

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

Lecture - 8a
Sequence Alignment: Online resources

In this lecture we will discuss about the online resources, which are widely used in the literature and which are available for aligning sequences. So, just to refresh what did we discuss in the last class.

Student: (Refer Time: 00:35).

PAM and BLOSUM matrixes.

Student: Dynamic Programming.

Right, what are the factors soon as to consider for the development PAM matrix?

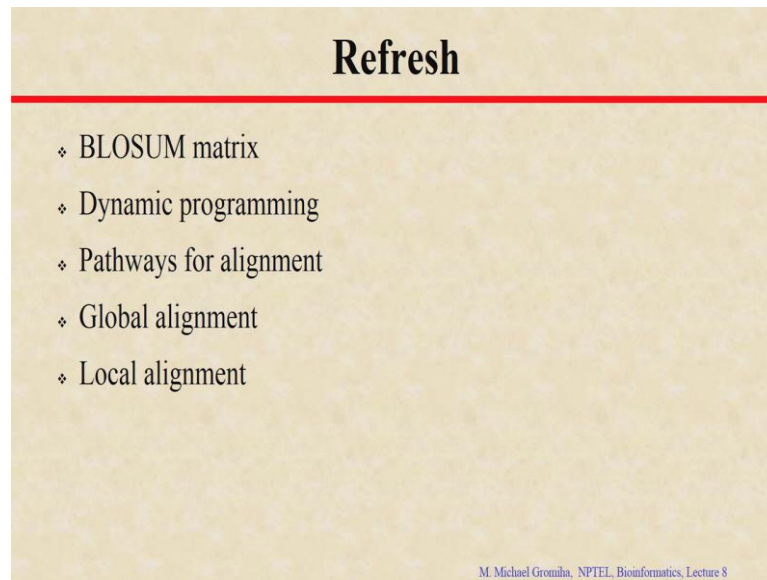
Student: Mutation frequency.

Mutation frequency.

Student: (Refer Time: 00:46).

Mutability, number of amino acids, number of specific amino acids, which are involved in mutation and so on and you can see the probable mutations right. It is mainly based on the substitution rates right, depending upon the probability of each amino acid residue to be replaced by other amino acids.

(Refer Slide Time: 01:06)



Then we discuss about dynamic programming right what is dynamic programming.

Student: So, we divide the alignment problem into a smaller and smaller part (Refer Time: 01:15).

Right smaller reliable problems and then finally, combine to get the overall result for a given problem right this is dynamic programming. Then we discussed about the pathways for alignment for example, if you have 2 sequences one is of 10 residues and another with eight residues, there are several possibilities to align using the inclusion of gaps right. So, we discussed about different pathways and how to obtain the optimal path.

So, then we try to apply scores for matching for mismatching and deletion insertions and we finally, try to see what are the probability of having different alignments and what is the most probable alignment. They will discuss 2 different types of alignments in protein sequences or nucleotides sequences right one is global alignment, another is local alignment; what is the difference between global and local alignments?

Student: Global is end to end alignment.

Global alignment are consists of all the nucleotides or the amino acids in 2 sequences local alignments, if there is a gap in the in the beginning or the ends. So, it omits the gaps

and try to identify motifs and the patterns right inside the sequences. What is the name of this global alignment put one of this algorithms?

Student: Needleman Wunsch.

Wunsch Needleman develop this algorithm right. So, what is the how what is the criteria using the algorithm to fill the metrics.

Student: 3 criteria match mismatch (Refer Time: 02:41).

Maximum of.

Student: Match mismatch (Refer Time: 02:43).

Match and mismatch.

Student: Or gaps.

Or the insertions.

Student: (Refer Time: 02:46) deletions.

Or deletions in the case of local alignments.

Student: (Refer Time: 02:48).

So, who develop this algorithm?

Student: Smith Waterman.

Smith Waterman develop this algorithm, they included one more criteria.

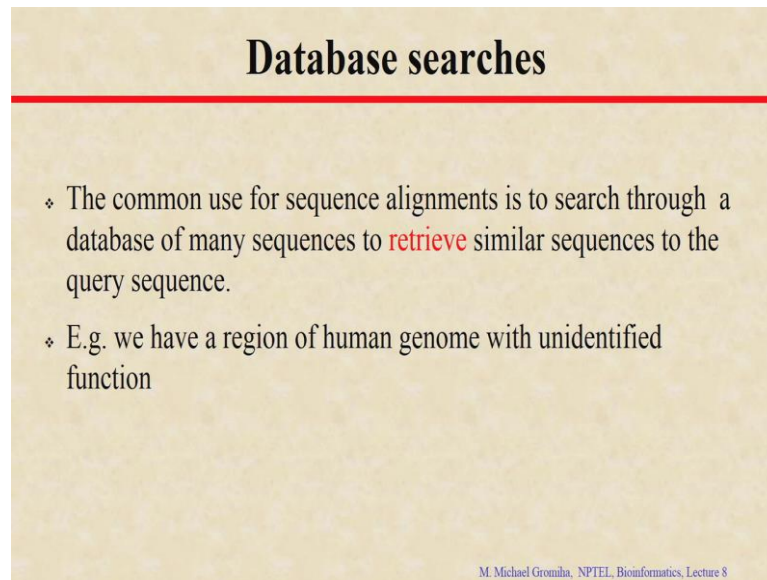
Student: (Refer Time: 02:55).

And maximum of the poor conditions, so 3 include in global alignments plus.

Student: 0.

0, right. So, avoid all the negative values. So, now in this lecture we will see how to use these alignments.

(Refer Slide Time: 03:11)



Database searches

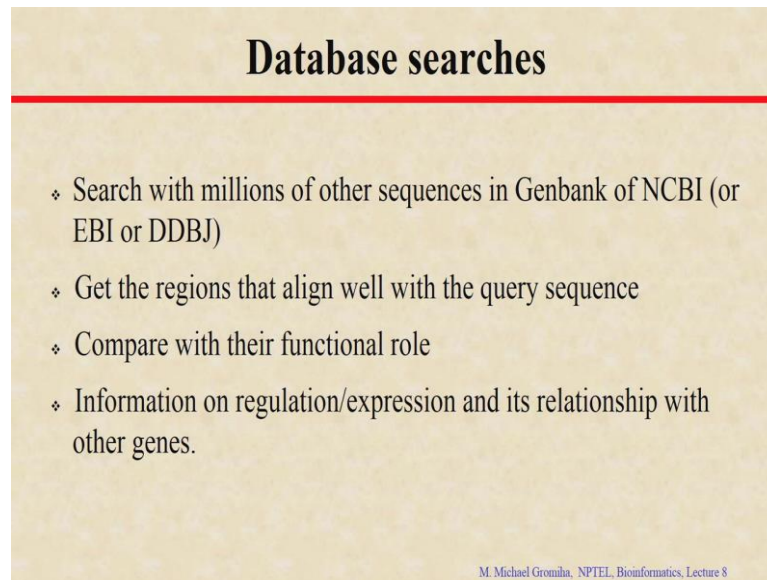
- ❖ The common use for sequence alignments is to search through a database of many sequences to **retrieve** similar sequences to the query sequence.
- ❖ E.g. we have a region of human genome with unidentified function

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

What are the software tools available in the literature, which are commonly used for aligning sequences or to identifying the match for any query sequence? So, why we need this and what is the use of the sequence alignment? The common use for a sequence alignment is, for a high specific sequence you have a query sequence. First you can see what are the other sequences similar to your query sequence right.

If you have a sequence which match with your sequence then you can try to understand how far they are similar depending upon the similarity or any functionally important regions, then you can infer that your protein has also similar functions. For example, if we have human genome, a region with unidentified function. So, you want to understand the function of these portions or some proteins with unknown function what to do guess?

(Refer Slide Time: 04:03)



Database searches

- ❖ Search with millions of other sequences in Genbank of NCBI (or EBI or DDBJ)
- ❖ Get the regions that align well with the query sequence
- ❖ Compare with their functional role
- ❖ Information on regulation/expression and its relationship with other genes.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

Student: (Refer Time: 04:03).

Well first you try to see whether any known sequences are matching with your query sequence right. So, we search with the millions of other sequences in different databases. So, we discussed about different databases for the DNA database or the protein database or the DNA databases.

Student: Genbank (Refer Time: 04:23).

EMBL DDBJ right Genbank and all right likewise you have the protein sequence databases called the UniProt right and also structure database called protein data bank.

So, we search with these sequences and see whether your query sequence is aligned with any of the sequences deposited already in this database, then if the align known sequences they have functions because we discussed in the uniprot and the other DNA sequences. So, you get all the information regarding the structure, regarding the function what are the links and the pathways and so on.

Now, with your query sequence, if you could find a sequence which is aligned with your query sequence then you can try to understand the function probable function of your query sequence. Then the information and regulation in the expression as well as the relationship with other genes right; that this is why we need the databases right to find the probable sequences which are aligned with your query sequence.

(Refer Slide Time: 05:19)

Database searches

Major points:

1. Size of the query sequence
2. Number of sequences in the database

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

So, when you align when you find the similar sequences, what are the major points one has to consider? Your query depends on the size of the sequence for example, if you have 100 residues, and you have another 1000 residues right depending upon the size of your protein it will take time but it has to search with a databases. Then second one is number of sequences in the database. So, your query depends upon the size of the sequence as well as number of sequences, you are trying to align, you are trying to compare right the time depends upon these 2 aspects.

So, now considering all these aspects, there are several tools developed in the literature.

(Refer Slide Time: 06:03)

BLAST

BLAST: **B**asic **L**ocal **A**lignment **S**earch **T**ool

BLAST finds **sub-sequences** from a **sequence database** for any **query sequence**.

Program name	Query sequence	Database type
Blastp	Protein	Protein
Blastn	Nucleic acid	Nucleic acid
Tblastn	Protein	Nucleic acid (translated)
Tblastx	Nucleic acid (translated)	Nucleic acid (translated)

Blastp: searches for protein sequence matches using PAM or BLOSUM matrices to score the ungapped alignments.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

One of the most widely applicable tools is BLAST right what is BLAST?

Student: Basic local alignment search tool.

Basic local alignment search tool; in the last class we discussed about the principle used in BLAST what are principle used in BLAST? First they divide this query sequence or the target sequence in the small bits right last time last class we discussed about the 2 residues right, we can use 2 or 3 residues and then see the alignment and get this score from the BLOSUM or the PAM matrixes, and where we have the value which is more than the threshold value then you can mark it then finally, you can see where you can see the marked regions you can align.

So, the BLAST it finds the sub sequences from the sequence databases of the query sequence. So, there are different program names for example BLASTp, there is query sequence is protein it will search with the protein database. Then BLASTn nucleic acids with the nucleic acids.

Because n and p stands for nucleic acids and the proteins we can also do the translated once the tBLASTn and the tBLASTx, so for using these queries. So, they use mainly the PAM metrics or the BLOSUM metrics to give scores for the ungapped alignments.

(Refer Slide Time: 07:21)

BLAST: Example

- ❖ Blastp first breaks down the query sequence into words or subsequences of fixed length
- ❖ All possible pairs are calculated using sliding windows

E.g. AILVPTVI → AILV, ILVP, LVPT, VPTV and PTVI

- ❖ Search for word matches (also called **High Scoring Pairs or HSPs**):

MVGQTIPKLAAILVGTVIAML ...
AILVPTVI

- ❖ Extend the match until the local alignment score falls below a fixed threshold
- ❖ It also allows gaps in the extended length.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

For example if you have a sequence right A I L V P T V I. So, BLAST splits this sequence into smaller pieces for example, if we have the 4 residue segments they took divide in AILV, ILVP, LVPT, VPTV, and PTVI. So, the overlapping segments now see this is the query sequence right. So, we have the sequence, now they see whether there is a match of this query sequence with any of these tetra-peptides in the sequence.

So, if you align these 2 you can see there is a match there is AILV. So, this type of match there is called the high scoring pairs, where all the residues will match in any subsequence for example, the 4 residue segments or the 5 segments sequence and so on. So, then you access the match right to further and see how long you can extend the same sequence to with the exact match. So, then once it fails then you can stop there and then you can see this will give you the high scoring pairs how many times you can get the exact matches between the query sequence and the target sequence.

This FASTA is also another program, they also use for the sequence similarity and the sequence alignment here they breaks into 4 to 6 nucleotides and 1 or 2 amino acids.

(Refer Slide Time: 08:43)

FASTA

FASTA is another program for sequence similarity search and sequence alignment.

FASTA breaks the words into 4-6 nucleotides or 1-2 amino acids

Eg. Query sequence: FAMLGFIKYLPGCM

Word	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Position	2	13			1	5		7	8	4	3		11							9
					6	12				10	14									

Target sequence: TGFIKYLPGACT

1	2	3	4	5	6	7	8	9	10	11	12
T	G	F	I	K	Y	L	P	G	A	C	T
3	-2	3	3	-3	3	-4	-8	2			
10	3			3				3			

F	A	M	L	G	F	I	K	Y	L	P	G	C	M
T	G	F	I	K	Y	L	P	G	A	C	T		

FAMLGFIKYLPGCM *Seq1*

TGFIKYLPGACT *Seq2*

Large number of sequences have the number 3;
Offsetting with 3 gives a reasonable alignment

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

We have the query sequence FAM and so on what they do? First take this sequence and identify the positions. So, F we see in the position number 1 right give me F is in position number 1 right I take this as position number 1. So, they put F they put in their position 1 and the second one is A is in position 2. So, we put 2 then M is 3. So, we put 3. So, likewise see this query sequence they have made in 2 different positions right and we can see another F here one 2 3 4 5 6 the position number 6 also, then take a target sequence then compare this target sequence with the query sequence if you take this G.

So, position number it is 5 here this is 2. So, $5 - 2 = 3$, then we got another possibility is 12, $12 - 2 = 10$ then we take this next one F, it is in 3 here this is 1; so $1 - 3 = -2$ and then $6 - 3 = 3$. Likewise for the target sequence they made this comparison and put the numbers. Then we look into this one they have found lot of numbers with 3. So, then this case they tell that if there is more number 3, if you offset one sequence with 3 residues right then you can get the probable alignment.

So, you offset the 3 sequences right if you see here. So, they are offside with 3 sequences then they get good alignment between sequence 1 and sequence 2.

(Refer Slide Time: 10:17)

Alignment score and statistical significance

- ❖ The primary indicator of how similar the search results are to a query sequence is the **alignment score (S)**.
- ❖ Score is given with **P or E value**.
- ❖ **E-value** is the expected number of sequences of score $\geq S$ that would be found by random choice
- ❖ **P-value** is the probability that one or more sequences of score $\geq S$ would have been found randomly.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

So, now how to compare the sequences and how to estimate which alignment is the best? There are various steps there various measures. To examine the scores for the best alignment or the bad alignment one of the indicators that is used as to measure the alignment that is the alignment score that is s . This score you can get the with the highly maximum pairs right and also with the P value or E value right.

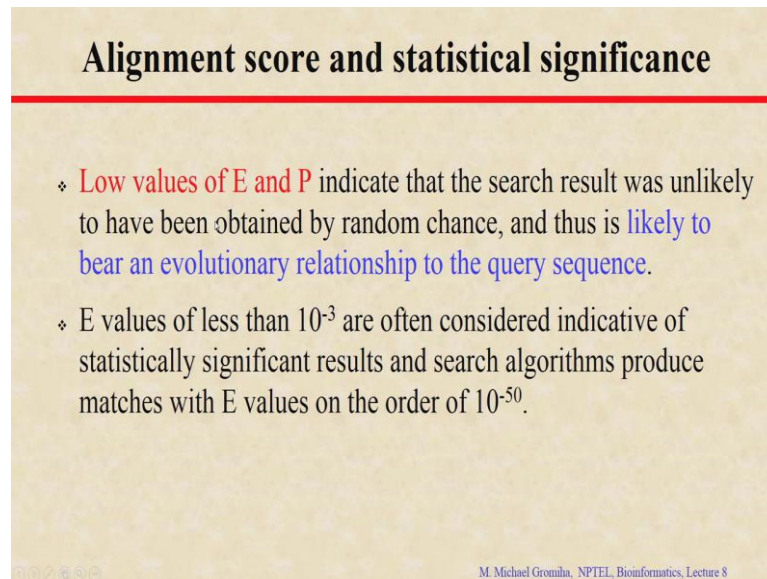
These P and E are very important in the sequence alignment, and to judge the sequences whether they are statistically significant or not. So, E value is defined as the expected number of sequences, which can get score more than S let it be found by random choice. For example, if you get your alignment score, say 10 or 15 what is the probability or what is the expected number of times you get that same number with any score of more than S that is the E value.

Then P value is the probability of having one or more sequences with the score of more than S . So, we get 2 numbers one is expected value right whether you get this number as expected or it is unique to your sequence. Then P value which is the probability of having this score is random or that has any significance these 2 numbers will give you whether your alignment score is statistically significant or not essentially what do you expect, the number should be high or less.

Student: Less.

Less, it should be less because you should not get randomly.

(Refer Slide Time: 11:59)



Alignment score and statistical significance

- ❖ **Low values of E and P** indicate that the search result was unlikely to have been obtained by random chance, and thus is **likely to bear an evolutionary relationship to the query sequence**.
- ❖ E values of less than 10^{-3} are often considered indicative of statistically significant results and search algorithms produce matches with E values on the order of 10^{-50} .

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

So, this E values and P values should be low, which indicate that your search results are unlikely to be obtained by random search. So, in this case you can say that your alignment likely to have evolutionary relationship with the query sequence; that means, what you obtain in the alignments between the query sequence and the target sequence right this statistically significant and you can rely on your alignment score.

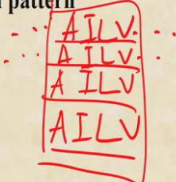
Usually we use the value of 10^{-3} less than 10^{-3} right to be indicative of these significant results, and if you align the sequences in many cases you will get the values in the order of less than 10^{-50} if you align the sequence you will get that.

(Refer Slide Time: 12:41)

BLAST: Features

- (i) identifying protein sequences similar to the query
- (ii) finding members of a protein family or build a custom position-specific scoring matrix
- (iii) finding proteins similar to the query around a given pattern
- (iv) finding conserved domains in the query
- (v) searching for peptide motifs

BLAST is available at <http://www.ncbi.nlm.nih.gov/BLAST/>



M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

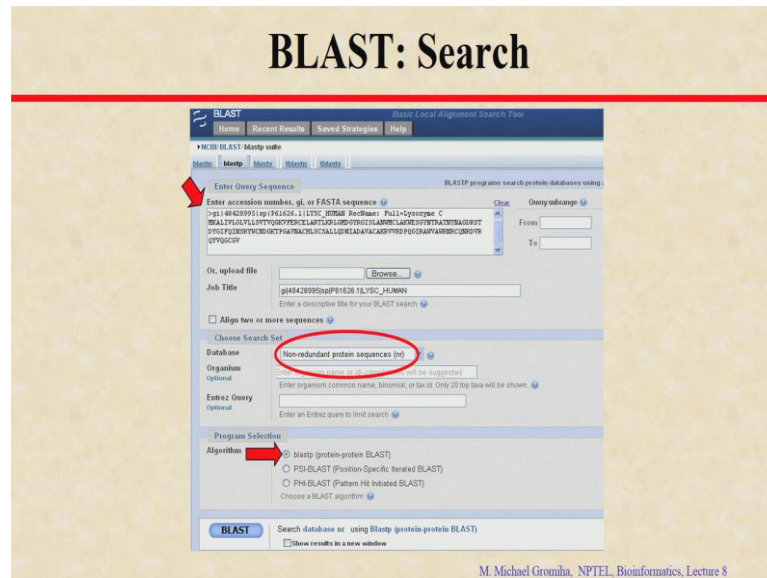
So, what are the features of BLAST, what are the information you obtain by using the BLAST, can tell some examples what are the information we obtain from BLAST, why the BLAST is used what are applications of BLAST right. First you have a query sequence you can identify the sequences, protein or nucleic acid sequence similar to your query right this is first one. Then the second one you can identify the members of the protein family right same different same protein family you can identify the members, this can also be used to derive some sort of matrices called the position specific scoring matrices and I will explain this after few minutes, then you can identify some patterns.

So, if you have a query. So, if you there are any patterns in your query with the available sequences, that you can identify the patterns or the motifs in your query. Then are also you can see any conserved domains any regions of the sequence, which is similar in all the sequences. For example, if you get several sequences right several sequences and you have a pattern right and this is maintained in all the sequences and you see that this pattern is very important for the structure and function.

You can see whether any patterns for your query sequence with respect to the sequences available in database. Then also you can search for the peptide motifs, where you have any specific motif in your protein as well as with the proteins available in the database. You can use BLAST to get all the information. So, this is a website in the NCBI. So, you

can use this website to identify any sequences, which are related to your query sequence ok.

(Refer Slide Time: 14:28)



Now, let us see how BLAST works. So, this is the website for the BLAST, and if you open the website it will ask you for the sequence and because that is essential. So, you need to give your query sequence to obtain the relevant sequences right. So, it asks the query sequence in different formats.

Either you can give the accession number, Genbank number or in FASTA sequence or Uniprot number or the PDB. So, it accepts several numbers, then if you give it then it will ask for the database which database you want to search right because as we discussed earlier 2 different aspects one we need your query sequence, and the second we need database to search right. So, first we give a sequence here and we need to give the database. There are several databases available in the literature as we discussed earlier, you have the Uniprot you have the protein databank. So, several databases are available.

So, you have to specify which database you have to search then they ask for which algorithms do you want to use, this is for the protein BLAST or the nucleic acid BLAST or you align different sequences and so on. So, if it is a protein BLAST, because this I give the protein name human lysozyme. So, it takes the protein BLAST and you can BLAST it.

(Refer Slide Time: 15:45)

BLAST: Options – gi number

Accepts gi number or FASTA format

gi : bar separated NCBI sequence identifier (e.g., gi|48428995).

Accession number : number allotted in Uniprot for each sequence (e.g. P61626)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

So, these are the different file formats you can ask the gi format, which is the bar separated NCBI identifier for example, if you see gi and then it is a bar separated and this is an identifier.

Student: (Refer Time: 15:56).

Then accession number you can give the Uniprot, then we discussed about the Uniprot database right. So, you can see this is a number P61626.

Student: (Refer Time: 16:05).

See if you give this number BLAST accepts this number, and takes the sequence directly from this Uniprot database right everything is mapped. So, if you give these numbers they can directly get the sequence from the relevant databases. So, what are the different file formats? The widely accepted file formats are most widely used formats in the bioinformatics problems right this is FASTA format.

(Refer Slide Time: 16:30)

BLAST: Options – FASTA format

FASTA format: begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than ("**>**") symbol at the beginning.

Example sequence in FASTA format:

```
>LYSC_HUMAN RecName: Full=Lysozyme  
CMKALIVLGLVLLSVTVQGVFERCELARTLKRLGMDGYRGISLANWMLAKWESGYNTRATNYNAGRSTDYGI  
FQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVRVQYVQGC
```

❖ Files in FASTA format have the extension “.fasta”

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

The FASTA format which begins with the single line description right followed by the sequence data, this is a single line description right and here this is a sequence data. So, in this line you can see this started with the greater than symbol.

This will distinguishes one sequence to the other sequence right and then you can use as a comment. So, you can write the description of their particular protein in that line. Then the second line followed with this sequence, now you end with this amino acid sequence. So, these FASTA files, they have the format and they also end with this extension of dot FASTA, that you can give the FASTA files with the extension of dot FASTA.

(Refer Slide Time: 17:07)

File format -2

2. NBRF/PIR (National Biomedical Research Foundation/Protein Information Resource)

First line begins with >P1 for protein sequence or >N1 for nucleic acid sequence.

```
>P1|LYSC_HUMAN
CMKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNA
GDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWR
NRCQNRDVRQYVQCGV
*
```

Files in NBRF/PIR format have the extension “.seq” or “.pir”

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

Then this is another format that is NBRF national biomedical research foundation or protein information resource right. So, in several cases even now they use pir formats. So, what is a pir format? It begins with greater than symbol and p1; p1 is for the protein sequence and n is for the nucleotide sequence right.

So, if we see here this is for the protein sequence, and followed by the sequence and then end with the star; that means, this is the one complete sequence. So, this is the format used in pir. So, in this case these files have the extension dot seq or dot pir, then the program will understand this is the sequence file or it is the pir file.

(Refer Slide Time: 17:51)

File format -3

3. GDE format, (essentially the same as FASTA, the difference is starts with %). The file format is “.gde

All three file formats ignore spaces and carriage returns.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

There is another format GDE format, this is also similar to the FASTA format and the difference is only this starts with percentage. It is the formats used earlier in literature, currently the FASTA format is widely used and in many bioinformatics problems. Here this have the extension dot GDE, all these formats they ignore spaces as well as a carriage returns.

So, if you have space automatically this will ignore spaces and take the sequence continuously.

(Refer Slide Time: 18:19)

Searchable databases

Peptide Sequence Databases

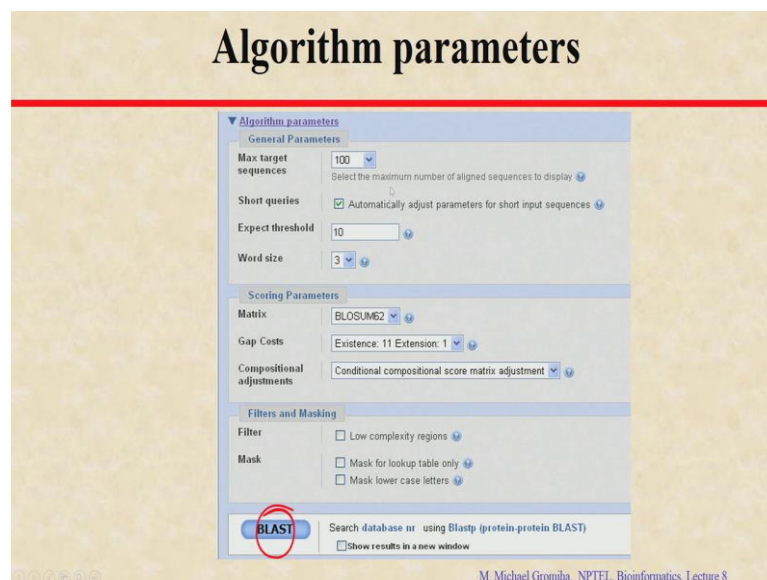
- nr** All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF
- refseq** RefSeq protein sequences from NCBI's Reference Sequence Project.
- Uniprot** Last major release of the Uniprot protein sequence database.
- pat** Proteins from the Patent division of GenPept.
- pdb** Sequences derived from the 3-dimensional structure from Protein Data Bank
- month** All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.
- env_nr** Metagenomic Protein sequences.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

So, now this is the query sequence. So, we have the query sequence you can give in different formats, either we give the sequence in specific formats or you can give accession number right Uniprot number or gi number and so on then the searchable database. So, we have to provide the database to search, there are different options available in the BLAST, the widely used one is the non redundant datasets, whether you say the rough seeking proteins or the PDB or the Uniprot and so on.

Or you can see the NCBI's reference sequence project, or you can use the uniprot data latest release or any data in the Uniprot database right. Likewise you can give different databases like PDB or any specific latest data specific in the month and so on.

(Refer Slide Time: 19:07)



So, now the first part is over; and second part is we need to define some parameters right the first part what is the information we gave?

Student: Query sequence (Refer Time: 19:17).

Query sequence and?

Student: Database (Refer Time: 19:19).

Database and then BLASTp right we need over the protein sequence; now we have to give some parameters. So, how many target sequences you need to find right for example, if we have one single sequence, and if you give the Uniprot. Uniprot has 75

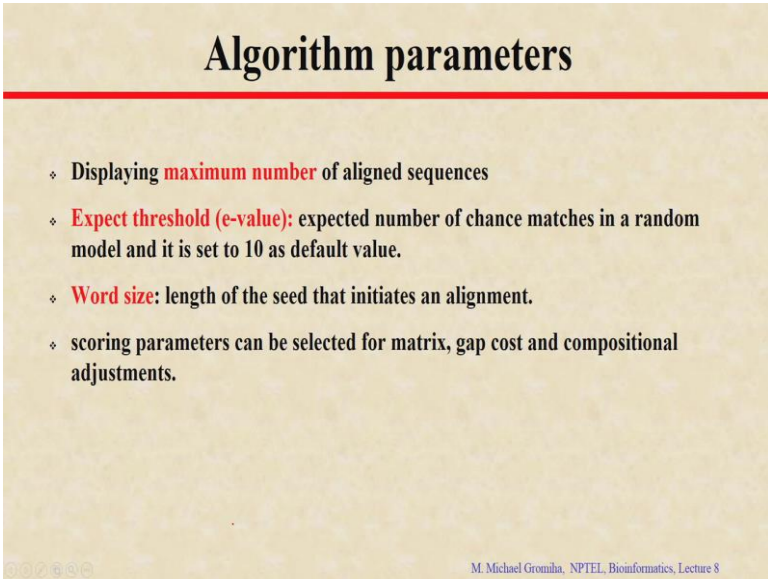
million 80 75 millions of sequences, it will give lot of sequences most of there is 100 percent it will show only the 100 percent is a highly redundant right. So, we can use the maximum target sequences that you want, then what is the expected threshold right what is the expected value to find this statistical significant the E value or P value? To the expected threshold is randomly default value is 10, but you can change right then the word size where you want to split the query sequence into 3 words or 4 or 5.

So, the random threshold is 3 right. So, you can change the word size to see whether we can get different types of alignments or not right. This is general parameters, then as we discussed earlier, we developed few matrixes what are few different matrixes you developed?

PAM matrix and BLOSUM matrix; so here, asking about the matrix which matrix you want BLOSUM62 or the PAM1 or 250 or PAM1000. We discuss for the closely related sequences which PAM matrix you use 250 is normally used right for the closely related sequences we use 1 or for distantly related we use 1000 all right for general case.

We use PAM 250 right. So, we can select which BLOSUM you want or which PAM matrix do you want to use for the alignment right. Then also there are some filters for example, your low complexity regions, you want to mask or you want include and so on.

(Refer Slide Time: 21:02)



Algorithm parameters

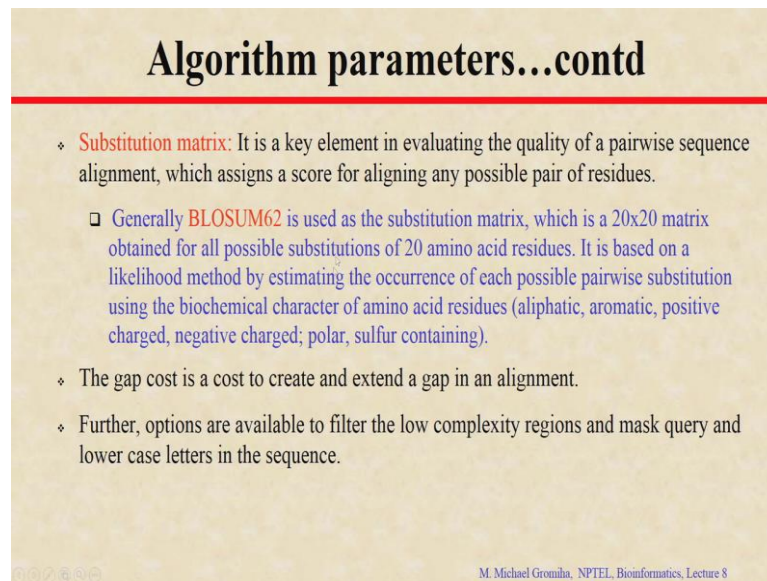
- ❖ Displaying **maximum number** of aligned sequences
- ❖ **Expect threshold (e-value):** expected number of chance matches in a random model and it is set to 10 as default value.
- ❖ **Word size:** length of the seed that initiates an alignment.
- ❖ scoring parameters can be selected for matrix, gap cost and compositional adjustments.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

Now you click on BLAST then it will give you the some data before showing the data. So, I will tell you the parameters which BLAST will provide. It will give you the maximum number of aligned sequences as you will give the expected threshold value right; that means, expected number of chance randomly it can match.

But we said 10 as a default value then word size we give and its scoring parameters also we finalized we gave as BLOSUM 62.

(Refer Slide Time: 21:32)



Algorithm parameters...contd

- ❖ **Substitution matrix:** It is a key element in evaluating the quality of a pairwise sequence alignment, which assigns a score for aligning any possible pair of residues.
 - ❑ Generally **BLOSUM62** is used as the substitution matrix, which is a 20x20 matrix obtained for all possible substitutions of 20 amino acid residues. It is based on a likelihood method by estimating the occurrence of each possible pairwise substitution using the biochemical character of amino acid residues (aliphatic, aromatic, positive charged, negative charged; polar, sulfur containing).
- ❖ The gap cost is a cost to create and extend a gap in an alignment.
- ❖ Further, options are available to filter the low complexity regions and mask query and lower case letters in the sequence.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

Then substitution matrix we discussed. So, we got 2 different matrixes one is a PAM and the BLOSUM. So, depending upon the alignment score we require. So, we can select any of these matrixes. Then gap costs is a cost to create an extend gap in the alignment, because if you introduce gap we need to give penalize right because insertion and deletions are not common compared with the mutation. So, we need to give penalty for the gaps fine.

(Refer Slide Time: 21:56)

BLAST: Output

gi|48428995|sp|P61626.1|LYSC_HUMAN RecName:...

Query ID	gi 48428995 sp P61626.1 LYSC_HUMAN RecName: Full=Lysozyme	Database Name	nr
Description	C	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid	Program	BLASTP 2.2.25+ P<3>tabon
Query Length	148		

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Related Structures](#) [Multiple alignment](#)

▼ **Graphic Summary**

Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq. 1 148

Lysozyme catalytic site Ca²⁺ binding site

Specific hits: LYZ1

Superfamilies: LYZ1 superfamily

Lysozyme catalytic site
Ca²⁺ binding site

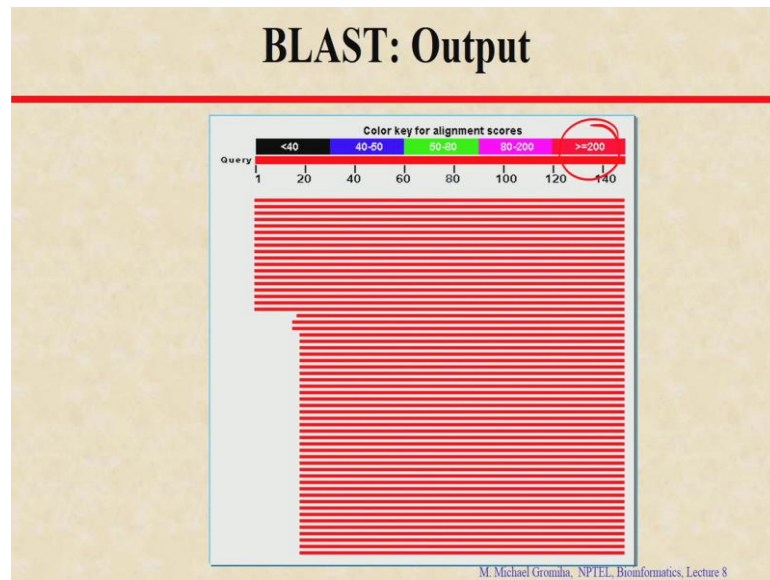
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

So, this is output. So, we give the input if you see. So, you have the sequence.

And here either we give a sequence or number or you can have the option to upload the file.

So, we select database, and this is the algorithm why what you want to perform and then we give the parameters right the BLOSUM metrics and other parameters we set, and if you click the BLAST it will give you the data. So, here this is your query sequence this is 1 to 148. Also it is possible if you want to align only few sequences, instead of the whole protein if you are interested in some regions for example, 10 to 100 then you can also give the information. So, you want to search only with this, say residue number 10 to residue number 100. So, here we search the whole sequence 1 to 148. So, there are several hits.

(Refer Slide Time: 22:48)



With the lysozyme super family if you click on that then you will get this number this diagram. This looks completely red why it is completely red; because we get many sequences which are highly aligned. So, if you look into the score, so most of the cases it will be more than 200, right. So, you get a lot of proteins which are highly aligned with the query sequence, this is why we get the completely red dash. If your query does not match with the data in the sequence database what will happen.

Student: Black.

It will black or the you can see the other colors right depending upon the alignments score you will get different other colors.

(Refer Slide Time: 23:26)

BLAST: Output 148 residues

Legend for links to other resources: [UniGene](#) [Gene](#) [GEO](#) [Gene](#) [Structure](#) [Map Viewer](#) [PubChem BioAssay](#)

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E-value	Links
NP_000230.1	lysozyme C precursor [Homo sapiens] >reflNP_001009073.1 ly	303	303	100%	4e-81	U G M
AA036188.1	lysozyme precursor (EC 3.2.1.17) [Homo sapiens]	303	303	100%	6e-81	U G M
AA67384.1	lysozyme [synthetic construct]	301	301	100%	2e-80	
P79179.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	301	301	100%	2e-80	
XP_002823550.1	PREDICTED: lysozyme C-like [Pongo abelii] >sp P79239.1 LYSC	300	300	100%	5e-80	G M
XP_003459554.1	PREDICTED: lysozyme C-like [Nomascus leucogenys]	298	298	100%	1e-79	G M
AA63078.1	lysozyme precursor [Homo sapiens]	297	297	100%	4e-79	G M
P79180.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	295	295	100%	2e-78	
P51633.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	283	283	100%	5e-75	
NP_001095293.1	lysozyme C precursor [Macaca mulatta] >sp P30201.1 LYSC_M	280	280	100%	6e-74	U G M
P79011.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	280	280	100%	6e-74	
NP_001106112.1	lysozyme C precursor [Papio anubis] >sp P61629.1 LYSC_PAPA	273	279	100%	8e-74	U G
P79006.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	273	279	100%	8e-74	
P79047.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	278	278	100%	1e-73	
P82729.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	278	278	100%	3e-73	
P82927.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	278	278	100%	2e-73	
P79087.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	277	277	100%	4e-73	
P61631.1	RecName: Full=Lyszyme C; AltName: Full=1,4-beta-N-acetylir	276	276	100%	5e-73	
I046_A	Chain A, Mutant Human Lysozyme With Foreign N-Terminal Res	272	272	88%	1e-71	S
I02P_A	Chain A, Crystal Structure Of Mutant Human Lysozyme With Fc	272	272	88%	1e-71	S
C4853144.1	lysozyme [synthetic construct] >g AAQ72800.1 lysozyme [s	271	271	87%	3e-71	
I10C_A	Chain A, Crystal Structure Of Mutant Human Lysozyme, E6ea-I	271	271	89%	3e-71	S
I17S_A	Chain A, Structural Changes Of The Active Site Cleft And Diffe	270	270	87%	4e-71	S
I58W_A	Chain A, Crystal Structure Of Mutant Human Lysozyme Substiti	270	270	87%	5e-71	S
I686_A	Chain A, Crystal Structure Of Mutant Human Lysozyme Substiti	270	270	87%	5e-71	S

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 8

So, now if you click into this one, so these each line represents each protein, which are aligned. So, this is a result let us see where your query sequence. So, you aligned with the different other sequences right you can give this maximum score right in the total score right here. This is only the highly aligned sequence pairs and this is the total score and you can see the coverage what is the query coverage.

Student: Amount of the sequence covered.

Covered for example, if you have 148 residues this aligned with another sequence with 148 residues; that means, you cover all this residue entire length in this case the coverage is 100 percent. If it is 88 percent what is the meaning of 88 percent.

Student: The 88 percent (Refer Time: 24:14).

Residues are aligned others are not aligned correct. So, that is not the internal gaps with there is outside you can see they are not able to align, then E value. So, you can see e^{-3} less than 3 significant, but in many cases you get this significant value of e^{-81} and so on, and link with the other databases.