

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

Lecture – 22b
Stability of Proteins Upon Mutations II

(Refer Slide Time: 00:16)

Prediction of Protein Mutant Stability

① Multiple regression technique (Protein Eng. 12, 549, 1999)

Knowledge based prediction (CBAC 30, 408, 2006)

Classification and regression tool (BPC 125, 462, 2007)

Average assignment method (Biopolymers 82, 80, 2006)

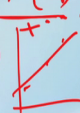
Residue pair potentials (Proteins 66, 41, 2007)

Sequence based prediction (Bioinformatics, 23, 1292, 2007)

ΔP $\Delta \Delta G$

$m \rightarrow$

$\Delta \Delta G = c + m(\Delta P)$



M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

If you have the relationship for example, here if you see the proper delta P it is related with the delta delta G you can relate delta delta G equal to some constant plus m * delta P. So, delta P is a property value and c is a constant because if you make a straight line if you see in the line is like this. You can use this equation to calculate delta delta G and then you can see whether you can predict.

See if you have some set of values and you frame the equation and then if you have a new mutation you know the c and m and subsequent value delta P you can calculate delta delta H you can predict stability. If you have this good correlation for example, if you single property cannot account the stability of this mutant what can be had, what the other options.

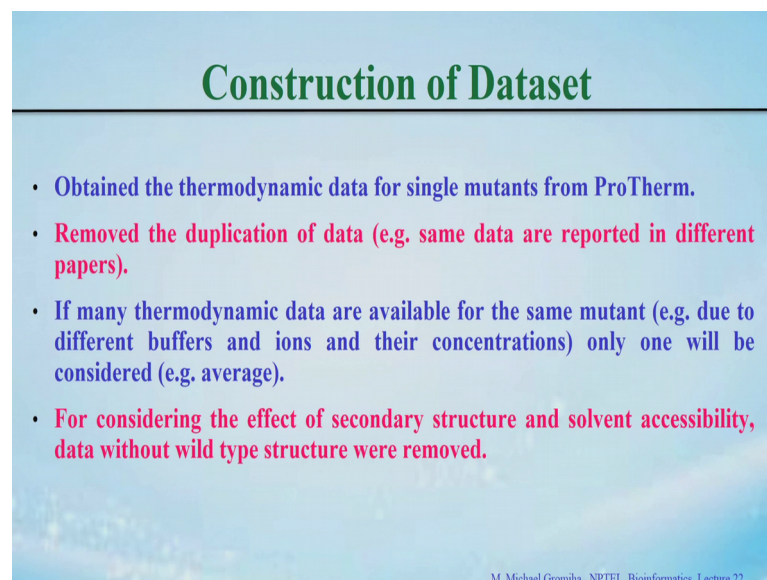
You can has many properties we discussed about 14 properties, then physical chemical energetic and conformational. You can use different properties based on multiple regression technique because extend this a variables. We try to fit this equations with known values is simplification of this squares and we get the constant values for

example, m_1 m_2 and so on we can add this. Once the constant values are known the coefficients are known then for any a mutant we calculate the properties and use this equation to predict the stability. This is based on multiple a regression technique.

Then also there different other techniques such as knowledge based prediction because you can obtain the information from the mutation and we can use the information for prediction. Then we can use a classification regression tool, the average assignment method, pair potentials, as well as sequence based prediction, is corraling the literature several methods are available for predict the stability change of upon mutation.

So, we will discuss the free methods mainly the average assignment methods, pair potentials as well as how to predict from the sequence. So, if you search that you ProTherm data you will get plenty of data. So, if you look into this data several, you can see several data you can see the repetition for example, if you see the alanine to valine for any particular protein. You can see this data can be obtained at different temperature or different pH or different buffers and different concentration. This case we can have different data for the same mutations.

(Refer Slide Time: 02:53)



Construction of Dataset

- Obtained the thermodynamic data for single mutants from ProTherm.
- Removed the duplication of data (e.g. same data are reported in different papers).
- If many thermodynamic data are available for the same mutant (e.g. due to different buffers and ions and their concentrations) only one will be considered (e.g. average).
- For considering the effect of secondary structure and solvent accessibility, data without wild type structure were removed.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

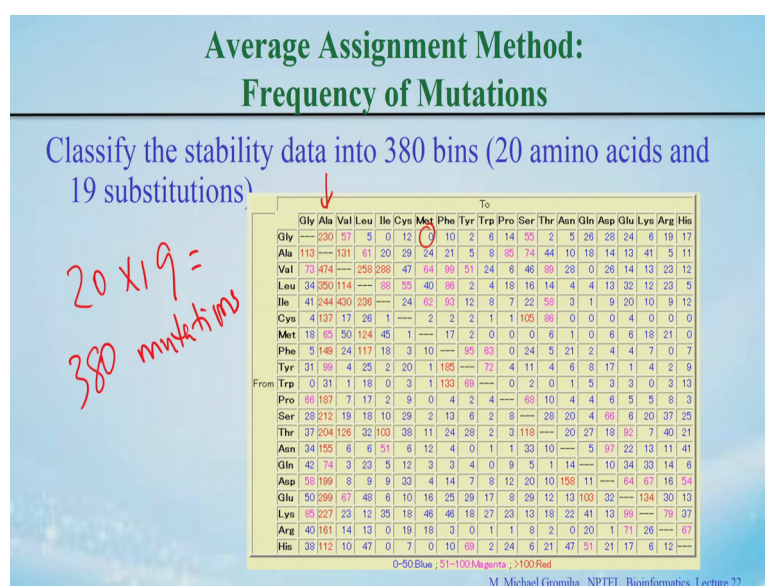
You have to check carefully and first we have to remove the duplication of data. It is also possible that same data could be reported from different papers because research is open everyone. So, you not necessarily that only one group should do the particular type of research because very competitive, so many groups at the same time work on same

proteins. In this case when they get the data they will publish with their information they will publish.

So, it is possible you have the same data from different groups in order to accumulate the data. So, in the ProTherm consider as all the papers published in literature because we are not sure which one we need to take. So, in this case we have to remove the duplication in different ways whereas unique you can take the data as it is. If it is duplicated you can check the data either you read the papers and see the method which method they use or see the average values. If the data are very close for example, say the deviation of less than 1 kilo cal per mole, in this case you can take the average for the particular mutant. If it is very deviated you can remove the outliers and get this average values and you can use that value is per constructing the model and you can also try to predict the extreme stability on what conditions, why it is extreme stability here also you can try to analyze.

So, now at then you can also consider secondary structure as well as solvent accessibility you can see this effect. If you want to do this, if the structure is not available then we can either you can predict or you can discard the data. So, you can both options one we can predict the secondary structure and ASA and you can use for prediction or if you do not have, if you need only the structure information then if it is not there then you can discard this data. So, finally, we can develop the data set.

(Refer Slide Time: 04:49)



So, if you have the data set we have all the data then we can construct the frequency of mutations. This data I show for all the mutations when you construct your data set, we are removing all the duplicate data and if you take only the unique data, so you will get the similar type of matrix with your data set. What is about the, how many data in this matrix? Usually we get.

Student: 20.

20 into.

Student: 19.

19.

Student: 380.

This equal 380 mutations. As I have discussed earlier, all the mutation not uniform some cases you have more number of data some cases you have less number of data depending upon this, the mutant type.

Mainly alanine mutations we have more number of data and other mutations for example, glycine to methionine is 0, so many cases we have less number of data. The average assignment method works based on number of data. So, if you have more number of data then we will get better results if it has less number of data in that case we don't know about the performance of that method for that particular mutant.

So, here you show you the frequency of stabilizing and destabilizing mutations and we use a particular set of conditions. For example, pH. If it is 7 we can take 5 to 9 and particular temperatures.

(Refer Slide Time: 06:03)

Frequency of destabilizing (and stabilizing) mutants upon denaturant denaturation ($\Delta\Delta G^{DSO}$)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0(0)	3(1)	1(0)	0(1)	4(0)	25(4)	1(0)	1(1)	2(0)	0(0)	1(0)	1(0)	1(3)	1(0)	1(0)	2(0)	3(1)	16(2)	1(0)	1(0)
C	4(2)	0(0)	1(0)	0(0)	0(0)	0(0)	0(0)	0(1)	0(0)	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)	5(2)	0(1)	0(0)	0(0)	0(0)
D	15(9)	0(1)	0(0)	1(0)	4(0)	10(1)	1(0)	0(0)	2(2)	0(0)	0(0)	7(2)	2(1)	0(0)	0(0)	2(0)	0(0)	0(0)	0(0)	0(0)
E	28(9)	1(1)	3(5)	0(0)	9(1)	20(5)	1(1)	1(1)	3(5)	2(2)	0(1)	2(1)	1(1)	12(4)	1(1)	3(1)	2(1)	3(1)	1(0)	1(1)
F	25(2)	0(0)	0(1)	0(0)	0(0)	3(0)	0(1)	2(2)	0(0)	5(0)	0(0)	0(1)	0(0)	0(0)	0(0)	2(0)	1(0)	3(0)	3(2)	2(4)
G	21(13)	2(1)	2(1)	0(1)	4(0)	0(0)	2(0)	0(0)	2(0)	2(0)	0(0)	2(0)	1(0)	1(0)	1(0)	2(2)	1(0)	12(0)	1(0)	1(0)
H	9(12)	0(0)	1(1)	0(1)	0(2)	8(1)	0(0)	0(0)	2(0)	0(0)	0(0)	2(2)	0(1)	6(0)	1(0)	1(0)	1(0)	1(0)	0(0)	2(2)
I	37(0)	2(1)	2(0)	0(0)	1(0)	5(1)	0(0)	0(0)	0(0)	8(0)	6(0)	0(0)	0(0)	0(0)	0(0)	1(0)	8(0)	28(2)	1(0)	0(0)
K	25(17)	3(0)	0(1)	1(5)	16(1)	26(6)	0(0)	1(0)	0(0)	1(0)	4(5)	1(1)	0(0)	0(1)	4(2)	0(1)	1(0)	1(0)	3(0)	1(0)
L	45(4)	2(0)	0(0)	1(0)	0(1)	13(0)	0(0)	12(0)	1(0)	0(0)	0(0)	0(0)	1(0)	1(0)	1(2)	1(1)	2(0)	14(0)	0(1)	0(0)
M	12(1)	0(0)	0(0)	0(1)	1(0)	4(0)	0(0)	1(0)	0(0)	2(0)	0(0)	0(0)	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
N	15(2)	0(0)	4(2)	0(0)	2(0)	5(5)	1(0)	0(2)	0(0)	0(1)	0(1)	0(0)	0(0)	1(0)	2(0)	2(0)	0(1)	0(1)	0(0)	0(0)
P	20(6)	0(0)	0(0)	0(0)	2(1)	9(1)	0(0)	1(0)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)	4(0)	1(1)	2(1)	0(0)	0(0)
Q	10(5)	0(0)	0(1)	0(1)	3(0)	10(3)	1(0)	0(1)	0(0)	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(1)	0(0)	0(0)	0(0)	0(0)
R	21(5)	1(0)	0(0)	1(0)	1(0)	8(5)	2(0)	0(0)	2(0)	0(2)	1(0)	0(0)	0(0)	2(0)	0(0)	1(0)	1(0)	0(0)	0(1)	0(0)
S	13(8)	0(1)	2(5)	0(1)	2(0)	8(0)	1(0)	1(0)	0(1)	2(1)	0(0)	2(0)	1(0)	1(0)	1(0)	0(0)	3(1)	2(1)	1(0)	1(0)
T	21(5)	7(1)	2(0)	2(0)	0(0)	12(0)	0(2)	6(5)	0(0)	0(0)	0(0)	0(0)	2(0)	1(0)	1(2)	14(1)	0(0)	11(1)	0(1)	0(0)
V	44(0)	4(2)	0(0)	0(0)	2(0)	11(0)	1(0)	8(0)	0(0)	10(5)	1(0)	1(0)	0(1)	0(0)	1(0)	9(1)	16(2)	0(0)	1(1)	1(0)
W	10(0)	0(0)	0(1)	0(0)	3(2)	0(0)	0(0)	0(0)	0(0)	1(0)	0(0)	0(0)	0(0)	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)	1(0)
Y	22(1)	1(0)	3(1)	1(0)	19(5)	9(0)	1(0)	1(0)	1(0)	7(1)	1(0)	2(0)	2(0)	2(0)	1(0)	3(5)	4(0)	1(0)	1(1)	0(0)

Mutations are from left to right (First column: wild-type residues First Row: mutated residues). The frequency of stabilizing mutants is shown in parentheses.

Destabilize: V→A, W→A, Y→A, I→A, L→G, T→G, I→T
 Stabilize: N→I, E→W, N→V and K→S
 Stabilize and destabilize: K→M, K→A, V→I, E→K, T→I

Handwritten notes: 44/48, 10/0, 22/23, 37/37, 8/8, F→A, 25/29, 4/9→22, 5/9→51, 2/4→50.

And finally, we collected all the data and we classify into 2 groups, one is stabilizing and one is destabilizing. So, here if you see the numbers they are given for the destabilizing mutants and the one in the parenthesis are they one for the stabilizing mutants.

If you see this table you can derivate some conclusions. For example, some mutations for example, valine to alanine. What is valine to alanine? Here, most of the mutants are destabilizing how many totally, totally how many mutants 48, 48, 44 are destabilizing; that means, more than 90 percent or above 90 percent.

Likewise if you (Refer Time: 06:50) find some other mutations which are mainly destabilizing.

It is W to alanine here totally 10 mutants and all the 10 are destabilizing these are the destabilizing. So, any other examples.

Phenyl alanine to.

Student: Alanine.

Alanine phenyl F 2 A yeah this is F 2 A, if you take F 2 A how many mutations 27 or 27 25 or destabilizing. Now, this Y to A this is 23, I to A 37 for 37, I to T 8 by 8. So, from this frequency table you can see that some mutations are always destabilizing or most of the times they are destabilizing. These are unique data you can get from these ProTherm

database because of some conditions, do you see the number of data are less, but you can see a trend.

Likewise you see stabilizing mutations where i, if you take a N to I. What is the case of N to I? 2 out of 2 by 2, this 2 by 2, the only 2 mutations stabilizing both of them are stabilizing. What is K to S? There is only 1, this 1 by 1. So, any other stabilizing mutations, if you see here this is 2 E to W this out of 2 both are stabilizing. But the numbers are very less because this is very difficult to stabilize a protein compared to the destabilizing protein because if you make any mutations they will destabilize the protein. So, because it is very important to enhance the mutations this was a stabilizing mutations are less. So, this is the reason why the numbers are less.

If you have some cases they will both stabilize and destabilize. For example, take K to M out of 9, then if you take K to M 9, 4 are destabilizing and 5 by 9 are stabilizing. In this case we are all know if you change the lysine to methionine it may stabilize or it may destabilize. So, then it will some other example which can act as both stabilizing and destabilizing.

Student: E to L.

E to L.

Student: E to S.

Or E to L, E to L.

Student: 2 come out

2 come out this is this is what are 4, 2 are a 50 percent, this is if there is we do not know.

So, if you do the average assignment method if you can see many mutant and destabilizing for example, valine to alanine, if you have new mutation valine to alanine you can assign this is destabilizing or if you find the mutation asparagine to isoleucine you can easily say that this will stabilize

And the other hand if some cases for example, V to I, E to K. What is a V to I? Then you has V to I, 8 and 10, 8 by 18 and 10 by 18. In this case we do not know whether stabilizing or destabilizing, with the maximum number we will take a stabilizing. In this

case we will have lot of negative values right. So, we assign this numbers and you can predict the stability change, whether this will increase or this will decrease this is very easy, make a matrix. Now, we know the stabilizing mutants and destabilizing mutants and you can make a table. These are the mutations which is stabilized and these are the mutations destabilized. If it is 0 0 then we do not know we cannot predict anything. For any new mutation if you take a protein each this location it can be 19 possibilities you can do this mutations and get the percentage accuracy with known data, you can get about 60 to 70 percent 70, up to 70 percent accuracy you can get depending upon the data because in this case you will get the very high performance, because this mutations are very frequently occurring mutations in this case you are accuracy is very high.

So, what will happened if the stabilizing and destabilizing effects? Even in this stabilizing effects some case some cases few of them are in the other end. This one we consider all the mutants together. Now, it is possible to classify this mutants based on the secondary structure or based on the solvent accessibility. If you classify this based on secondary structure we can make 3 classes helix, strand and coil. If you classify based on accessibility there also we can make 2 classes or 3 classes for example, buried or partially buried or exposed and exposed right.

We can make 3 groups of secondary structures, 3 group for accessibility then we have 3 matrix based on secondary structure and 3 matrix based on accessibility. Then if new mutations come you will check the secondary structure, if this is helix then pick up from the matrix from helix, if it is strand then pick up matrix from strand or you can take this is buried you from the buried and the exposed from the exposed right. Then you can combine make 9 matrices for the helix based on solvent accessibility 3 or the likewise strand and coil 3 multiplied by 3 we can get 9 matrices if you classify based on both secondary structure and ASA.

Now, if you have any new mutants first you check what is secondary structure what is solvent accessibility, then accordingly pick up the data right. So, what is the disadvantage of having this classification, what is advantage of having this classification?

Student: More accurate prediction.

Yeah, advantage is because you can even if the same mutation can stabilize and destabilize depending upon different location you can classify. In this case you are accuracy will improve. Disadvantage if you see you do not to get sufficient number of data all the here itself many 0 if you classify then all these 20 will be in an 9 groups similarly less number of data. So, data this their problem. If you have more number of data then it is easy to classify and you can get better predictions this is your classification. Now, if you want to predict the real values. What can we do?

Student: Accurate.

(Refer Slide Time: 13:23)

Average $\Delta\Delta G^{H2O}$ values for the destabilizing (and stabilizing) mutants

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.00	-1.32(0)	-0.40(0)	0.01(1)	-1.10(0)	-1.40(4)	-0.70(0)	-0.72(4)	-0.90(0)	0.03(0)	-0.10(0)	-0.30(0)	-1.11(6)	-0.40(0)	-0.20(0)	-1.00(0)	-1.60(7)	-1.70(4)	-1.10(0)	-0.70(0)
C	-1.92(4)	0.00	-1.40(0)	0.00	0.00	0.00	0.00	0.04(0)	0.00	-1.80(0)	0.00	0.00	0.00	0.00	0.00	-1.91(0)	0.23(0)	0.00	0.00	0.00
D	-1.81(4)	0.00	0.00	-0.30(0)	-3.00(0)	-1.30(3)	-4.40(0)	0.00	-0.30(0)	0.00	-1.01(5)	-1.62(9)	0.00	0.00	-0.80(0)	0.00	0.00	0.00	0.00	0.00
E	-1.10(6)	-0.72(4)	-2.10(6)	0.00	-1.62(9)	-1.90(7)	-1.02(5)	-2.76(6)	-3.11(0)	-2.14(0)	0.03(0)	-0.20(4)	-1.70(7)	-1.00(4)	-3.40(9)	-1.30(5)	-1.21(0)	-1.92(9)	0.00	-2.40(5)
F	-2.90(3)	0.00	0.00	0.00	0.00	-5.20(0)	0.01(0)	-3.52(3)	0.00	-2.00(0)	0.02(9)	0.00	0.00	0.00	-4.30(0)	-5.30(0)	-2.10(0)	-0.40(4)	-1.11(0)	0.00
G	-1.90(4)	-1.00(1)	-2.90(2)	0.01(2)	-1.70(0)	0.00	-1.20(0)	0.00	-1.80(0)	-0.60(0)	-1.70(0)	-0.10(0)	-0.90(0)	-2.80(0)	-2.80(5)	-3.20(0)	-2.60(0)	-1.00(0)	-0.20(0)	0.00
H	-1.10(7)	0.00	-2.21(3)	0.01(2)	0.00	-3.51(6)	0.00	0.00	-2.10(0)	0.00	-2.00(6)	0.07(0)	-1.10(0)	-1.00(0)	-2.10(0)	-0.20(0)	-3.80(0)	0.00	-1.10(7)	0.00
I	-2.80(0)	-2.01(4)	-3.20(0)	0.00	-1.20(0)	-4.70(0)	0.00	0.00	-0.40(0)	-1.10(0)	0.00	0.00	0.00	-4.30(0)	-2.80(0)	-1.10(3)	-0.40(0)	0.00	0.00	0.00
K	-1.00(5)	-0.50(0)	0.01(5)	-4.91(0)	-0.80(1)	-1.20(5)	0.00	-1.20(0)	0.00	-1.80(5)	-1.20(3)	0.00	0.01(6)	-0.30(0)	0.01(0)	-1.30(0)	-1.40(0)	-1.10(0)	-1.10(0)	0.00
L	-2.31(0)	-0.60(0)	0.00	-0.80(0)	0.00	-4.10(0)	0.00	-1.40(0)	-1.40(0)	0.00	0.00	-1.50(0)	-1.40(0)	-0.71(2)	-0.40(7)	-2.60(0)	-1.70(0)	0.01(0)	0.00	0.00
M	-1.90(2)	0.00	0.00	-1.60(7)	-1.60(0)	-3.40(0)	0.00	-0.60(0)	-1.80(0)	0.00	0.00	0.01(0)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	-1.71(7)	0.00	-1.90(4)	0.00	-1.60(0)	-1.90(4)	-1.70(0)	0.01(2)	0.00	0.04(7)	0.04(0)	0.00	0.00	-0.80(0)	-3.90(0)	-1.10(0)	0.01(9)	0.02(0)	0.00	0.00
P	-1.81(7)	0.00	0.00	0.00	-0.50(3)	-1.01(3)	0.00	-1.70(0)	0.00	-1.70(2)	0.00	0.00	0.00	0.00	-0.80(0)	-1.71(1)	-3.02(0)	0.00	0.00	0.00
Q	-0.40(3)	0.00	0.01(5)	0.00	-0.70(0)	-1.30(9)	-0.80(0)	0.01(4)	0.00	0.04(0)	0.00	0.00	0.00	0.00	0.00	0.02(0)	0.00	0.00	0.00	0.00
R	-1.30(3)	-2.70(0)	0.00	-0.80(0)	-3.10(0)	-2.30(5)	-1.70(0)	0.00	-5.50(0)	0.00	-3.60(0)	0.00	-1.30(0)	0.00	-2.00(0)	-3.50(0)	0.00	0.00	0.00	0.00
S	-1.40(3)	0.01(0)	-1.70(3)	0.00	-1.50(0)	-1.10(0)	-1.30(0)	-1.40(0)	0.01(1)	-0.70(0)	0.00	-0.50(0)	-0.30(0)	-1.50(0)	-0.20(0)	0.00	-0.90(1)	-1.21(3)	-0.20(0)	-2.30(0)
T	-1.40(3)	-0.90(4)	-0.50(0)	-1.70(0)	0.00	-1.70(0)	0.01(2)	-1.01(3)	0.00	0.00	0.00	-2.70(0)	-4.40(0)	-0.10(0)	-1.10(1)	0.00	-1.30(4)	0.00	0.00	0.00
V	-2.10(7)	-1.80(4)	0.00	0.00	-2.50(0)	-4.30(0)	-1.70(0)	-0.70(4)	0.00	-0.70(9)	-1.10(0)	-0.80(0)	0.02(4)	0.00	-1.20(0)	-3.90(3)	-2.30(3)	0.00	-3.51(0)	-1.30(0)
W	-2.50(0)	0.00	0.00	0.00	-1.90(9)	0.00	0.00	0.00	-2.40(0)	0.00	0.00	-3.60(0)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-1.40(0)
Y	-2.91(9)	-2.90(0)	-4.81(7)	-5.00(0)	-1.11(8)	-3.60(0)	-1.60(0)	-2.50(0)	-3.90(0)	-2.50(2)	-2.00(0)	-3.90(0)	-2.10(0)	-2.90(0)	-2.41(17)	-2.00(0)	-3.00(0)	-0.60(2)	0.00	0.00

The average $\Delta\Delta G^{H2O}$ values for the stabilizing mutants are given in parenthesis

$$r[\frac{1}{N}] = \frac{\sum (Exp - Pred)^2}{N}$$

$$MAC = \frac{\sum |Exp - Pred|}{N}$$

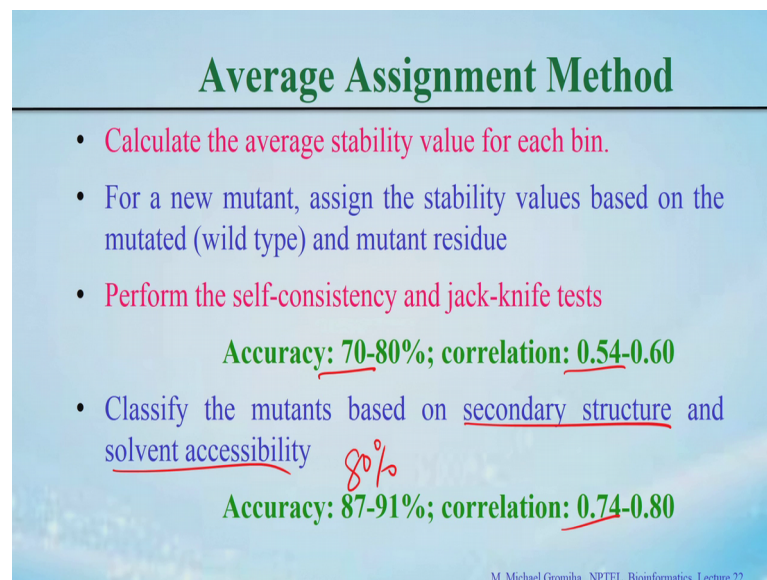
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

For example, if you have these mutants for this is C to A its 4, all 4 we know the values take the average now this is the value it is -1.9. This is what the destabilizing and for the stabilizing we get the value is in the parenthesis if you have the values.

So, now if you have any mutation for example, it is alanine to valine, alanine to valine. So, you can say this is -1.7 because this is a highest value, put -1.7. Then assume all the values and you can calculate the error, then how to calculate the error that is experiments minus predicted you will take the absolute you will because either we actual will be higher under predicted will be higher and you add up all these values and divided by N is the total number of mutants.

You take the predicted values take the absolute values and you can get this a N. Also you can calculate the correlation because we have the experimental values here and the predicted values also here. So, you can this is the experimental this is a predicted. So, you can get the correlation.

(Refer Slide Time: 14:31)



Average Assignment Method

- Calculate the average stability value for each bin.
- For a new mutant, assign the stability values based on the mutated (wild type) and mutant residue
- Perform the self-consistency and jack-knife tests

Accuracy: 70-80%; correlation: 0.54-0.60

- Classify the mutants based on secondary structure and solvent accessibility

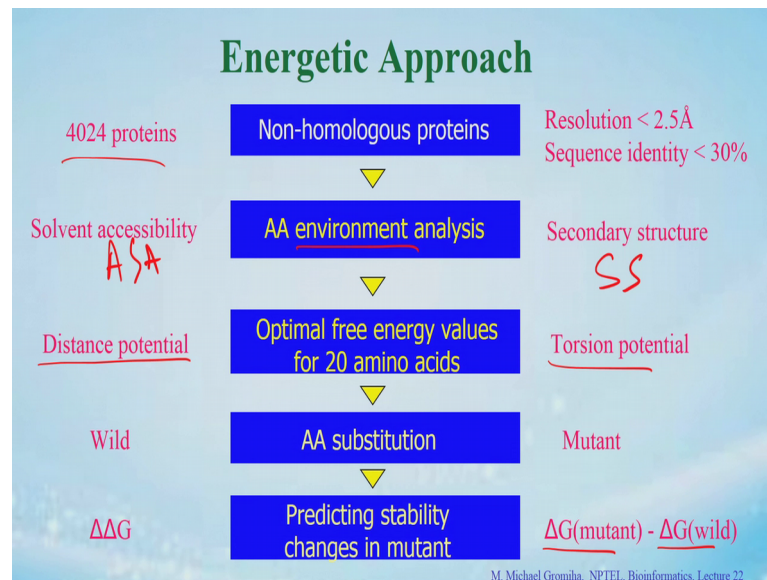
80%
Accuracy: 87-91%; correlation: 0.74-0.80

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

So, what is the status if you do like this? You can get around after up to an accuracy of about 70 percent. You can discriminate the stabilizing and destabilizing, but if you classify with the secondary structure and ASA you can increase this up to 80 percent, this is the if you take the different tests then you can get up to 80 percent.

Take the correlation, this will be about the around 0.5 and here also you can see the correlation about 0.7. So that means, if you make any method this is a minimum accuracy and correlation you can get, we get from the assigning a method. Now, we can add more information to increase the performance. Let us see what are other by different methods.

(Refer Slide Time: 15:13)



So, one of the most popular methods, this is based on energy because energy is very important for the different interactions and if you mutated a specific residue this will spoil the energy which will disturb the energy or improve the energy this can tell you whether this will increase stability or decrease the stability. For getting this energetic approach, what can we do? First we need to know what will happen if your particular residue is mutated to other residues that is means if you have a residue “A”, how far this is interacting other residues how these residues making contacts, if you mutates how the contacts will change.

Second aspect is what is torsional angle for this particular residue, if you change the confirmation how this will change the confirmation, how this effects the stability of the mutations. Based on these aspects first we need to know we collected a set of proteins for example, around 4000 proteins. They are with high resolution may be less than 2.5 angstrom and less sequence identity. Then we classify into different groups location of each residues based on ASA and the secondary structures this will give you the environment. Then see, then finally, we need to do the potentials how far each residue in each secondary structure are depending on each accessibility they are making contacts with the different residues. Likewise you can see the torsion potential how far they conformational change, how the conformational effect this stability.

(Refer Slide Time: 17:15)

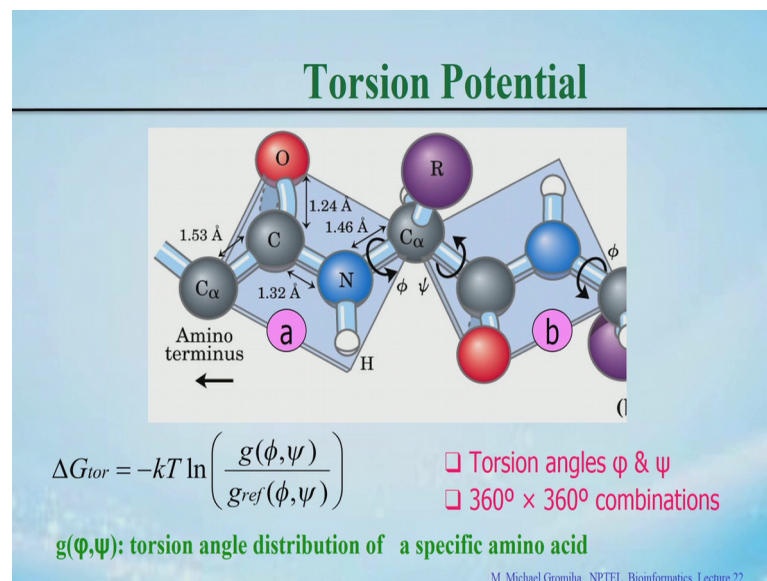
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

Likewise you have different types of atoms we treat separately. So, if you made it into 4 different types of atoms based on location and the connectivity and the chemical nature. So, what is the chemical nature, where this atom is located and where is connected right. So, we have 40 different types of atom types 36, 37, 30, 40. So, now for all these 40 types of atoms we can calculate the distance potential using the preference of residue pairs. You can see the g rd this is for any particular distance i and j are any specific residues alanine and valine, alanine to aspartic acid.

So, what the preference of having these two residues in contacts with the any specific distance say for example, 6 angstrom or 8 angstrom and here we have the same distance

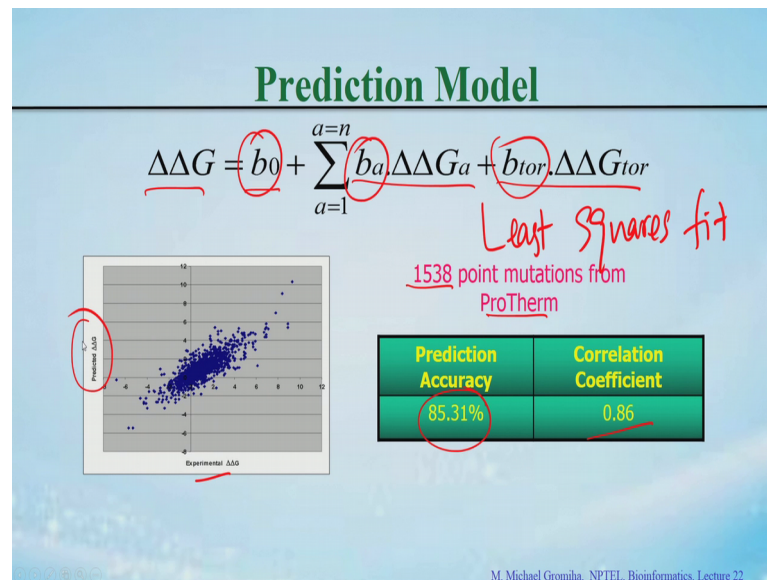
totally how many contacts. Using this information you can calculate the torsion potentials is the minus kT logarithmic of this probability to get the torsion potentials. This will give you the distribution of residue pairs i and j at any distance d . We can optimize the distance we will using different distances finally, how this will relates with the experimental data we can optimize a distance fine. So, you get the torsion distance potential is done.

(Refer Slide Time: 19:10)



So, second aspect you can see the torsions you can use the say similarly you can use same equation ΔG_{tor} , this is a g of ϕ and ψ for any reference points and normalize with the any reference points. We can use various combinations and see what is the probability of having the particular confirmation it compared with the all random conformations. Then we convert this in to potential by multiplying with minus a kT logarithmic of this one we can see torsion potentials.

(Refer Slide Time: 19:42)

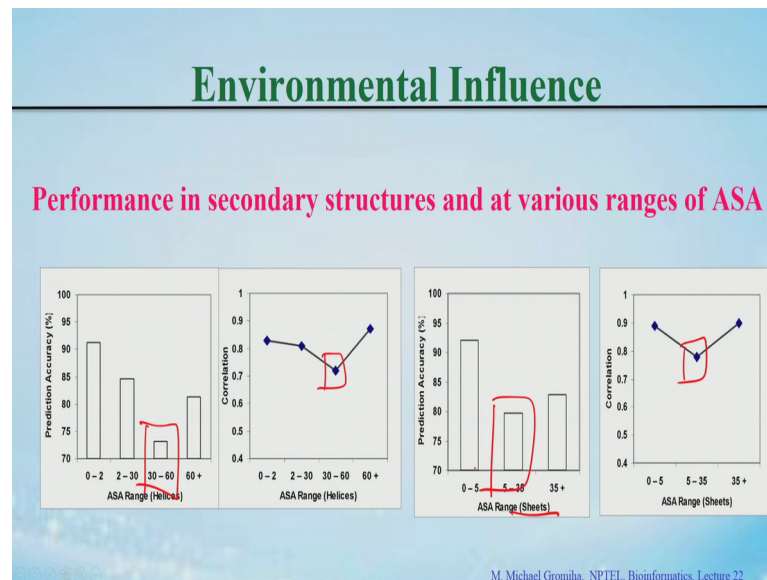


So, 2 different aspects one is based on distance, one is based on torsion then you compare these two see here you can see the torsion potential and the distance potential these are constants. This is a kind of multiple regression technique. For all the cases we know the values because we know the delta G torsion or delta G, the distance potential delta delta G we know. Then we use the principle of least squares. So, use the least square fits. This will give you this b_0 , b_a and then b_{tor} .

Once we get this values then for any data you know the G and G tor substitute then values and you can get the delta delta G. Here show the results for example, if you take the 1538 mutations from ProTherm database because here you get the reliable database. So, you consider the relationship between experimental delta G and the predicted delta G. So, you can see the accuracy of 85 percent. Accuracy means distinguishing stabilizing and destabilizing mutants and the correlation is exact relationship between experimental and the predicted value is this is 0.86.

Now, the question is how these performance varies with respect to secondary structure or with respect to solvent accessibility. So, I show the data.

(Refer Slide Time: 20:58)



So, we classify the ASA in different ranges 0 to 2, 2 to 30 and so on. So, these numbers also we made a classification such a way that their data can be related with the stability with the highest performance. So, accuracy it is about a 90 percent if it is a buried, even the case of beta strand if it is buried then you can check it a more than 90 percent.

We discussed earlier its buried mutations you can easily relate with the hydrophobicity and single a property can also explain the stability. But the case of exposed single properties very difficult to explain this is a reason why we include the sequence information or structure information. Even if you add this structure information the performance is only limited.

So, if you see here this 60 percent 60 plus solvent accessibility in helix the performance about 80 percent and the correlation is around 0.85 and so on. So, for the buried mutation you can see it is higher. This is for the sheet the buried one you can higher performance here also you can see. And importantly if we notice here that the partially exposed case or partially buried case these values are less compared with the completely buried or completely exposed cases.

(Refer Slide Time: 22:16)

Web Server CUPSAT

Predict Mutant Stability from Existing PDB Structures

Protein (PDB ID): (4 letters)

Amino Acid Residue No.: (including code, if present)

Amino Acid (Native): (Ref)

Experimental Method: ☒ Thermal ☐ Denaturants

Details of Mutation Site

PDB:	2LZM
Wild Type Amino Acid:	Ile
Chain:	0
Residue ID:	3
SS Element:	Helix
Solvent accessibility:	11.89%
Torsion Angles (φ, ψ):	-52.1°, -40.3°

Comprehensive Prediction Results

Mutation Site		
PDB	Chain	Wild type AA / Residue ID
2LZM	0	Ile / 3

Structural Features		
SS element	Solvent accessibility	Torsion angles (φ, ψ)
Helix	11.89%	-52.1° -40.3°

Amino Acid Mutations			
Amino acid	Overall Stability	Torsion	Predicted ΔG (kcal/mol)
GLY	Destabilizing	Unfavorable	-2.25
ALA	Destabilizing	Favorable	-1.39
VAL	Destabilizing	Unfavorable	-1.39
LEU	Destabilizing	Favorable	-2.15
MET	Destabilizing	Favorable	-2.15
ARG	Destabilizing	Favorable	-2.15
TRP	Destabilizing	Favorable	-1.3
ASP	Destabilizing	Unfavorable	-2.98
THR	Destabilizing	Unfavorable	-2.91
PRO	Destabilizing	Unfavorable	-1.12
GLN	Destabilizing	Favorable	-1.77
LYS	Destabilizing	Favorable	-1.1
TYR	Destabilizing	Unfavorable	-2.78
ASN	Destabilizing	Unfavorable	-2.64
CYS	Stabilizing	Unfavorable	1.24
GLU	Destabilizing	Favorable	-2.96
ASD	Destabilizing	Unfavorable	-2.48
APG	Destabilizing	Favorable	-2.1
HEI	Destabilizing	Unfavorable	-2.11

Note: Overall stability is calculated from atom potentials and torsion angles. In case of unfavorable torsion angles, the atom potentials may have higher impact on stability which results in a stabilizing mutation.

<http://cupsat.uni-koeln.de>

V. Parthiban, MM. Gromiha and D. Schomburg (2006)
Nucl. Acids Res. 34, W239-242

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

Now, we can have a server this is the CUPSAT server. It takes the PDB ID and you have to give the residue number and the which is the residue either we need thermal or denaturants. We put it stability first should be automatically calculated, the secondary structure and solvent accessibility will calculate, torsion angles also it will give you, if there is any problem in the torsions it will tell you the torsion is a problem in this case the results are not reliable. If it is fine then we proceed to a next step then finally, here it will give this is the prediction results. For this mutation for this isoleucine is mutated to all the 19 cases and here overall stability was stabilizing or destabilizing torsions favorable non favorable and you can see delta predicted delta delta G.

So, one search you can get all the values, for all the 19 mutations. This case it requires the structure because we need to calculate the distance potential and the torsion potentials that require structure based on that the check the available database and then see how they are attributed in terms of distance and torsion potentials and we can get the delta G values. So, the structure is not available then what to do? There are two different ways one is you can do the modelling to get the structure and use the server or we are trying to get a new method just from sequence information because in the average assignment method I mentioned that if this mutation is the alanine to valine it can be stabilizing or destabilizing depending upon the values. Then the sequence information we can tell if alanine to valine and the neighboring residues aspartic acid and then this could be stabilizing residues.

(Refer Slide Time: 24:06)

Sequence Based Prediction

Variables used for discrimination and prediction

1. Wild type residue
2. Mutant residue
3. pH
4. Temperature
5. Neighboring residue information (three residues on both sides of the mutant)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

So, we can make rules, way this case we can take all the information wild type residue mutation residues and neighboring residue information all this information we take that.

(Refer Slide Time: 24:15)

Sequence Based Prediction

QA →

- If the wild-type residue is Ala and the neighboring residues contain Gln: destabilizing (accuracy = 97.1%)
- If the wild-type residue is Glu and its second neighbor at N-terminal is Met: the mutation stabilizes the protein (accuracy = 100%)

N ← E → M

L-T. Huang, M.M. Gromiha and S-Y. Ho (2007) Bioinformatics 23, 1292-93.

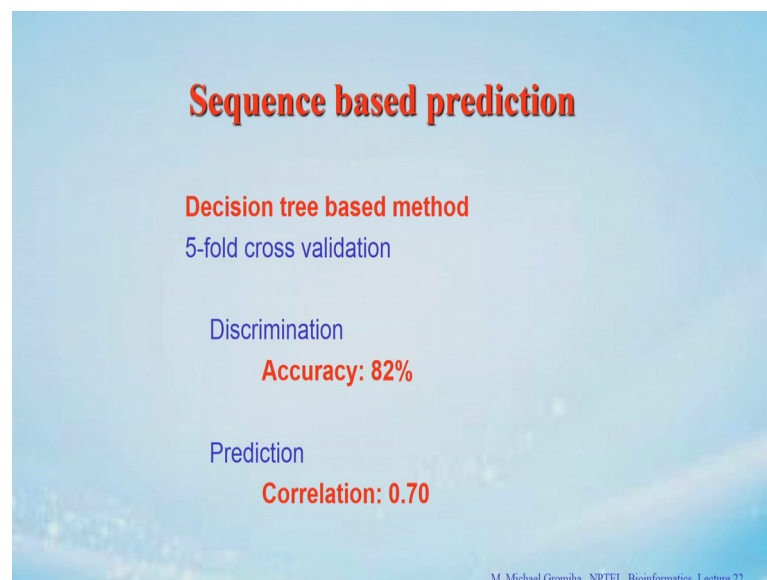
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 22

And then we can develop the rules. For example, wild type residue is alanine right. So, A is the wild type residue, A is mutated to any other residue and if the neighboring residue contains glutamic acid then this is destabilizing. In this case you use various options available. So, it get the it could this rule can predict up to accuracy the accuracy of 97 percent.

Second condition with the wild type residue is glutamic acid this mutate the other residues and the second neighbor this is the N terminal this is the C terminal second neighbor is this E M, then this will stabilize the protein and because set of data. So, we could get an accuracy of 100 percent.

So, here the problem is if we have more number of mutations and we could accommodate all the mutation with less number of rules, we can get better accuracy right. So, we consider about 3000 mutations and made about 100 rules, that means approximately 30 mutations. Some case we have we have 100 sometime may be less, if your say data are very high like the valine to alanine mutation then we can predict with high confidence.

(Refer Slide Time: 25:28)



So, we have the server this can get an accuracy of 82 percent and the correlation of 0.7.

(Refer Slide Time: 25:37)

Sequence Based Prediction

the protein sequence of single point mutation you have submitted is **W405G** with the following details:

- 1) Mutational residue is wild type is **W**
- 2) Mutational residue is mutant position is **405**
- 3) The ΔG value of experimental condition is **5.98**
- 4) The temperature value used in the experiment is **27.8**.

The predicted value of thermal stability change is **-0.997** kcal/mol.

(The composition information of the neighboring residues)

Figure 2: A pie chart showing the percentage of residues based on priority. The chart is divided into two segments: 50% (Hydrophobic) and 50% (Hydrophilic).

Legend:

- Hydrophobic: Ala (A), Ile (I), Leu (L), Met (M), Phe (F), Trp (W), Tyr (Y), and Val (V)
- Hydrophilic: Arg (R), Asn (N), Asp (D), Glu (E), Gln (Q), His (H), Lys (K), Pro (P), Ser (S), and Thr (T)

D225Q in TIM: Experiment: $\Delta T_m = -3^\circ\text{C}$ (Lopez et al. 2008)
 Prediction: $\Delta\Delta G = -0.97$ kcal/mol (Destabilizing)

<http://bioinformatics.myweb.hinet.net/iptree.htm>

M. Michael Gromiha - NPTF, Bioinformatics Lecture 22

So, these a server it will take the mutation residue and this wild type and the mutation residue, this is the wild type this is mutant right. So, we have neighboring residues. Now, we can predict. This will tell you the whether the stabilizing or destabilizing, now this will tell you this is the stability change this is the stabilizing and destabilizing and this only give some information how many hydrophobic residues nearby based on this mutation and so on. So, one example I give D225Q, here the experiment T_m is 33 degrees and our prediction is also minus 0.97 both are destabilizing, but if you are we know the delta only the T_m values, but there is anyway higher than 2 to 3 times higher than the delta G values.

So, you can use this sever also if you the structure is not available. We can this server for predicting the stability upon mutations. So, summarizing, what did we discussed today.

Student: Stability upon mutation.

Stability upon mutation, if you have, which database we use - ProTherm. So, if the data is available first we can try to relate amino acid properties with stability. Stability values we can get from ΔG , $\Delta\Delta G$ or ΔT_m . Properties, which are various properties?

Physical, chemical, energetic, confirmation properties. So, we can relate with the correlation coefficient because ΔG we know, ΔG_{mutant} minus ΔG

wild. Properties we know ΔB a mutant minus ΔB wild, for all mutations you can write with the correlation. So, this will tell you which properties are important where explaining which type of mutants.

Then we can develop different methods either from regression techniques or average assignment method or different potentials to predict the stability. Then also you can use different rules for predicting the stability of this mutations. So, we derived various structure based parameters, sequence based parameters and these parameters of potential applications because we discussed about regression techniques right. So, we can use all these parameters for predicting the stability of mutants, as well as it has other potential applications for example, to understand the folding rates or to understand how the proteins interact with other molecules and so on. So, that I will explain in the subsequent classes.

Thanks for your attention.