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

Lecture - 7a
Sequence Alignment

In this lecture we will discuss about the different algorithms for alignment protein sequences. So, already we discussed in the last lecture, you have different types of alignment you have two sequences, what are the information you can directly get from the two sequences, immediately what can we observe one of the easiest one is.

Student: Dot plot.

Dot plot compares two sequences if the sequences is a same in amino acid it is a same, then you put a dot. Then if you make plot you can observe with any matching sequences exact matches or we can see deletions or insertions in any of the two sequences. Then we discussed about the aligning sequences with some scores; the its only we can see the plot and look at the regions, where you can see the similar residues or same residues then we can give some scores.

For example if here is a matching amino acid and matching nucleotide, we give a score where you work orbits mismatching we give penalty, then we also introduce gap. So, what is the many of gaps in alignment?

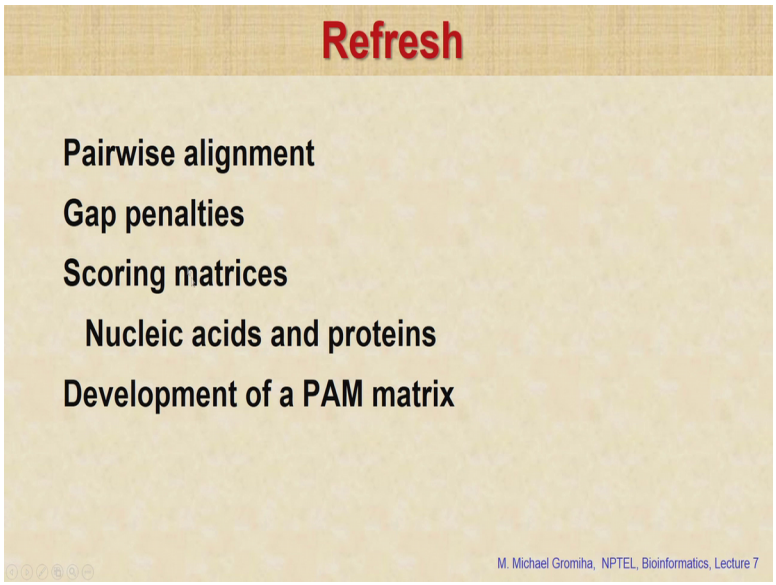
Student: Insertion or deletion.

Insertions or deletions, right. So, in this case comparing the sequences which you have mutations or substitutions, the insertion and deletions are rare. So, we give penalty when we introduce a gap. So, now, we have 3 different aspects one is a match, mismatch and the gap. So, we give a reward for a match and we was score for the mismatch, and the penalty for the gaps. Then the gaps we have two different types of gaps what are two type different types of gaps?

Student: Gap.

Origination penalty plus gap penalty how many times gap is introduced and we have totally how many gaps. So, if you have different originations. So, we have more penalties. So, we use these course 12 and 2 sequences.

(Refer Slide Time: 02:18)



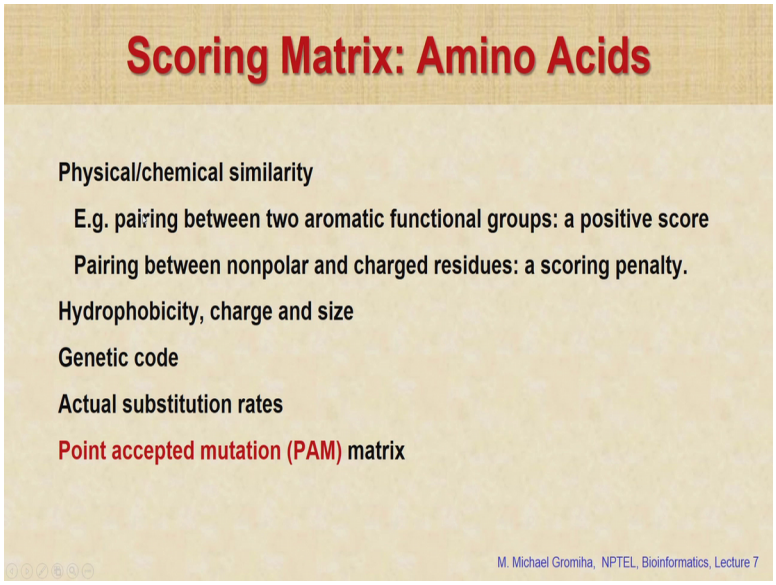
Then we try to construct a scoring matrices to align the sequences; so different matrix for nucleic acids and sulfide proteins, but the glycine nucleic acids how many bases.

Student: (Refer Time: 02:30).

4. So, in this case based on a substitution try to either you want or the penalty either small penalty or similar penalty.

For example if we have purine to purine or pyrimidine to pyrimidine or other way purine to pyrimidine or pyrimidine to purine. So, we give the penalty accordingly so that you can give preference to match similar sequences. So, when you look into proteins, there are various ways to align the protein sequences, to give weightage to the amino acids it is the one of the physical chemical properties.

(Refer Slide Time: 03:06)



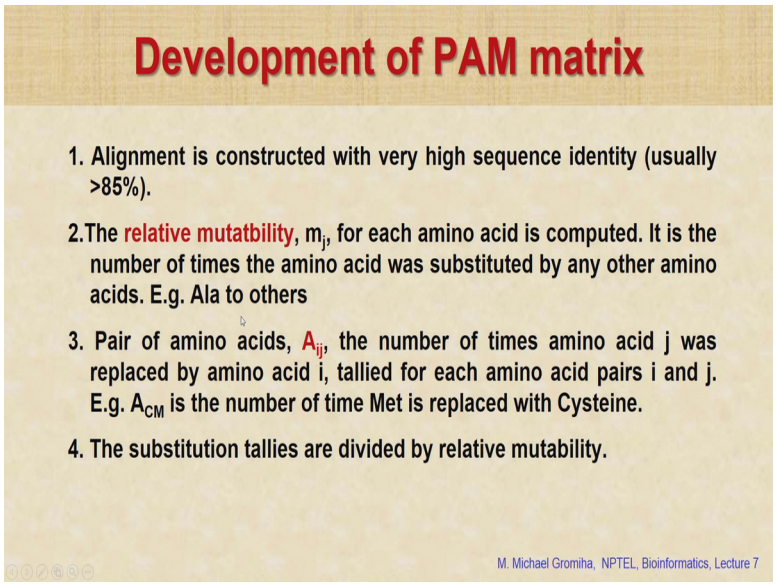
For example if the two residues are charged for example, aspartic acid and glutamic acid. So, in this case they are similar this we give the positive score compared with aspartic acid is replaced by valine or any hydrophobic residue. So, this is the pairing of this similar functional groups like either the aromatic groups or the nonpolar and charged groups and so on either you give the score or give the penalty; and we also compare the hydrophobicity or the charge as well as size.

For example if you alanine and the serine this is similar in size. So, if you compared with size. So, they are similar accordingly you can align this sequence you can give this score. Then also be discussed about a genetic code how to give score based on genetic code? Number of mutations in the DNA then the codons right, how many mutations in the codons. So, we which type of mutations in the codons, then we can give weightage based on the genetic codes. Among all these things how we derive the matrix.

Then finally, we take the actual substitution rate for example, if we have set of sequences get the sequences with the high sequence homology, from these sequence similarities we see what are the possible mutations; from that mutations then we can derive the matrix you can see what are the changes actually happen and based on that we derive the matrices. So, this matrix is called the point accepted mutation matrix for example, if you have 100 amino acid residues and NDR same.

So, how many variations 10 percent this can 90 percent of similar. So, what are 10 variations? We get the information and then we see the substitutions what are substitutions probability substitutions.

(Refer Slide Time: 05:52)



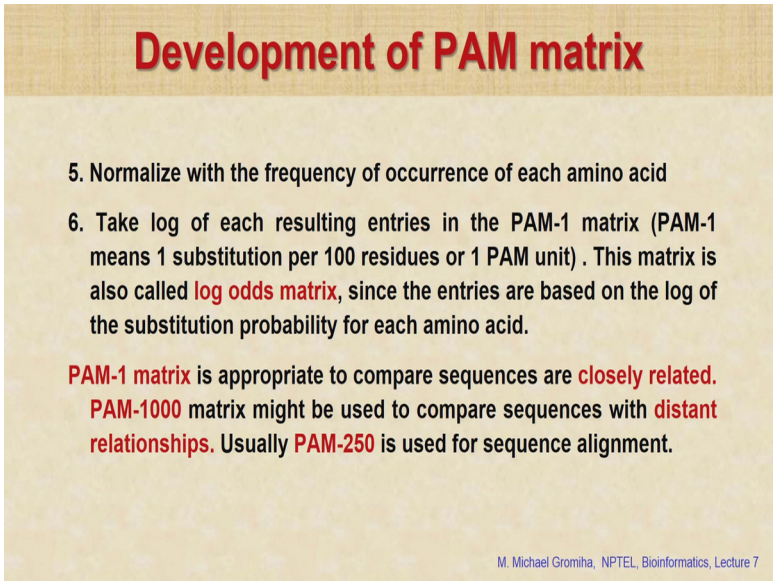
So, here first we take in alignment with the may be at the sequence identity, say more than 85 percent. Now what are the various factors one has to consider to derive a PAM matrix?

Student: (Refer Time: 05:05).

First we need to see for example, a how many alanines in the sequence, total number of residues in the sequence, how far alanine is mutating to other residues, what is the probability of mutating alanine to a specific residue for example, valine. So, here I show that for a relative mutability for each amino acid. How many times its alanine its mutated to others and the second one we need to see the exact mutations for example, aij how many times i is mutated to j.

For example you see A CM number of time methionine is replaced with 16. Likewise we can see the relative mutability of uni amino acid.

(Refer Slide Time: 05:48)

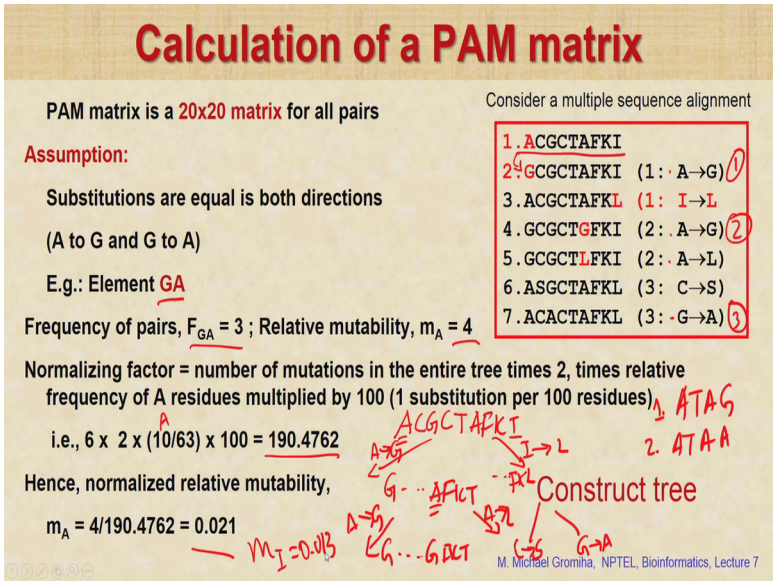


Then one has to consider the number of amino acids and the total number of residues in the alignment then you take the log of these entries to get these PAM matrixes this is we also called this PAM matrix as log odds matrix.

We can derive the PAM matrix based on various sequence homologies; we can see 80 percent, we can see 90 percent or we can see a very less homologies. So, do we see this PAM one matrix? That is to compare the sequence as is a close related and use PAM 1000 for comparing the sequences with the distant relationships. That means, they are very not homolog to each other.

Usually, in the literature we use PAM 250 for the sequence alignment for example, blast or FASTA you use PAM 250 for the sequence alignment. So, I will show an example and how to construct a PAM matrix.

(Refer Slide Time: 06:38)



So, we see one sequence here ACGCT AFKI in the first case A is mutated to G. So, we can construct a tree. So, the first one we have this sequence ACGCTAFIK here A is mutated to G first one is A is mutated to G.

So, here this A comes to G and the second one I is mutated to L this I is mutated to L right. So, everything is here. So, I instead of here a of FKL, then from this sequence then we have next mutation underline it is mutated to glycine; so here this alanine.

This alanine is mutated to glycine. So, it is with starts G. So, the sequence here also GFKT then again A is mutated to leucine. So, here look at this a is mutated to leucin then go with the second one here we have two mutations one k is C to S and the second one G to A. So, we have the tree. So, we can make the alignments from this tree and the number of substitutions, we will see the probability of residues to be mutated to another residue.

For example if you see A to G or G to A as we discussed in last class we take first sequence and second sequence, if a is mutated to G for example, if have this one here the mutation is G to F this is sequence one, this is sequence two this is G to A. If it second one as sequence 1 and the first one as sequence 2 then the mutation is A to G right. So, we have to consider these substitutions are equal in both directions whether A to G or G to A; fine take the element GA. So, how many frequency of pairs GA, how many mutations involved G and A?

Student: 2 3.

1 2.

Student: 3.

3. So, FGA equal to 3, what is relative mutability ma how many times A is mutated?

Student: 4.

4 1 2 3 4; so it is equal to 4 1 2 3 4. So, it is a normalizing factor, this is a number of mutations in the entire times multiplied by 2 times relative frequency of A residues multiplied by 100. So, is the totally number of mutations in the entire tree. How many numbers of mutations? 6 mutations in the entire tree multiplied by 2 then times relative frequency of A what is relative frequency of A; totally how many alanines.

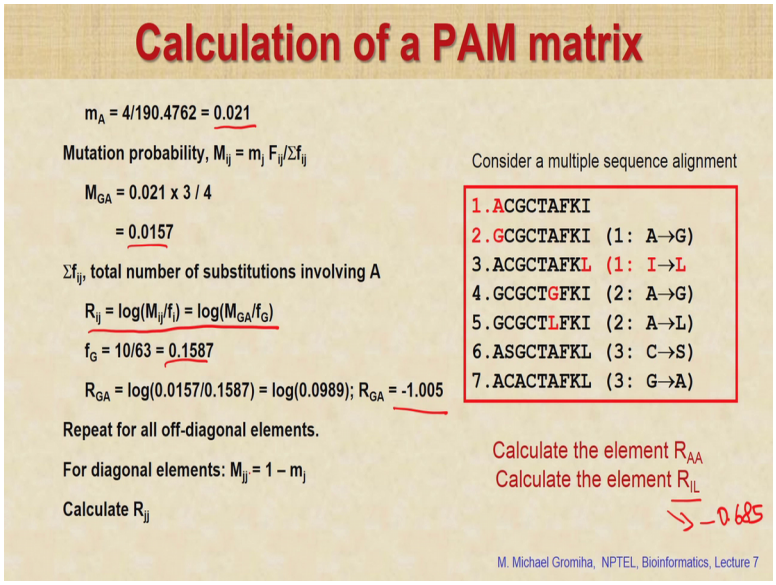
Student: Totally 10.

10 alanines, right. So 10 A totally how many residues?

Student: 63.

63; 1 2 3 4 5 6 7 8 9; 9 in to 7 is equal to 63. So, we will get this number 190.4762 then we get the normalized mutability see because this ma is 4. So, 4 divided by this number we will give you the normalized relative mutability. So, this is 0.021 so this here.

(Refer Slide Time: 10:12)



So, now we get the mutation probability M ij, this is given as m j and F ij divided by sigma F ij. So, here F ij that is MGA this is equal to 0.021 multiplied by what is F ij?

Student: Frequency of.

That is equal to 3 divided by sigma F ij this equal to 4. So, we get this MGA its equal to 0.0157. So, F ij is the total number of substitutions involving alanine its 4 right. So, this is equal to 0.0157. Now we get the R ij this is the value we get logarithm of this M ij divided by fi this frequency of G. MGA is given as 0157 right. So, fG we get this is out of 63 10 glycines earlier we take the value of alanine now we take value of glycine.

So, is equal to 0.1587 then RGA substitute value is here log of M ij equal to 0.0157 divided by 0.1587. So, we give the value of minus 1.005. So, we can repeat this for all the off diagonal elements.

For example if you want to get the value of RIL what is the how to calculate RIL, how many times I to L mutations.

Student: 1.

Only one and the how many I involved in this a mutation? One only one right. So, now, we can calculate normalizing factor, that is same as here 6 into 2 multiplied by how many out of 63 how many is.

Student: 4.

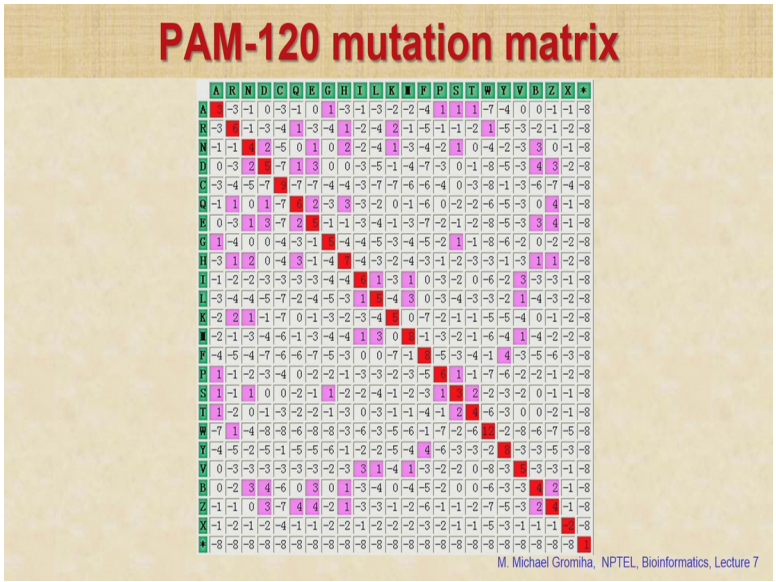
Four right. So, normalizing factor equal to 6 multiplied by 2 multiplied by 4 by 63 divided by 100. So, we get the value of 76.2, now we get these normalized relative mutability. So, it is only one a solution mutilation 1 by 76.2. So, this is equal to m I equal to 0.013, then we get this m I solution to solution MIL we can get this numbers right. So, we get 0.013 multiplied by only 1 divided by only 1. So, this is equal to 0.013. So, now, we get the R ij. So, we take the logarithm of these values finally, you get this value as minus 0.685 we can work out in the free time.

This is for the off diagonal elements for the diagonal elements you will get the M jj; M jj is here for example, (Refer Time: 00:00) AA alanine will get one minus m j m j will get from this a formula and we can get this values then we get the values M jj then we will get the value of R ij we can get that. So, when you derive this matrix right. So, I get large number of data in the protein sequence database what is the protein sequence database?

Student: Uniprot.

Uniprot you can get that data with any sequence identity finally, you derive matrix this is a PAM 120 mutation matrix.

(Refer Slide Time: 13:12)



So, if we see this matrix can you tell something from this matrix, there are some letters are red right. So, all these are.

Student: Diagonal.

The same residues there are no mutations.

Student: No.

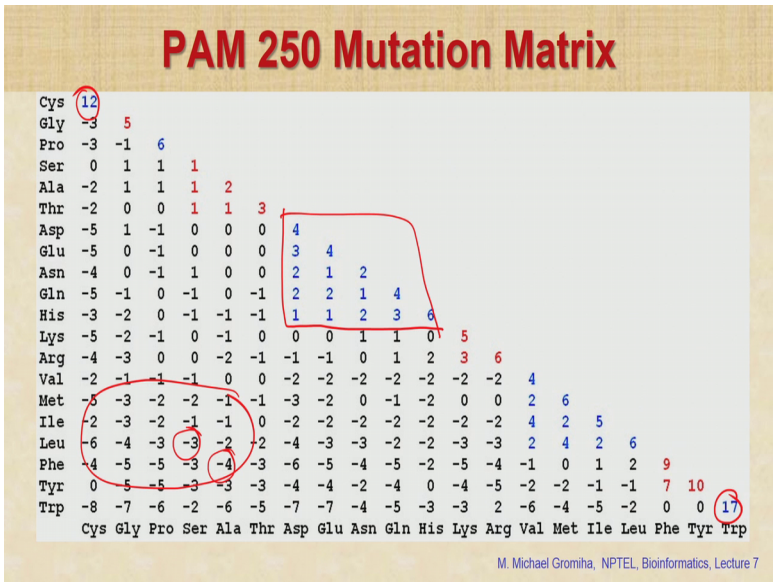
These residues have highest values because the mutations the proteins usually they do not want to mutate at certain specific positions right. So, the same residue they like to be same specific position for some functions or some (Refer Time: 13:42), then next question is for the different cases numbers are different.

For example if you take alanine that is 3; if you take tryptophan the value is 12, you take 16 it is 9. So, there are some amino acids which are rarely (Refer Time: 14:00) amino acids. So, in this residues are very very important to maintain the same position for the structure as well as function for example, 16 is important for the formation of.

Student: Rights of a bonds.

Rights of a bonds. So, if you mutate the 16. So, it give me a adventure adverse effects. So, we show the red is very high compared with other amino acids like alanine or valine.

(Refer Slide Time: 14:24)

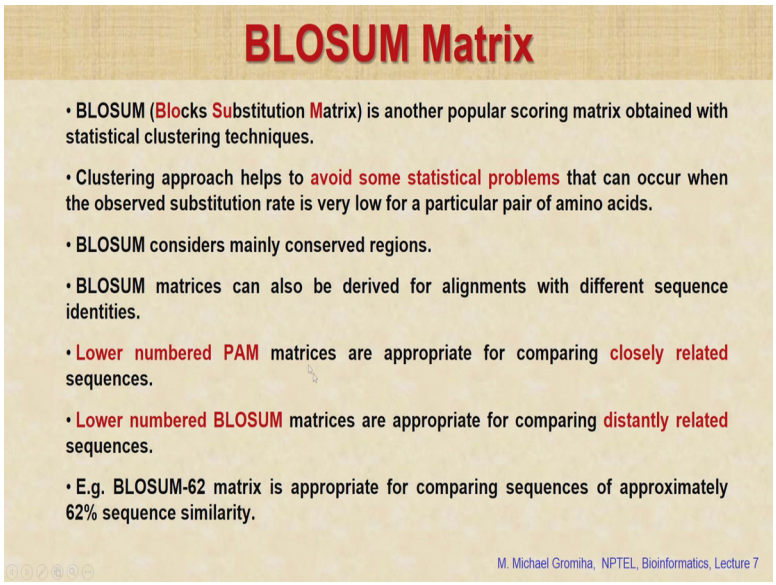


So, we can show another matrix PAM 250 matrix, here also we can see the 6 into 6 it is 12 and the tryptophan frequency it is 17 and also the same amino acids have the high values, other than the same amino acids. If you do look into mutations, some mutations

have positive values some mutations have negative values. Can you see the positive values for example; you see here they are in blue. So, we see they are similar type of substitutions for example, asparagine to aspartic acid or asparagine to glutamine. So, you can have the positive values or the hydrophobic residues or some small residues, likewise if you see some cases we have very adverse effects minus.

For example: this region for example, if you substitute serine valuation or alanine by phenylalanine. So, if you see some mutations are acceptable by nature, some mutations are not accepted by nature. So, based one the real frequency of substitutions, now it we derive the PAM matrix likewise that is another matrix this is called BLOSUM matrix, if you construct the PAM matrix if you align the make the alignments sometimes we can see several gaps; in the case of BLOSUM matrix these also another popular matrix.

(Refer Slide Time: 15:47)



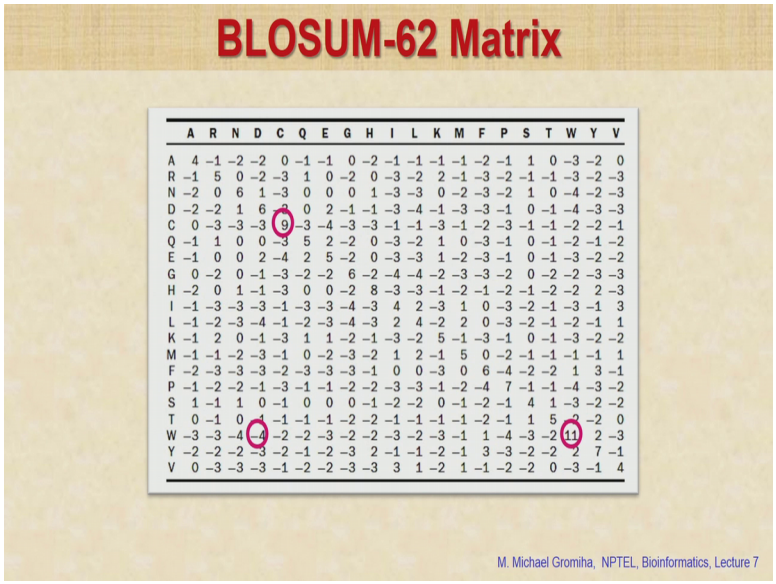
Here they use only the places where they highly conserved, where you can get the proper alignment. In this case they avoid the regions where they have lot of gap. So, here we can see mainly the conserved regions, this can avoid some of the statistical problems whereas, substitution that is very low where in the particular pair of amino acids that reduces some sort of a like PAM matrices.

So, in the case of PAM matrix, lower number of PAM matrix is appropriate for comparing which type of sequences.

Student: (Refer Time: 16:18).

Closely related sequences in the case of BLOSUM they did other way around the lower numbered BLOSUM matrices are appropriate for distantly related sequences. And generally we use BLOSUM 62 for comparing sequences of about 62 percent sequence similarity this is very commonly used in the alignment programs.

(Refer Slide Time: 16:42)



Now, we show the BLOSUM matrix and the compare this data with the PAM matrix. If we see here what is the value goes C to C 9 tryptophan.

Student: 11.

11. Likewise if you have some mutations for example, tryptophan to aspartic acid.

Student: Minus 4.

Minus 4. Here if you see here tryptophan to aspartic acid.

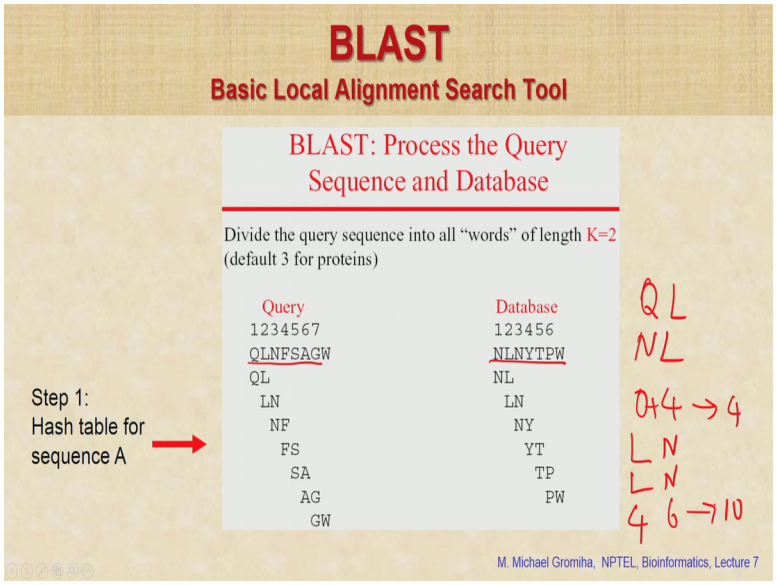
Student: Minus 7.

Minus 7 here are the errors adverse effect adverse effect; so if you look into these two matrices. So, qualitatively you can see that both are similar. So, now, we derive the matrices. So, what is the purpose of deriving the matrices? So, what is usefulness of this matrix?

Student: Alignment.

For alignment for example, now we start the alignment. If we have two sequences if you are not sure how to align then we can use these matrices to compare the similar amino acid residues and score the alignment, I show one example.

(Refer Slide Time: 17:33)



So, now one example is blast, it is a program this stands for basic local alignment search tool. So, they develop an algorithm and it is also available online. So, we can use a tool to get your alignment how it works that I will explain this. So, first if we gather query sequence and we have database. How to map they divide this sequences in to small bits small watts of length k. So, usually they use K equal to 2 or K equal to 3 for the case of proteins ok.

So, now we have the query sequence QLNFSAGW these we compare this with the database NLNYTPW. For example, if you take into word length of 2. So, how to divide this query sequence QL LN and NF and FS and SA AG and GW? So, we make into overlap in segments, then go to the database here also you made into overlap in segments NL LN NY YT TP and pw then where we match first we see this YL QL and NL. So, if we take the QL and LN what is the score for if you take q L and here NL. So, you compare this values what is Q and NQ and N0 L going L L to L this 4.

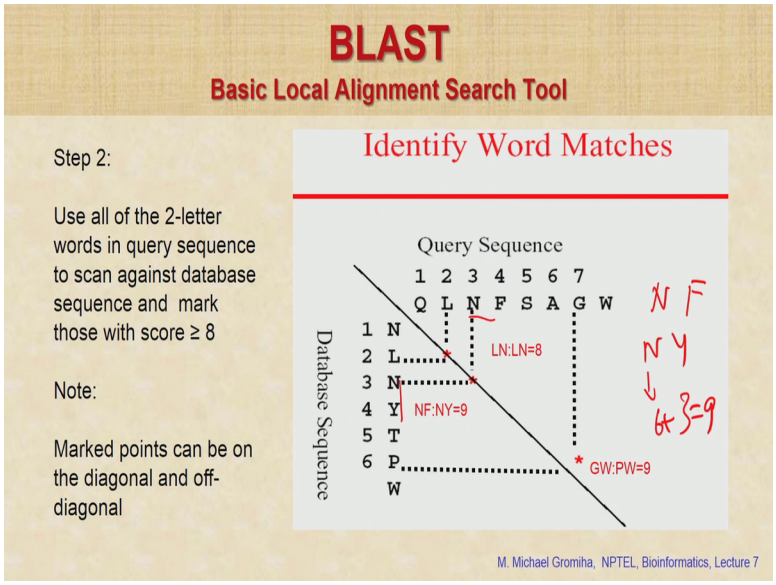
Student: 4.

Right, 4. So, total will be 4. So, if you use LN LN second one LN LN what is score for LL? LL is 4 what NN.

Student: 6.

6 NN 6, 6 and this is equal to 10. So, now, we can use these numbers.

(Refer Slide Time: 19:44)

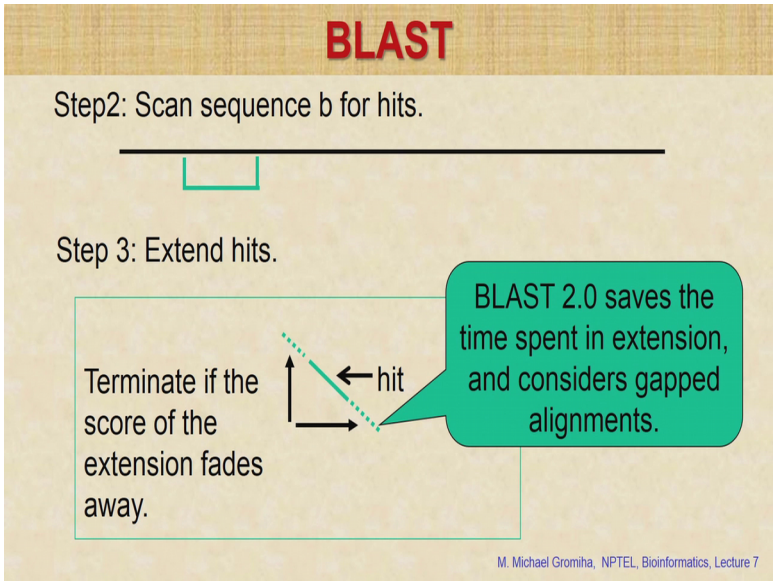


So, we this matrix and you make any cutoff score 8 this is this can be adjustable. So, you can put the example its score of more than greater than equal to 8, and see where we have the values which are more than this greater than 8 here. if it is LN and LN if you get 8 and NF and NY gets score of 9 NF NY what is NN.

Student: 6.

NN is 6 FY Y F is 3 3. So, 6 plus 3 is equal to 9. So, wherever we get the values which are above this is your threshold, then you put a star and then continue this, and try to connect these dots.

(Refer Slide Time: 20:35)



So, you can do this first start we in its small segment and you give this give the dots if it is connect, where connected and if the score is less blast will fades away. Then you can extend it and you can get for the full alignment.