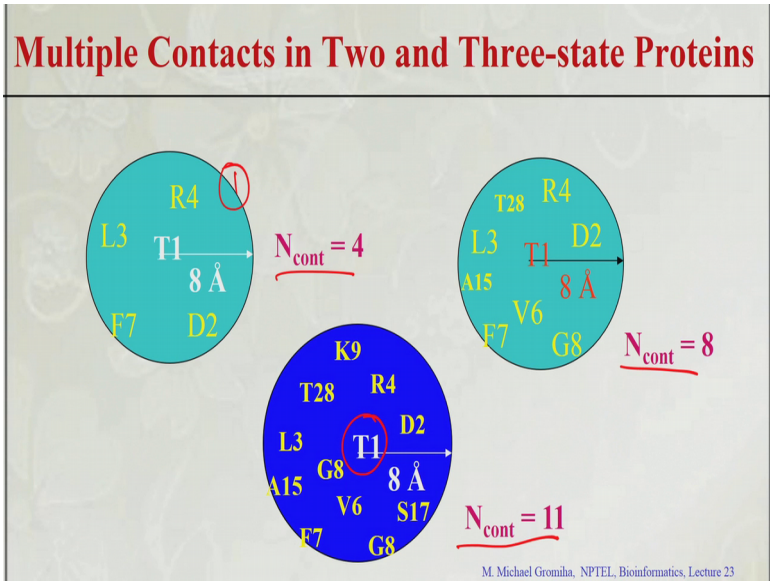
BioInformatics: Algorithms and Applications
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 23b
Protein Folding Rate II

Now we propose one more parameter, this will constitute the residues which have more number of contacts. For example, if there are many clusters right one residue is forming many long range contacts. So, then see in a protein how many clusters you can find right based on long range contacts and there are more clusters are formed by the long range contacts then this proteins take time to fold under this assumption.

(Refer Slide Time: 00:47)



We see the residues which can have more number of contacts for example, take this as an example how many contacts

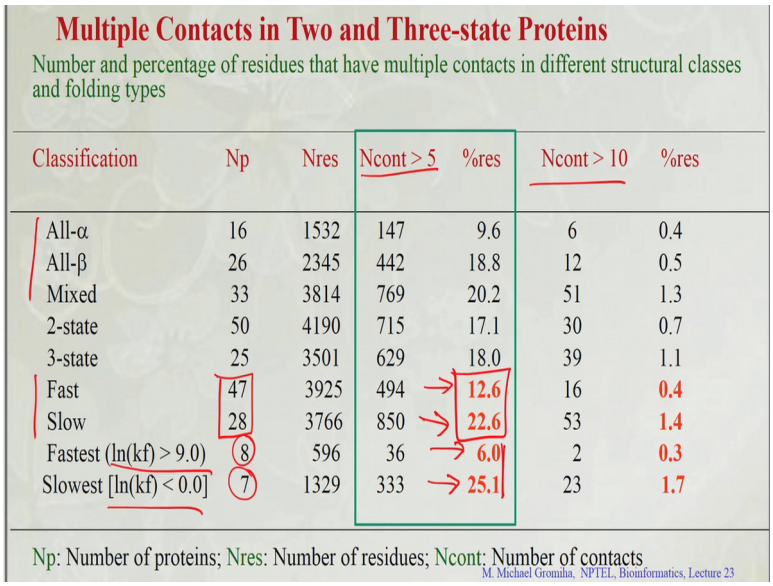
Student: 4.

4 contacts, right the only 4. So, if you take this one 8 contacts with this one; 11 contacts this is for the residue T1; for example, a protein one case only 5 residues have this multiple contacts one protein they have 20 residues which have multiple contacts which will fast fold fast which will fold slow, right. One with only 5 residues; there are having

multiple contacts another one 20 residues have the multiple contacts which will fold fast the 5 one fold fast because only the 5 one, they have to get more number of contacts.

In this case all the 25 different residues we are making the contacts from different residues. So, it takes time to get all the contacts.

(Refer Slide Time: 01:39)



So, you get a several example. So, you have the different classifications alpha, beta, mixed and the fast and slow folding proteins. So, here this is are the percentage of residues which have multiple contacts. So, if you see the fast and a slow folding proteins the with a condition of 5 residues we see in here we see the clusters only if the a residue contains more than 5 long range contacts we do so, fast folding proteins there are only 12.6 and the slow folding proteins you can see the 22.6 percentage of the residues they are present in the case of slow folding proteins.

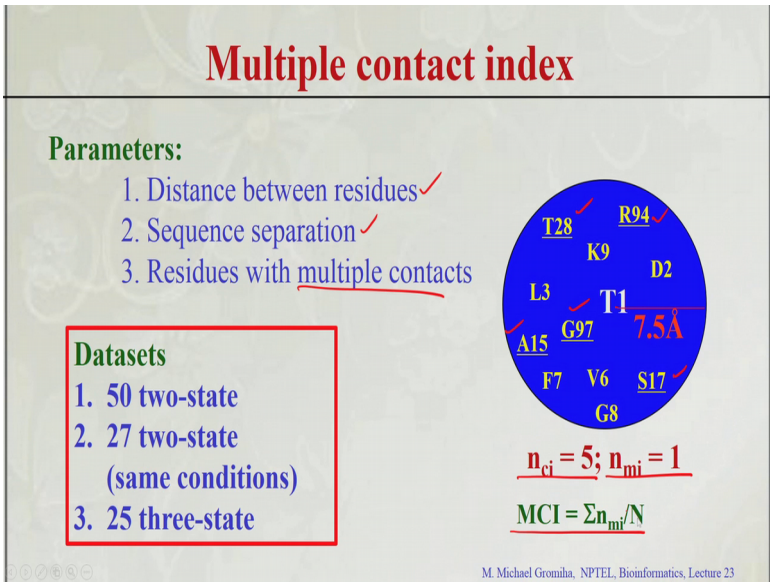
This will tell you that the proteins with the more number of the multiple contact residues are slowing down in the folding process, then I show another example we take the fastest; that means, ln kf is more than nine and the slowest this is less than 0. Now we have the 8 examples for the fastest and 7 for the slowest.

So, here we see only 6 percent in the fastest and more than 25 percent in the case of the slowest proteins this will clearly tell you the proteins right which containing more number of residues with multiple contacts right they are slow folders. Here we did with

the 47 fast and 28 slow proteins, here the fastest and slowest we can see the difference between the percentage of residues with the multiple contacts.

So, based on this information we can derive a parameter just we can relate the folding rate with an because we need to quantify right here this will tell you some numbers how many residues which you have a threshold of more than 5 contacts in the here we give the more than 10 we can change, this number then also in this case we need to quantify.

(Refer Slide Time: 03:25)



So, based on that there is another parameter called the multiple contact index have been developed right; here they take distance between residues is already used in the previous methods right; for example, contact order long range order and the second one is sequence separation this was done earlier.

Now, here in addition to that we added the one more feature what is this feature.

Student: Number of residues.

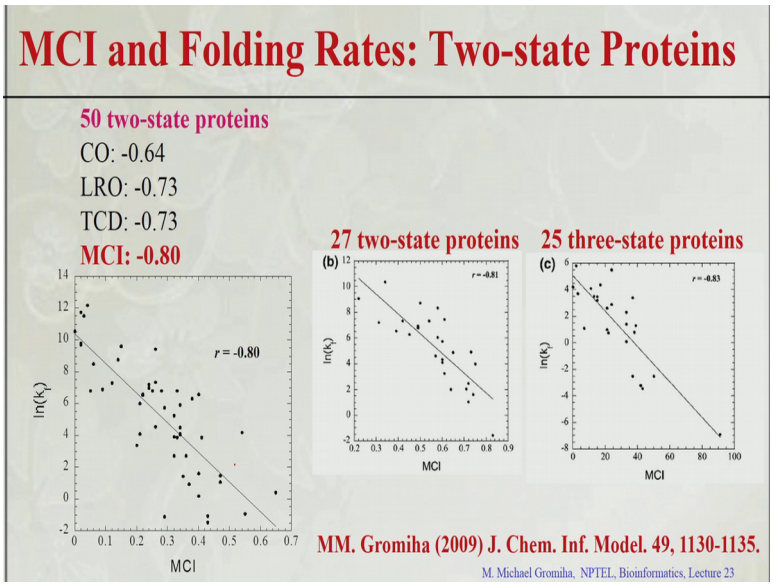
A residues with multiple contacts. For example, if you see this figure here the how many this you take the T1, how many long range contact this can form one.

Student: 5.

1, 2, 3, 4, 5 right. So, in this case n_{ci} equal to 5 if this is the case then you can you can define n_{mi} equal to one because it is more than 4 or 5 we can put any cutoff right and if this is a case MCI we can calculate sigma n_{mi} by n for all the protein and get the number.

Now, you see whether this has any influence in the folding rates compared with long range order and the contact order right, here we have different data sets right 2 state proteins and 3 state condition 3 state proteins right.

(Refer Slide Time: 04:36)

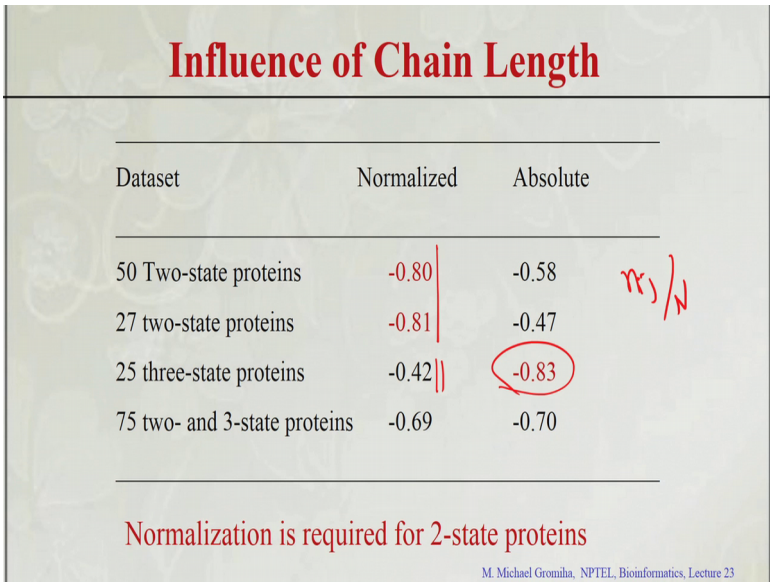


So, this is a data you can see this is a 2 state proteins a fifty proteins contact order 0.6, long range order is 7 3, this TCD a total contact distance this is the multiplication of the CO and LRO right. So, 0.73 and MCI is 0.8 because this will this information will enhance the correlation between the MCI and the folding rates.

So, now we try to see whether this is correct or not how far; this can be trusted. So, we set another 27, 2 state proteins same experimental conditions. The beginning on the lecture, I discussed about the data determined by the different experimental groups with same conditions. So, for that if you do that, but there also we get the about the similar correlation minus 0.81; that means, MCI is a good measure to relate protein folding rates, then we use the same with the 3 state proteins. Here also we could get the same correlation of 0.83.

So, then this this can be a good measure for predicting or relating the folding rates. Then we need to compare the 2 state and 3 state proteins there is one important information between 2 state and 3 state proteins the 2 state proteins.

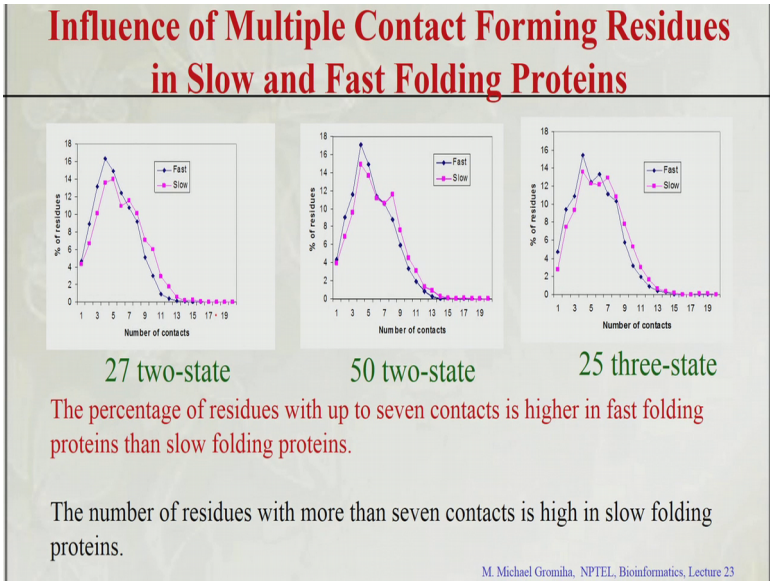
(Refer Slide Time: 05:49)



We need to normalize with n for example, in their long range order. So, what is long equation of a long range order n_{ij} by n, but the case of 3 state protein right if you normalize the values is very less, but if you do not normalize, then this is 0.83 still it is very clear why this normalization is very important for the 2 state proteins and it is not required for the case of the 3 state proteins.

3 state proteins if you normalize the values are very less 2 state proteins, if you do not normalize the values are very less. So it is very important normalize the 2 state proteins right they to get the folding rates with better performance. So, till now we discussed about the different features contact orders; long range order as well as the multiple contact index. Now if we see whether any difference in the folding rate and the number and the fast folding proteins and the slow folding proteins based on number of contacts.

(Refer Slide Time: 06:45)



Here if you see the x axis the number of contacts y-axis a percentage of residues right. So, the percentage up to the 7 residues for example, up to here. So, see this is high in the fast folding proteins after that it is higher in the case of slow folding proteins the same in the case of 27- 2 state proteins and the 3 state proteins as well as another set of 52 state proteins; that means, you take the slow folding proteins right these proteins have more contacts than the fast folding protein. Because if we see the slow folding proteins the values are very high compared with the fast folding proteins right that is we can see the relationship between the slow folders, fast folders and the contacts right they can be have a picture regard related with all these parameters ok.

These are we did with the completely with 3D structures. Now is it possible to get this information from the sequence the topological parameters contact order, long range order multiple contact index, we have need a structure is also depends upon the type of the protein alpha beta mixed class proteins right can we predict or can we relate the sequence based parameters with the folding rates last class we discussed about the stability right.

So, how many properties we used last time which type of properties: physical, chemical energetic and confirmation properties right we discussed about all the properties. They are available and developed right, those values are there. So, how can we relate these properties with the folding rate right, there are various aspects to predict the folding rate

from the sequence; either we get the sequence based parameters right what are sequence based parameters we discussed in the earlier classes?

Amino acid composition occurrence.

Student: Average.

Average property values conservation.

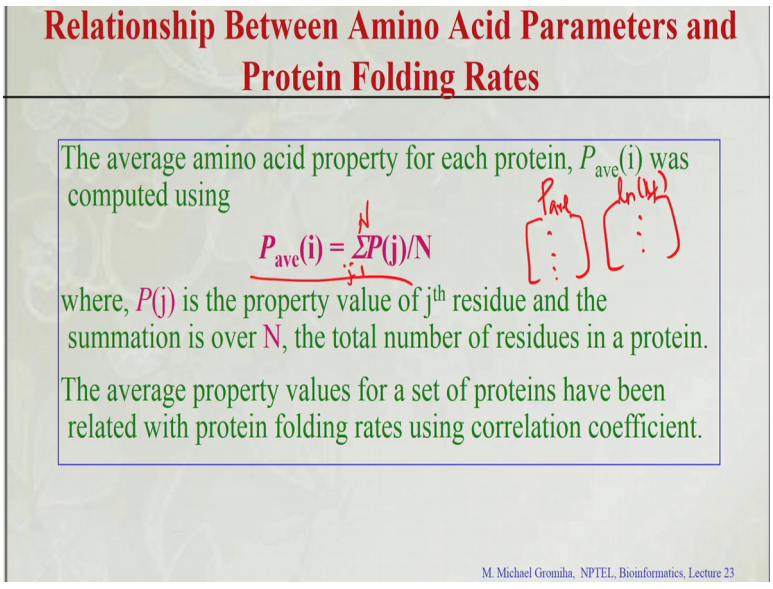
Student: Hydrophobicity

A hydrophobicity profiles right. So, we can see if we have many properties, we can see the average property values, if we take a full protein right we can get the average property values. How to get the average property values.

Calculate total divided by ligand normalize with the chain length. So, then we see whether it can be related with the folding rate or we can do because contact order and long range order based on contacts. So, if one program can predict the contacts then we can use their contacts to predict the folding rate right that is also possible or we have to model the structures and then you have to use a structures to get the contacts because we a long range order depends upon which atom c alpha atom.

So, crude model is enough; we do not need the refined model. So, then we can calculate the LRO and we can predict the folding rate right with the various ways.

(Refer Slide Time: 09:37)

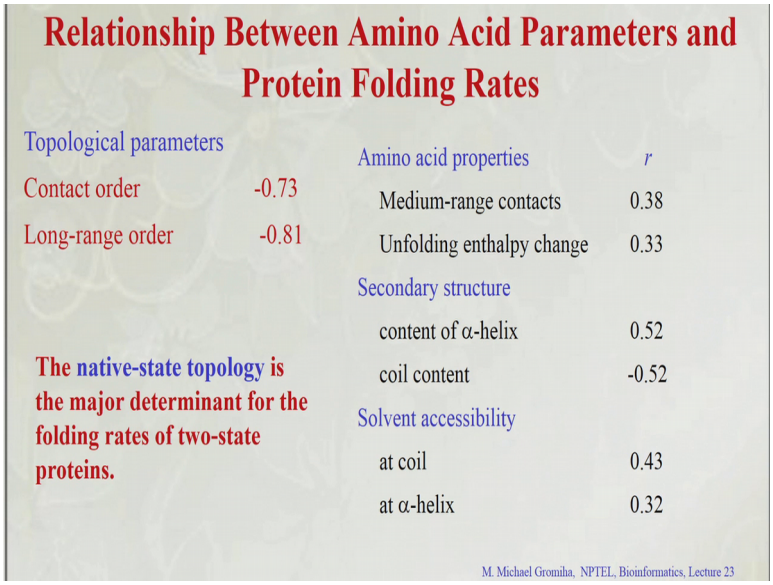


So, let us see whether any specific property right can relate the folding rate right. So, in this case first we calculate the average value right, for example, if we take the hydrophobicity how to calculate the average hydrophobicity; here we take the hydrophobicity of values; all the residues right j equal to 1 to n, right j equal to 1 to n what is n? n is the;

Student: Number of residues.

Number of residues in a protein right. So, compute n. So, they normalize with n then we get the for any protein I we will get the average value. So, we have the for example, we have the hundred proteins for all the hundred proteins we can calculate the average property right we can gave the average property here. So, here the P average. So, here we have the ln kf. Now we relate how to relate these 2 using correlation coefficient you get the correlation coefficient right then see whether any correlation between the average property value as well as the folding rate. So, this is the data.

(Refer Slide Time: 10:38)



So, we used various properties right physical properties chemical properties confirmation properties energetic properties and these are the properties we showed the highest one the values are very less there is only 0.38; that means, any single property they cannot relate the folding rate right that is also correct; that is fine because if the single property can do it right then there is no use of these topological parameters and. So, on then we try to a secondary structure right. What are different secondary structures.

Student: Alpha helices.

Alpha helices beta strand and the coil turns, right.

You can see the contents a commonly residues or the percentage of residues in alpha helix or the beta strand or the coil right or a helices and strands. So, we if you compare all these secondary structure information we can get up to 0.52 and then we use solvent accessibility right, either you can predict the solvent accessibility from sequence or you can use these exact data, right. So, with what are different locations based on solvent accessibility.

Student: Exposed, burried.

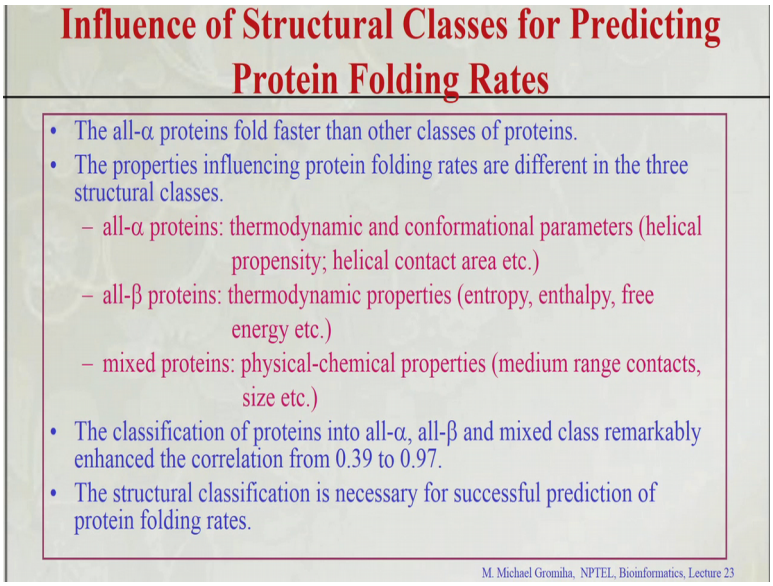
Exposed, burried, partially buried and partially exposed right. So, we try to use various aspects even their correlation is up to 0 point four; that means, amino acid properties or secondary structures or solvent accessibility they could that some limitations. So, they can get up to 0.5. So, on the other hand if you use a contact order or the long range order you can get up to 0.8 or multiple contact index right these from these analysis what can you infer we can see that a native state topology is very important and that is a major determinants in the folding rates of these 2 state proteins ok.

So, native state topology means this will tell you the folding type of a protein right. So, in this if so, what are the different types of classes?

Student: Alpha.

Alpha beta and mixed class right. Mixed class means containing both alpha and beta right. So, if you classify the proteins based on structures and then use these properties then we include the some of the topology aspects. So, because for different classes or different folds.

(Refer Slide Time: 12:46)

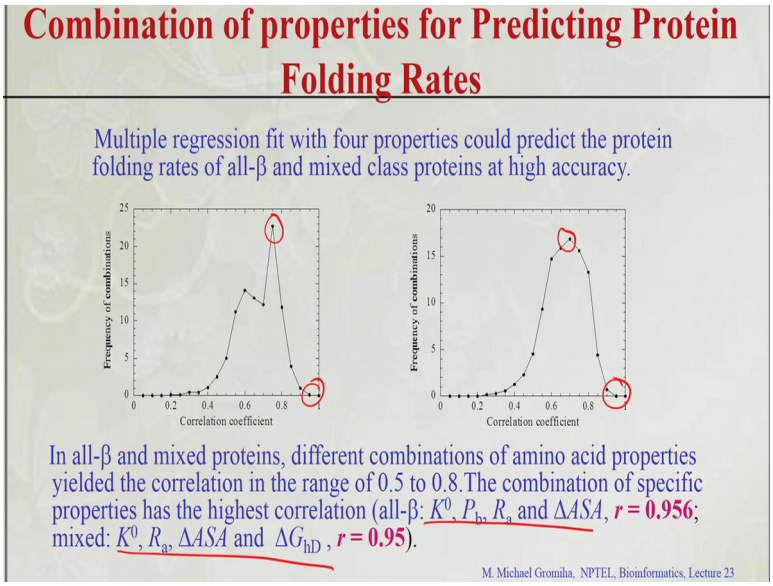


So, then we did that. So, in this case we see that all alpha proteins there are some specific parameters like the conformational parameters are important in all alpha proteins. The all beta proteins mainly thermodynamic properties they influence the all beta proteins.

The mixed proteins we could see that if some physicochemical properties right they play an important role in the case of mixed proteins. So, if you classify the proteins into alpha, beta and mixed class then you could see that that the correlation enhance right very significantly right from the using all the datasets together. Now how to relate right in this case we can we have the folding rate values and we use multiple regression technique right we combined a the different values and you can relate with a folding rates.

Now, the question is when you make the combinations there will be millions of combinations right for example, if you have the 5 residue combination right, we have 2 residues means only 2, if you make more number of combinations right there will be lot of combinations. So, then how do you see whether any particular combination shows the highest performance?

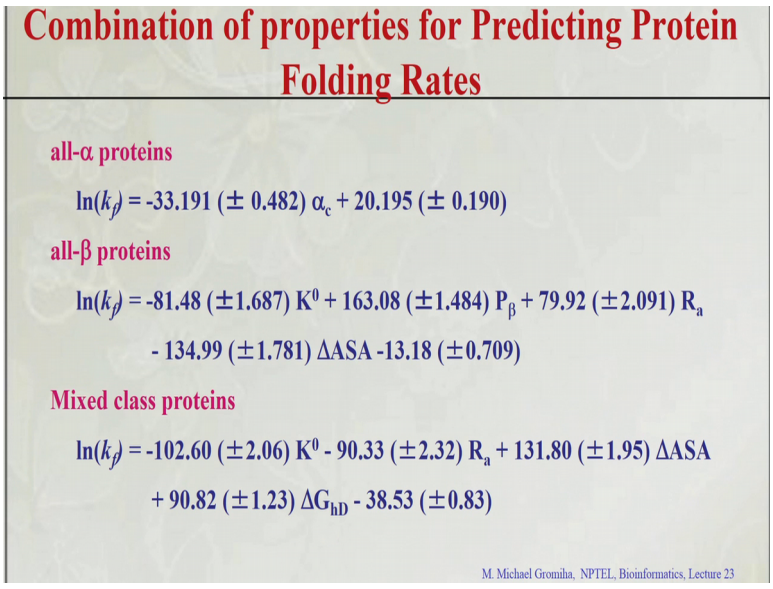
(Refer Slide Time: 13:54)



So, we did with the different combinations and see that this is the frequency of combinations. Most of the combinations you can see this is the performance is very less right and here this is the average performance more than 25 percent about 25 percent.

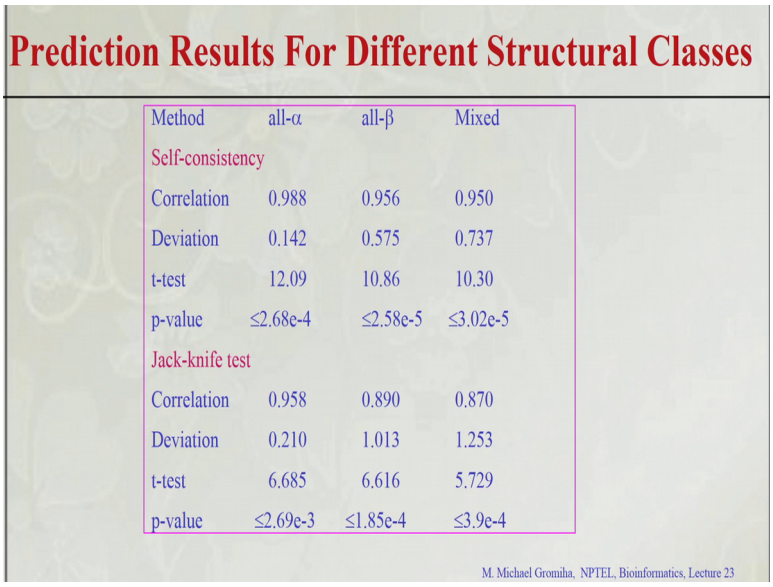
In the case of all beta here also, you can see only for a particular combination it shows the highest value right in this case we could see only specific combination for example, all beta these are the properties and the for the case of mixed class these are the properties to get the highest correlation.

(Refer Slide Time: 14:25)



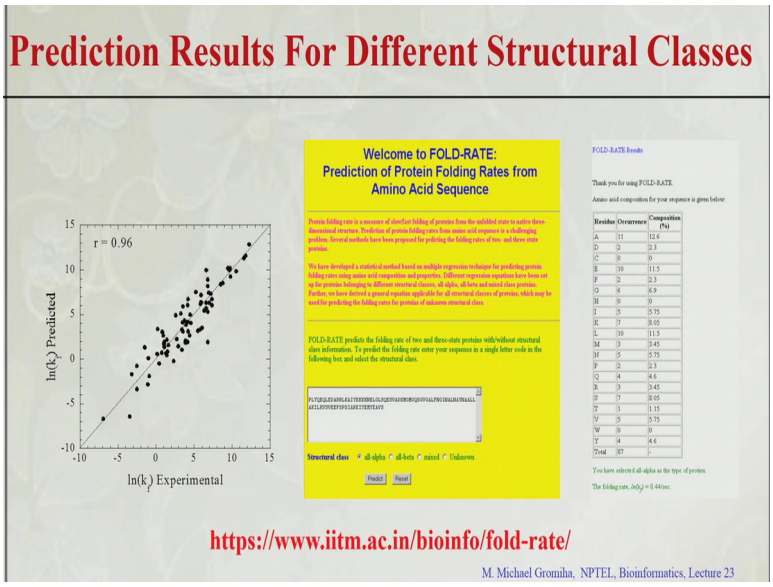
And from that you can derive equations right for alpha because less number of proteins. So, it can be derived from just one property called alpha c helical contact area all beta proteins they used 3 4 properties and the mixed class 1, 2, 3, 4 properties.

(Refer Slide Time: 14:43)



We use these equations right then you can get the correlation of 0.9 for each of the classes and the deviation is also comparatively less and we checked with this the T test and the p value. So, the values are also statistically significant. Once it is done,

(Refer Slide Time: 14:59)



Then we can develop your program right. So, this is the relationship between experimental and the predicted values. So, you can see the good correlation because you can see most of the proteins right the values are on the 45 degree slope. So, when if you get this then we can make a server right. So, here it accepts the amino acid sequence single letter code and then here the important thing is we need to give the structure class.

If you do not use the structure class, if it is unknown then you can predict the folding rate, but this case the performance is not so good and you can have some limitations. If you use the unknown values, but if you know the structure class where it is all alpha or all beta all mixed class proteins then you can take the particular class and appropriately use the equation and this all alpha proteins and you can see this is the 8.44 per second. If you look into these structures how many structures are known in the PDB.

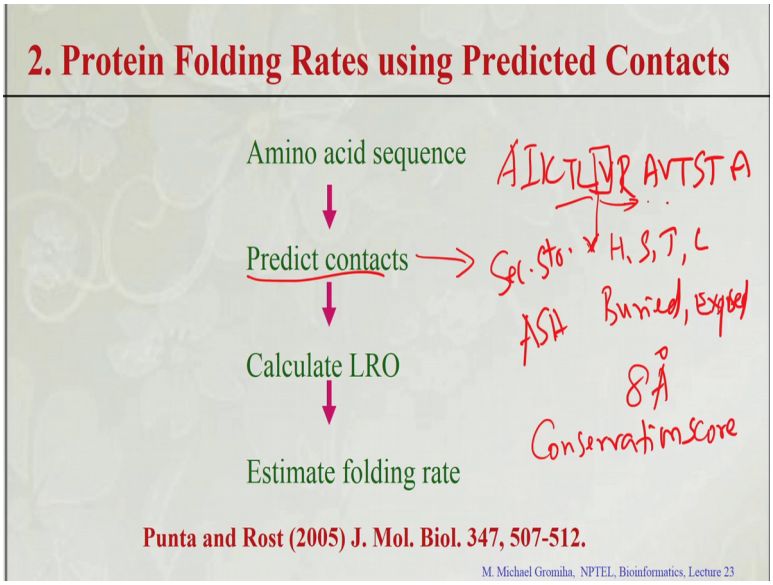
Student: 1, 30,000.

One lakh one lakh thirty thousand structures are known, but with the folding rates we know the folding rates only for about hundred proteins hundred plus proteins. So, many proteins we know the structures. So, in this case we know the all alpha or all beta or the mixed class. Even with the known sequences right many sequences they have high sequence identity with the known structures.

In this case you can get more number of sequences with known structure information. If the structure information known then we can use the server to predict this folding rate even if the structure is not known right. So, you can use the sequence information. This is a one section you can use the amino acid properties combine together and you can predict the folding rate. So, second aspect for example, if we have a sequence and how to predict the folding rate. So, there is an indirect method to predict the folding rate using contacts because we know that topological parameters are important right and these parameters explain well about the folding rates and the parameters are mainly depending on contacts.

And if we are able to get the contacts, then we can convert these contacts into folding rates.

(Refer Slide Time: 17:05)



So, its indirect methods they developed in 2 thousand 5 right Punta and Rost from Columbia university proposed this one. So, what they did. So, first they get the sequence. For the sequence they get the contacts how to get the contacts they use this amino acid sequence information for example. So, take any of this residue for example, take this valine they have identified the residues which are close to valine right.

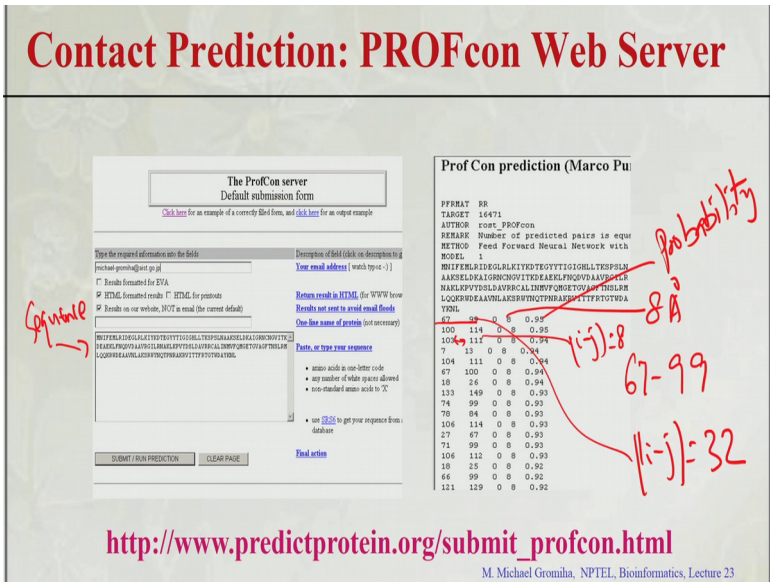
For example, this lysine as well as on the right side towards n terminal and towards e terminal and they located the position of this valine based on secondary structure whether it is helix or strand or turn or coil and based on these solvent accessibility it is burried; this is secondary structure and a solvent accessibility surface area right burried or partially burried or exposed.

And they compared with the known structures then they see this valine is in contact with whate type of the residues take any cutoff for example, 8 angstrom compare with the experimental data right. So, they use these information and they use the model for predicting this contacts in this case, they do not know the structure for any sequence they give the sequence information, neighboring residues and secondary structure solvent accessibility as well as the conservations, right, we were discussed all the score the parameters in the previous classes.

So, take all the information to predict the contacts once the contacts are known can we predict the LRO? Yes. we can predict the LRO, right because this we only we need the

cutoff twelve residues then you calculate the number of contacts the residues with the cutoff of more than twelve residues once we get the LRO, then we can get the folding rate, right. This is the indirect method to predict the folding rate using predicted contacts here, this is a server which can predict the contacts.

(Refer Slide Time: 19:18)



So, here this amino acid sequence right you can take the amino acid sequence in a single letter code. So, then they use all the other information, they use the what are information they use for the above prediction.

Student: Secondary structure.

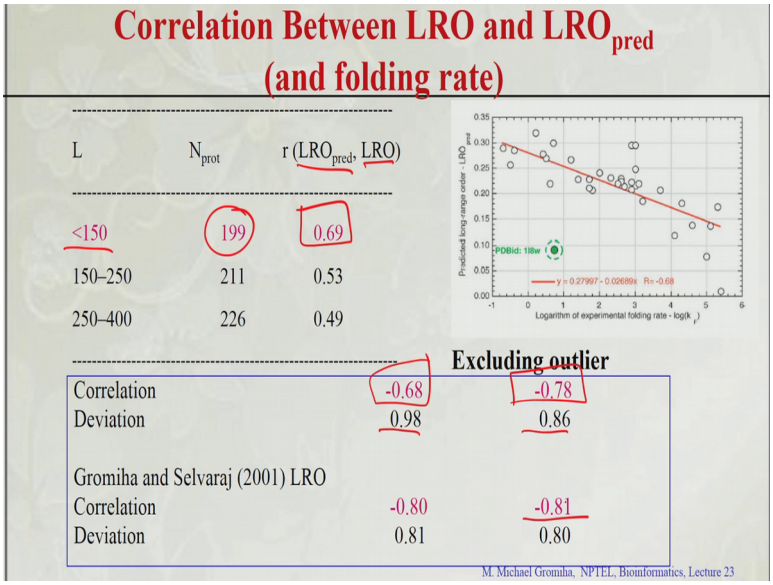
Secondary structure, solvent accessibility and the neighboring residues and the central residue information, conservation score. So, you take all the information finally, they give you the output if you see here if you take this is 8 angstrom distance the state means 8 angstrom and these are the residue 67 and 99, right 67 and 99 are in contact right with the probability of this is the probability of 0.95.

So, they give the probabilities I can take the contacts with high probably and if you get these contacts you can convert this information to get this long range order ;for example, if you see 67- 99, what is the distance operation.

Student: 32.

32, right. So, you see i minus j 32, if we see this one right if you see this one; i minus j equal to 8. So, you have all the contacts. Now the next step is once you download the data, you can get the i minus j and compute how many contacts are within the limit of 12 residues is very easy to do that now the question is whether these contacts right are at least the residues with the separation of twelve residues right are related with the known long range order. If you have known the structures we can get the long range order likewise they can do that.

(Refer Slide Time: 21:17)



So, they read that for example, if they predicted the LRO and this is the LRO obtained directly from the structure and see the correlation.

So, specifically if you see the small proteins right that they concerned 199 proteins and the correlation is not bad its about 0.7. In this case if you predict the contacts and the predicted contacts to help to get this long range order, but with the correlation about 0.7; now they converted this information into predicting the folding rate, right. So, what they did. So, they use all the predicted LRO and use the same equation and to predict the folding rate. So, here the various set of proteins, we give the folding rate right.

. So, you get the correlation is 0.68 minus 0.68, the deviation of 0.98, this is out layer right this outlier mainly because it contains disulfide bonds. So, it is not able to take into account. So, now, if you exclude this outlier then the here the correlation is minus 0.78; the deviation of 0.86 when we compare the data with the experimentally calculated the

LRO; that means, using known 3D structures, if that also here is comparable it is 78, this is 0.81; that means, if you know the sequence and if you do not know structure indirectly you can calculate the long range order or eventually you can get the folding rate from this long range order right which can be predicted from the contacts.

So, this is how they can do that. So, if we summarize right, what are the various aspects we discussed today.

Student: Protein folding rate.

Protein folding rate right what is a protein folding rates.

It is a measure of how fast a slow if protein can fold from its amino acid sequence.

To its negative 3 dimensional structure. So, what is the order of magnitude of this folding rates?

Student: Slow or fast.

Slow to slow to fast may be the microseconds to one hour. So, it also varies up to 8 orders of magnitude. So, we compare with the all alpha proteins and all beta proteins right which are fast folders.

Student: all alpha.

Right all alpha is proteins fold faster than all beta proteins right if we can compare this information with the contacts. If you compare the all alpha proteins and the all beta proteins which one have more number of long range contacts.

Student: All beta.

All beta proteins can long range contacts and also these contacts influence with which residues hydrophobic residues in the case of a fast folders you can see mainly the medium and short range interactions and some of the polar residues are dominant. So, there is related with the folding rate right, then we discussed about different topological parameters based on the 3D structures and contacts right what are the different parameters we discussed? contact order?

Student: The long range long

Long range order.

Student: Multiple contact index.

And the multiple contact index. So, if you see these indices all the indices right they showed good relation with the folding rates.

Right about 0.7 to 0.8 right, inverse relationship between any of these indices like contact order or long range order or multiple contact index with the folding rates this will tell you the topological parameters are important for determining the folding rates right. Then we discuss whether the sequences or any properties can relates these folding rates right and we compare the sequence base information secondary structure and accessibility right as well as the topological parameter right, we found the difference among all these features the topological parameters could explain up to 0.8, but in the case of the other parameters like the properties or secondary structure or accessibility like up to 0.5.

Then we discussed about whether it is possible where you can predict the folding rate from sequence right is it possible to predict the folding rate from sequence there are various methods available right either you can a directly use the sequence or you can indirectly use the sequence to predict the folding rate how we directly predict the folding rate from sequence.

Student: Physicochemical property.

So, we can use physical chemical properties, right because we derived various parameters right and then we discussed in the earlier classes. So, we get the average values about these physio chemical properties and we can combine together at using linear regression, we can predict the folding rates right indirectly; how can you predict.

Student: We will predict the contact

Now, predict the contact first contact information we convert it to the indices right. For example the contact order or a long range order and use that information to predict the folding rate. So, that is fine. So, we have a standard equations we can predict the folding rates fine. So, we discussed the applications of sequential structure based parameters and various aspects like structure prediction folding stability and the folding rates right.

In the subsequent classes we will see how these parameters can also be used to identify the binding sites in complexes for example, protein-protein or protein-RNA or protein-DNA or the protein ligand complexes and whether we are able to understand the recognition mechanism of how this complex is formed and how they maintain this binding affinity and how these parameters are useful in structure based drug design for example, to identifying the lead compounds or potential inhibitors right for any specific targets addressing any specific diseases. Then we will discuss about some of the steps we have to follow to develop any of these applications right for your projects or anything right that we will discuss in the next few classes.

Thank you for your kind attention.