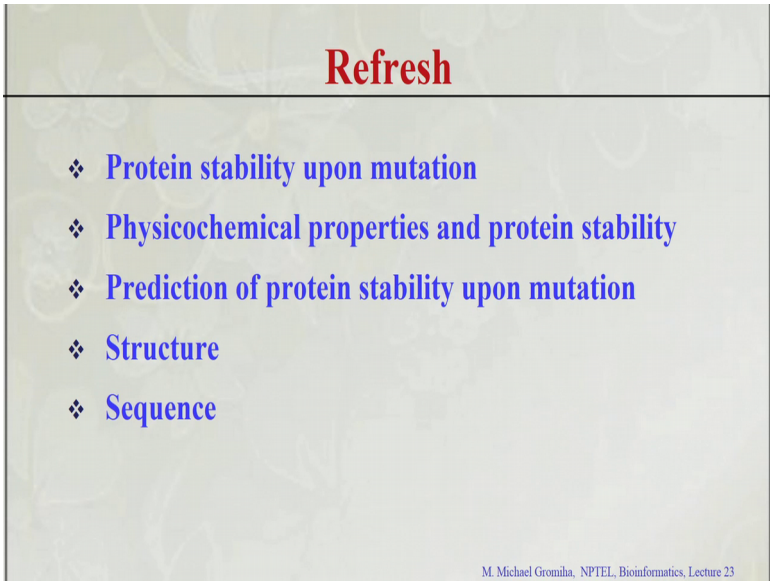
Bioinformatics
Dr. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 23a
Protein Folding Rate I

In this lecture, we will discuss about how a protein folds from its sequence to 3 D structures and how long it will take to go from the sequence to the structure; that is called the protein folding rate.

(Refer Slide Time: 00:31)



In the last; we discussed about protein stability specifically on stability upon mutation right; what did we discuss in the last class.

And we will discuss about database which database deals with the thermodynamic stability.

Student: ProTherm.

ProTherm right? So, using the protherm database or we try to analyze the important factors which influence the stability of the mutants at different locations right we related the physicochemical properties on one side and the experimental stability on another side and we try to relate using correlation coefficient.

So, we using correlation coefficient. So, we identified their some properties right which are important for stability. For example, in the burried mutation which properties a highly correlating right the hydrophobicity right hydrophobicity has high positive correlation with stability, but if the mutation is at the surface right then what did we observe.

Student: Reverse inverse.

Inverse relationship right there is called the inverse hydrophobic effect; that means, the change in increase in hydrophobicity you will decrease the stability.

Then we try to predict the stability upon mutation right. So, we discussed about various methods. So, the one is the average assignment method right; how the average assignment method works?

Then we will be we get the values for all the mutants how many combinations how many mutations;

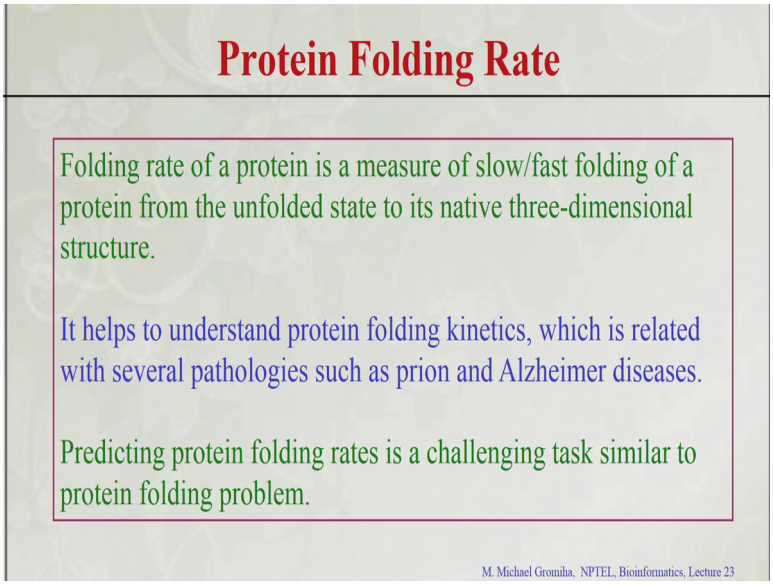
Student: 380.

380 mutations, right, then for new mutation you assign the value from the table let us say kind of look up tables right, then we can evaluate the sensitivity, specificity, accuracy right and correlation so on

So, then we discuss about the inclusion of the sequence information or the structure information right with the different neighboring residues or surrounding residues right and then we will discuss about the potentials right 2 types of potentials right one is distance potential and torsion potentials and we combined these 2 potentials to predict the stability change these are all from the structures if you predict with the sequence, right; we derive several rules right from the known information regarding the mutant residue mutated residue right as well as the residues which are neighboring the sequence we derived set of rules, then we use the rules for predicted stability whether this is a stabilizing or destabilizing.

So, we discussed about the structure prediction right from the unfolded state to the folded state, right.

(Refer Slide Time: 02:59)



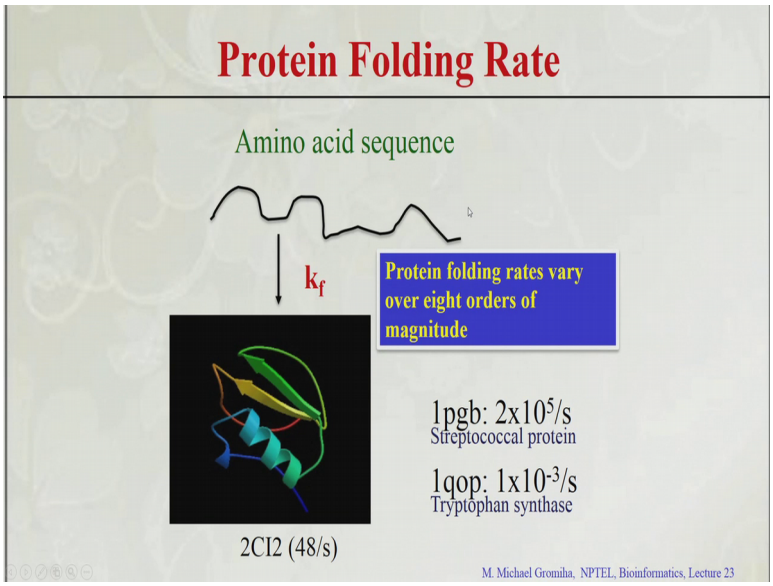
And then we have discussed about a stability today we will deal with the folding rate. So, what is a folding rate of a protein means right it is the measure of slow or fast folding of a protein from its unfolded state right the random coil conformation to its the native 3 dimension structure right how long you protein will take to fold in your native 3D structures.

Why it is important because this folding rate helps to understand the kinetics which is related with several pathologies like the prion diseases, Alzheimer diseases and so on. So, it is very important to understand the folding rate of any particular mutants as well as for the mutants right you can see for proteins as well as what will happen if you mutate a specific residue and the predicting the folding rate right; it is also a challenging task similar to protein folding problem, then what is protein folding problem.

Student: Final protein folding (Refer Time: 03:48).

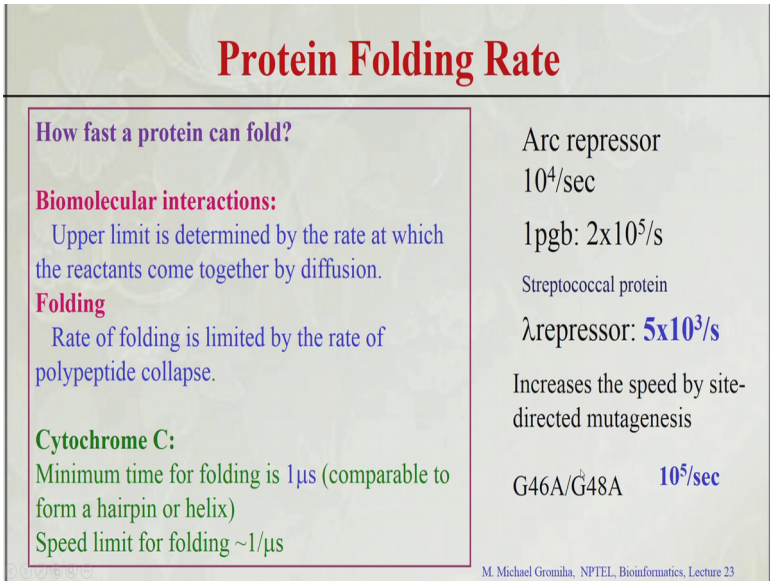
Right you can decipher in a native conformation of protein from its sequence like still it is a challenging. So, next folding rate is also important to see how a protein can fold from the unfolded state to the folded state regarding the time.

(Refer Slide Time: 04:02)



So, next one example here this is a amino acid sequence may be it takes time to folder specific 3D structures. So, these folding rate vary over eight orders of magnitude right you can you can see these folds from microseconds to an hour. So, you get 2 examples one is a streptococcal protein here the rate is 2 into 10 to power 5 per second. This is very fast folding and if you see a tryptophan synthase,t its very slow folding it takes as 1; 1 into 10 to the power minus 3 per second.

(Refer Slide Time: 40:30)



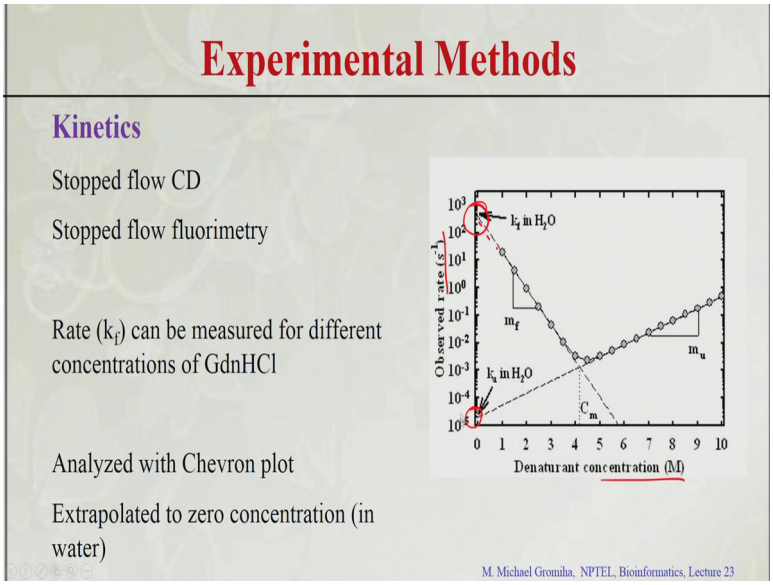
So, now the question is how fast your protein can fold. So, I will discuss would be 3 different types of examples; for example, if you take the biomolecular interactions therefore, 2 different to proteins or the protein and DNA. So, whatever they want to make a complex, you can see the upper limit now determine by rate at which the reactants what are the molecules which are involved in the reaction they come close together by diffusion.

So, how much the rate they take then if you talk about the folding here you can see the rate of folding is the rate the polypeptide collapse right to form these structures. So, from this you can see if it one example say you say cytochrome c right, here if the minimum time for folding is about one microsecond this is comparable to the hairpin loop or the helix formation; so, is about one microsecond.

In this case, you can see the rate right for the folding is about one per microsecond. Now I show few examples like arc repressor is 10 to the power four per second and lambda repressor; it is 5 into 10^3 per second and so on. These are for the proteins likewise you can also accelerate or decelerate the protein or increase or decrease the folding rate right by amino acid substitutions like what we discussed about the stability right when introducing a mutation; we can enhance the stability of proteins right.

So, we are identify the mutants which can enhance the stability and we can design proteins likewise in the folding rate, we can identify a some mutations which can accelerate the folding process, in this case, we will go to the proper folding because if it takes more time then this way end up with this missfolding, right, due to aggregation right and some other aspect. So, you can get some diseases right this is the reason why we need to see the particular protein which folds into a particular a time into stable 3D structures.

(Refer Slide Time: 06:30)

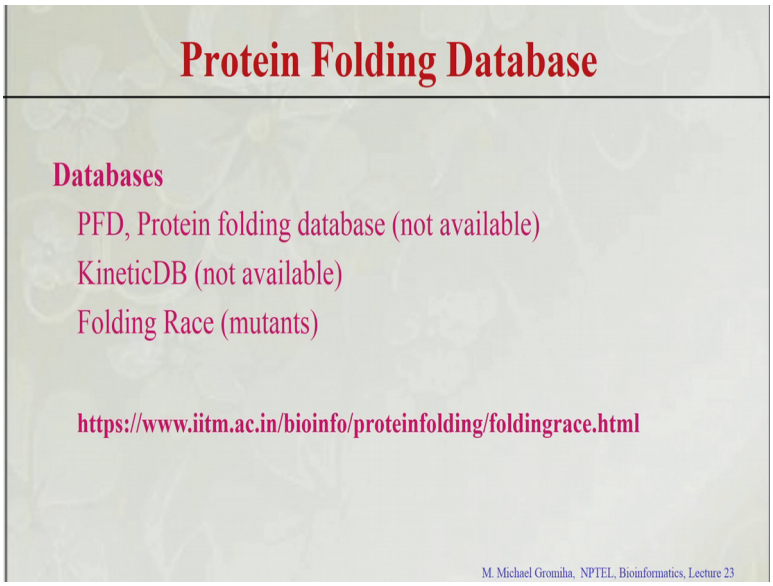


So, how to measure this folding rate; so, there are various measures to get this folding rate mainly the stopped flow CD and the fluorimetry; this will give you the rate right a different concentration for example, here the x axis. We give the different concentration of the guanidine hydrochloride on the y axis. We show the observed rate right. So, you see the concentration one to the different concentrations up to 10 molar and if you see this is the folding region and here this is unfolding region and we extrapolate to 0 to get the folding rate or unfolding rate or 0 concentration right because to get the uniform data for different proteins and mutants, right.

We extrapolated here in the folding region it goes to here. So, here now this is the folding rate in water likewise because here we extrapolate this one right likewise we extrapolate the unfolding a region here. So, finally, you get here this is the unfolding rate right in water, water environment right this can be analyzed using this chevron plot right.

Now, once we have the data, right for doing the new bioinformatics analysis right we require a sufficient number of data and reliable data right. So, it is very important to construct databases right on these aspects, right.

(Refer Slide Time: 07:46)



First they started to develop few databases. So, one is called the protein folding database. So, another one is after few years started they developed kinetic DB, but unfortunately both the databases are not available at the moment.

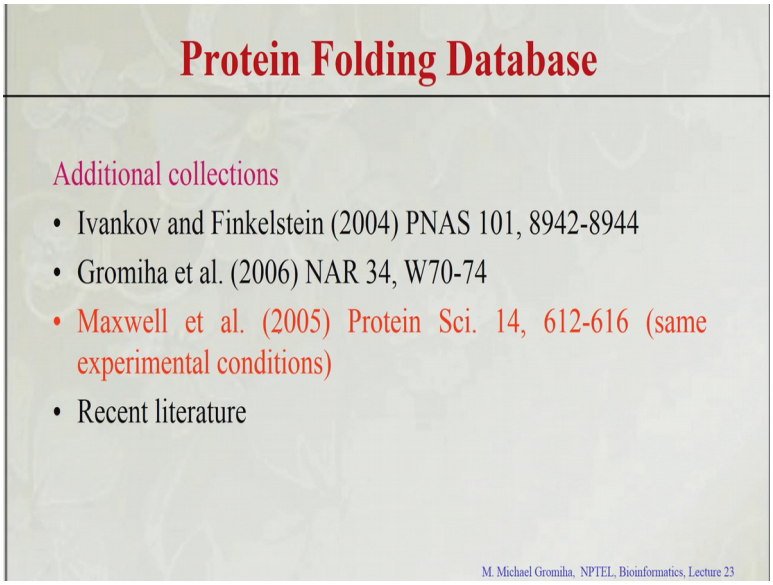
Now, folding data folding database they have tried to establish using a collaboration among different experimental list they assign the different proteins to a different experiment experimental list to get the data same concentrations right and then they compiled all the data and they developed by database.

Now, how can you get the data like the various aspects to get the data either you can get from literature or you can see other databases. For example, the folding race this will deal with the mutants to get the folding rate upon mutation. So, you can use this folding race right this is server come database, right. So, there is available in our website right to get the folding rate upon mutations, right, you can get or download the data and you can use for analysis.

So for the data for the proteins right I listed some of the papers which give the data actual data we can get from this literature. So, in 2005, right, the Maxwell's they found a consortium and they try to see the folding rate and same concentrations and same concentrations, right or same experimental conditions in earlier days we get the data for the different experimental conditions because different groups they use the proteins at different a conditions and they measured the folding rates and the published literature.

So, same groups they assigned these temperature pH and conditions, they ask them to get the folding rate.

(Refer Slide Time: 09:26)



In this case, you get the data for different proteins at same concentration this is how they published this paper in 2005 with the same experimental conditions. Then currently there are several other literatures are reviews right, they discussed about the folding rates or proteins and currently we get more than 100 proteins, we know the folding rates of the proteins.

So, we will have this data, then we can analyze right what factors at what influence the proteins to fold fast or to fold slow right earlier we discussed about the different types of interactions also they are influence with different structure classes what are different structure classes we discussed earlier.

Student: All alpha.

All alpha proteins.

Student: All beta proteins.

Student: Alpha plus beta.

Alpha plus beta proteins.

Student: Alpha beta.

As well as alpha by beta proteins right. So, what is the difference between all alpha or all beta.

Student: Predominantly alpha

Predominantly alpha predominantly beta which interactions are dominant in all alpha and all beta.

Student: Medium; medium short.

Medium and short range interactions right the influence the all alpha proteins and the longer interactions are dominant in the all beta proteins. So, if we look in to these which type of proteins fold fast which type of proteins takes time to fold.

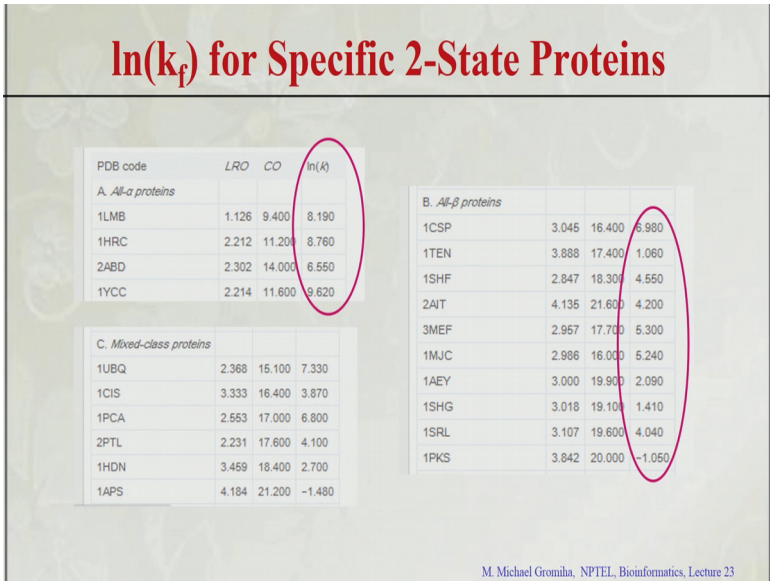
Student: All alpha.

All alpha proteins.

Student: Will fold faster.

Fold fast right because the mainly medium range interaction short range interaction they fold fast. Now I give you the one examples.

(Refer Slide Time: 10:42)

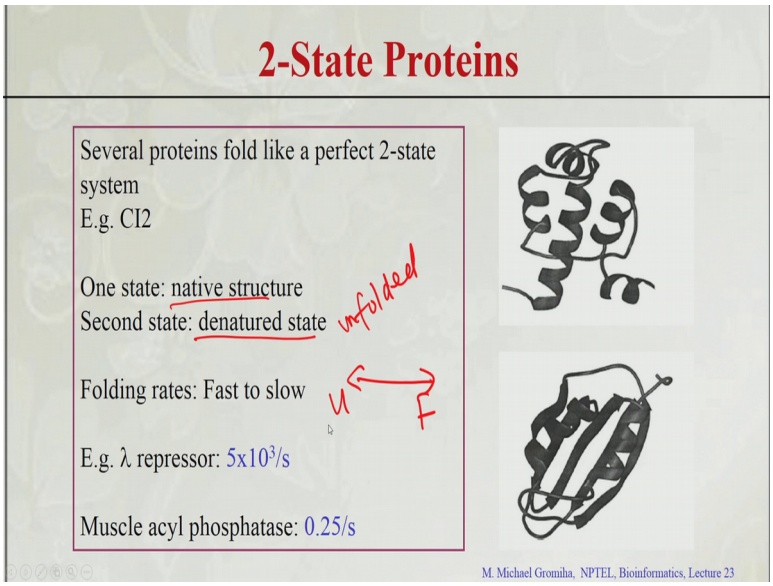


So, this should be example for the all alpha proteins right you can see here and this is for the all beta proteins all these proteins are 2 state proteins and small proteins; what is the meaning of 2 state proteins?

Student: Intermediate have a unstable major conformation

Right, we have 2 states one state the native structure.

(Refer Slide Time: 10:59)



And another state the unfolded state or a denatured states, right. So, you can have the unfolded state to the folded state there is no intermediate.

So, several proteins they fold perfect to state and also; there is small proteins, we know the experimental data for the folding rates. So, they can also have fast folders and slow folders; even if all the proteins are 2 state proteins and these proteins are small about hundred residues. So, now, I show the values for the alpha protein all beta proteins and we look into the all alpha proteins this is the logarithmic of this folding rate the one of ln kf. So, this is very high. So, it can these the rate is 10.6 right per second over logarithmic of these; you take the all beta. So, you have the values 1; 2 as well as the -1 right because that is it takes time.

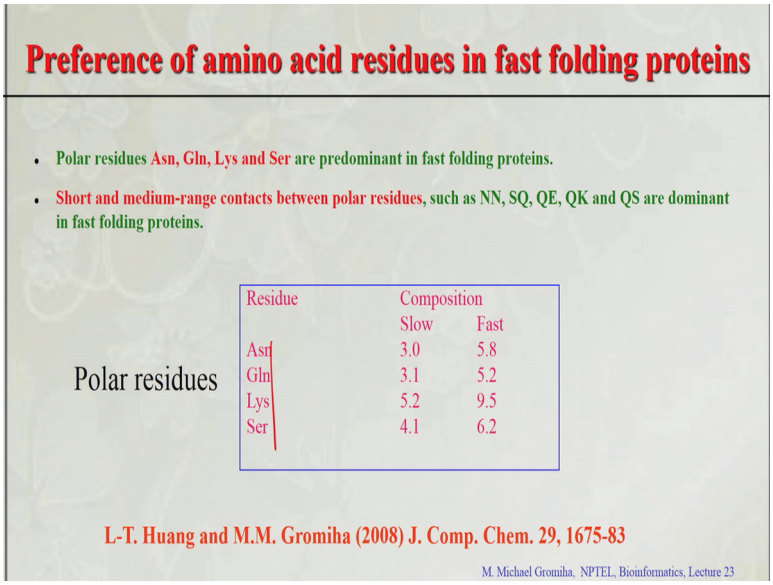
So, if you look into these numbers generally you can say that all alpha proteins fold faster than all beta proteins, but that is start very strict because sometimes even all beta

proteins can also fold fast because the one CSP the cold shock protein; it folds at 6.98 right similar to this one.

But generally you can see that all alpha proteins fold faster than all beta proteins. So, we discussed earlier all alpha proteins are influenced with the short medium interactions right and all beta proteins with the long range interactions which residues influence long range in the contacts and which residues influence medium range contacts right I will tell you. So, if you see the fast folding proteins one; one group and the slow folding groups another group and see the amino acid compositions or the preference or the contacting pattern.

Then you could see some patterns for the slow and fast folding proteins.

(Refer Slide Time: 12:42)



For example if you see the polar residues. So, asparagine, glutamine lysine and serine. These residues, right are predominantly present in the fast folding proteins if you see the difference between slow and fast fold right. This is 5.2 and fast folding, it is 9.5 right, it is about 2 times.

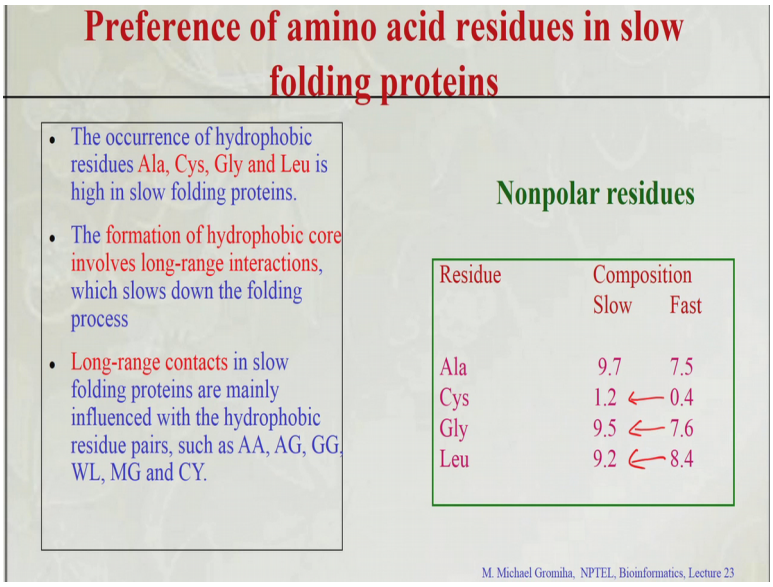
Likewise these residues they are occurring highly and fast fold proteins right compared with the slow fold proteins then also we have see these medium range contacts right, you can see the contact between the polar residues are dominantly fast folding proteins, right, this earlier we discussed about these interactions contacts as well as the composition.

So, that that agrees with the experimental data for the fast folding proteins. Likewise if you see the slow folding proteins what do you expect which residue should be dominant.

Student: Hydrophobic.

Hydrophobic residues right then about the contacts long range conducts made by these residues right.

(Refer Slide Time: 13:38)



This is what we did right with the. So, higher occurrence of alanine, cysteine, glycine and leucine, these residues are predominantly occurring in the slow folding proteins than fast folding proteins for example, there is 3 times, right, here also, you can see the difference right, in the case of slow folding proteins.

Then the formation of this hydrophobic core involves long range interactions right this is how which slows down the folding process then we see the contacts in these slow folding proteins and fast folding proteins you can see the long range contacts and slow; slow folding proteins right, they are mainly influenced with the hydrophobic pairs see alanine, alanine, tryptophan, leucine, alanine, glycine, right, cysteine, tyrosine, right you can see the contacts between the hydrophobic residues are more in the case of the slow folding proteins compared with fast folding proteins.

So, we see the slow folding and fast folding proteins generally all alpha proteins right fold faster than all beta proteins and if you classify into 2 groups, right which one is fast

folding which one is slow folding right and see the preference of residues right the mainly the polar residues in the fast folding and the non polar residues in the slow folding proteins likewise the interactions or the contacts right; mainly the slow folding proteins are dominated with the long range contacts using these hydrophobic residues.

Now, the question is whether we are able to predict or relate these folding rates with any of the parameters right we discussed about various parameters obtained from sequence or structure what are the various structure based parameters we discussed earlier.

Student: structure; sequence based.

Structure based parameters.

Student: Accessibility.

Accessibility.

Student: Contact orders.

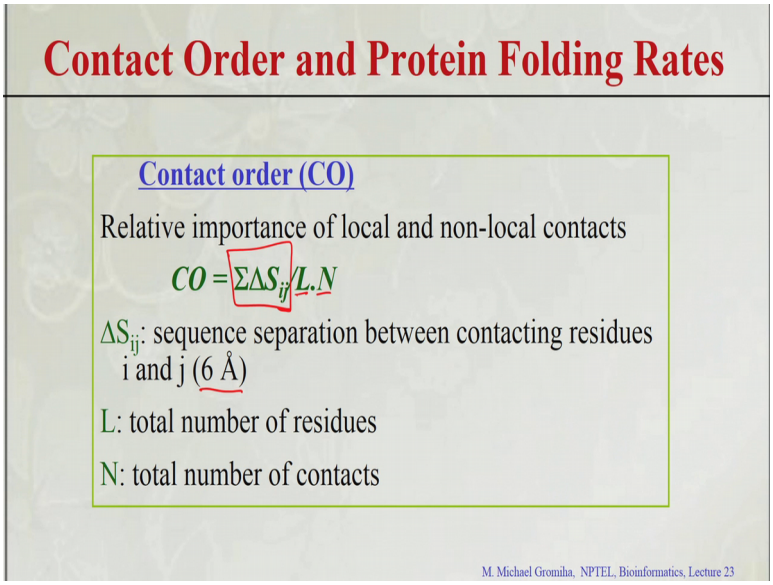
Conduct order long range order hydrophobicity so, various parameters right. So, also if you see the slow and fast folding proteins the contacts are very important. So, we convert the contacts right the 3 d structures you convert into contacts using in the form of;

Student: maps

Contact maps right look at the contact maps and from these contacts you deduce various parameters like contact order long range order and so on.

So, now we will see; how these contacts it can be used for predicting or understanding protein folding rates whether there is a relationship between the contacts and the folding rates right basically yes right because if you see the more number of long range contacts that will slows from the process, right, we will see it.

(Refer Slide Time: 16:10)



So, what is the contact order we discussed earlier how to define the contact order; this is given as delta sigma S_{ij} by L and n; what is delta S_{ij}.

Student: Sequence separation.

Sequence separation between the contacting residues. So, we can set this sphere of radius 6 angstrom and look at the contacts, right, then the change check the distance separation and that will give delta S; what is L?

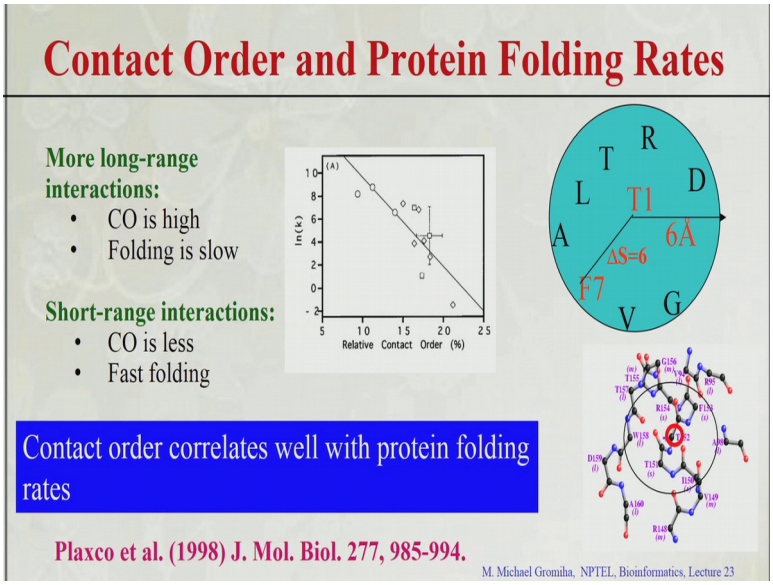
Student: total no of residues.

A total number of residues there is the length of this protein and the N.

Student: Number of contacts.

Its total number of contacts right. So, if we do it for each residue to get the delta S_{ij} add up together and then normalize with the length and the number of contacts right.

(Refer Slide Time: 16:48)



So, here I show the data for example, if this is sphere one and seven delta s is six likewise if all the residues you compute right get for the t one then you repeat it for all the residues in the protein right, then we will get the value. So, then we relate that value with the folding rate if there are more long range interactions, what will happen to the contact order. So, more long range interactions means what will happen in this equation.

Student: Difference is high.

This will be high because delta s will be high if this is high, then what will happen the contact order?

That is high because that will be proportional right. So, in this case we get the contact order is high if the contact order is high what do you expect the folding.

Folding is slow because more number of long range contact then the there is folding is slow for example, they are more short range interactions or more medium range interactions then compared with the long range contacts. So, you can see contact order is less because delta S_{ij} is less right, then in this case, it can fold faster.

Now, I show these figure, why x axis is the contact order and y axis is the logarithmic of the folding rate you can see the inverse relationship between the contact order and the folding rate so; that means, what we assumption what we made that is correct if you have

more contact order right; that is less folding rate if it is less contact order is folding rate is it folds fast right fine.

So, the correlation is also about to 0.7 to 0.8 that you can see inverse relationship between the contact order and the folding rate. Now the second parameters for the next parameter we discuss.

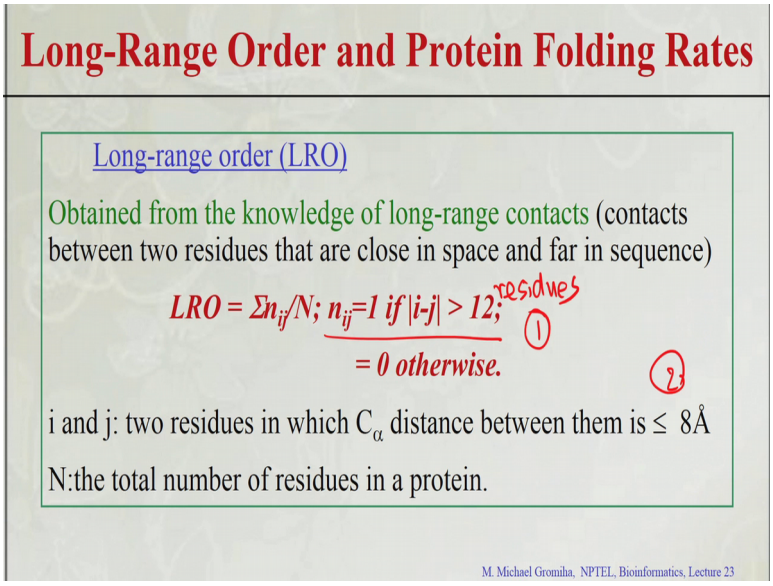
Student: Long range order.

Long range order; right. So, what how long range order is derived; what is concept used for right so; that means, the contacts.

Student: (Refer Time: 18:26).

Which are close in space, but they are far in the sequence right. So, in this case you see all the contacts which are close in space right and far in sequence.

(Refer Slide Time: 18:36)



In this case, nij equal to one if the separation is 12 residues.

So, here you can see there are how many adjustable parameters in long range order.

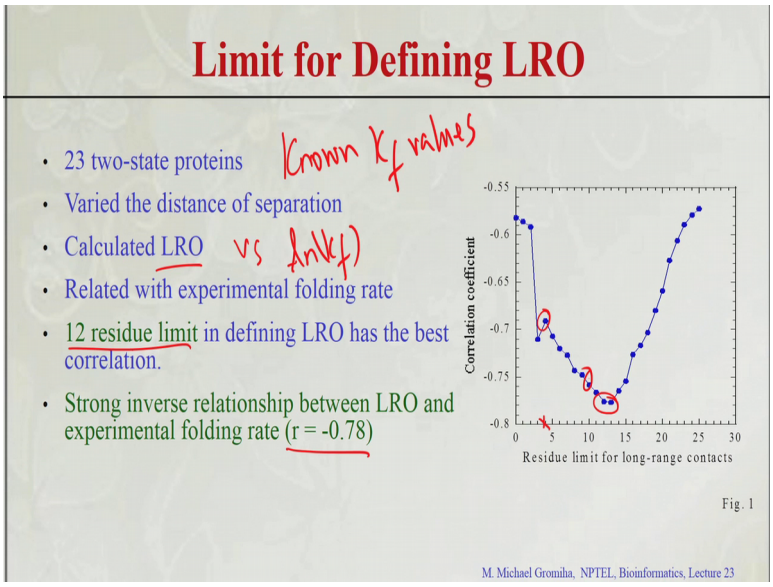
Student: Distance cutoff.

Right one is a distance cut off.

Student: Sequence; sequence separation.

This one and then this one you can say distance is eight angstrom and the sequence of percentage, right that is 12 residues. So, you have 2 adjustable parameters you can adjust these parameters to get the long range order and how far this will relate the folding rates this is how we get the numbers 12, right earlier we use number 12 right. Now I will explain; how we get this number 12.

(Refer Slide Time: 19:23)



So, we take a set of proteins with non folding rates experimental folding rates right. So, I can see the set of 2 state proteins with non folding rates. So, in this case now we can change the distance of separation we can change from 1 to 25 residues. So, how many residues which are within this limit one residue or 2 residue 3 residues, four residues so on.

Then from that we can calculate the LRO because nij; you can put i minus j that we change right 1 to 25, right, for each case, we will get the value. So, then we relate that LRO with the folding rate and look into calculate the correlation for each distance separation we will calculate the LRO and you have to get this correlation there is LRO versus l and kf.

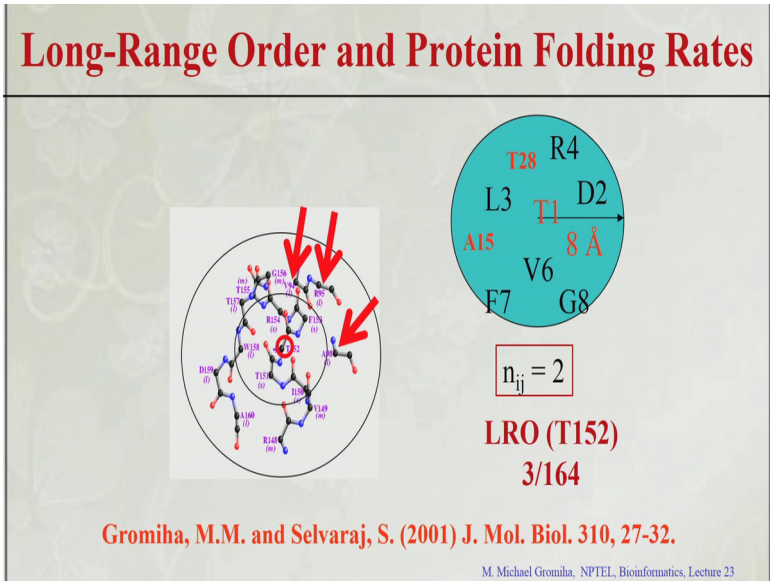
For example if it is 5, right we get the correlation of minus 0.69 approximately then you go with the 10. So, you can see the correlation is negative correlation (Refer Time:

20:27) increases with respect to the distance separation of 2 here 12 or 13; almost the same.

After that if you increase more then again will decreasing. So, finally, it go up to 25 now get goes from the minus 0.58. This is similar to what we get from the residues separation 2 or 3 residues; that means, there are some distance contacts which are influential for the folding rates right the contact order considers both short medium and long range contacts, but in the case of long range order we consider only the long range contacts because there is one which shows the highest performance.

If add the long range short range contact then the correlation is less; that means, there is something to do with only for the long range long range contacts right. So, from here, we get this value 12. This is how we get the 12 residue limit and with these 12 residue limit we can see the correlation of minus 0.78 with the experimental folding rate.

(Refer Slide Time: 21:26)

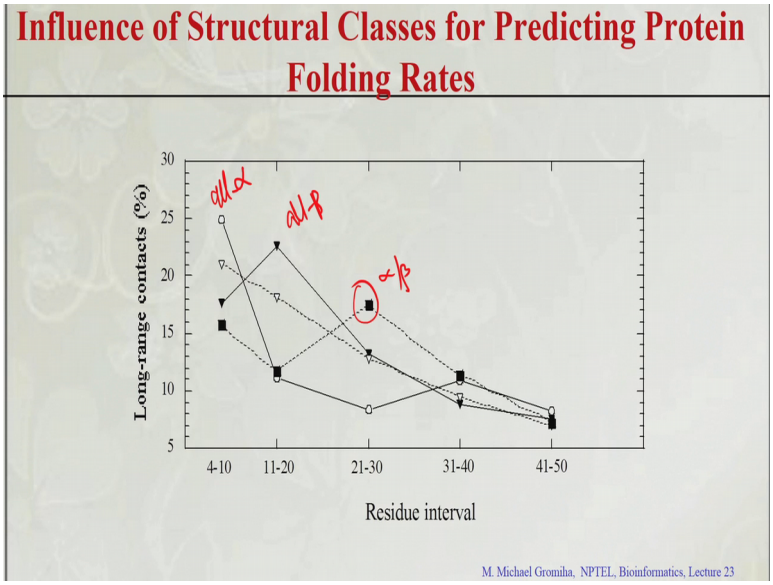


How to calculate the long range order? So, here we give the sphere of a eight angstrom look at the residues which are occurring within the limit and see the residues right which are having the distance of at least 12 residues.

So, if we take the Threonine 152, I discussed last class right how many long range contact this can make;

Right, 1, 2, 3, right, in this case; 3 by 164; you can get the number right for this long range order, right. So, so do it for all the residues and finally, we divided by n right this is the this; this is how you will get this for we do the summation of n ij for each of these residues sum up together and then we get this a value right you normalized with the n you will get LRO.

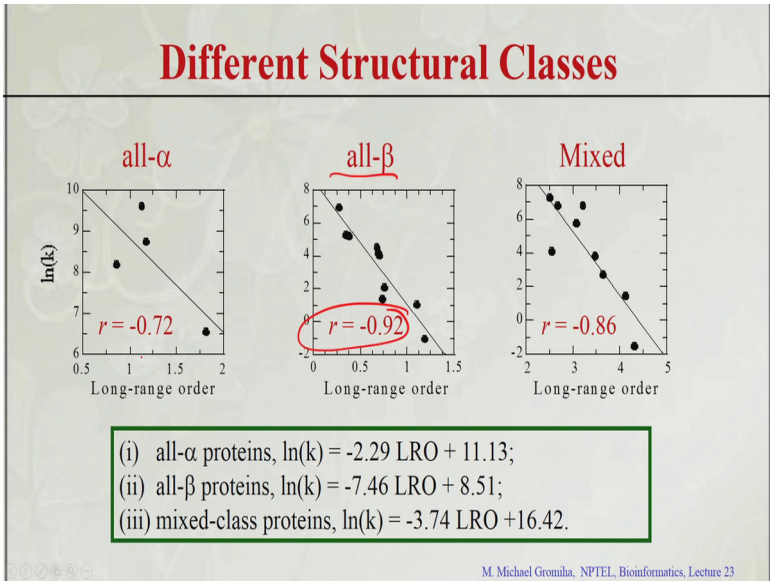
(Refer Slide Time: 22:14)



So, now I show the long range contacts, right at the different structure classes, right as I discussed earlier four different structure classes all alpha, all beta, alpha plus beta, and alpha beta. So, here in the case of the all alpha, this is mainly here in the 4 to 10 range and the all beta, so, it is mainly in the 11 to 20 range and the alpha plus beta right you can see here right in this between the all alpha and all beta and alpha beta right; this is the range.

So, this can be reflected in the long range order because you can see the contacts are different in different lecture classes.

(Refer Slide Time: 22:59)



So, we classified the proteins at 3 groups all alpha, all beta and mixed class and we derived the equations for these 3 classes and see whether it is improvement in the correlation if you do like this mainly long range order deals with which type of interactions;

Long range interactions. So, in this case, which type of which class of proteins perform better with long range order?

Student: All beta protein.

All beta proteins right because influence with long range interactions. So, I show the data. So, here if you see all beta. So, you can see the correlation is very high -0.92 compared to the all alpha and mixed class proteins right this make sense.

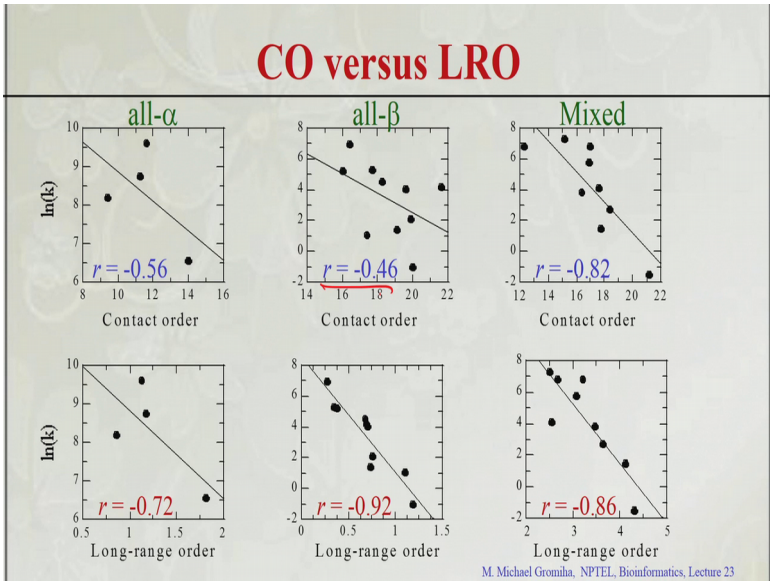
So, if you see the long range order which we involves mainly long range interactions right. So, you see all beta proteins right you can a get better correlation then the all alpha and the mixed class proteins. So, we derive the equation. So, this is the equation for getting the ln kf for all alpa proteins all beta proteins and the mixed class proteins

Here why is the ln k right and the x axis, we can see the only the long range order. Now we compare with the contact order long range order, right, some classes you can get the comparable results, some classes this long range order will perform better than contact order which class LRO performs better.

Student: Beta.

All beta right. So, we see the same all beta as we see this very less.

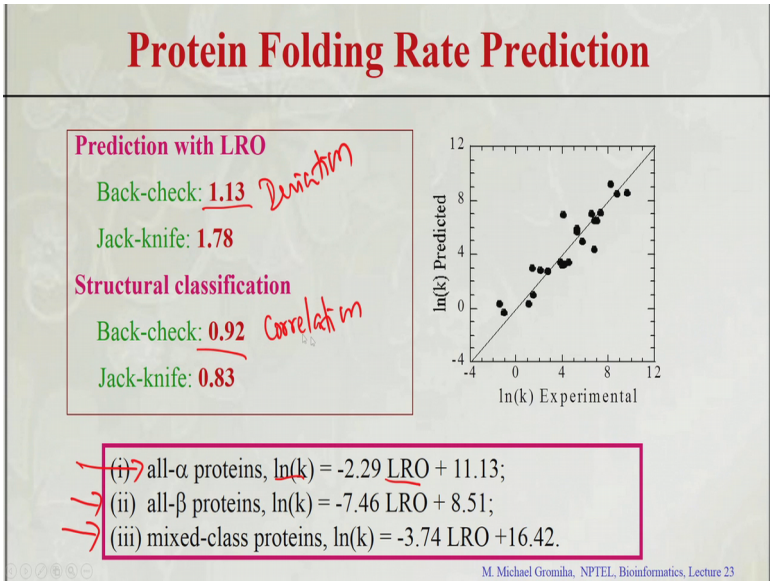
(Refer Slide Time: 24:18)



It is because they include the short and medium range contacts; that means, short and medium range contacts include some noise in calculating the folding rate this is the reason why if you including these medium range contacts the correlation is only 0.46, but if we exclude this we will get up to 0.92. So, in the case of mixed class proteins, you can see similar level 0.82 and 0.86 right the number is almost similar in both contact order as well as for the long range order. So, now, we can predict the folding rate right how to predict the folding rate.

No because we have the relationship between LRO and the ln kf.

(Refer Slide Time: 25:03)



So, we have the equation y equal to mx plus c straight line equation. So, this is a equation is all alpha all beta and mixed class proteins. So, we calculate for any all alpha proteins we calculate LRO and substitute the values in this equation right and this you can get the ln kf. So, we did this with the LRO right. So, you will get the deviation of just 1.13 rate per second.

Then we classify the proteins in 3 different groups we derived four equations, one equation by considering all the proteins together, right only one common equation for all the all the proteins in this case the deviation is one point one 3 then what we did. So, we classify the proteins in 3 different groups right mainly all alpha all beta and mixed class proteins and we derived the LRO from these equations and calculate the folding rates.

If you do like this then you can see this is the correlation the correlation is 0.92 with the different structure classes. This is a deviation here I give the correlation right. So, if you do with the classification you can see the correlation of 0.92 for the different structure classes alpha and beta with your class.

So, here is with this figure right I show the comparison between experimental and the predicted ln kf. So, you can see there is a good correlation between the experimental and the predicted folding rates right most of them are very close to these straight line right this slope.

So, we discussed about the contact order and the long range order contact order considers the distance of six angstrom and take all the distances long range order, we have the one more conditions based on the distance separation.