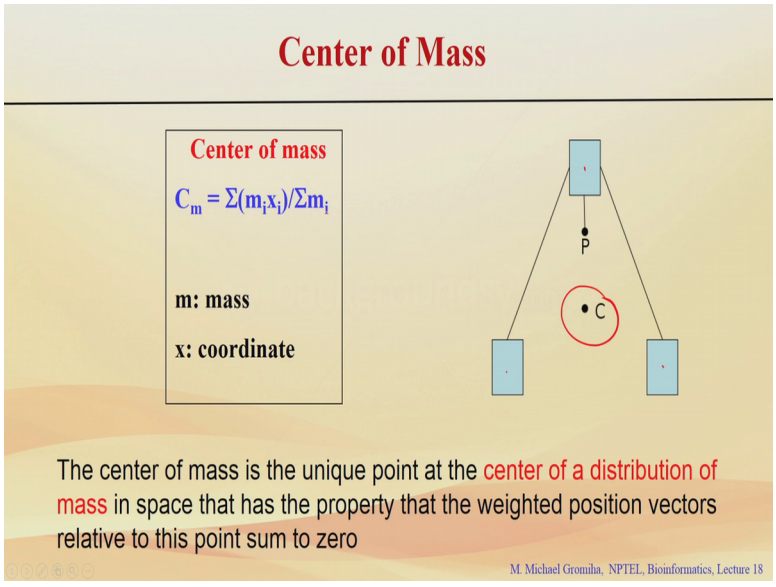
Is average of the all these this is a numbers. So, you can get the average. So, this is equal to B mean then we get the B sigma or B sigma.

Student: Standard deviation.

The standard deviation, you can see the deviation from all these numbers. So, you will get the B sigma. So, we get this formula B minus B mean of B sigma, right, we get the normalized values, if it is more than 1 it treat this as a flexible and if it is less than 1 we take this rigid; because depending upon this value of this B, if it is more than the mean right or less than the mean. So, you can see there it is flexible or rigid with more than the mean right it is more flexible right if the less than the mean then it is rigid. So, using this formula you can see which residues are flexible and which residues are rigid.

So, this is a physical quantity and then also several physical quantities for example, radius of gyration, center of mass, moment of inertia all the information you can calculate from protein 3 D structures right. So, one of the center of mass, what is center of mass right?

(Refer Slide Time: 02:01)



This is the center of a distribution of the mass in space dsf property it is here right that the weighted position vectors relative to the point sum is equal to 0.

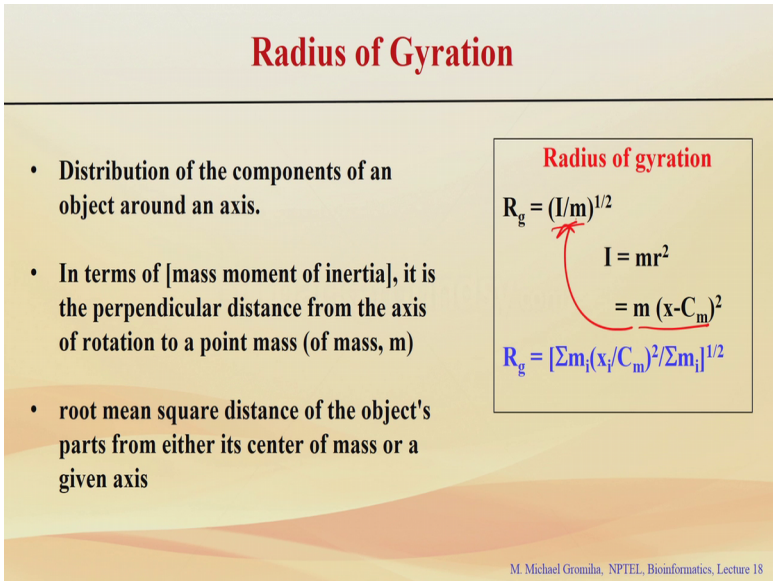
For example if there are 3 different objects here. So, we can see the center of mass right it is located at position number C. So, how to calculate the center of mass? So, can use

the cm here this is for the any coordinates x coordinate or y coordinate z coordinates you can calculate. So, mi into xi where is mi? Mi is the mass and the xi is the coordinates normalize with the total mass right if use this formula you can calculate center of mass for the a different coordinate system or you can see where is the center of mass for the whole protein; how far this will be helpful?

Student: (Refer Time: 02:49).

Because when we compare the different protein structures to small proteins or big proteins and all, and the distribution of the residues right any in any different shell. So, in this case you better to find the center of mass where it is located, and based on that you can see how far how the each residue is located with respect to the center of mass.

(Refer Slide Time: 03:08)



Then we can also derive the radius of gyration, what is radius of gyration? It is a distribution right of the components of an object around any axis for example, if we take the moment of inertia right it is the perpendicular distance from the axis of rotation right to any point mass. For example, say mass m. So, you can calculate radius of gyration using this equation, I this is the moment of inertia m is the mass how do equation for the moment of inertia.

Student: Mr square.

I equal to mr square, right. So, m is mass and r square you can calculate from the center of mass right x is any coordinate from the center of mass. So, x minus m whole square, now substitute these values in this equation here get the Rg there will be sigma mi right multiplied by xi into Cm to whole square divided by this m, this is sigma mi then take this square root.

You can calculate radius of gyration this can also you can tell the information regarding the different positions right for the atoms in any particular protein. When you discuss about the molecular dynamic simulations right how the atom positions vary right that you can determine using this term about the radius of gyration. Now we discussed various aspects then we can also use the distance between the different residues or atoms to convert into the type of interactions. There are various interactions in protein structures what are the various interactions in protein structures?

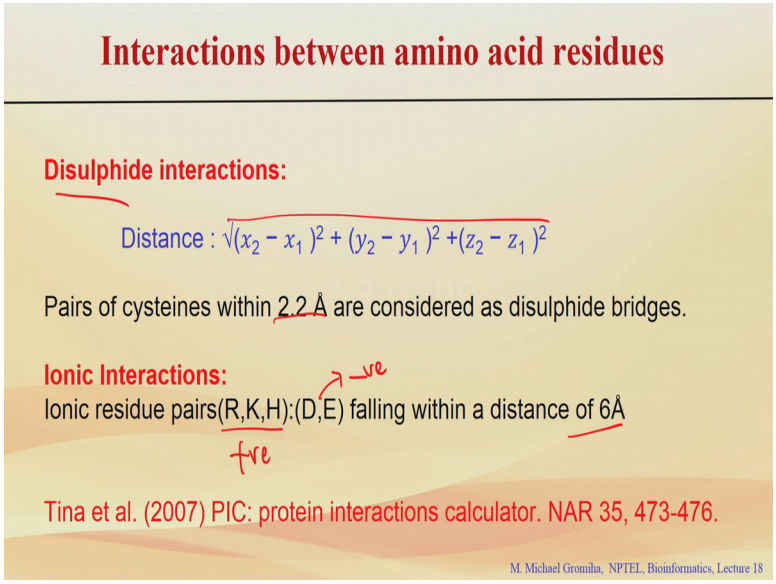
Student: Hydrogen bonds.

Hydrogen bonds electrostatic interactions.

Student: Hydrophobic interactions.

Hydrophobic interactions, Van Der Waals interactions, right. So, really you can calculates the values from the 3 d structures, and also it is possible to see the contacts to identify what are the probable residues which can make this type of contacts. How many residues in a protein which have the potential to form any specific type of interactions. Instead of doing the complete calculations you can identify the residues or residue pairs which are involved in any specific type of interactions.

(Refer Slide Time: 05:19)



For example in the case of disulphide, here it take these distance how to calculate the distance?

Student: Square root.

Square root of x 2 minus x 1 the whole square, plus y 2 minus y 1 the whole square plus z 2 minus x 1 the whole square. So, you can calculate distance; now for the case of disulphide bond interactions you can see the pair of cysteines, these cysteines from the disulphide bonds which atom in cysteine sulfur right. So, we can see whether any sulfurs which are within the limit of 2.2 angstrom, they are considered as disulphide bridges to you have the high tendency to form disulfide bonds.

In this case for example, if you have protein structure right how to identify the disulphide a bridges whether any a disulphide bridges in the particular protein? You have the structures first you search for cysteine and the cysteine we know the atom position for this sulfur atoms and you get the distance if the distance is less than 2.2 angstrom, then we see that these 2 cysteines they have high potential to form the disulphide bonds right this is disulphide interactions, then you can see the ionic interactions essentially what is the ionic interaction?

Student: Columbs law.

Columbs law; this is the interaction between attractive interaction between.

Student: 2 charge residues.

Charge residues 1 positive charge 1 negative charge what are the positive charge residues in a protein?

Student: Arginine lysine and histidine.

Arginine lysine and histidine these are positive charge residues what are negative charge residues.

Student: Aspartate and glutamate.

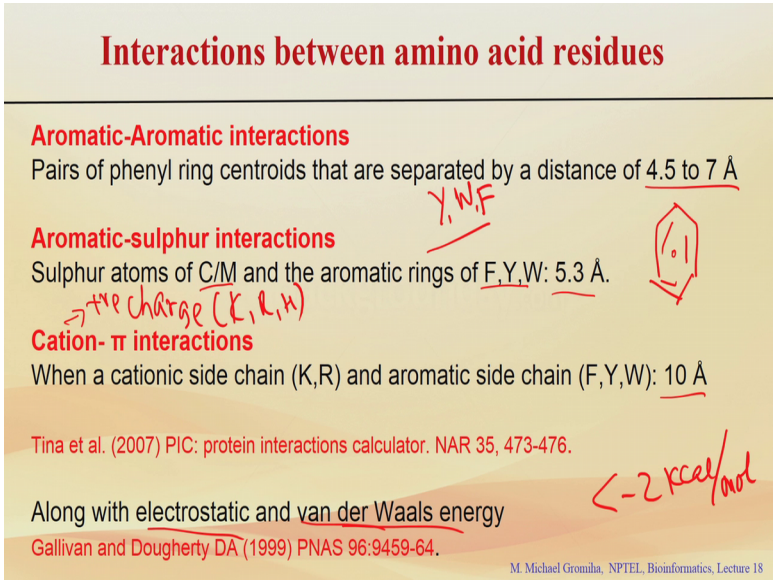
So, this is negative charge residues right then you can see any pairs which are within the limit of 6 angstrom. So, in literature different distance have been used for example, sometimes they use 4 angstrom, but in this paper they try to use 6 angstrom because they do not want to lose any of these ion pairs. So, can we consider all atoms within the limit of 6 angstrom in the arginine and aspartic acid or glutamic acid?

Student: Only if charge (Refer Time: 07:18).

Right. So, n if we take the arginine right many several atoms also we take the aspartic acid or glutamic acid we get several atoms right a main chain atoms N C-alpha CO and several side chain atoms. So, it also if the take the arginine or lysine you can see NH 2 group. So, they have the positive charge likewise in aspartic acid or glutamic acid right you can see OE1 or OE2 they have a negative charge. You get the distance between these atoms and check whether the distance is less than 6 angstrom. Specifically, in the case of lysine and arginine it has long side chain. C alpha will be somewhere and if you see this NH2 group, this will be very far right because several CH 2 groups right.

So, in this case you see the distance between the positive atom and the negative charge atom and the distance is less than 6 angstrom, this can be treated as right high potential to form ion pairs.

(Refer Slide Time: 08:13)



The likewise you can see aromatic-aromatic interactions. What are the aromatic rings in the protein structures?

Student: Tyrosine.

Tyrosine.

Student: Tryptophan.

Tryptophan.

Student: Phenylalanine.

And phenylalanine. So, you can see the centroid of the phenyl ring, how to get the centroid phenyl ring?

Student: Center of the mass (Refer Time: 08:33).

Right you can see the coordinates for all the carbon atoms in the ring right and then see the centroid; that means, you can get the average values like kind of center of mass you can see these centers for the different rings. Then check the distance from different rings and if the distance is 4.5 to 7 angstrom, then you can see that these 2 residues which are involved in aromatic-aromatic interactions right. In this case we take a protein look at these rings and then they get the centroid, and get the and calculate the distance and the

distance is 4.5 to 7 angstrom, then you can say these residues are involved in the aromatic-aromatic interactions.

Likewise aromatic sulfur interactions here sulfur atoms of cysteine and methionine because 2 residues they got sulfur atoms, but if you consider the disulphide bonds we consider only cysteine if in the aromatic sulfur interactions we take both the residues cysteine and methionine and aromatic rings. So, they take the distance of 5.3 angstrom, likewise in cation pi interactions, what is cation pi interactions?

Student: Aromatic and (Refer Time: 09:46).

What is cation?

Student: Positive charge.

Positive charge right positive charge mainly lysine and arginine right then they pi what is the pi system?

Student: Tyrosine.

Phenylalanine, tyrosine or tryptophan.

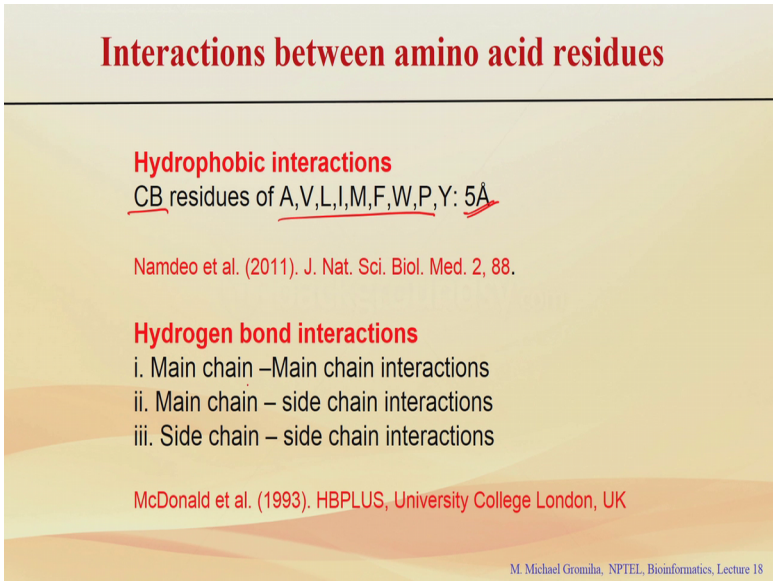
So, you can see the interaction between the charge residues, positive charge residues and this a ring systems right. So, there is the distance of 10 angstrom then we can treat this as cation pi interactions. In fact, the in the original paper when gallivant dougherty developed the concept of cation pi interactions right this 10 angstrom distance is so long.

So, they imposed constraints with the interactions energy, they computed the electrostatic energy and van der Waals energy, how to get this electrostatic energy? q1q2 by r. So, we can see electrostatic energy right we will discuss one of the later classes about the details and the van der waals energy given by the leonard jones potential potential and get the energy and the energies also there is some cutoff for example, this is less than minus 2 kilocal per mol, right.

In these things we are confident that these 2 residues. Are they forming the cation-pi interactions? If you take the only the distance right, maybe they are very far. So, energetically may not be favorable, but they have high tendency to form the cation pi

interactions. When you really see the cation pi interactions, you need to consider the energy along with these distance criteria.

(Refer Slide Time: 11:20)



So, we can get the hydrophobic interactions, see here you can see the C beta residues right because C alpha is the main chain. So, we go to the side chains, they we need the C beta. So, this is why they considered the C beta residues of these the specific hydrophobic residues, and they consider the cutoff of 5 angstrom, and then see if the c alpha c beta residues which are in the limit of 5 angstrom, then you can see that these residues are involved in hydrophobic interactions.

Then go for the hydrogen bond interactions right what are the requirement for the hydrogen bond?

Student: 2 electronegative atom.

2 electronegative atoms.

Student: 1 with the hydrogen.

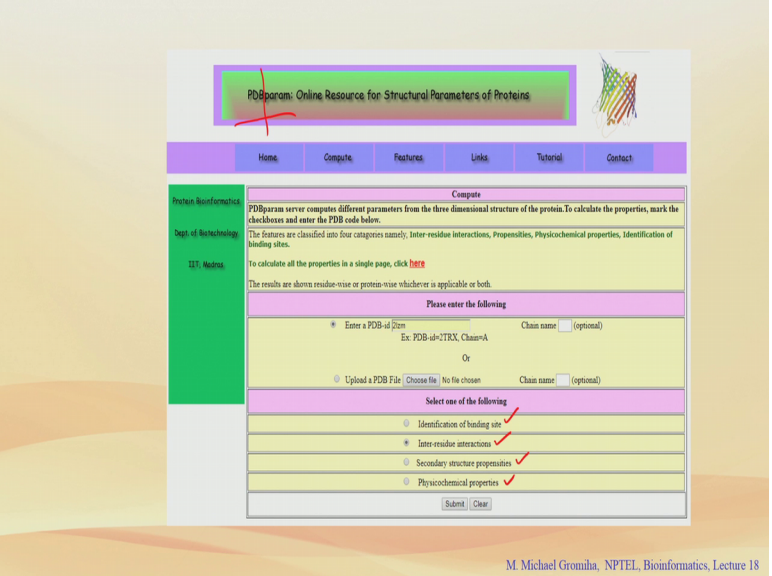
1 with the hydrogen.

Student: (Refer Time: 11:54).

With the attach with the hydrogen and the distance right that is the 2.5 to 3.2 angstrom. So, we can get the hydrogen, main chain main chain interactions, main chain side chain interaction side chain side chain interactions you can see all the hydrogen bond distance right if you have the calculate the distance right between these heavy atoms.

So, we discussed about various types of interactions, and various parameters properties which we can derive from protein 3 D structures. Doing computationally all the parameters requires time right and again there is a program called PDB param.

(Refer Slide Time: 12:31)



This is an online resource for the structural parameters of proteins right this is for they use the PDB structures this way they put PDB, they derive various parameters this is why they use the term param.

So, PDBparam tells you and to provides you the numerical values for several parameters, which can be obtained from protein 3 D structures. If you go to the PDBparam. So, there are 4 different types of values you can get right and for 4 different aspects we can get the binding sites that I will explain later when we will discussed about the binding sites, and you can inter residue interactions what a inter residue interactions we discussed?

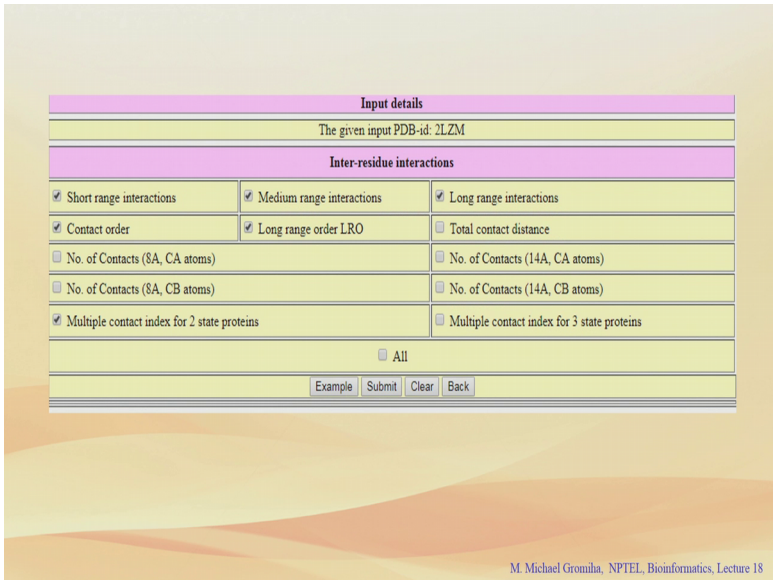
Student: Aromatic interaction.

Right you can see the short range interactions, medium range interactions, long range interactions and you can get this contact order, you can get the long range order, all these

things you can get. The secondary structure properties you can see the tendency to form helix strand and so on and the physicochemical properties, if you see there are various types of interaction radius of gyration, center of mass and the different types of interactions right electrostatic; cation pi and the hydrophobic interactions and various type of interaction.

If you are interested in the inter residue interactions for example, contact order or long range order or short range contacts or long range contacts, you go with the inter residue interactions.

(Refer Slide Time: 13:56)



If you click on inter residue interactions right then you will get a several options which parameter you want. So, to decide these parameters right we also required the PDB file. So, in this case in the previous slide you can have an option.

So, to enter the PDB id right. here I gave the 2LZM right either you can give the chain name or without chain name also you can give the PDB id. We give the chain id it will calculate only for the particular chain. With the chain id is not specified then will calculate for all the chains either you can use to PDB id or also you can upload the file right in the PDB format. So, if you have your own file for example, if you obtain any structure from md simulations right the structures are not available in the protein bank. So, we want to calculate the parameters then you can upload the file. So, then you can get the parameters; so both options are available in PDB param.

So, now we give the PDB id and you want to get some parameters right. So, for example, the inter residue interactions right it tell you this is the id PDB id you can check, and there are various options available there are various properties what is short range interaction? 2 or 3 residues right the 1 and 2 residues and the medium range interactions 3 or 4 residues right and long range interactions some more and more than 4 residues right in this programs. So, they have a little bit difference in this a limit that I will show in this slides.

Then you can calculate a contact order, how to get the contact order?

Student: The number of 1 type of (Refer Time: 15:28).

Right number of distance separation normalized with the.

Student: Number of residues.

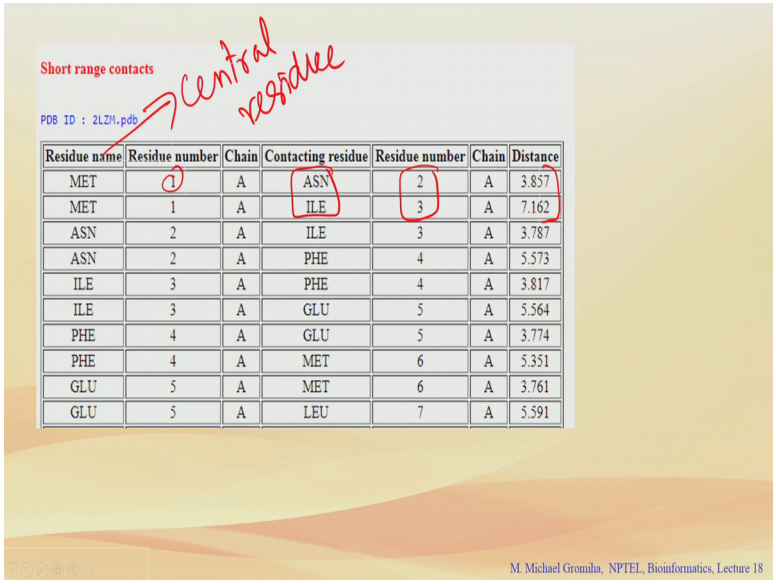
Number of residues and number of contacts. Long range order?

Student: Number of long range contact.

Right you can see consider only if the 2 residues are separated by many of 12 residues right. Total contact distance is another parameter, which is the combination of contact order and long range order they combine these 2 and then develop this parameter total contact distance and also you can see the number of contacts within 8 angstrom either from C alpha atoms or C beta atoms, and 14 angstrom the C alpha and C beta atoms why they give 8 angstrom 14 angstrom because these distance are widely used in literature. 14 angstrom covers almost several regions in a protein and 8 angstroms well used in several aspects.

Then we can get the multiple contact index right for different types of proteins. If you want to get all the data just click on all on otherwise if you are specifically interested on the long range order or contact order you can get the data. So, now, if you submit we will get output. So, this is the residue number.

(Refer Slide Time: 16:29)

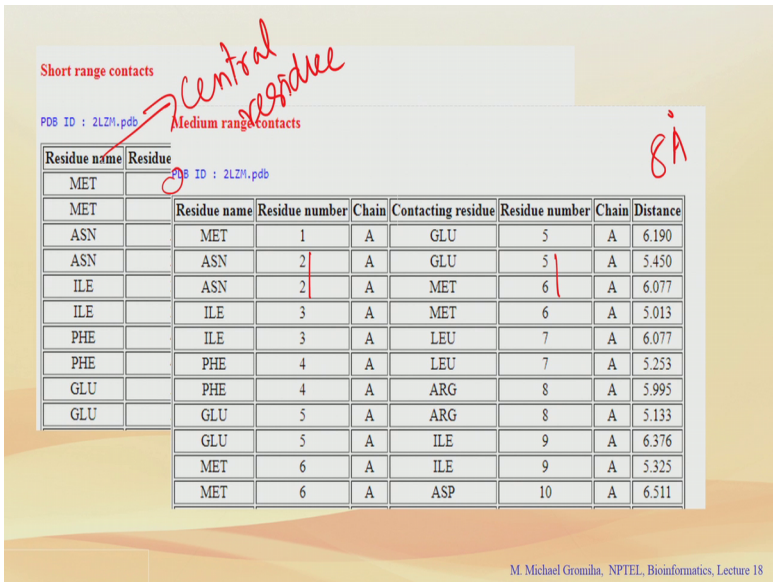


Residue name	Residue number	Chain	Contacting residue	Residue number	Chain	Distance
MET	1	A	ASN	2	A	3.857
MET	1	A	ILE	3	A	7.162
ASN	2	A	ILE	3	A	3.787
ASN	2	A	PHE	4	A	5.573
ILE	3	A	PHE	4	A	3.817
ILE	3	A	GLU	5	A	5.564
PHE	4	A	GLU	5	A	3.774
PHE	4	A	MET	6	A	5.351
GLU	5	A	MET	6	A	3.761
GLU	5	A	LEU	7	A	5.591

One its contact with 2 and 3, 2 is contact 3 and 4 right 3 with 4 and 5. So, here this is the residue numbers with the chain information this is your central residue. So, these are the contact in residue.

So, for example, this one is contacted with these 2 residues, they are located in position 2 and 3 and this is the distance right. So, you can get all the; a short range contacts likewise go with medium range contacts right.

(Refer Slide Time: 17:02)



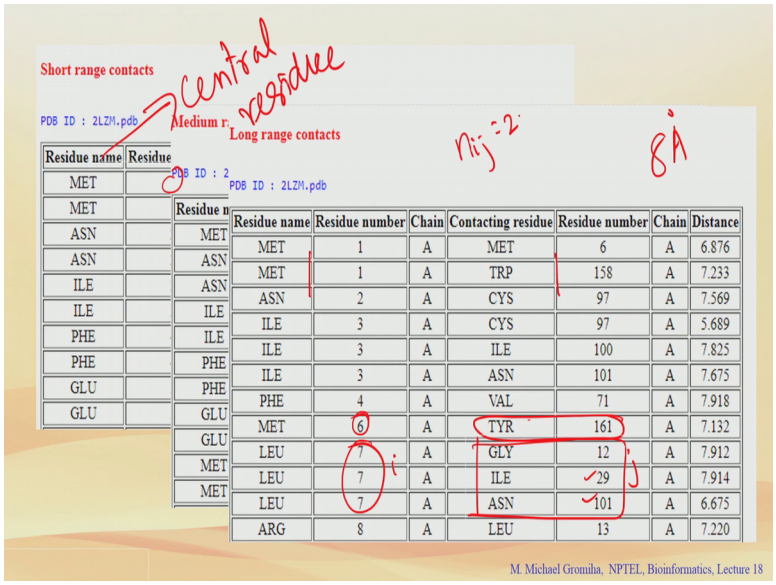
Residue name	Residue
MET	
MET	
ASN	
ASN	
ILE	
ILE	
PHE	
PHE	
GLU	
GLU	

PDB ID : 2LZM.pdb

Residue name	Residue number	Chain	Contacting residue	Residue number	Chain	Distance
MET	1	A	GLU	5	A	6.190
ASN	2	A	GLU	5	A	5.450
ASN	2	A	MET	6	A	6.077
ILE	3	A	MET	6	A	5.013
ILE	3	A	LEU	7	A	6.077
PHE	4	A	LEU	7	A	5.253
PHE	4	A	ARG	8	A	5.995
GLU	5	A	ARG	8	A	5.133
GLU	5	A	ILE	9	A	6.376
MET	6	A	ILE	9	A	5.325
MET	6	A	ASP	10	A	6.511

You can see the distance here they use the distance of 5 and 6 right for the case of 2. So, they use 5 and 6 right they use the I plus 3 and I plus 4. So, then we get the; these are the central residue; this is the contact residues and you have the distance. Distance we give the cutoff 8 angstrom, right. So, we get the distance of 8 angstrom it can get the numbers.

(Refer Slide Time: 17:33)



Likewise we can see sort long range contacts for example, if we see these residues a 6 right with the tyrosine 161 and see the residue 7, how many long range contacts for residue 7.

Student: 3 3.

Right 3 right these 3. For example, if you want to get the long range order right what is nij?

Student: 3.

3.

Student: No/

Because we need to have the cutoff of twelve residues. So, this is the i and here this is j totally 3 contacts how many of them have the distance separation 12 residues 2 right this

is 5 and this and this right this case nij equal to 2 right you can get these values right. If you see the long range order you can see the numbers.

(Refer Slide Time: 18:30)

So, now, we go with the contact order. So, for this case you have the number 9.102. We do the calculations for some all alpha proteins and some all beta proteins you can see a difference right how far the contact order right varies based on different structure classes of proteins. Then long range order we can get for all the residues. So, here the total value is equal to 0.744 right and for all each residues you can calculate.

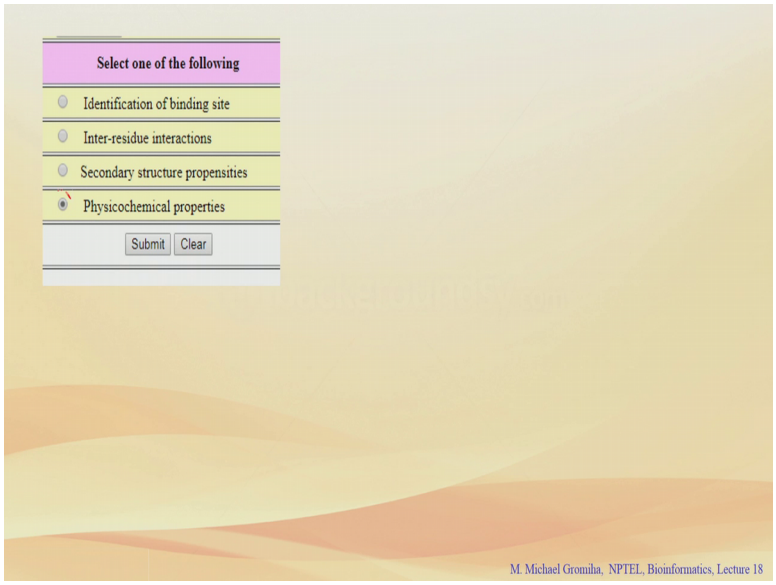
So, for example, take residue number 1 long range order equal to 0.006 because we got 1 contact. So, we can see is one contact in this case 2 contact right. So, this is this 7 this will be 0.012 number residues equal to 164 this will be divide by these 2 by 164 you will get this number right you can then this is fine.

Then multiple contact index you also you can calculate here the value is 0.018, compare with the contact order long range and multiple contact index you can see the number is in which order decreasing order, why is decreasing order? Because we impose more constraints first one no constraint, in the contact order the contact take all contacts and they give weightage 2 distance separation and long range order what is the constraint we give?

Student: Greater than 12.

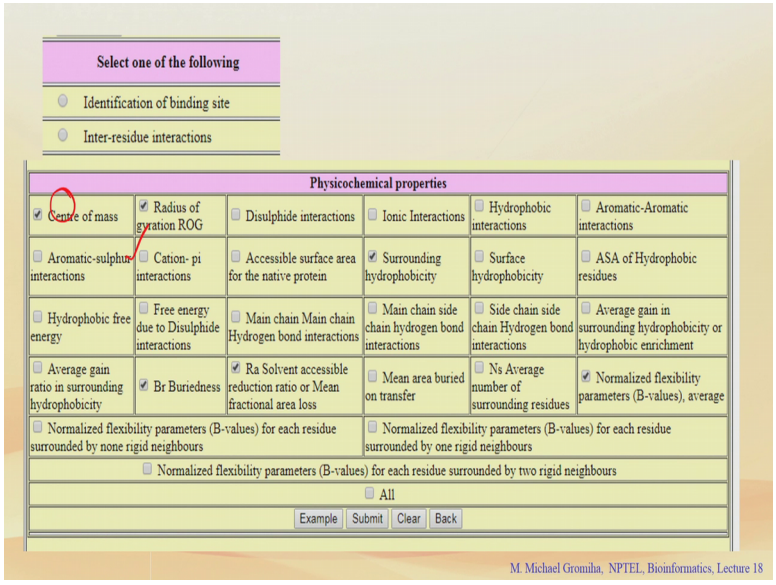
Greater that 12. So, in this case is if we reduce several contacts and you go with the multiple contact index, we again reduce because we see the residues which have at least 4 contacts, this way number is less. But if you have a set of proteins then you can compare, on the second one right if you want to get different types of interactions or the center of mass or radius of gyration.

(Refer Slide Time: 20:09)



In this case you select the physicochemical properties and if you submit right.

(Refer Slide Time: 20:14)

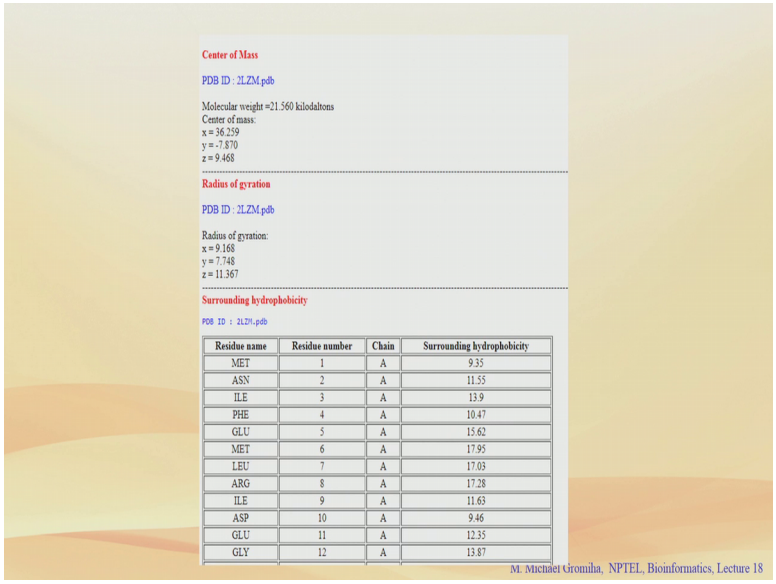


It will show you all the property, there are various properties you can calculate; center of mass, radius of gyration all types of interactions, disulfides ionic interactions and several interaction.

But several case I did not explain the details, but if we go to the PDB param that is a file right the tutorial file as well as you can see the compute file, there if you click on this files you can get all the details what is the criteria used for different interaction right as well as the formula and the all the details you can get.

So, here now if several physiochemical properties if you get to all these things, you put all and if you submit I show you 1 or 2 examples right.

(Refer Slide Time: 20:52)



This is center of mass this molecule weight is 21 kilodaltons and this is center of mass for all the 3 coordinates xyz. Radius of gyration this is the radius of gyration and the 3 different axis and surrounding hydrophobicity right what is surrounding hydrophobicity it is the.

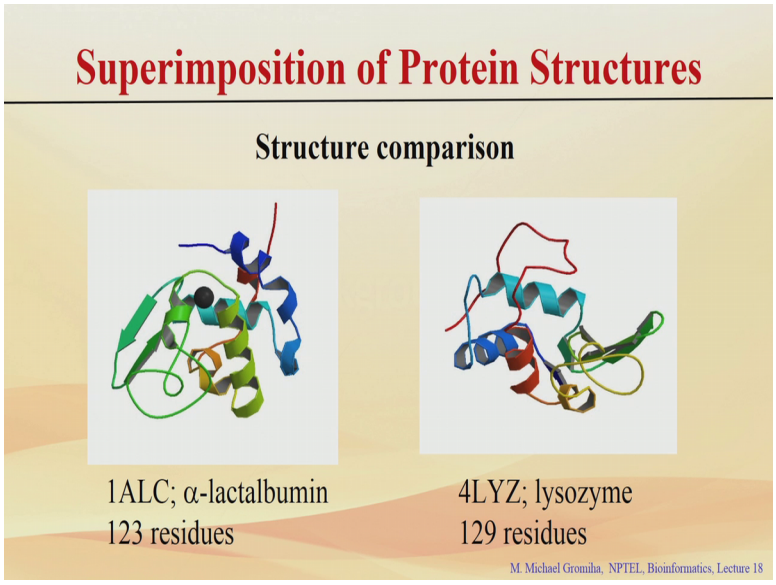
Student: Average of hydrophobicity.

Average the total hydrophobic values of all the residues which are occurring within the limit of 8 angstrom. See if you see some residues you know the contacts right. So, you can calculate the values then this case for all the residues we have the surrounding hydrophobicity. For example, if you take some of the residues which are high say 15 or

17 and some residues which are very less say 9 and 10 and so on. Even some polar residues which are highly hydrophobic, if they are surrounded with the hydrophobic environment hydrophobic residues.

Non-polar residues also if you are in the midst of the polar residues they have less, but you take the average and then you can see the tendency of these residues right specifically between the polar and un-polar residues. So, now, if you get 2 proteins, we discussed about the various parameters and we can also see how far the 2 structures they can be super imposed and how far they are similar to each other.

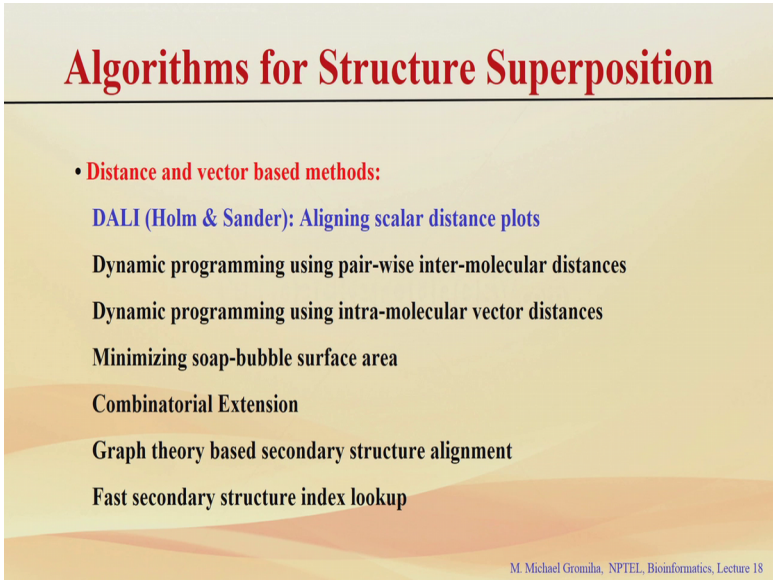
(Refer Slide Time: 21:59)



Here I show 2 proteins one is the lactalbumin and lysozyme, they are from different families. So, 2 different proteins this have 123 residues and 129 residues. Sequence identity is less, but if you compare these residues look into these residues you can see some sort of similarity in structures. See this helices here you can see some similar type of helices and the strands here similar type of strands here.

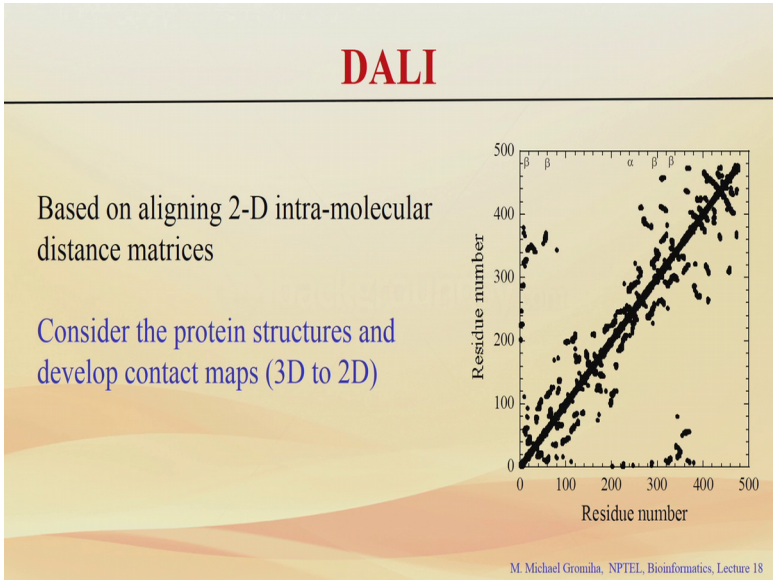
So, how far they are different structures right in this case put one structure and put on the other and the top, and measure the distance between the different locations right this will give you the root mean square deviation, this will tell you how far these 2 proteins can be aligned in the structures.

(Refer Slide Time: 22:38)



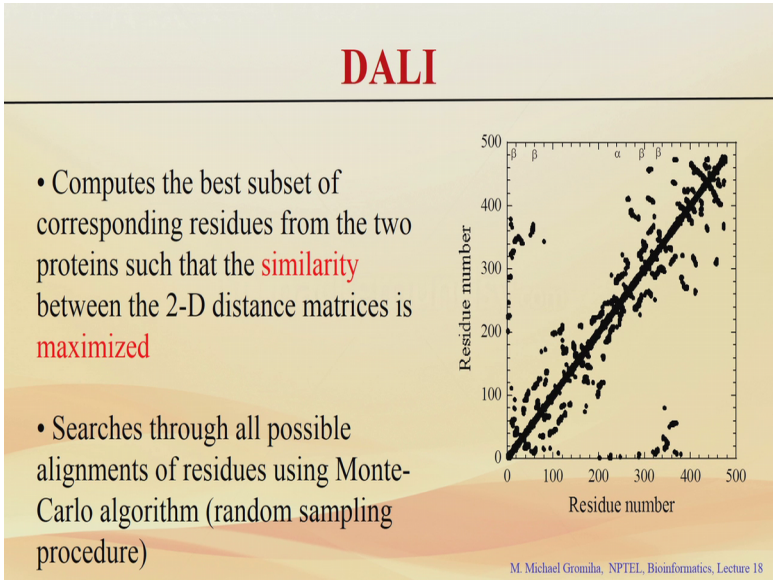
There are various methods to align the structure for example, distance based methods and vector based methods I will explain briefly about one methods called DALI right distance matrix alignment developed by the Holm and sander right. So it is mainly based on the contact maps.

(Refer Slide Time: 22:51)



So, we discussed about the contact maps, this is the representation of the 3D structure information in 2D form. So, we get these contact maps and we compare how far the contacts maps are similar.

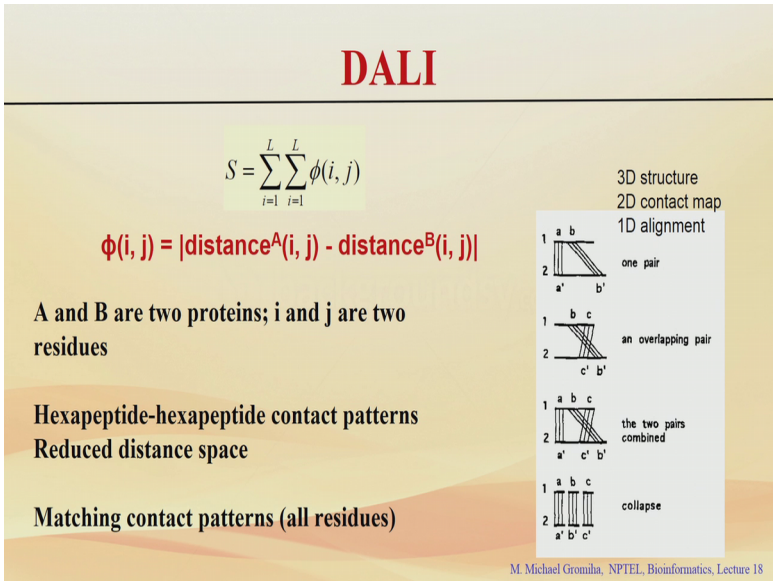
(Refer Slide Time: 23:06)



How to do this. First, we set the subset of the residues they have the hexapeptides cut into different pieces and then see how these hexapeptides matches.

So, for example, there are some helices in the N terminal in the first protein, and the C terminal second protein if they have same contact map or similar contact map then they can be aligned together in 3 D structures. They use the search using Monte Carlo algorithm right using the greedy algorithm it will find most probable matches and then combine together and put it together in the 3D structures.

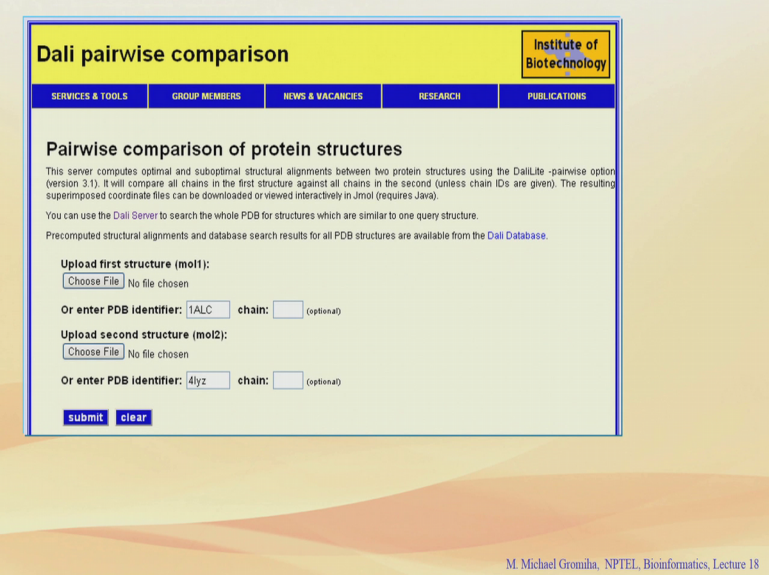
(Refer Slide Time: 23:38)



How they do? So, for example, this a and a dash if one is in a and the second is a dash the same place then directly we can align. Something is the second is b and the b dash is very far then we see the contact map and this can be aligned together put it together. In second option is an overlapping pair ab with b dash and c with c dash and the third one you can combine these two right a is directly mapping that exact match and here b with b dashed and c with c dash.

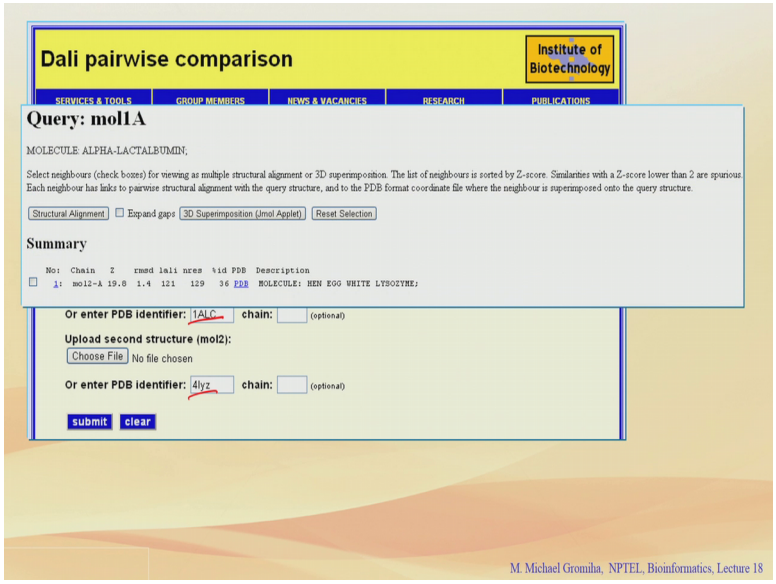
So, value will collapse and then make similar to each other b with b dash, and c with c dash. So, now, they calculated the score they have the score this phi, this is the distance of protein a and protein b between any 2 residues. Using this formula right they can calculate the distance between i and j in protein c and b and they get the based on the score they can match the residues, they can map the residues. Once you map the residues then they align together then see the RMSD and see the best possible match and finally, they put together the two structures.

(Refer Slide Time: 24:47)



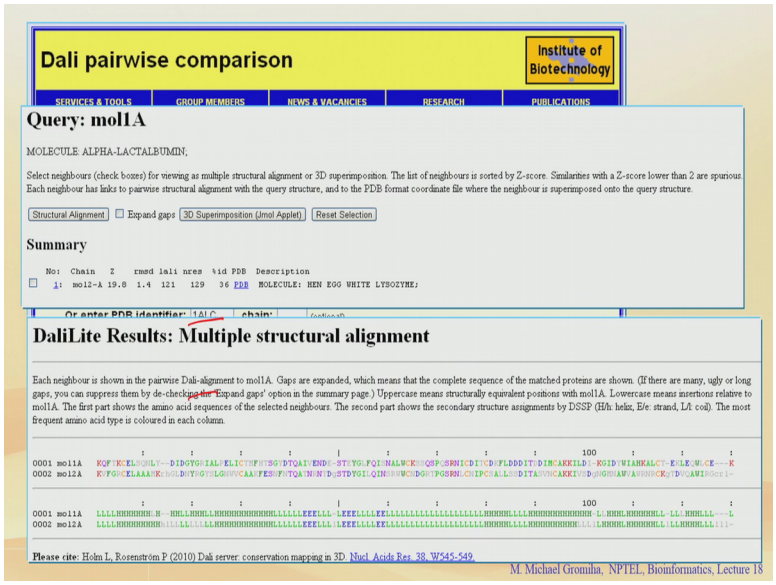
So, this is the server for the Dali. So, it asked for the proteins one is this is first protein one, protein 2.

(Refer Slide Time: 24:54)



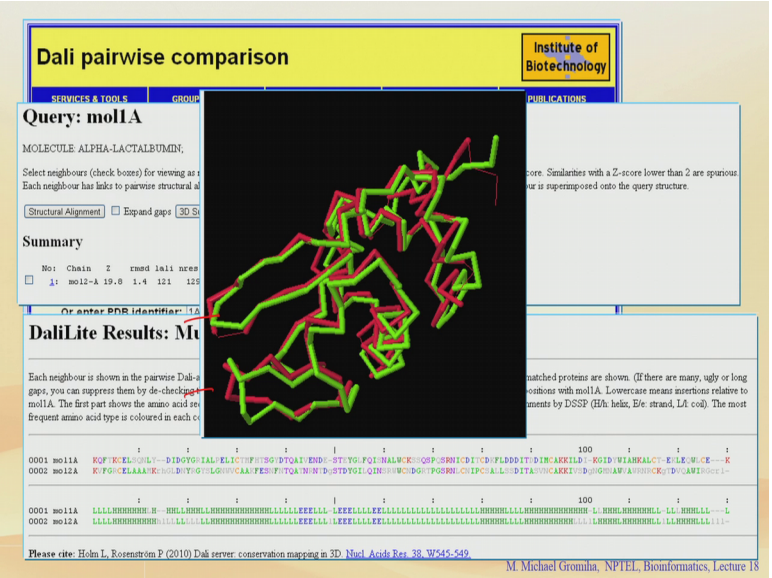
So, then take this protein and finally, it asks for the superimposition get the similar structures.

(Refer Slide Time: 25:01)



And do the structure based sequence alignment in this case they do not align the sequence wise they get the similar structures they align together.

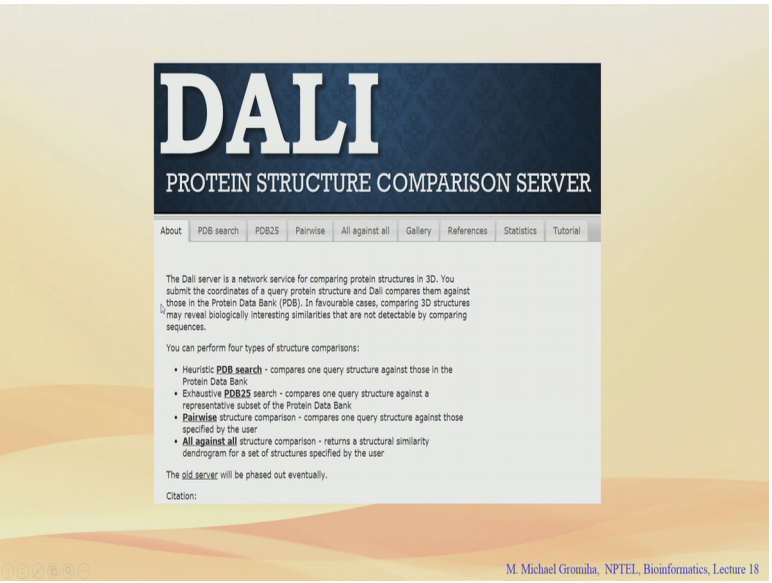
(Refer Slide Time: 25:12)



So, based on the structures they align together and using this information they will construct this superimposition.

If you see the red the magenta one and the green one right the red one is aligned very properly right the RMSD is a around 1.4 angstrom, right between these 2 protein structures.

(Refer Slide Time: 25:29)



So, recently they develop this serve you can see the new server, in this case they already calculate the similar structures for several cases and the develop database for the similar

structures right, in this case you do not have to calculate again. So, if you see these 2 proteins which is available in database, they then you can easily extract these 2 residue proteins which are close to each other and so on.

So, over all if you see the structure based parameters, what are the various parameters we discussed?

Student: (Refer Time: 25:55).

Starting from the beginning is start from the contact maps, solvent accessibility.

Student: Buriedness.

Buriedness.

Student: Buriedness transfer a free energy.

Transfer free energy.

Student: Reduction in accessibility

Reduction in accessibility, contact order.

Student: Long range order.

Long range order.

Student: Multiple contact index.

Multiple contact index.

Student: Surrounding hydrophobicity.

Surrounding hydrophobicity.

Student: Flexibility parameter.

Flexible parameters we discussed plenty of parameters right then after that different type of interactions what are the possibility of the 2 specific residues, which can form electrostatic ion pairs or electrostatic interactions or disulfide bonds cation pi interactions and so on.

So, then discussed about the program called PDB param, it takes the PDB id and calculates more than 50 parameters right. This can be useful for any project and we can take different types of proteins and calculate the parameters and see how the parameters vary in different classes with different functions and see why these parameters are attributed for different functions, then we discussed about the alignment. So, we how can align these different structures.

In the following classes we will discuss about the structure prediction whether we are able to predicts the 3 D structures from the amino acid sequence or the applications of these parameters, what we derived in these 2 3 classes right to understand protein folding rates or protein stability or protein interactions some sort of applications point of view, we will discuss in the later classes. Then we will explain about the development of algorithms and what are the factors we need to consider, when you develop your own algorithm for example, for if you want do any projects and so on. That will end up with that type of analysis.

Thank you very much.