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

Lecture – 11b
Protein sequence analysis II

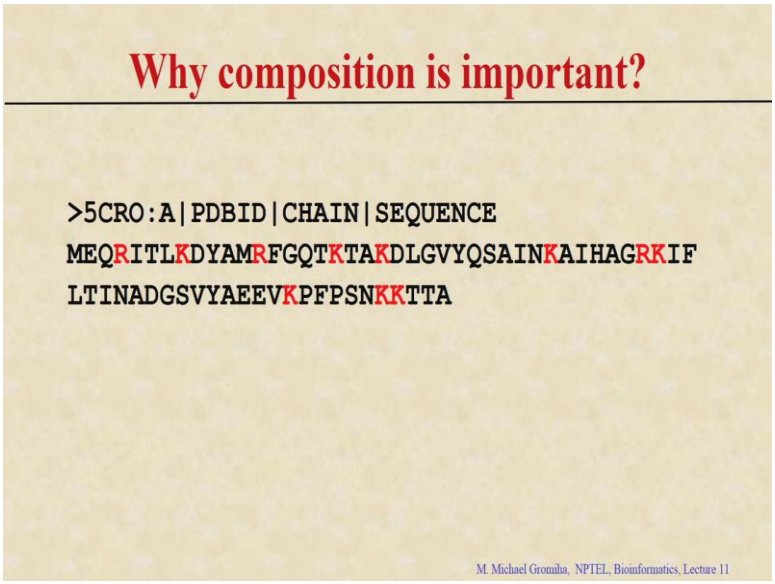
(Refer Slide Time: 00:17)

So I'll show some examples, this is one amino one sequence for a particular protein and you give something on the red font. So, can you see this red font? So, can you guess the region with red font?

Student: Hydrophobic.

So, it is completely hydrophobic if you see the residues most of the residues are alanine valine, phenylalanine, isoleucine and so on. In this case this protein, this region is predominantly with the hydrophobic residues.

(Refer Slide Time: 00:56)



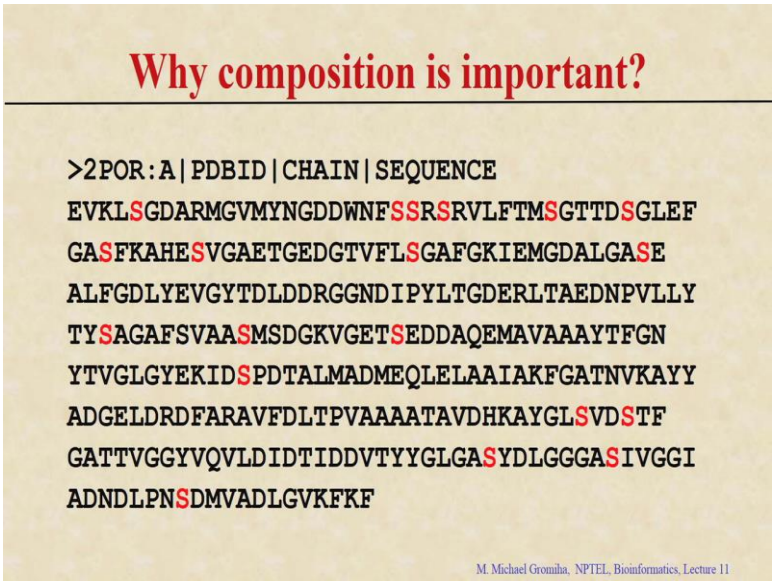
So, if you have another sequence this is another sequence called 5CRO.

So, if you see this sequence is short sequence, when short sequence you will get the predominant of some specific residues, which residues are dominant?

Student: Positively charged.

Positively charge residues. So, how many residues are here? 1 2 3 4 5 6 7 8 9 10 11 residues, 11 residues, 11 positives are residues in this sequence here.

(Refer Slide Time: 01:11)



You have another residue sequence, here if you see there is a higher occurrence of serines you can see may be several serine residues in this case.

So, when you scan the literature, if you see the literature about the functions of different proteins right. So, we can guess these positives charged residues are important to interact with negative charge moieties.

In this case you can think of a protein, which can have high tendency to interact with negative charge moieties, we say which molecules have negative charge?

Student: (Refer Time: 01:46).

Mainly DNA has a negative charge because the phosphate group.

So, in this case it requires the opposite side positive charge residues. So, this protein could be at DNA (Refer Time: 01:54) protein have high tendency to interact with DNA. In fact this because 5CRO is the cro repressor, these interact with the DNA likewise here if you see. So, here you can see the dominants of the structure of hydrophobic residues, they this residues is preferred to be in the membrane right.

So, in this case this chain could be membrane protein actually that is correct because this is the photosynthetic reaction center of m chain. So, in this case this is a membrane protein you have several transmembrane segment this is one of the segment. So, likewise here you can see one example, here this is the sequence for one beta barrel protein. Beta barrel protein have high dominance of serine residues to form the hydrogen bonding network to maintain the stability and the function was a structure of these barrels right. So, this why it is highly occurring in this particular protein fine.

Now, the question is, if I give these 3 sequences, is it possible to discriminate A, is this type B, is this type and C, is this type, yes or no? Let us see. So, there are various features we discussed, what various features we discussed till now?

Student: Composition.

Composition, occurrence.

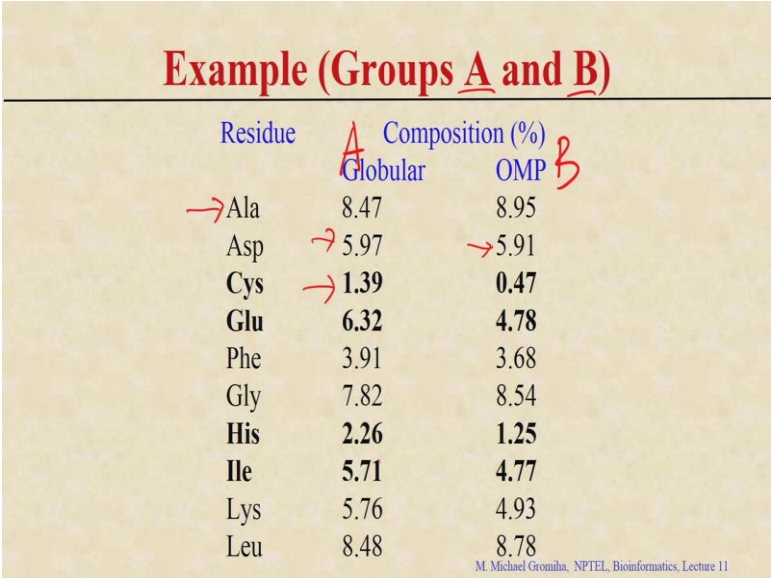
Student: Molecular weights.

Molecular weights.

Student: Hydrophobicity.

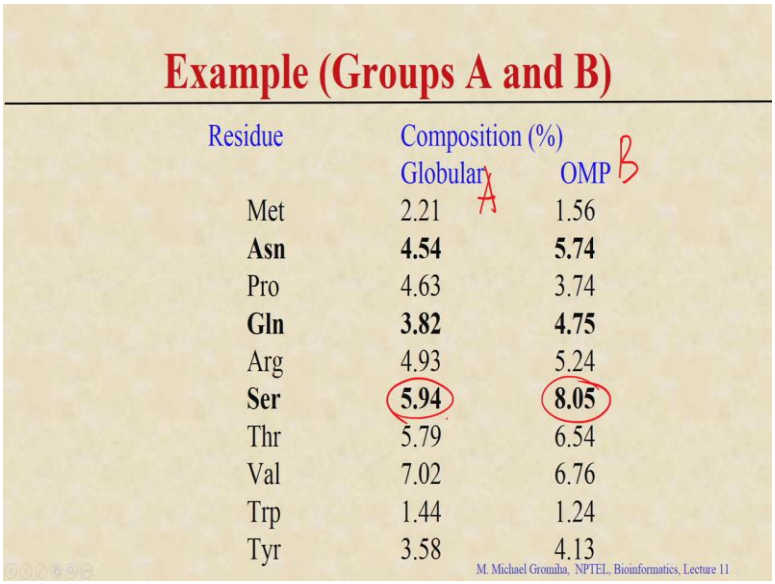
Hydrophobicity and different amino acids properties right. So, for example, if you have 2 groups of proteins, group A and group B. If I give a new protein, can we able to tell this protein belongs to group A or this protein belongs to group B? ok.

(Refer Slide Time: 03:27)



So, if I here, I take 2 groups group A and group B. So, I get the, for example, I collected all the proteins in the UniProt in group A and all the proteins belong to group B, then if you give to all these sequences I can calculate the composition right. Because just you can calculate n(i) by N, take all the residues, count the residues of each type, normalized by N, you will get the composition right. This group A for example, this is group B. You get the composition for 20 different residues, here 10 residues and here 10 residues.

(Refer Slide Time: 04:00)



And when you compare these 2 values for example, if you take alanine or the group A if the value is.

Student: (Refer Time: 04:10).

8.47 and the group B.

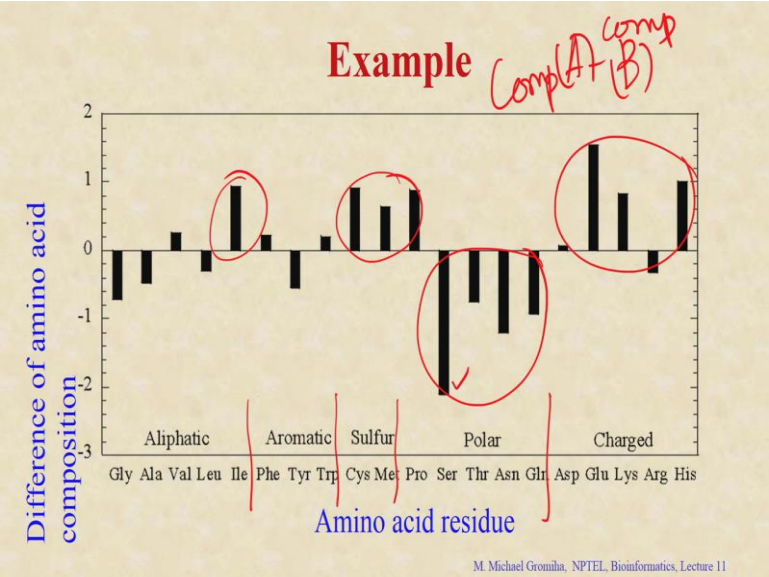
Student: 8.95.

8.95; if you get the deviation then you can this is the almost similar we use the average value, but if you consider standard deviation. So, this may plus or minus 0.1 or 2 right. So, in this case you will get the similar numbers some cases yes for example, aspartic acid there is no change at all here 5.97 here is 5.91 there is no change. The difference is only 0.06, but some residues if you look in the details, that are highly different for example, if we take cysteine, here it is 1.39, but here in this case is 0.47, there is 2-, 3-times difference, 3-fold difference, is 0.5, in the 1.4. Likewise if you take glutamic acid, histidine, this is 1.25, is 2.32 times different, isoleucine.

I will show some more data, this is the other residues. Here also you can see the differences is other way around. Here it is highly dominant in the case of group B, this 8.05, this is group A it is only 5.9, this is 69. So, some residues which are similar in both the cases and some residues which are significantly different, if they are different then

we can use these residues as features or the parameters or the properties to discriminate or distinguish between these to 2 groups, group A and group B.

(Refer Slide Time: 05:45)



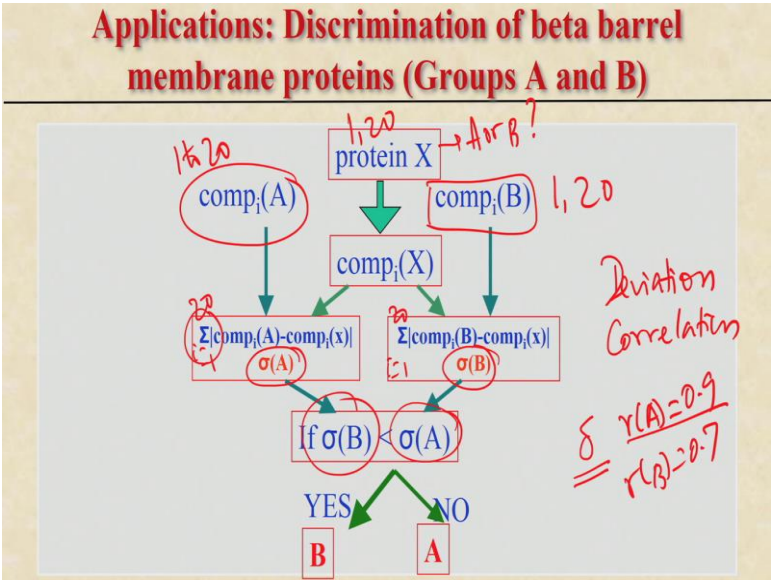
So, how to do this? First we see the important residues and some residues this is just close to the 0, this a difference of composition, I calculate the A - B, comp(A) - comp(B), right. So, we have the average values.

So, if we take some cases, it is very high, then will be less than 0, and some cases they are above 0. So, if you look into these different residues, I made in few groups, this is the aliphatic and the second one aromatic here, we have sulfur-containing residues and polar residues and charged residues and we could see some patterns, some cases you can see pattern for example, polar residues. All the polar residues which are highly preferred in the case of this group B, in case, minus. Some of this, the sulfur-containing residues they are dominant in the case of group A, here you can see some cases, also mainly charged residues in group A.

So, in this case, you can try to relate you can try to understand why some a residues are dominant in group A and some residues are dominant group B. Actually we know what is group A what is group B, the structures we know, the functions we know. So, now, we can explain the importance of these residues, why its occurrence is high in some cases why it is low in some cases we can find. And we use this information, if it is very highly different for example; this is highly different, even this plays important role for

discrimination. So, if you take these residues and the combination of these specific residues, we can able to distinguish between the proteins group A as well as group B. Fine, how to do this?

(Refer Slide Time: 07:32)



Since, a very simple algorithm right, but you can complicate the algorithm when you refine the results; simply what we do, we have the set of proteins from group A and group B. Take the all the proteins from group A and the values of group B composition you can calculate. Then take all the proteins from group B get the composition, now you have 2 set of values for example, here you have the data one is A one is B.

Now, what you will do is, we get the new protein x, we do not know this belongs to A or this belongs to B, A or B we do not know. Then what to do first? Get the composition of this particular protein. So, we get 20 numbers here. So, here you get 20 numbers, we are here, 1 to 20 per 20 residues and composition B, 1 to 20 and protein x also we have 1 to 20.

Now, we take compare each composition for example, A, D, C or 20 residues get the differences, comp(A) - comp(x) respectively, we get the difference, take the absolute and repeat for 20 times 20 residues, then get the summation, i equal 1 to 20. And finally, we get a number. This says we can see this σ(A), deviation from composition of the group of proteins A. Repeat the same, take the same comp(x) and compare with B, do it for 20 residues, i equal to 1 to 20 right.

So, get the difference now, you can see the total difference will be σ(B). So, with the comparing comp(A), we get the σ(A) and comparing with the comp(B), we will get the σ(B). Now compare these 2, σ(A) and σ(B). Which one has less deviation? Which one is less deviation here? B is less, then this belongs to group B, if A is less, then the protein x belongs to group A.

So, here just we compare the deviations, we can also apply some error functions you can apply error function δ, like 0.1 or 0.2 on 0.3 so on. And you can optimize the performance. You can, I will explain later about the how to assess the performance, there are various measures to assess the performance like sensitivity, specificity and accuracy. I will discuss in later classes. So, you can add this δ and see in a set of 100 proteins, we are able to predict or discriminate exactly for the maximum number of proteins, then we can define this error function. In this method, we use the composition and the deviation. Instead of deviation, you can also try to use some other measures. What other measures you can use? Here we use deviation, we can also use the correlation, we got 20 numbers here, 20 numbers here, you can get the correlation with A, likewise you can get the correlation with B, then how which one we need to select? Which one belongs to A? The correlation will be?

Student: lower

Higher; if this correlation or the correlation with A is 0.9, and r with B is 0.7, then the protein belongs to?

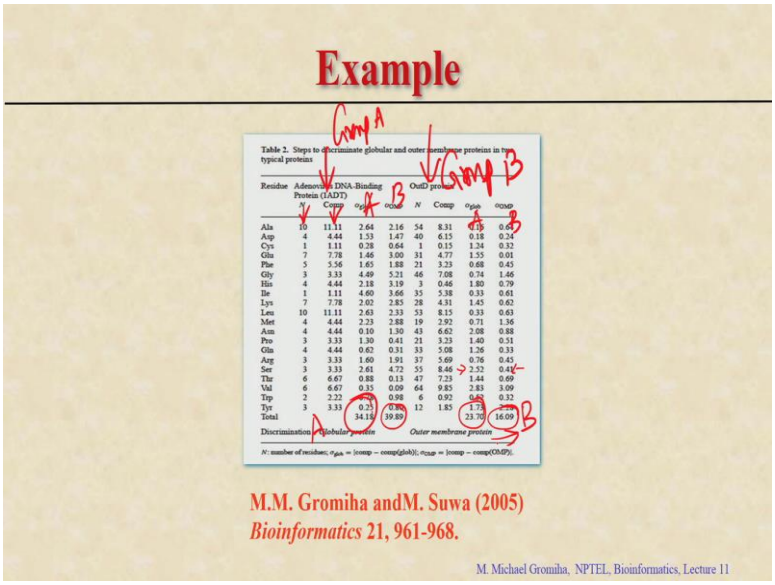
Student: A.

A right.

So, you can use correlation, you can use the deviation and you can use any error function and so on. Also here now use the composition of all 20 residues, but if we look into this figure all 20 residues are not showing very high difference, only some residues the difference is significant. Then you can try with the residues which are showing only significance between this group A and B and see whether we can able to discriminate better or not.

Secondly when you develop any method, it is important to reduce the number of parameters number of properties. Instead of 20 if you get the same performance with 10, then 10 is better. Unnecessarily we do not have to include noise in the prediction performance ok.

(Refer Slide Time: 11:51)



Now, I will show an example here this is I got 2 proteins, one is the from group B this is group B here, this is from group A. So, I have a sequence, now what to do? Get the composition, occurrence, here this is N, is here from this N, you can calculate the composition, this composition, then from this composition we know the composition from the globular protein or group A or group B you can see here this is the composition of group A and composition of group B.

So, to get the, subtract the values. So, this is the σ, σ(A) and σ(B); do it for all the residues then finally, sum up; here this is a 34 and here is 39 which one is small?

Student: This

This is small, this is small. So, this is belongs to group A, then we take the example group B here also we have to calculate N, you can get the composition, and here we can see, this is A this is B, get the deviation, the original values we have here, these are original values, it take the composition and subtract from either A or B for 20 residues.

So,8.47 is here, and you can see this is 8.31. So, the difference is 0.16, or the other case it is 0.64.

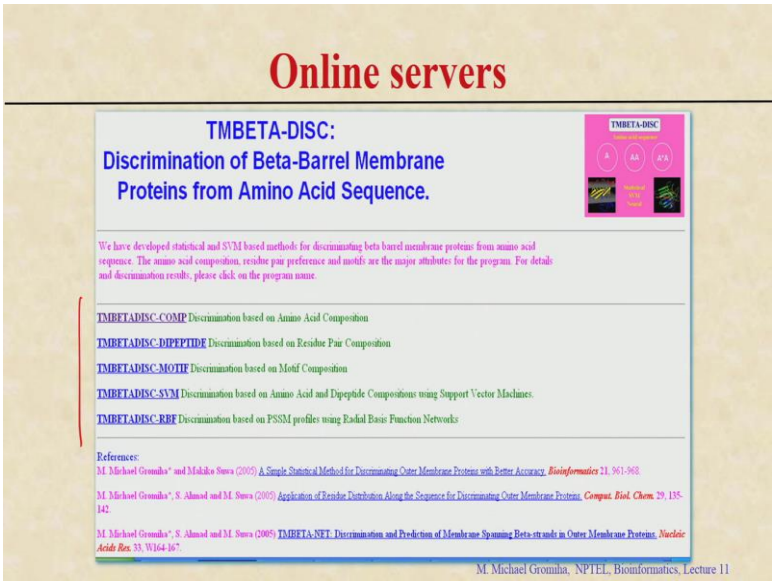
Likewise you can see the difference only if you see the serine is here is 8.46, 8.46; if you look into this A it is 5.94 and the B, it is 8.05; you can see the difference is 2.52 and B, it is 0.41. Now take the summation, this is 23, this is B, which is small? B is small right?

Student: Small.

So, B is small. So, this is identified as group B right. So, this is simple algorithm, but we need to refine this, adding various other factors and then estimating the performance and validating, that everything we need to do, but simple algorithms or simple features we obtain from the sequence, you are able to classify the proteins from different structures or different functions.

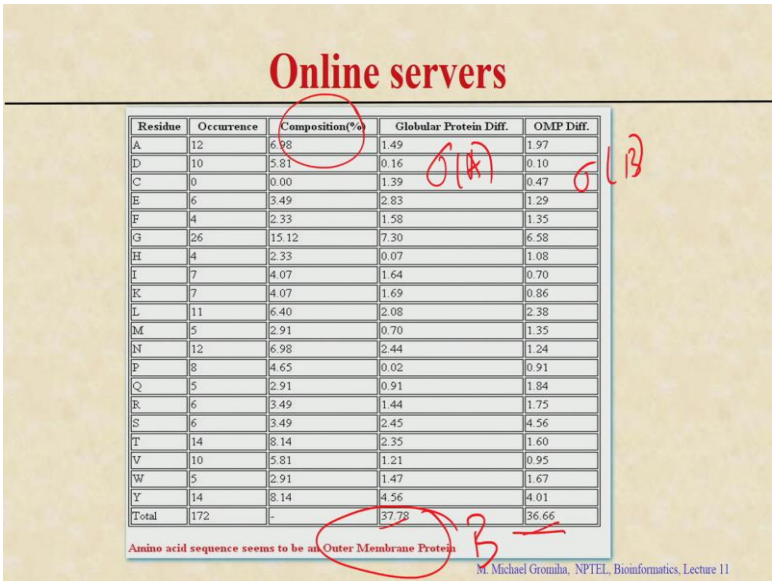
So, then once done, then we can do several online methods. So, here these are the various methods you can discriminate these types of proteins group A and group B.

(Refer Slide Time: 14:28)



So, if you go with the composition you can try with the composition right. So, if you give the sequence here you give the sequence here, so if you click on this, then you can give the sequence.

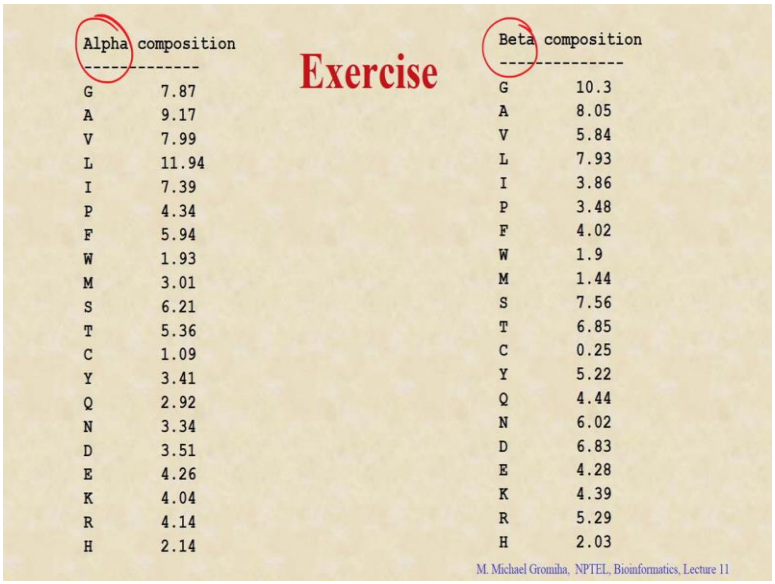
(Refer Slide Time: 14:34)



We get the sequence, then we will automatically calculate the occurrence from the occurrence. We will get the composition, from the composition you will get the σ(A) here, σ(B) and from this number we can see this is group B. So, it belongs to group B, I put the membrane beta barrel membrane protein as group B right.

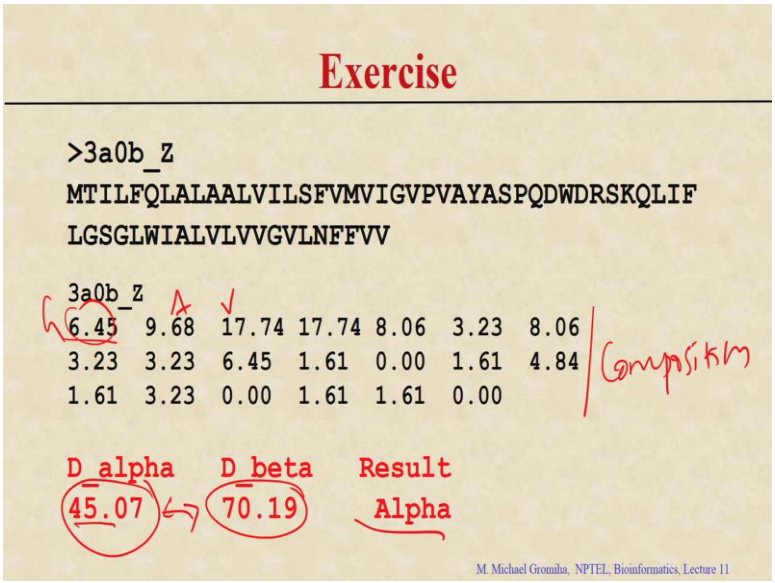
So if you see this one, the difference is very less, because if you add several residues and if you add only with use the data only for the specific residues, you can improve your performance that we need to try again and again. So, improve the performance and I have question. So, this is the data.

(Refer Slide Time: 15:14)



So, I have the composition for alpha and I have the composition for beta, I have 2 values, this is alpha composition and this is beta composition.

(Refer Slide Time: 15:23)



So, now I have a question. So, this is the question, I give this sequence. I want to know whether this belongs to alpha or beta. How to proceed, what to do?

Student: Composition.

Right first we use the sequence and calculate the.

Student: Composition.

this is the composition. I can calculate, you can take the number of, as number of Ds and normalize with the number of residues, this is for either same order, we give like GAV, this is G, this A, this V and so on, now what we do?

Student: (Refer Time: 15:59).

Take these number.

Student: (Refer Time: 16:02).

And subtract.

Student: (Refer Time: 16:04).

With this numbers alpha right. So, if we add up for all the residues, we get 20 numbers when we take then absolute values and the add up. If you add up everything then we will get these numbers then what we have to do?

Student: repeat

Repeat the same.

Student: Beta.

With this number, this beta number, and for 20 residues we calculate and sum up the absolute values finally, we get this number, then what we want to do?

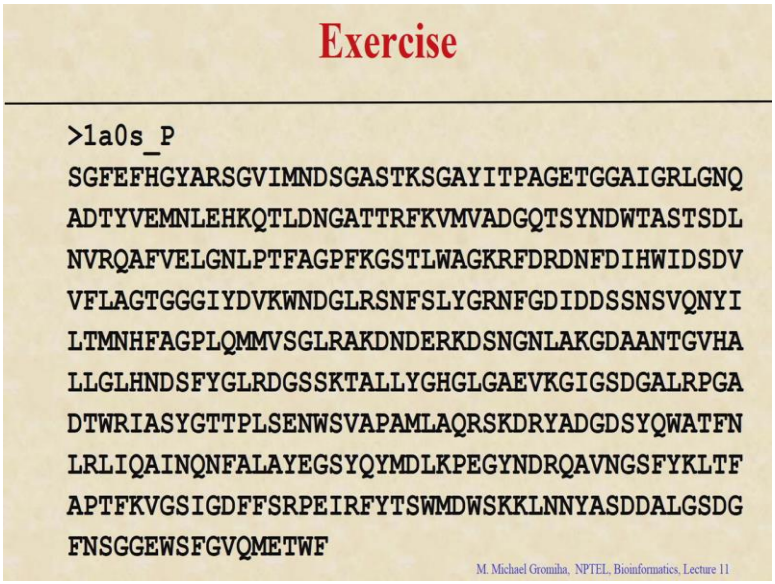
Student: Compare.

You have to compare these two. So, and then how to decide these alpha or beta?

Student: Whichever less.

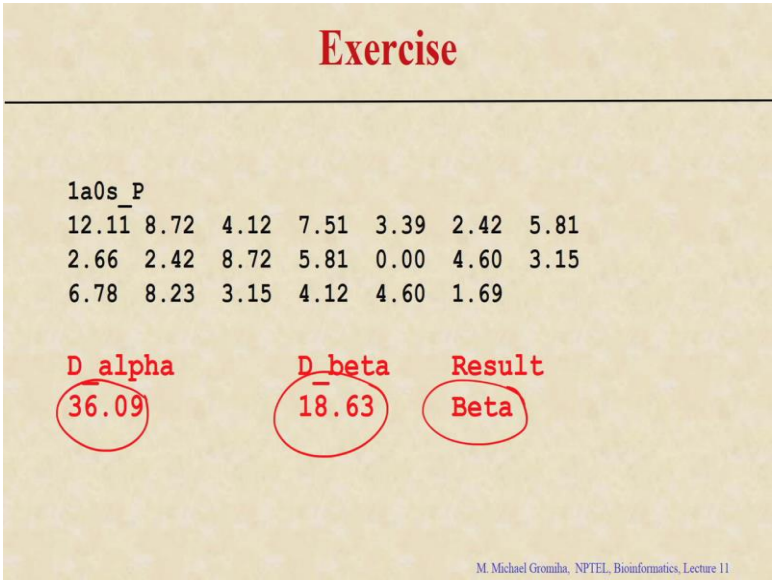
Whichever lower because the deviation lower that is biased with that particular protein. So, this is lower. So, you can protect this is alpha.

(Refer Slide Time: 16:46)



I give another quest another question here you can you can do the exercise. So, these also you can you have a sequence is a long sequence, there is small sequence you can easily count long sequence you have write to your code, you can calculate as we discussed earlier about the algorithm; you can write the program to calculate the composition.

(Refer Slide Time: 17:03)



So, this is a result. 20 residues you have the composition now you subtract these from either alpha or from beta, then we have to get absolute values and take that the

summation, you will get the values alpha is 36.09, beta is 18.63. So, if you compare these 2 this is less. So, this belongs to.

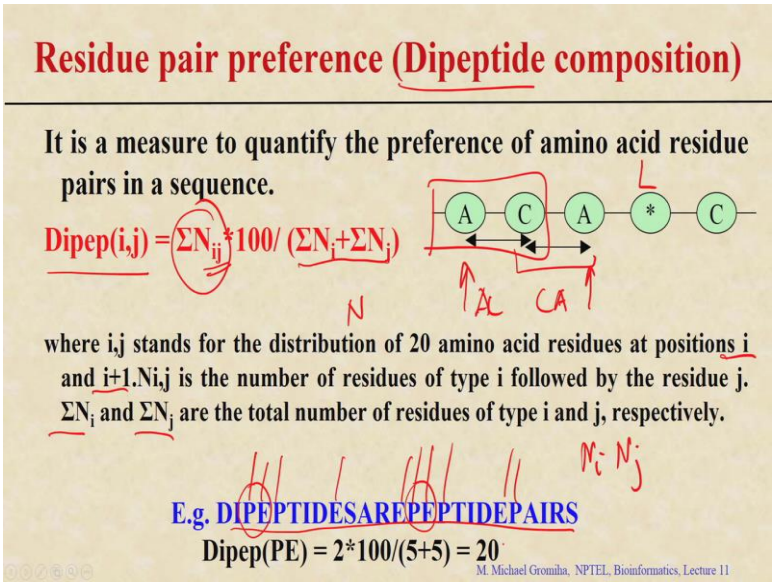
Student: Beta.

Beta see you for the different is significant very high 10 15 difference then easily you can discriminate. If it is very close then it is difficult. Some cases this is why all the prediction methods sometimes fail in some aspects, you cannot get 100 percent prediction accuracy because you can see some type of overlapping likewise see in this composition, you can see some cases it is very close see is very close. May be either group A or group B this is the reason why it is find a difficult to discriminate.

For that cases you have to identify some features not exactly this one, which can discriminate in some other way only at least to that situation. In this case you can combine only if you why is the large difference in composition, you can use this if it is less difference use some other features. Then you can incorporate this case we can improve your performance. One give 70 percent and another aspect it will another is 70 percent if you combine both you can increase to 80 percent you can do that.

So, now the another features there is pair preference, in the composition we use only one residue and see how many times A how many times D how many times D E all these things.

(Refer Slide Time: 18:44)



But then the question is the occurrence of A always close to A always next to D always next to E here there are 2 A's this A is just before C this a is may be something else for example, this L. So, this A is after C this C is after A with this C is after L.

So, another question is how many times A comes next to A, how many times A comes next to C or E and so on. In this case you can get the pair preference because we had 2 dipeptides. So, this is so, we call this as like dipeptides if we take these two. So, we can say this is like dipeptide like on these 2 this is AC and here this is CA. So, in this case there are various way is to get the preference.

So, if we see the dipeptide of i coma j where is you can see Nij with number of times of residue i which comes next to j, i and j are the distribution of the 20 residues and to get the percentage 100 right, but you can normalize with the different factors, either you can normalize with the N as we did in the case of composition, here I normalize with Ni plus Nj; that means, totally how many residues which are involved in this i and j if it is i 5 times and j 9 times, then there are 9 cases in the sequence.

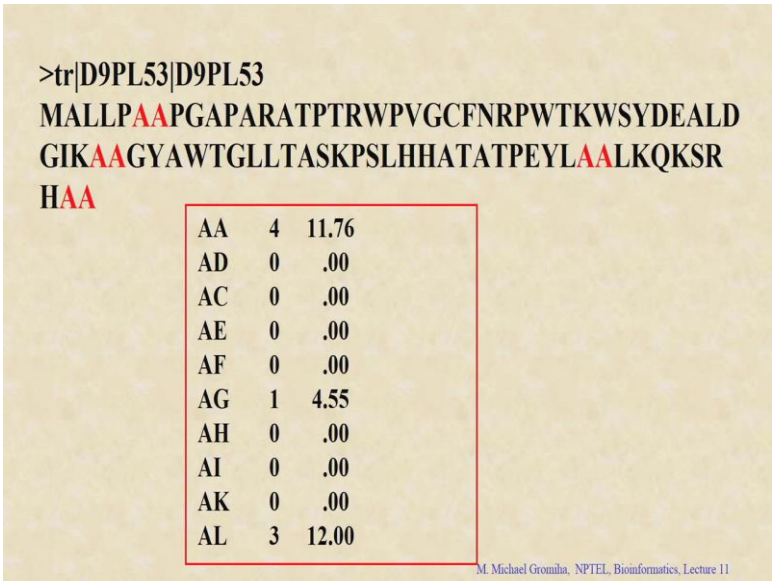
So, i normalize Ni plus Nj, but if you want to get the probability, your i residues of type i i now a this is of type j and this type j then you can use as Ni into Nj. So, there are different ways to normalize, but they are related to each other right, but eventually the one which gives the importance is Nij number of times the residue i which is come close to j. This is distribution of residues and the questions i and i plus 1; here Ni is the number of residues of type of i and Nj is the number of residues of type j.

For example if we use this equation and I give a dipeptides see if you all for example, if you see dipeptide preference for PE how many times PE occurs here? 1 2 that is it two times how many P's and how many E's.

Student: 1 2.

1 2 3 4 5 6 7 8 9 10 5 P's and 5 E's is equal to 10. So, we get values will be 20.

(Refer Slide Time: 21:25)



Likewise if you see this real protein, here you can see now number of AA's totally how many times A comes? 1 2 3 4 4 times AA for total AA's is 17 times. So, if we get this we will get the value of 11.76.

If you see interestly if you see other case is totally 0, keep some 2 classes one class it is 0, another class if you can get some numbers then we can say that that pairs are important that pairs can be able to discriminate different classes. This case it can perform better than the specific composition, just we use only one residue preference for any particular protein.

So, in this case dipeptide proteins preference have more information, but the drawback is there we have 20 values, but here we get 400 values. So, if your data series is very high then it is good to use the residue pairs data series is less because we get several zeros, but that is not so good for that discriminations. So, in this case you better reduce the number of pairs or reduced amino acid to amino acid composition.

So, summarize what did we discuss today?

Student: (Refer Time: 22:44) lot of features.

A different features because sequence what can we do what are the different features we discussed.

Student: Composition.

Composition.

Student: Occurrence.

Occurrence.

Student: Molecular weight.

Molecular weight.

Student: Hydrophobicity.

Hydrophobicity different properties pair preference and so on. There are also we explain one example to distinguish between the 2 types of proteins group A and group B using amino acid composition. There are several ways we can use the several features, you can use 2 for the distinguish in different types of proteins based on the structures or function and so on.

Next class we will discuss about the more details of the hydrophobicitys and the how to construct profiles, and you can use the potential applications of the different aspects from the features obtained from the primary sequence, then we go with the secondary structures secondary prediction and so on.

Thanks for your kind attention.