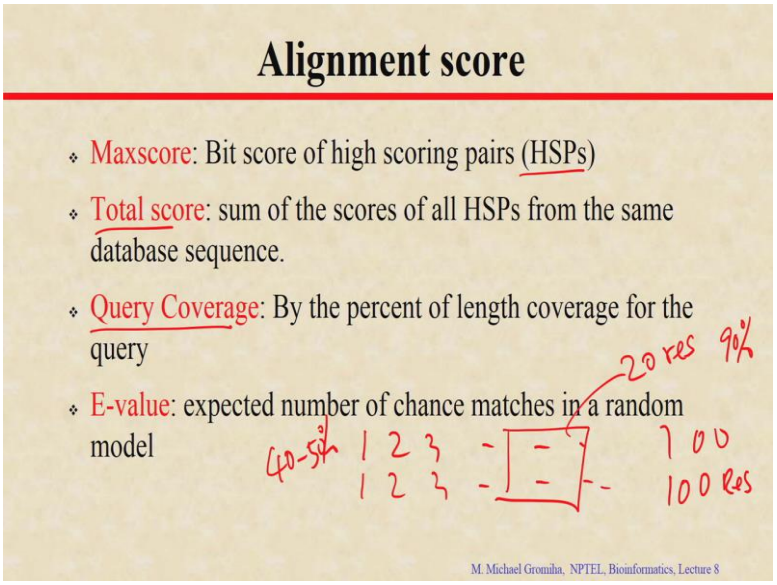
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 8b
Sequence Alignment: Online resources II

(Refer Slide Time: 00:17)



So now what are the difference scores; what are the various parameters used to assess this score? What are the maximum score? This is discussed earlier; this is the high scoring pairs; how many high scoring pairs in the alignment, right? This is a bit score of high scoring pairs. Then the total score, this will give you all the HSPs from the same database sequence, then the query coverage as we discussed; now the percentage of length coverage for the query. If you have 100 residues and the alignment you obtained only for 60 residues, then 40 percent you do not cover, right. In this case, the coverage is only 60 percent, then E-value; what is E-value.

Student: (Refer Time: 00:57).

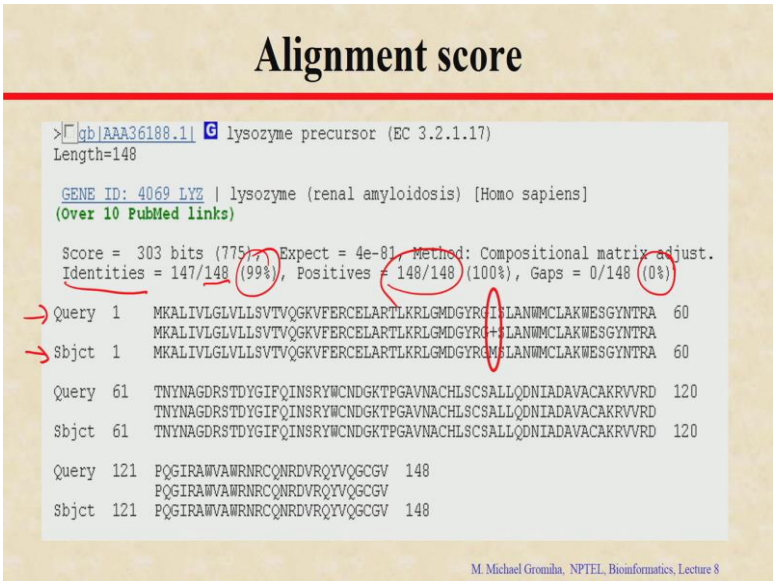
It is expected number of chances, right you randomly you can get with a same alignment, why the query coverage is very important?

Student: Sometime partial match can be used.

Sometimes you will partial match right if you have 100 residues right if you use all the 100 residues, if another match with the 100 residues, then your score will be less number of similar residues or identical residues maybe 40 or 50, but on other hand, if you have some only some segments for example, 10 residues or 20 residues and this have high matching. For example, we have the matching about ninety percent.

Now, if there is a bias, if you have this one; the random choice to get the specific residues is very high right compared with the complete alignment. In this case, you have to take into consideration not the just score, but also the query coverage; there is a reason why they give the query coverage also in the alignment score.

(Refer Slide Time: 02:04)



So, now, align the 2 sequences, this your query sequence and this is a subject one. So, this is the first sequence and second sequence what is the middle if what the residues are the same, then we put the same letter right MKA and so on at one place you put plus what is the meaning of plus.

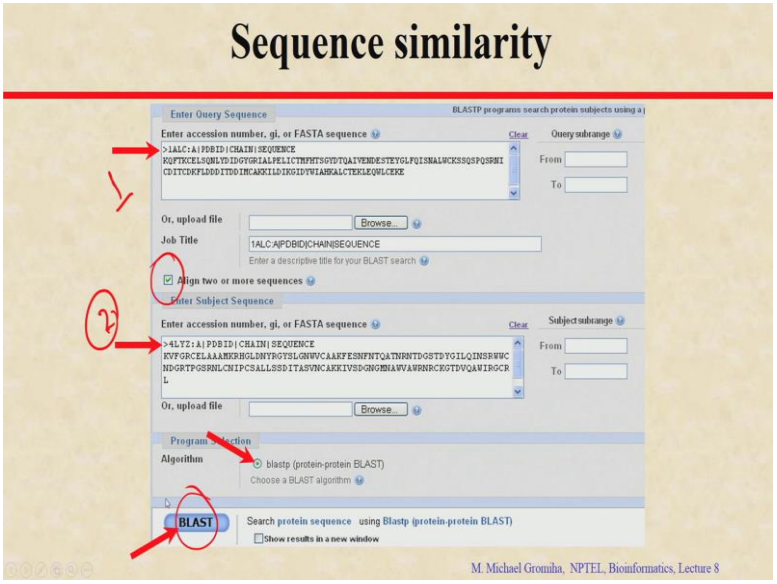
Student: Similarity.

Similarity right similarity from the matrix you use either the PAM matrix or BLOSUM matrix. So, here I and M, they are similar right they are a positive score. So, they put positives. So, based on that they have to calculate the identities; that means, out of 148, the 148 sequence right completely everything is aligned. So, gap is 0 percent. So, in this

case query coverage is 100 percent. So, out of 148; 147 are matching right the identity means exactly the same. So, this is 99 percent and then if you see the positives it contains identities plus the plus that means similar residues.

So, at everything out of 148, all the 148 or positives; so this is 100 percent right they give along with this scores as well as the E-values.

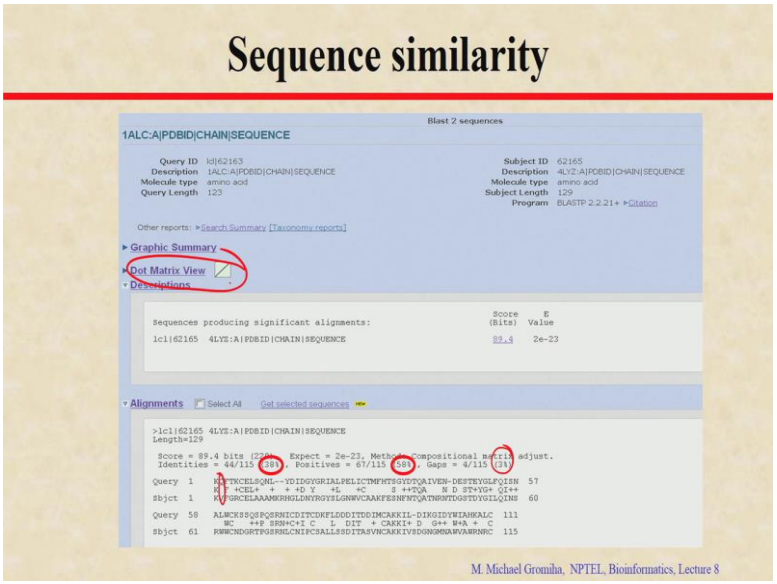
(Refer Slide Time: 03:16)



So, now here if you give the 2 sequences right, the earlier one we give one sequence we have try to find the match all the matches and then see which will has the best match that would be similar to your query sequence.

Now, if you have 2 sequences. So, I gave sequence number one, right and the sequence number 2 and how far these 2 sequences are similar in this case, BLAST can do handle the situation and enter the first sequence right in the first in the first place and here you have to click on this align 2 sequences if you align 2 sequences, then the next box will open. So, in the next box, if you give a second sequence right and then BLAST, to do this.

(Refer Slide Time: 03:57)



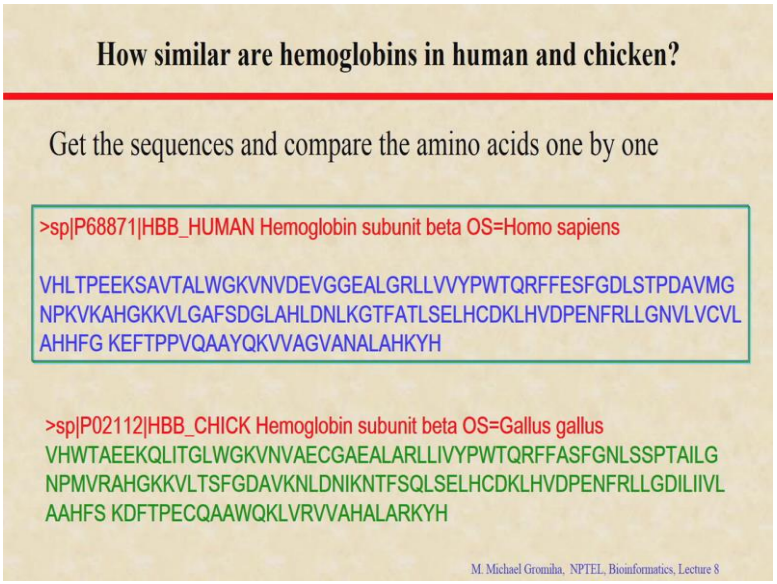
Right, you get this is a query one and this is the subject one.

So, you have the alignment here. So, there are 115 residues; among 115, 44 are same and among this 115, 67 are positives, if you compare this alignment and the previous alignment what is a difference right the previous one it is identities 99 percent and the positives are 100 percent almost complete alignment. So, you cannot say that every all for all sequences you get this type of alignment.

So, if you see here lot of gaps right the residues are totally different you can see Q and V right they are not they have negative score in the BLOSUM matrix or the PAM matrix. So, here their gaps; so some cases if it is same they put the same residue and a plus means their positives in the case of the BLOSUM or a PAM matrix. Now, we here see there are 3 percent gaps right and 58 percent positives and 38 percent identities.

So, be sure always the positives are higher or lower than identity higher than the identities because positives include the identities plus the similar ones.

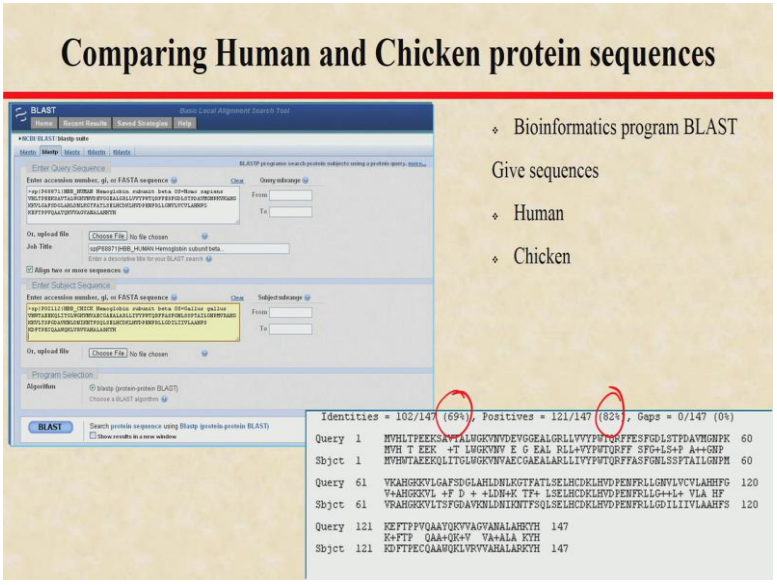
(Refer Slide Time: 05:08)



So now, in the previous class, I showed these 2 sequences as an example to get a dot matrix, if you want a dot matrix here we have the dot matrix view, right. So, if you click on this then you will get a dot matrix for these 2 sequences we did the dot matrix and we found some matches how far they are similar how many percentage this human and chicken hemoglobin are similar. So, what to do?

BLAST.

(Refer Slide Time: 05:32)

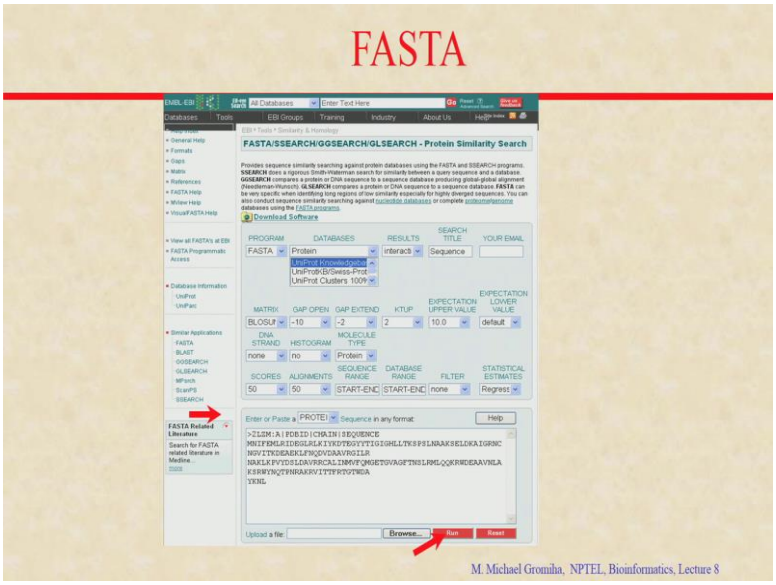


So, here if you give the first human sequence and the chicken sequence; so now view the alignment; how much is the identity.

Student: 69 percent.

69 percent identity and the positives are 82 percent, there is a completely aligned without gaps they aligned right. So, this is the sequence identity between human and chicken protein sequences.

(Refer Slide Time: 05:56)

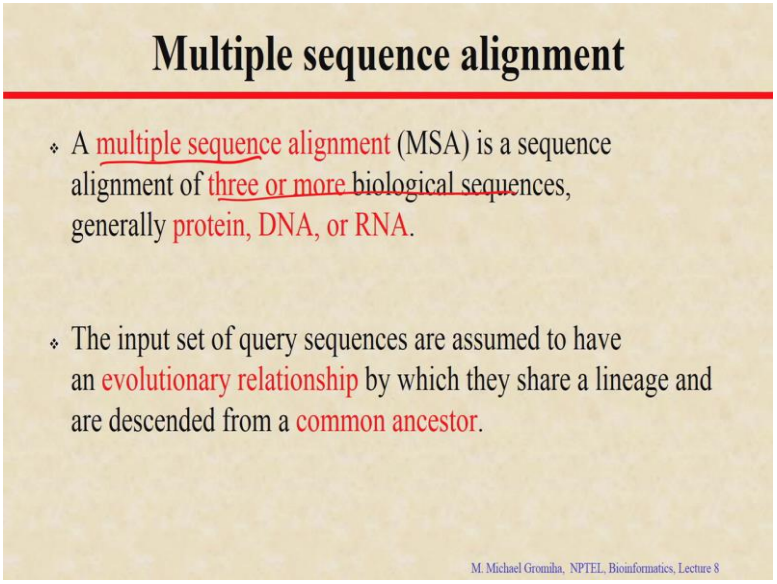


So, we discussed BLAST and we have another program called FASTA I discussed earlier. So, FASTA also can do this if you give the sequence here right this is the FASTA sequence and if you run then this will give you the sequences. And you will give alignment right we can get the values similar to the BLAST.

(Refer Slide Time: 06:08)

So, till now we discussed about 2 aspects; one is if you have query sequence you can get the similar sequences in the databases and the second one pairwise alignment, if you have pairs, you can align if you have more number of sequences hundreds of sequences or thousands of sequences then how to make the alignment?

(Refer Slide Time: 06:37)



So, in this cases, we do the multiple sequence alignment right the name itself tells this is a multiple sequence we have more than 2 sequences together and you can see whether there is any residues which are similar in different sequences; see the sequence alignment

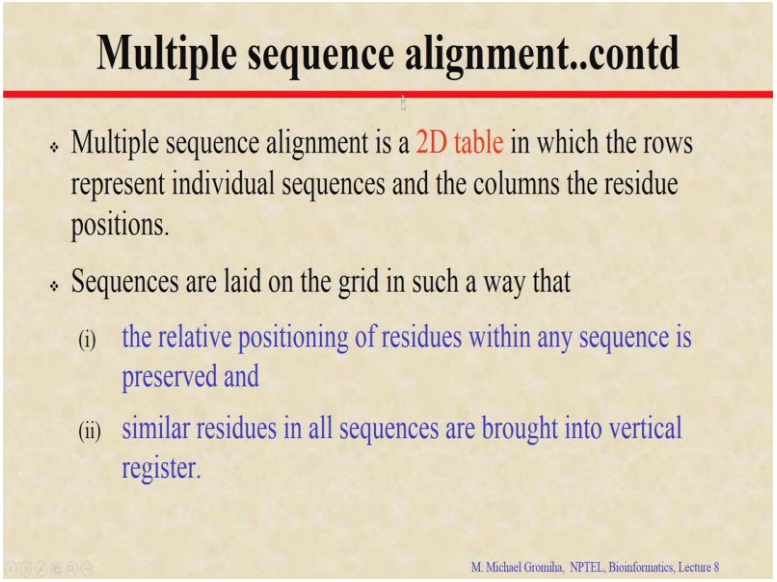
of 3 or more sequences right generally proteins or DNA or RNA or whatever. So, what will it provide which information we get for multiple sequence alignments.

Student: Conservation (Refer Time: 06:59).

Conservation. So, these sequences whether they have an evolution relationship right. So, see the input of query sequences we assume that these sequence are similar and where they are similar and where they are distant right for example, just we show an example like chicken and human hemoglobin right what is a sequence identity 62 percent. So if you have different organisms and if you align these sequences then you can see which region are similar which region you have maintained the same residue and where we have the variations and variations happen in which type of substitutions.

So, then we can see whether they have share a lineage or they have any common ancestor right you can see the relationship if you align different sequences using multiple sequence alignment how to do this right this is a 2D table, because we give the sequences one by one right and put the same sequences in different in same residues positions in different sequences of same place.

(Refer Slide Time: 07:58)



So, here first we see the relative position of residues, if any residues preserved; for example, some residues when the PAM matrix, we have high score for some residues right can you list few residues which have high score in the PAM matrix.

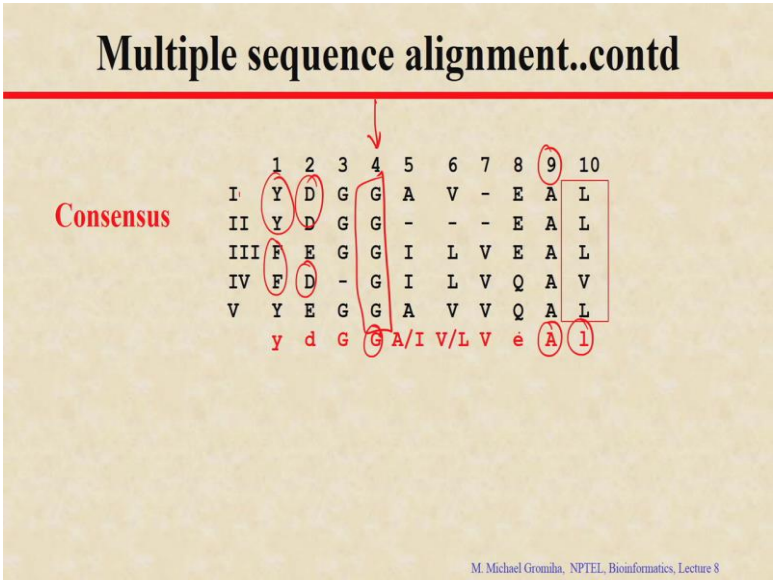
Student: Cysteine.

Cysteine, tryptophan.

Student: (Refer Time: 08:16).

Where these residues are very conserved residues right; they prefer to be not changed. So, if we change you may have adverse effects likewise if you see the sequences at any positions or any set of regions, why they preserve to be the same in different sequences. Then we have similar residues in all sequences there placed one way each other. So, that they can see vertically the same residue this is what they want to maintain when do in the multiple sequence alignment.

(Refer Slide Time: 08:48)



For example, here 5 sequences 1, 2, 3, 4 and 5, if you see the different positions right give 10 positions if we consider first position right. So, it is accommodated with which residues.

Student: tyrosine ok.

Mainly Y.

Student: Aromatic residue

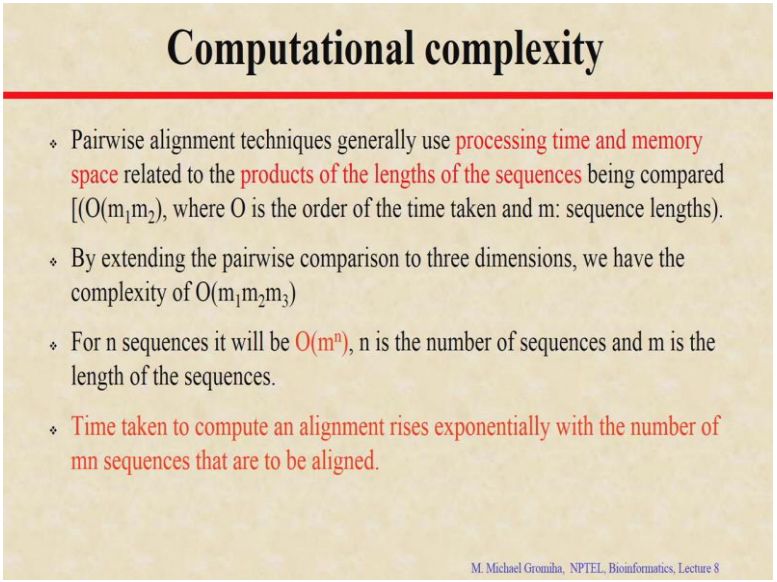
And F

Student: Phenylalanine.

Right so there are 2 residues. So, they put small y because it is mainly dominated by y and to other residues if you do a second one, so mainly with D. So, they put small d and the fourth position if you see G is maintained because G is the same in all the sequences they put G here and some cases, if you see the tenth one this is small l like because mainly L and V and some cases they put the equal preferences right for example, if you see position number 5 or 6 they put either A or I or either V or L.

So, in this case you can see the positions some residues are maintained so they like to be the same residue or different position. For example, number 4 the glycine is maintained likewise position number 9 its alanine is maintained. So, this will help you to see the importance of some residues which are functionally important or structurally important.

(Refer Slide Time: 10:07)



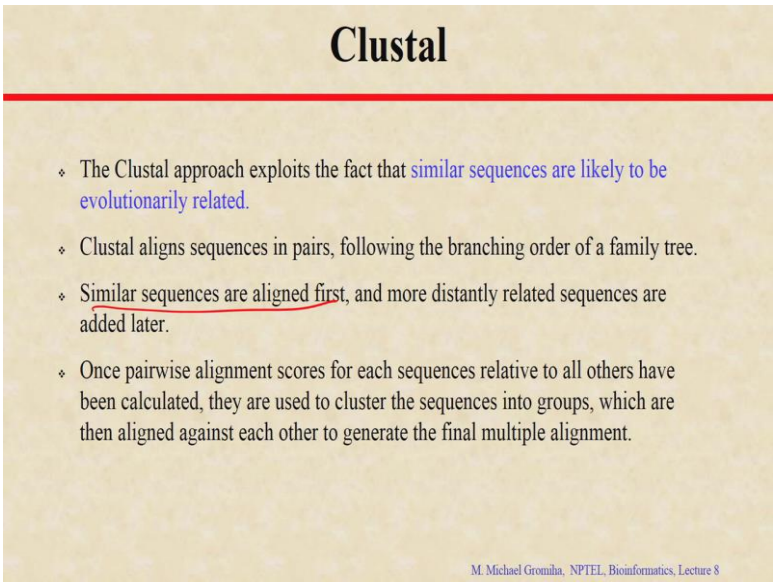
So, what is a computational complexity right because if you align 2 sequences it takes some time which depends upon the length of the sequences right depend of the length of the sequence it will depend on this size for 2 sequences, if you increase the sequences the complexity will;

Student: Increase.

Automatically increase right because the processing time or the memory space this will related to the product of the length of the sequences if you increase more number of

sequences it will take time to align all the sequences and make the consensus one. So, we extend this comparison right pairwise comparison to 3 sequence, 4 sequence, 5 sequence so its complexity increases if there is n sequences here the complexity is very high right. So, in this case the time taken to compute the alignment is also very high then how to do this how to tackle the situation, right. There are several approaches to tackle the situation Clustal is one of the most popular approaches.

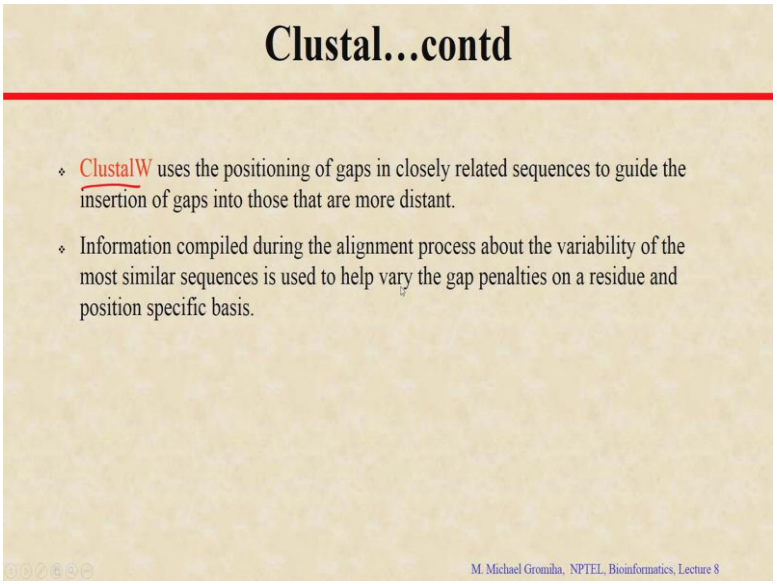
(Refer Slide Time: 11:59)



To align these similar sequences which are evolutionary related how it works first they try to sequence in pairs right, then similar sequences or small sequences they put together right then they form the Clusters of sequences which are similar to each other and they put one by one each other and the distantly related sequences, they do later. Likewise, they can see the correspondence between similar sequences and distance related sequences.

So, first what they do similar sequences are aligned first and the distant related sequences they are later once the pairwise sequences are calculated, they use the sequence into groups and then from that they make the final sequence alignment there are different ways to construct this multiple sequence alignment.

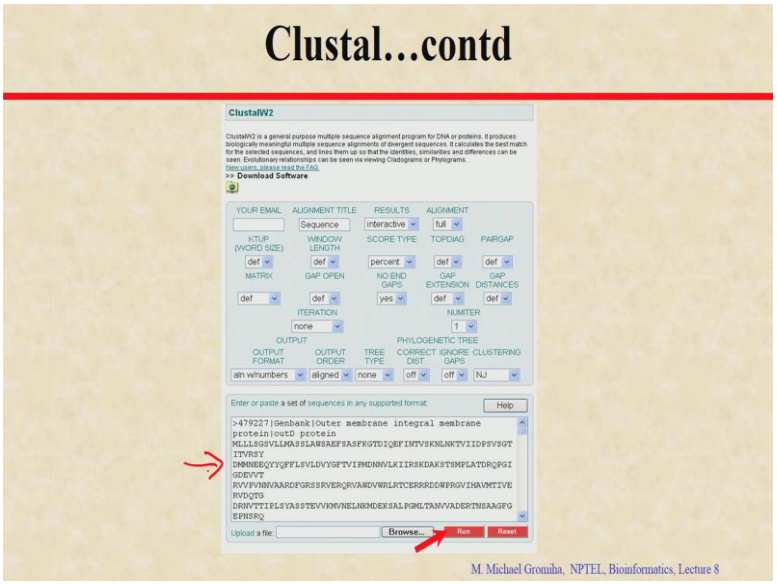
(Refer Slide Time: 11:47)



So, latest one is the ClustalW, this is a program which uses the positioning of the gaps also in the closely related sequences compared with the more distant ones.

So, the information we obtained in the alignment process about the variability of the similar sequences which will help to introduce gap penalties because if we have similar sequences right then you have less gap penalties compared with the one which are having the distant each other from this query sequence. So, this is a program ClustalW right currently we are using Clustal-omega right to align sequences of the several organism with multiple sequence alignment.

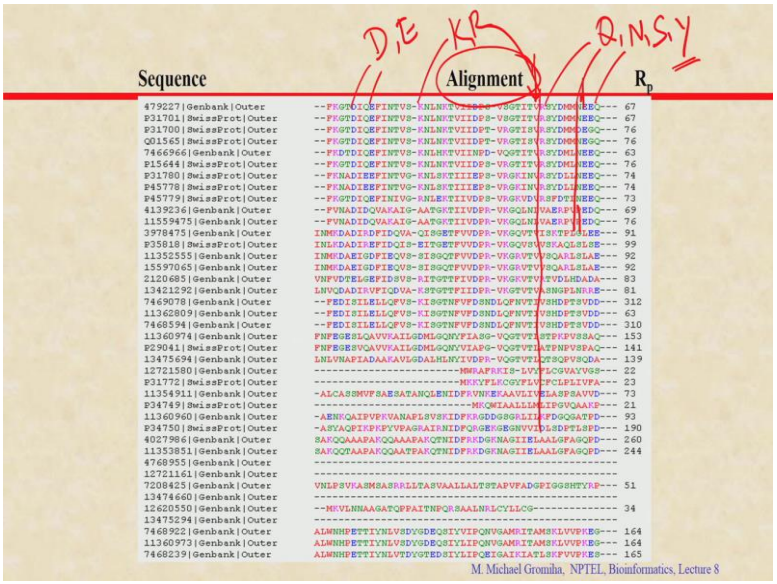
(Refer Slide Time: 12:26)



How to do this we go to the ClustalW. So, here you enter the sequence right of these multiple sequence alignment give one by one.

So, if we give all the sequences in any format because it except several formats then you run the program.

(Refer Slide Time: 12:45)



So, then we get the data. So, from this figure you can see how different sequences are aligned. So, if you see this one this is the sequence different sequences we gave and the

next one is the alignment how they are properly aligned and then see the residue position in this each sequence.

So, if you look into this alignment you can infer some information and what are information you can obtain from these?

Student: Conserved regions.

Right first you can see the different color codes right; what is the meaning of different color codes; what is the meaning of blue color here?

Student: Charged.

Charge right blue is D and E.

Student: negatively charged.

Right, there aspect in simple term because it is negative charge you have the green; green one.

Student: Hydrophilic.

Q, N, S, Y, right.

Student: Polar residues.

Polar residues; they have listed why as polar residues because it contains.

Student: OH;

OH group right this is aromatic, but even it contains OH group and hence the ClustalW consider this as a polar. So, they give green colored for that then what is the magenta one.

Student: Positive charge.

K.

Student: Positive.

Or here is R. So, positive charges then we have the red one mainly the hydrophobic residues right valine isoleucine phenylalanine and so on. So, if you see the colors you can also you can see the same colors are aligned the same position for example if it is the E aligned up to from here to here. Likewise if you see this position hydrophobic residue the valine, leucine, isoleucine; so it is aligned for several sequences.

So, if you do the multiple sequence alignment you can see there are any specific residue positions are maintained in different organisms here interestingly I give the sequence that is mainly the membrane proteins there are 2 different types of membrane proteins I used the beta barrel membrane proteins I try to align. So, even these proteins maintain some conservation this is why they are having some specific functions for this type of proteins.

So, you can see the residues are conserved. So, you can see the; structure function relationship for any types of proteins. So, now, I have a question how to get the multiple sequence alignment for hemoglobin A chain from different organisms how to do this?

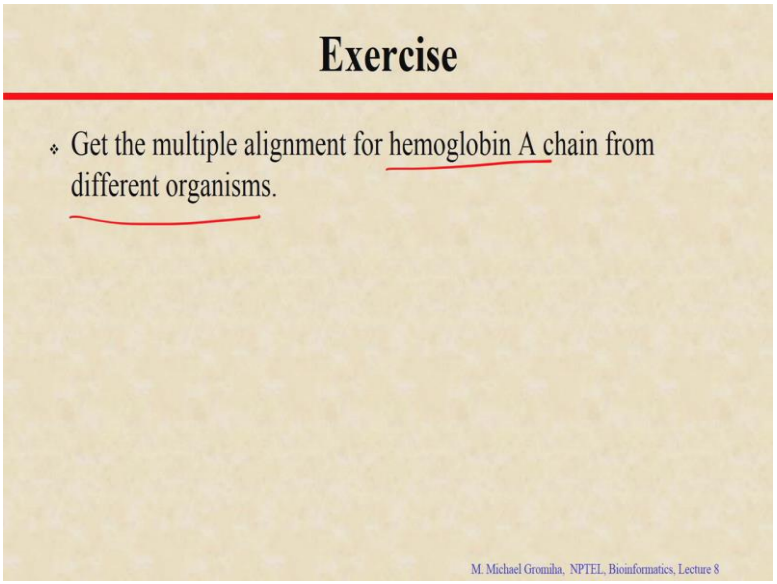
Student: So, we can get the hemoglobin A chain sequence, we can do BLAST, from BLAST we can obtain the different (Refer Time: 15:18).

Right first we get.

Student: sequence.

The sequence of hemoglobin A.

(Refer Slide Time: 15:23)

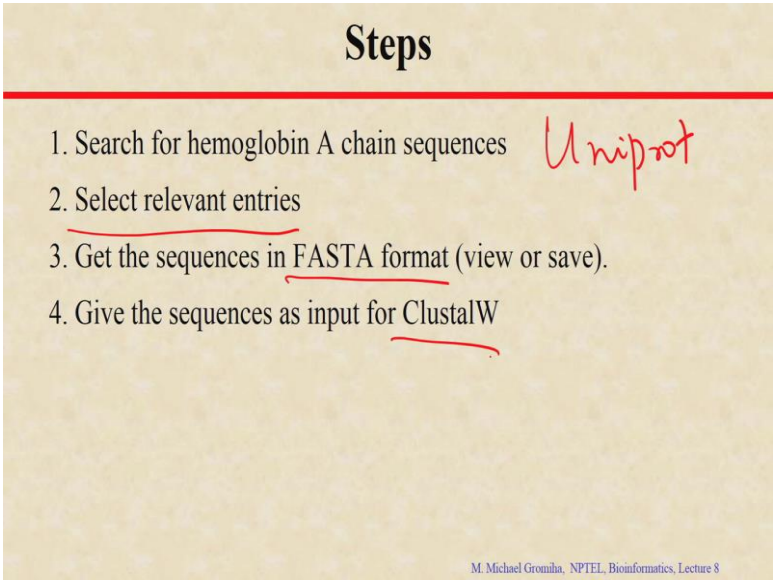


where shall we get the hemoglobin A chain sequence.

Student: UniProt.

UniProt right, first you go to UniProt.

(Refer Slide Time: 15:29)



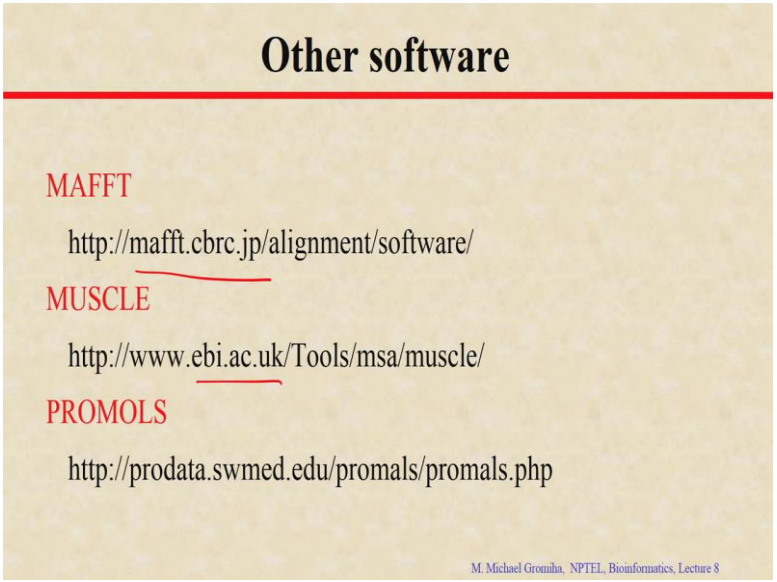
You check with the hemoglobin A you will get a list of sequences then list of sequences, then what to do? you have to identify relevant entries the question is you are to get the alignment for different organisms what if you do; UniProt, you search with the

hemoglobin A you will get several entries from same organisms as well as different organisms what is the possibility of having same organisms why you get same organisms different data.

Student: Mutation data.

Mutation data; so in this case, you will get a same human you will get 2 3 times, but the question is to get from different organisms then you have to select relevant entries then if you can select relevant entries and then you can save in FASTA format I have discussed in one of the previous classes, right; how to obtain the data using interpreted database. So, you can get the sequence FASTA format and you give these sequences in ClustalW. If you do this right for take the sequences, now you give this in this place you can paste sequences and now you can get the multiple sequence alignment.

(Refer Slide Time: 16:34)



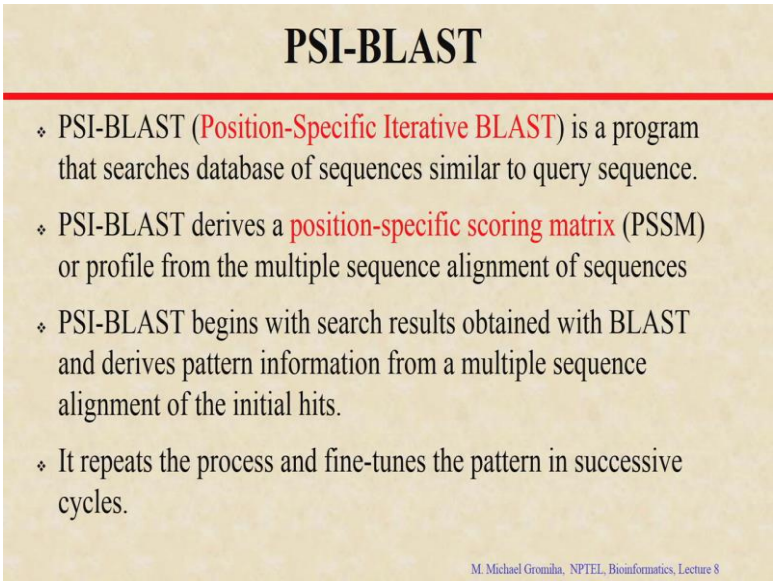
So, this is we discussed about the ClustalW; there are several other software which are available in the literature which you can also be used for the multiple sequence alignment. So, for example, MAFFT right this is available in these website cbrc.jp right and this MUSCLE this also one of the widely used softwares. So, you can see in uk. So, you can use this MUSCLE and the PROMOLs; this is another software; this also you can use for the multiple sequence alignment.

So, why do we need different software to use a multiple sequence alignment? So, what do you expect you get similar results or different results.

Student: Slightly different (Refer Time: 17:08).

Almost you will be a similar results, but they try to improve the alignment values mainly to align maintained in the conserve positions and second option is a speed right we need to get the data very quickly, then to handling more number of sequences and currently the sequence are of different lengths. So, there are several complexities in multiple sequence alignment. So, they developed different algorithms right to handle these situations. So, you can use any of the software to align the multiple sequences. So, in the last few minutes I discussed about psi-BLAST right. So, what is psi-BLAST?

(Refer Slide Time: 17:44)



Student: (Refer Time: 17:45).

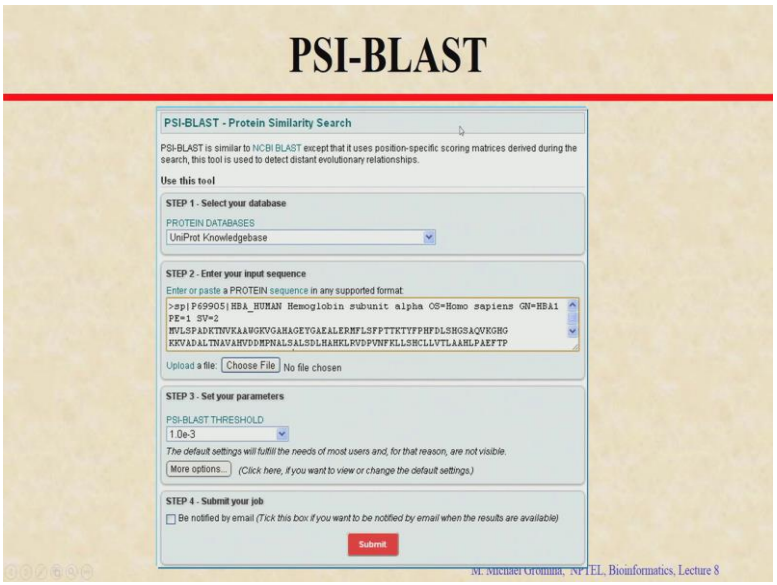
It is a position specific iterative BLAST right is a program which searches database sequences similar to query sequence then they get the BLAST sequences right and then we try to derive the position specific scoring matrix based on the alignment you can compare the positions and the residues and different positions, then they try to align these positions and derive the position specific scoring matrix. For example, we have your own sequence; what is the possibility of having the same residues or different positions depending based on the aligned sequences in your homologous sequences how

to do this first you search the results with the BLAST and derive the patterns derive the pattern from the multiple sequence alignment from the initial hits.

Then they repeats the process right because for example, if you have 2 alignments right they are very perfectly matching and the third one is matching if the another for example tenth position or eleventh position they iteratively changed. And see if this is fine tunes with the best match then keep the top then again the rest they tune again. So, reiterative process finally, they make the perfect alignment this is how the psi-BLAST works.

Once they make these alignments then they try to see; what is the probability of having a specific amino acid residue in a protein to maintain in different organisms of same type.

(Refer Slide Time: 19:10)



So, this is the site you can use the psi-BLAST. So, here we give your input sequence once you give the input sequence what it will do it will do the BLAST it will collect the sequences and do the multiple sequence alignment do different iterations you can see the threshold value.

So, depending upon the expectation value; so iterate again and again and finally, if you click here then you will see your alignment multiple sequence alignment plus you will get these position specific scoring matrices now this position specific scoring matrices you can use it for further applications for example, if you have a sequence of unknown function and if you want to see which residues are binding with somewhere molecules.

So, what you what can we do you will get a sequence right, then you can see the similar sequences and then we can make this matrix and see where it will matches with the known sequence and mapping with the known sequence with your own sequence you can see this region could be a probable binding site for your particular region likewise you can have several potential applications by using these psi-BLAST profiles and the development of these position specific scoring matrices.

So, by summarizing what did we discuss today.

Student: BLAST different kinds of algorithm for example, BLAST multiple sequence alignment.

Alignment, yes.

Student: Then psi-BLAST. So, these FASTA, Clustal .

Fasta, ClustalW, right we can give the started with the pairwise sequence alignment

Student: Yeah.

Right the 2 different algorithms like the BLAST and FASTA then we extended with the multiple sequence alignment and see what are the various parameters you can use to get the score how about the sequence identity? what about the sequence similarities right and how the 2 sequences are similar or different from each other the next is the multiple sequence alignment then using the multiple sequence alignment you can see the position where we have the similar residues maintained in different organisms and then you can also use these alignments for deriving these position specific scoring matrices that I will explain in later classes.

So, in the next coming classes, I will discuss about the use of this multiple sequence alignment to get the conservations score how for how to identify which residues are conserved are there any matrix to get the numbers using some numerical values or you can have this alignment based on any trees whether you can relate form of this evolutionary trees right as well as what can we do with these specific alignments and so on.

Thanks for your kind attention.