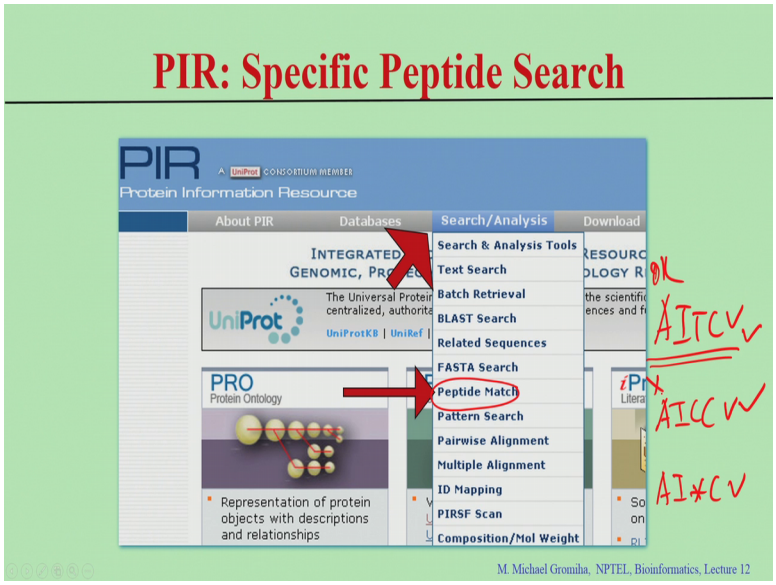
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 12b
Patterns and PSSM Profiles

(Refer Slide Time: 00:16)



So now this PIR, we can also use this to get the peptide match, in the patterns and peptide, what is the difference?

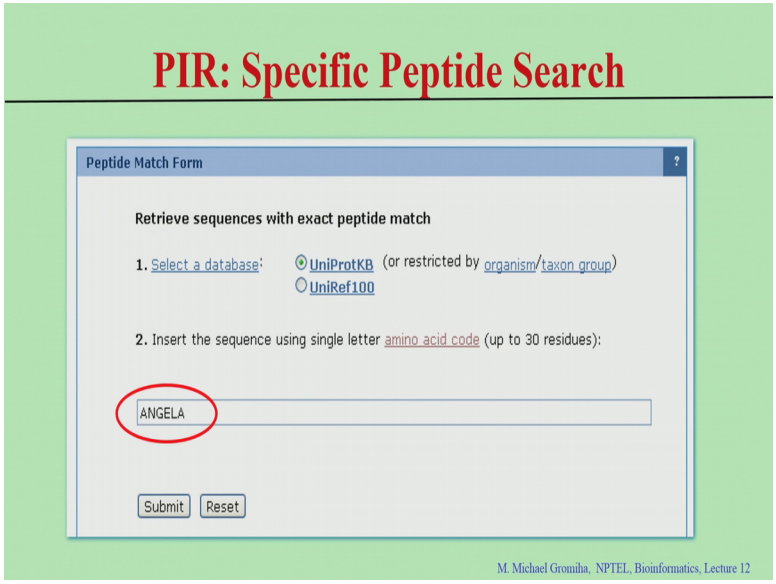
Student: peptide will be specific

Peptide means exact match.

Student: Exact.

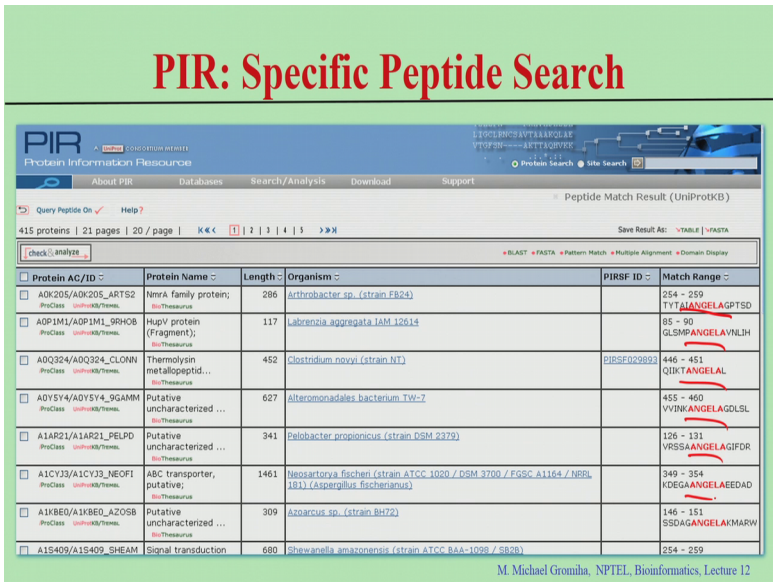
Right for example, some peptide tend to aggregate right. So, even if you make a small change, once residue change, the character will change. In this case you require the exact match for example, if you have AITCV, this is yes, but if you get AICCV this is not. So, if this is the case, we need to find exact match. So, here you can do the peptide match. So, it gives the peptides and exactly you can match. In the pattern search you can write in this way, AI*CV, this you can include, this and this, but in the case of peptide match this is, but this is not.

(Refer Slide Time: 01:09)



So, for example, if you give any specific peptides and you search in this any database.

(Refer Slide Time: 01:16)



And if you see these are the proteins which have exactly the same peptides. For example, if you are interested to find the peptides which aggregate right. So, you can give the peptides and check all the sequences and see which sequences contain that particular region and are there any similarities or differences, you can do analysis, then you can do lot of projects on this aspect.

There, once this is done then we can also construct some type of matrices. We discussed about some motifs, and discussed about some patterns,and discussed about the exact match. You look into different sequences and then you can try to develop some sort of matrices. Earlier we discussed about multiple sequence alignment and conservation score. What is multiple sequence alignment?

Student: Aligning two or more (Refer Time: 02:06).

Aligning 3 or more sequences, if you have 3 or more sequences, you align together and finally, you find the alignment. How to use the alignment for conservation?

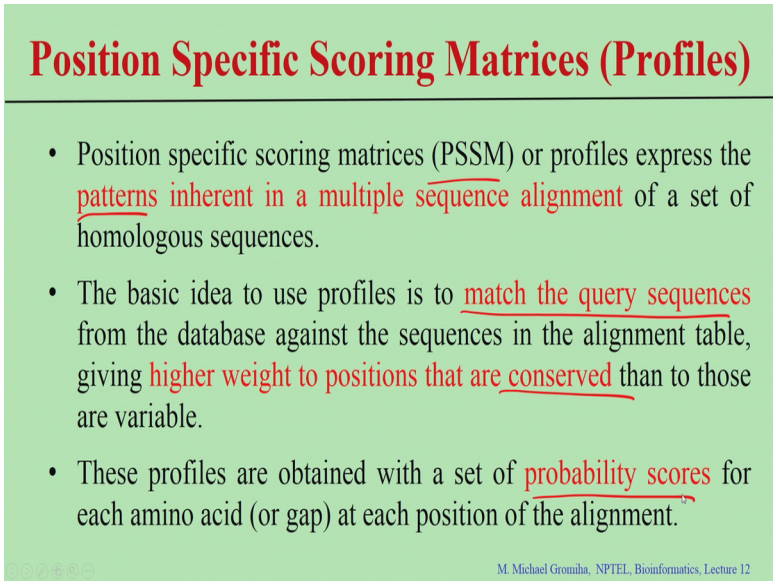
Student: find the residues.

Yeah we can see the residues.

Student: Which (Refer Time: 02:21).

Which occur in the same position in different sequences. Based on that, we can calculate the conservation score and see which residues are conserved.

(Refer Slide Time: 02:27)



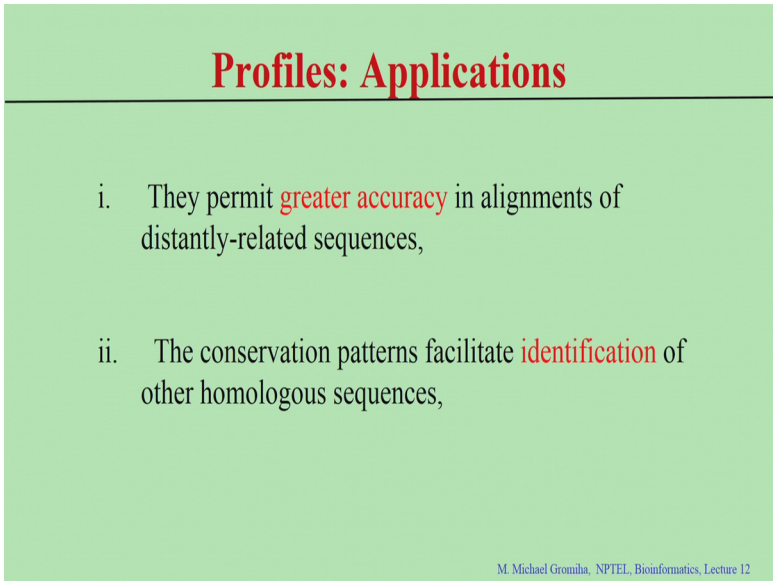
So, here PSSM. This is the position specific scoring matrices, this also provides the patterns which are inherent in the multiple sequence alignment, in these homologous sequences. How to do this? I will explain. In a multiple sequence alignment, you can see whether you can find any inherent patterns. So, how to do this? The basic idea is to

match the query sequences against the database sequences and the alignment table and give that high weightage to the conserved position. Giving high weightage for the positions which are highly conserved than their variable, then you can see some type of matrices. This will tell you, for example, if in position number 15 Ala is present, what is the probability of having Ala at position number 15.

This is fine or not, if it is good then you can exactly match then for many prediction algorithms, they use, they construct a matrix, then if going to new sequence, and see what is the probability of particular residue at particular position, if this is the case that can be part of helix. We can use this matrix for many prediction algorithms, how to obtain these profiles?

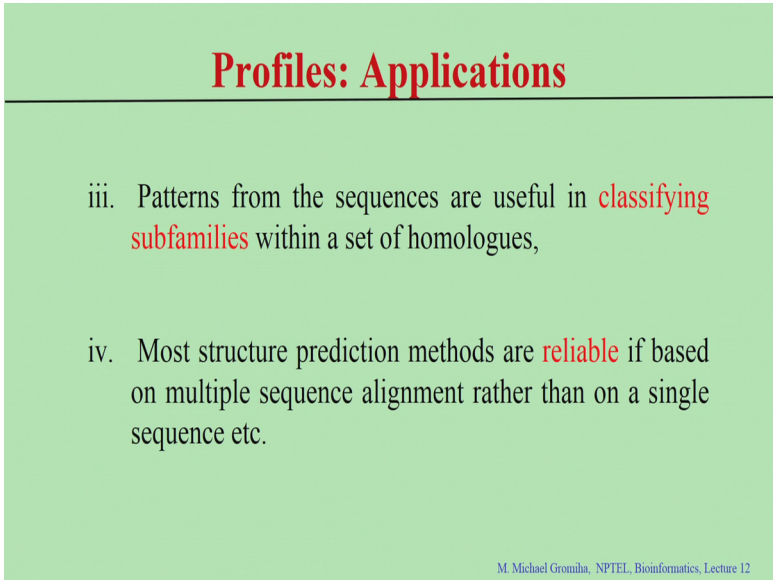
So, these profiles we can obtain with a set of probability scores; for each amino acid, for new protein we have a probability score, that a particular position in the n during the alignment.

(Refer Slide Time: 03:45)



So, in this case what are the various applications compared with other measures for example, conservations and all? So, it gives greater accuracy, when you align the different distantly related sequences and the conservation patters also facilitate the identification of the homologous sequences because we have the high score in the case of the homologous sequences.

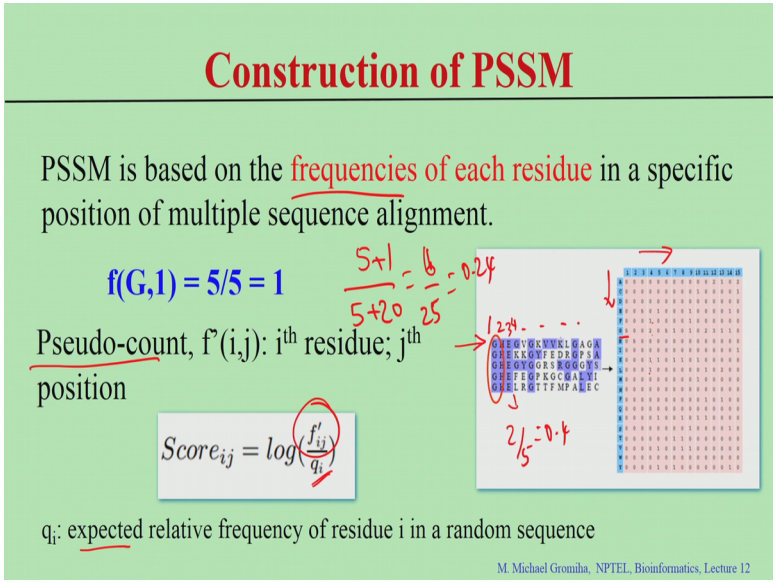
(Refer Slide Time: 04:04)



And you can see the patterns which are useful to classifying the subfamilies within a set of homologous sequences.

Then if you look into the various secondary structure prediction algorithms or predicting the accessibility or predicting the binding sites. Different prediction algorithms, they are reliable if they have the multiple sequence alignment, rather than a single sequence. Getting a single sequence they can construct profiles like hydrophobicity profiles or you get the propensities, rather than single sequence if you use the multiple sequence alignment, I will discuss in the next classes, when we discuss about secondary sequence prediction. You can get that the prediction results are more reliable than using just a single sequence.

(Refer Slide Time: 04:46)



So, now to construct PSSM matrix, I will give the examples how to construct the matrices. So, what is the basis for the PSSM?

Student: Sequence alignment

Sequence alignments. So, it gives you the frequencies of each residue at each position, in the multiple sequence alignment for example, if you see one sequence here GHEG VGKVVKLGAGA. So, I have few sequences for example, 5 sequences, they are used in the multiple sequence alignment. From this, based on the first sequence, the one which are the same residues or conserved residues they are highlighted. So, we take this as position number 1 2 3, take the position number 1, what is a probability of having G in position number 1, what is the frequency of occurrence of glycine in position number 1.

Student: 1.

So, you can say 1 because there are 5 cases 5 sequences, and among the 5. So, 5 if all are glycine. So, 5 that is equal to one likewise, what is the occurrence frequency of glycine in position number 4?

Student: 2 by 5.

That is 2 by 5, this is 2 divided by 5, basically 0.4. So, now, we have the 20 different amino acids residues here and the sequence here. For the first case it is glycine, here it is

5 5 and the second one you can have a residue, have the 5 in the case of H and the third one we have 5 for E and the fourth one distributed, G 2 times, F 1 time, and K 1 time, L 1 time distributed. So, now, if you have the number sequences and make the matrix, several cases, if their numbers are 0.

If you see there are many zeros; in the case of zeros they add few numbers, add a number, this is called a pseudo count, to avoid these zeros because when you take the logarithmic. So, what they do for example, if you take the amino acids sequence, what is the probability of each residue to be in a particular position?

Student: 0.05

0.05, 1 be 20. So, what they do? So, they add one at the top then 20 at the bottom and then they find the pseudo count for example, in this case.

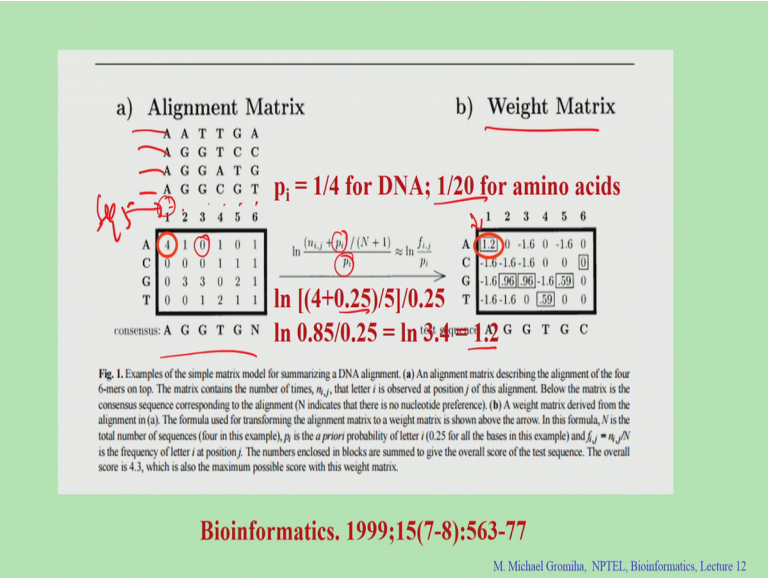
They put 5+1 divide by 5+20, what is the value? Then 6 by 25 right. So, we get the number 0.24, and then this is the one f'(i, j) then they normalize with this q, it is expected frequency of a residue i at the particular position, like in a random sequence. These are shuffled sequences, then what is the expected frequency of a particular residue i at a particular position, because all the residues are not randomly distributed in amino acid sequences. If all the residues are randomly distributed, all the sequence, whatever the sequence you take.

Each residue should have 0.05, but this is not the case, which I showed the data from the UniProt some residues are biased for example.

Student: (Refer Time: 08:04).

Leucine, valine they are highly dominant, and tryptophan, cysteine are less frequent in the in amino acid sequences. So, if you take any sequence, what is a random sequences, what is the frequency. Now divide by q_i then take the logarithmic, that will be a score.

(Refer Slide Time: 08:21)



Like the PSSM there is a another matrix, that is called weight matrix. This is also frequently used in the prediction algorithms, how to get the weight matrix. For example, if you have 4 different sequences, this is a DNA sequence. So, first we construct the matrix, all occurrences of which nucleotide at different positions. So, how many positions? 6 positions. So, it is 1 2 3 4 5 6. So, how many sequences 4 sequences right.

So, you can put first A to all the 4 sequences, you can have this A. So, you can have 3 Gs, and 1 A and the third 1 T, and 3 Gs, you can fill this matrix based on the sequence alignment. Then you convert this into scores in due weight matrix; if you look into this alignment you can see consensus series A, this is G because this is 3 times, here G 3 times, here G 2 times, here R, everything has equal probabilities. So, they put n. So, they have put this consensus. So, here also they make a pseudo count. So, they add the probability for example, if you have one more sequence now there is 4 sequence if you add one more sequence here what is the probability of having A at this particular position.

Student: One

One.

Student: First position (Refer Time: 09:48).

Yeah for example, if you add any sequence here, sequence, this is sequence 4, and you put sequence 5. I can write anything here, what is the probability of having A in this position?

Student: 4, 1 by 4

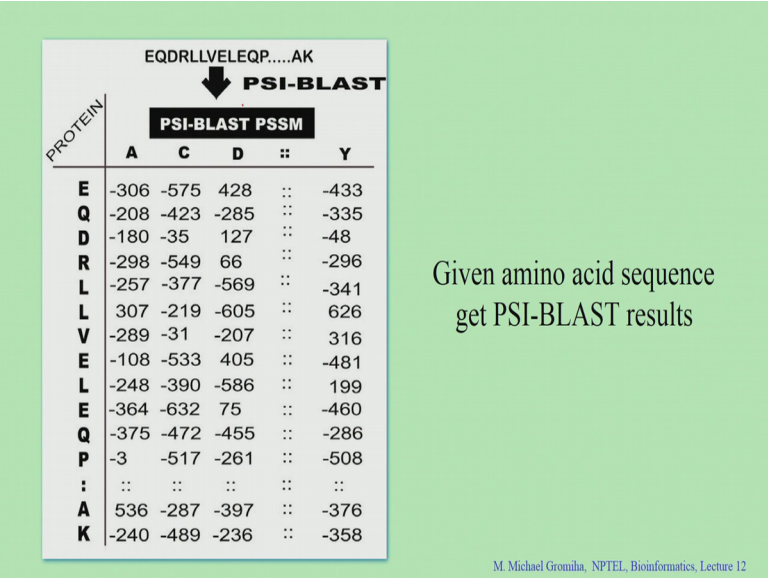
1 by 4 because it can be A, it could be the T or C or G. So, probability of having A is 0.25. In the case of protein sequences, if you have a several sequences and if you add 1 and the probability is 1 by 20 right. So, here is a p_i. So, this is equal to 0.25, in the case of DNA that is 1 by 4, this is equal to 1 by 20 for the case of amino acids. So, now, they use n_{ij} here n_{ij} equal to 4 plus the probability if you add one more what is the probability of having that nucleotide? This is 0.25 divided by the number of sequences, will be added to 1, 4 plus 1. So, here 4. So, now, add one more sequence. So, 4 plus 1 equal to 5. So, in this case you can get the values, this equal to 4 log(4) plus 0.25 divided by 5 right.

Normalize this to this p_i that is 0.25; you do the calculations, this is equal to 0.85 due to 0.25, this equal to log(3.434), this equal to 1.2. So, this is the number. For example, if you take the 0, what could be the value the weight matrix, this is 0.

Student: 0.

0 plus 0.25 divided by 5 then the whole thing normalized with 0.25. So, finally, you get the value of minus 1.6 right. So, we can convert this any alignment matrix you have lot of sequences or the alignment data. So, you can convert this matrix into weight matrix. Now from this weight matrix, now if we have new sequence or you can have to identify the second structure or any specific binding site of any new sequence. You can compare if this belongs to a binding site, then if you have the same A in this particular position then you can see that this A could be probably a binding site residue with any particular protein right. So, likewise they use this weight matrix for identifying any binding sites or any applications.

(Refer Slide Time: 12:09)



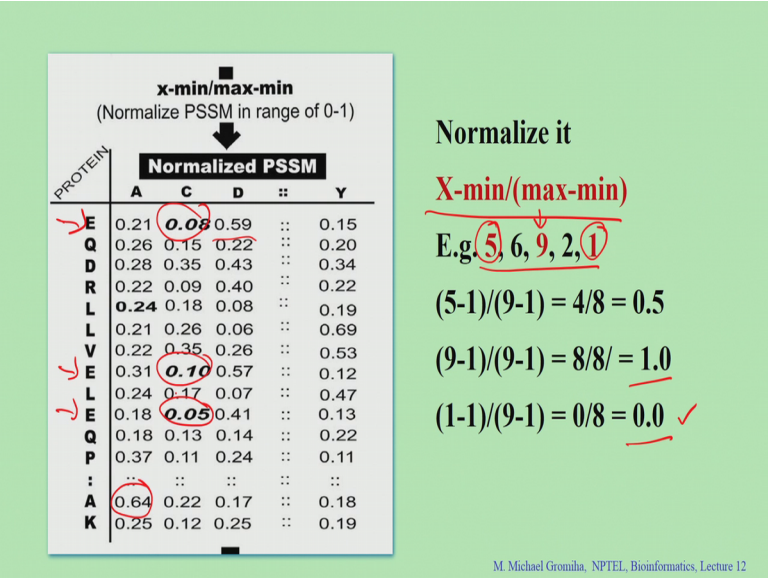
And there is an one more case. So, you see PSI-BLAST, what is PSI-BLAST?

Student: position specific iterated

Position specific BLAST right. So, in this case, we align several homologous sequences and then use different iterations for example, if you this is a protein and this is the possible 20 different amino acids, they gave some big numbers. So, big number. So, first we need to normalise these numbers. So, then we know which one is the preferred one which one is not preferred one because it has you know positives and negatives scores depending upon the preference of these residues at any particular position. So, this case you can normalise between 0 and 1. So, how to normalise the values between 0 and 1?

Student: Maximum minus minimum.

(Refer Slide Time: 12:50)



So, you can use this equation.

X minus minimum.

Student: Um.

Divided by maximum minus minimum for example, this is the 5 numbers for example, this is the lowest one. So, what is the expected value for the lowest one?

Student: (Refer Time: 13:05).

0. If you use this equation, 1-1 divided by 9-1, this equal to 0 by 8, this equal to 0. So, that is fine. So, if you have the highest one.

Student: (Refer Time: 13:15).

Right what is the expected value?

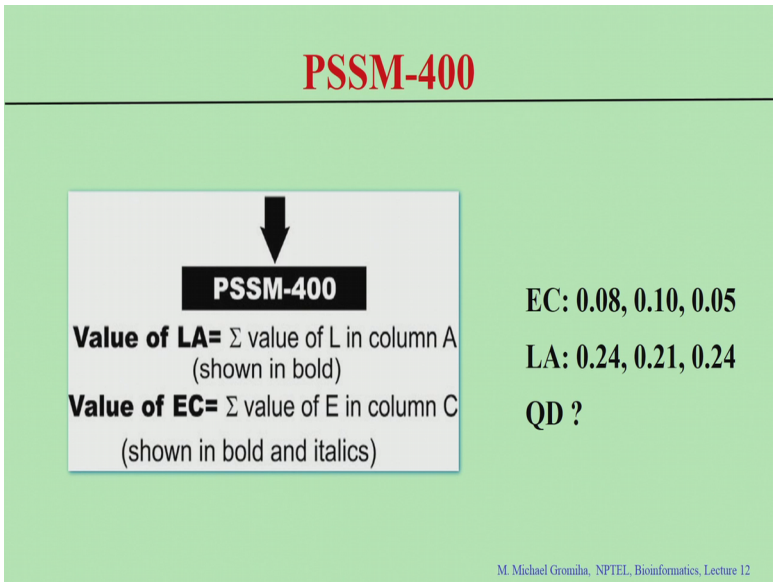
Student: one.

That is equal to 1. So, if you use this one 9-1 divided by 9-1 this is equal to 8 by 8, that is equal to 1. So, now, if you take any other values, 5 or 6 or 2, what is the expected value?

Student: Between 0 and 1

Between 0 and 1. So, we take this 5. So, we get this numbers 5-1 divided by 9-1, this equal to 4/8; this equal to 0.5, that is fine. So, we can use this equation to normalise. So, now, if you see these numbers, very high numbers, they are normalised between 0 and 1. We can see some numbers, such very high and some numbers are very less, for example, if you see any high numbers this is high 0.64. So, here this is a 0.59, some numbers are very highly preferred and some residues, this is the less preferred ones.

(Refer Slide Time: 14:08)



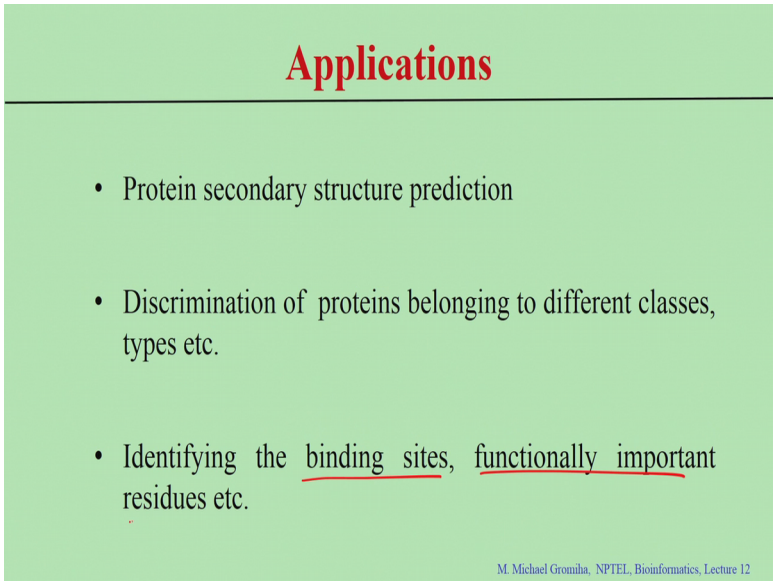
So, now if you got different numbers, it is also possible to combine these numbers and make a 20×20 matrix. Instead of using different numbers, it is also possible to take 20×20 matrix, this case you can uniformly use this for different cases. How to do this? For example, if it is EC, they take all the E's here, here is a E, here is a E, here is a E and C, C here, here, here. So, take all the values, taken together, finally, you will get the average value of E and C; likewise if you see LA, combine all the LAs and you got the numbers; if you take QD, what is the value for QD? This is Q here, yes, another Q here.

Student: (Refer Time: 14:54).

Right. So, D is here, this one, this is one, this is one. So, we add up these numbers, and we will get the average value, we will get the matrix. So, finally, we will get the 20 by 20 matrix this also you can use for several applications or several prediction algorithms. So, we discussed about the position specific scoring matrices and weight matrices right. So,

there are several potential applications for example, the major applications if you see protein secondary structures.

(Refer Slide Time: 15:15)



when you have a new sequence, if you give the amino acids sequence and there are several secondary structure elements for example, alpha helices, beta strands and the coils, and all we'll discuss in the next class right. So, how to predict which regions can should alpha helix and which region is responsible for this formation of beta sheets, you can use the PSSM right.

You can get the predicted secondary structures then you can also discriminates proteins from different classes and different functions and based on different structures, that you can also use this PSSM to get the information. Then also you can also identify the binding sites or functionally important sites. So, here also you can see which residues are functionally important in any given protein. Likewise we has several applications of bioinformatics is here, yes you can relate this data with experimental data and we will see this PSSM as potential applications in various prediction methods. Till now we discussed on various aspects. So, we will summarize what are the various aspects we discussed today.

Student: hydrophobicity profile.

Hydrophobicity profile. So, what is hydrophobicity profile?

Student: Sequence versus.

This is sequence versus hydrophobic plot, this is a plot connecting amino acid sequence versus hydrophobicity values. So, what are the applications of constructing hydrophobicity profiles?

Student: identifying

You can get some patterns, and these patterns resemble specific regions in a proteins based on structure or based on function, typically for example, helices.

Student: Helices.

Strands.

Student: Strand.

Transmembrane regions, where you have binding sites, we can see different patterns of hydrophobicity and we can use these patterns to identify these regions. Then we discussed about specific patterns right. So, in PIR, you search for patterns; how to write a pattern?

Student: Using (Refer Time: 17:13).

Square bracket.

Student: curly braces.

That represents what?

Student: Any residue.

Any residue, any allowed residues and then any conserved residues, that is how to find the conserved residue.

Student: (Refer Time: 17:27) write the name.

Write the name write the name of the same amino acid then any residues.

Student: X.

The x number of times.

Student: N.

Put n. So, a a any number and curly bracket represents.

Student: excluded

If the residues are not allowed in that particular position. So, you can write the patterns right. So, what is the importance of a identifying patterns or the peptides.

Students: (Refer Time: 17:48) identify motif.

We identify motifs, but you can see any inherent residues.

Student: (Refer Time: 17:52).

Right for any specific features we can check the UniProt sequences and see whether they're very specific for any type of proteins based on structure or function. Then what did we discussed after that?

Student: PSSM.

PSSM.

Student: (Refer Time: 18:08).

Or position weight matrix, what are the principle based on PSSM or weight matrix?

Student: Positional conversation.

Positional conservation, it will be the positional conservation for that multiple sequence alignment, they take the alignments and we get the frequency and they convert to the frequency into score.

Student: (Refer Time: 18:26).

Right you can have a matrix. So, you can use this matrix and the matrix has several potential applications. So, in the next class, we will discuss mainly about using several features and making these features requires a dataset, and how do we construct a dataset,

what is redundancy, what is non redundancy, how to define this type of redundancy and what are various methods, which are available in the literature, to obtain these non-redundant sequences and so on. And later on we will see in the potential applications and you can merge these features and the data set and find your problem and then you can use to see; what are the potential applications of using different Bioinformatics tools.

Thank you very much.