

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 of protein sequences. So, what 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 sequence is a same, and amino acid 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; visually you 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 or matching nucleotide, we give a score a reward or if it is a mismatching we give penalty, then we also introduce gap. So, what is the meaning of gaps in alignment?

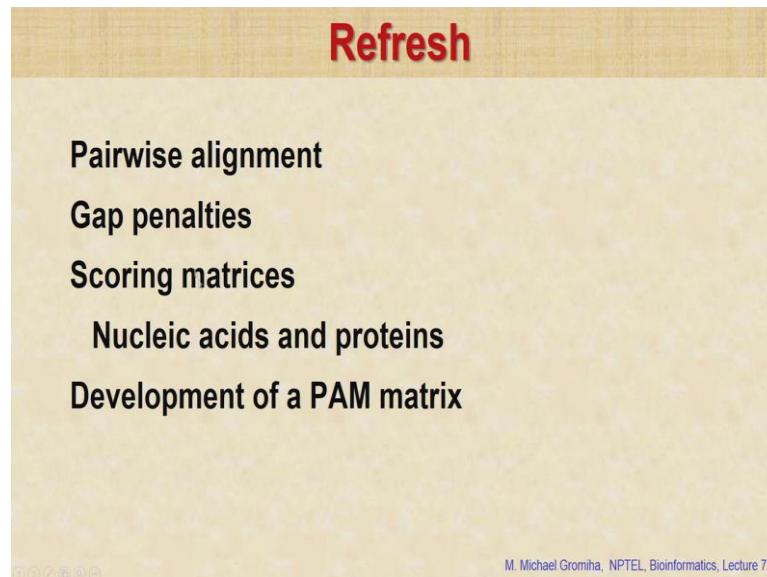
Student: Insertion or deletion.

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 have less 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 open

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 score to align 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 for proteins, for the case of nucleic acids how many bases?

Student: 4(Refer Time: 02:30).

4. So, in this case based on a substitution try to either reward or the penalty either mild penalty or severe 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)

Scoring Matrix: Amino Acids

- Physical/chemical similarity
 - E.g. pairing between two aromatic functional groups: a positive score
 - Pairing between nonpolar and charged residues: a scoring penalty.
- Hydrophobicity, charge and size
- Genetic code
- Actual substitution rates
- Point accepted mutation (PAM) matrix**

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

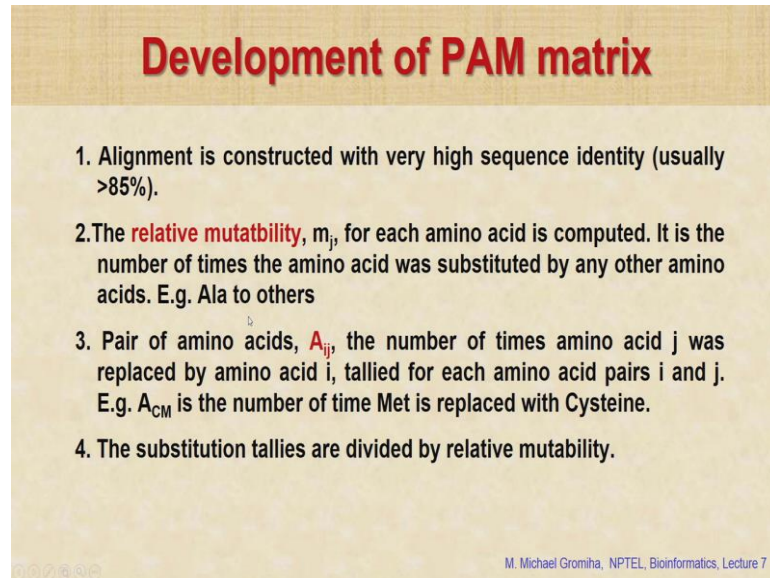
For example if the two residues are charged for example, aspartic acid and glutamic acid. So, in this case they are similar this we give the positive score compared with aspartic acid is replaced by valine or any hydrophobic residue. So, this is the pairing of this similar functional groups like either 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 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 we discussed about a genetic code how to give score based on genetic code? Number of mutations in the DNA, in the codons right, how many mutations in the codons. So, 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 90 are same.

So, how many variations? 10 percent this case 90 percent are 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)



Development of PAM matrix

1. Alignment is constructed with very high sequence identity (usually >85%).
2. The **relative mutability**, m_i , for each amino acid is computed. It is the number of times the amino acid was substituted by any other amino acids. E.g. Ala to others
3. Pair of amino acids, A_{ij} , the number of times amino acid j was replaced by amino acid i , tallied for each amino acid pairs i and j . E.g. A_{CM} is the number of time Met is replaced with Cysteine.
4. The substitution tallies are divided by relative mutability.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, here first we take the alignment with 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 alanine is mutated to others and the second one we need to see the exact mutations for example, a_{ij} how many times i is mutated to j .

For example you see A_{CM} number of time methionine is replaced with cysteine. Likewise we can see the relative mutability of any amino acid.

(Refer Slide Time: 05:48)

Development of PAM matrix

5. Normalize with the frequency of occurrence of each amino acid
6. Take log of each resulting entries in the PAM-1 matrix (PAM-1 means 1 substitution per 100 residues or 1 PAM unit) . This matrix is also called **log odds matrix**, since the entries are based on the log of the substitution probability for each amino acid.

PAM-1 matrix is appropriate to compare sequences are **closely related**.
PAM-1000 matrix might be used to compare sequences with **distant relationships**. Usually **PAM-250** is used for sequence alignment.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

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 why 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 homologous. So, to do this we use PAM one matrix. That is to compare the sequence which are closely related and use PAM 1000 for comparing the sequences with the distant relationship. That means, they are very not homologous to each other.

Usually, in the literature we use PAM 250 for the sequence alignment for example, BLAST or FASTA 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)

Calculation of a PAM matrix

PAM matrix is a **20x20 matrix** for all pairs

Consider a multiple sequence alignment

Assumption:
 Substitutions are equal in both directions
 (A to G and G to A)
 E.g.: Element GA

Frequency of pairs, $F_{GA} = 3$; Relative mutability, $m_A = 4$

Normalizing factor = number of mutations in the entire tree times 2, times relative frequency of A residues multiplied by 100 (1 substitution per 100 residues)

i.e., $6 \times 2 \times (10/63) \times 100 = 190.4762$

Hence, normalized relative mutability,
 $m_A = 4/190.4762 = 0.021$

1. ACGCTAFKI
 2. GCGCTAFKI (1: A→G) ①
 3. ACGCTAFKL (1: I→L)
 4. GCGCTGFKI (2: A→G) ②
 5. GCGCTLFKI (2: A→L)
 6. ASGCTAFKL (3: C→S)
 7. ACGCTAFKL (3: G→A) ③

Construct tree

M.I. = 0.013

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

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 ACGCTAFKI 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 FKL, then from this sequence then we have next mutation alanine is mutated to glycine; so here this alanine.

This alanine is mutated to glycine. So, it starts with G. So, the sequence here also GFKT then again A is mutated to leucine. So, here we have this A is mutated to leucine then go with the second one here we have two mutations one case 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 A 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, F_{GA} equal to 3, what is relative mutability M_A 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, this is the total 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; $9 * 7 = 63$. So, we will get this number 190.4762 then we get the normalized mutability, because this M_A is 4. So, 4 divided by this number will give you the normalized relative mutability. So, this is 0.021 so this here.

(Refer Slide Time: 10:12)

Calculation of a PAM matrix

$m_A = 4/190.4762 = 0.021$
 Mutation probability, $M_{ij} = m_j F_{ij} / \sum f_{ij}$
 $M_{GA} = 0.021 \times 3 / 4$
 $= 0.0157$
 $\sum f_{ij}$, total number of substitutions involving A
 $R_{ij} = \log(M_{ij}/f_i) = \log(M_{GA}/f_G)$
 $f_G = 10/63 = 0.1587$
 $R_{GA} = \log(0.0157/0.1587) = \log(0.0989); R_{GA} = -1.005$

Repeat for all off-diagonal elements.
 For diagonal elements: $M_{jj} = 1 - m_j$
 Calculate R_{jj}

Consider a multiple sequence alignment

1. ACGCTAFKI
2. GCGCTAFKI (1: A→G)
3. ACGCTAFKL (1: I→L)
4. GCGCTGFKI (2: A→G)
5. GCGCTLFKI (2: A→L)
6. ASGCTAFKL (3: C→S)
7. ACACCTAFKL (3: G→A)

Calculate the element R_{AA}
 Calculate the element R_{IL}
→ -0.685

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

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 M_{GA} 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 M_{GA} its equal to 0.0157. So, F_{ij} is the total number of substitutions involving alanine that is 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 f_i this frequency of G. M_{GA} is given as 0.0157 right. So, f_G 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 R_{GA} substitute value is here log of M_{ij} equal to 0.0157 divided by 0.1587. So, we give the value of -1.005. So, we can repeat this for all the off diagonal elements.

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

Student: 1.

Only one and the how many I involved in this 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 I's.

Student: 4.

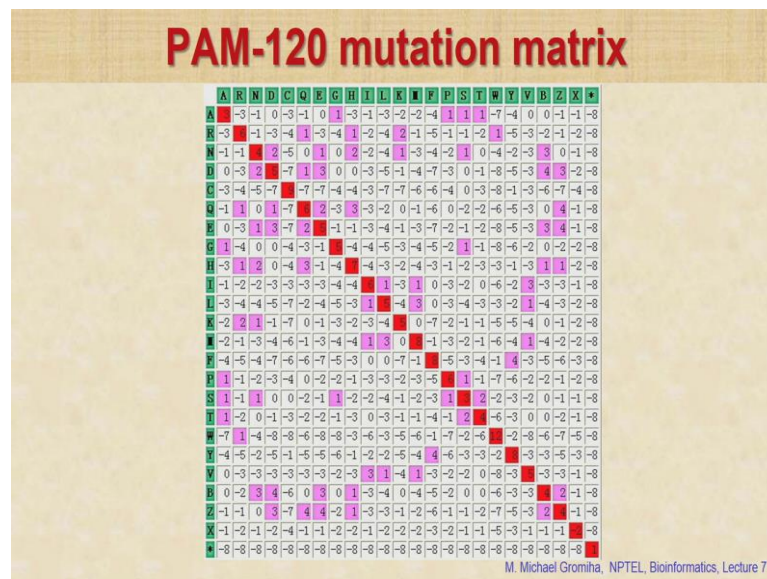
Four right. So, normalizing factor equal to 6 multiplied by 2 multiplied by 4 by 63 divided by 100. So, we get the value of 76.2, now we get these normalized relative mutability. So, it is only one isoleucine mutation 1 by 76.2. So, this is equal to m_i equal to 0.013, then we get this m_i solution to solution M_{IL} 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 you can work out in the free time.

This is for the off diagonal elements for the diagonal elements you will get the M_{ij} ; M_{ij} is here for example, R_{AA} alanine will get one minus m_j , m_j will get from this formula and we can get this values. Then we get the values M_{ij} then we will get the value of R_{ij} we can get that. So, you can derive this matrix right. I take a 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)



PAM-120 mutation matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	A
A	1	-3	-1	0	-3	-1	0	1	-3	-1	-3	-2	-2	-4	1	1	1	-7	-4	0	0	-1	-1	-8
R	-3	1	-3	-4	1	-3	-4	1	-2	-4	2	-1	-5	-1	-1	-2	1	-5	-3	-2	-1	-2	-2	-8
N	-1	-1	1	2	-5	0	1	0	2	-2	-4	1	-3	-4	-2	1	0	-4	-2	-3	3	0	-1	-8
D	0	-3	2	1	-7	1	3	0	0	-3	-5	-1	-4	-7	-3	0	-1	-8	-5	-3	4	3	-2	-8
C	-3	-4	-5	-7	1	-7	-7	-4	-4	-3	-7	-7	-6	-6	-4	0	-3	-8	-1	-3	-6	-7	-4	-8
Q	-1	1	0	1	-7	1	2	-3	3	-3	-2	0	-1	-6	0	-2	-2	-6	-5	-3	0	4	-1	-8
E	0	-3	1	3	-7	2	-1	-1	-3	-4	-1	-3	-7	-2	-1	-2	-8	-5	-3	3	4	-1	-8	
G	1	-4	0	0	-4	-3	-1	1	-4	-4	-5	-3	-4	-5	-2	1	-1	-8	-6	-2	0	-2	-2	-8
H	-3	1	2	0	-4	3	-1	-4	-4	-3	-2	-4	-3	-1	-2	-3	-3	-1	-3	1	1	-2	-8	
I	-1	-2	-2	-3	-3	-3	-4	-4	1	-3	1	0	-3	-2	0	-6	-2	3	-3	-3	-1	-8		
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	-4	3	0	-3	-4	-3	-2	1	-4	-3	-2	-8		
K	-2	2	1	-1	-7	0	-1	-3	-2	-3	-4	0	-7	-2	-1	-1	-5	-5	-4	0	-1	-2	-8	
M	-2	-1	-3	-4	-6	-1	-3	-4	-4	1	3	0	-1	-3	-2	-1	-6	-4	1	-4	-2	-2	-8	
F	-4	-5	-4	-7	-6	-6	-7	-5	-3	0	0	-7	-1	-5	-3	-4	-1	4	-3	-5	-6	-3	-8	
P	1	-1	-2	-3	-4	0	-2	-2	-1	-3	-2	-3	-5	1	-1	-7	-6	-2	-2	-1	-2	-8		
S	1	-1	1	0	0	-2	-1	-1	-2	-2	-4	-1	-2	-3	1	2	-2	-3	-2	0	-1	-1	-8	
T	1	-2	0	-1	-3	-2	-2	-1	-3	0	-3	-1	-1	-4	-1	2	-6	-3	0	0	-2	-1	-8	
W	-7	1	-4	-8	-6	-8	-8	-3	-6	-3	-5	-6	-1	-7	-2	-5	2	-8	-6	-7	-5	-8		
Y	-4	-5	-2	-5	-1	-5	-5	-6	-1	-2	-2	-5	-4	4	-6	-3	-3	-2	-3	-3	-5	-3	-8	
V	0	-3	-3	-3	-3	-3	-3	-2	-3	3	1	-4	1	-3	-2	-2	0	-8	-3	-3	-3	-1	-8	
B	0	-2	3	4	-6	0	3	0	1	-3	-4	0	-4	-5	-2	0	0	-6	-3	2	-1	-8		
Z	-1	-1	0	3	-7	4	4	-2	1	-3	-3	-1	-2	-6	-1	-1	-2	-7	-5	-3	2	-1	-8	
X	-1	-2	-1	-2	-4	-1	-1	-2	-2	-1	-2	-2	-2	-2	-1	-1	-5	-3	-1	-1	-1	-8		
A	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8

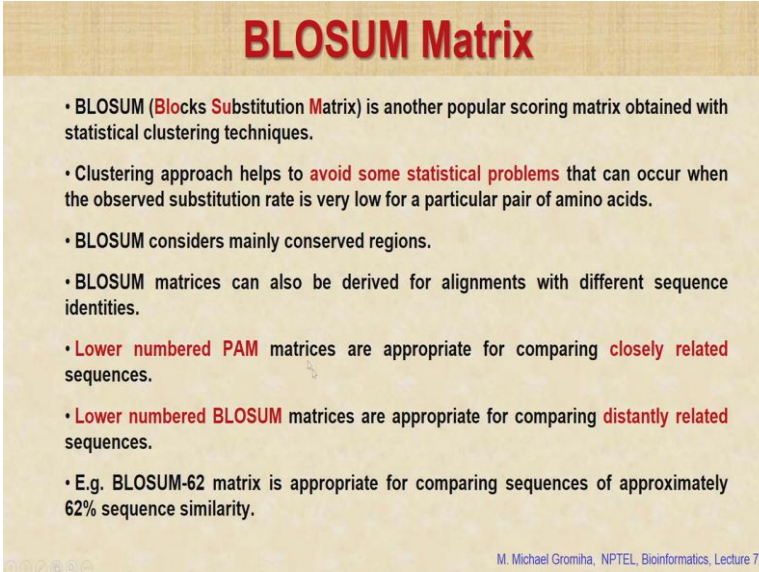
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

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

mutations have positive values some mutations have negative values. Can you see the positive values for example; 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 have minus.

For example: this region for example, if you substitute serine by leucine or alanine by phenylalanine. So, if you see some mutations are acceptable by nature, some mutations are not accepted by nature. So, based on the real frequency of substitutions, now we derive the PAM matrix likewise there is another matrix that 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 this is also another popular matrix.

(Refer Slide Time: 15:47)



BLOSUM Matrix

- BLOSUM (**B**locks **S**ubstitution **M**atrix) is another popular scoring matrix obtained with statistical clustering techniques.
- Clustering approach helps to **avoid some statistical problems** that can occur when the observed substitution rate is very low for a particular pair of amino acids.
- BLOSUM considers mainly conserved regions.
- BLOSUM matrices can also be derived for alignments with different sequence identities.
- **Lower numbered PAM** matrices are appropriate for comparing **closely related** sequences.
- **Lower numbered BLOSUM** matrices are appropriate for comparing **distantly related** sequences.
- E.g. BLOSUM-62 matrix is appropriate for comparing sequences of approximately 62% sequence similarity.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Here they use only the places where they are 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 rate is very low where in a particular pair of amino acids that reduces some sort of bias 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 the other way around the lower numbered BLOSUM matrices are appropriate for distantly related sequences. And generally we use BLOSUM62 for comparing sequences of about 62 percent sequence similarity this is very commonly used in the alignment programs.

(Refer Slide Time: 16:42)

BLOSUM-62 Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-2	-1	1	0	-3	-2	0	
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	0	-1	-1	-1	-2	-2	-1	-1	-1	-2	-1	1	5	-2	0	0	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	11	2	-3	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

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

Student: 11.

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

Student: -4.

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

Student: -7.

-7 there also adverse effect here also adverse effect; so we 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)

BLAST
Basic Local Alignment Search Tool

BLAST: Process the Query Sequence and Database

Divide the query sequence into all "words" of length $K=2$ (default 3 for proteins)

Query	Database
1 2 3 4 5 6 7	1 2 3 4 5 6
<u>QL</u> NFSAGW	<u>NL</u> NYTPW
QL	NL
LN	LN
NF	NY
FS	YT
SA	TP
AG	PW
GW	

Step 1: Hash table for sequence A →

Handwritten notes on the right:
QL
NL
0+4 → 4
LN
LN
4 6 → 10

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, 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 the tool to get your alignment, how it works that I will explain this. So, first, if we have the query sequence and we have database. How to map they divide this sequences in to small bits small words of length k. So, usually they use K equal to 2 or K equal to 3 for the case of proteins.

So, now we have the query sequence QLNFSAAGW 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 made into overlapping segments, then go to the database, here also you made into overlapping 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 QL and here NL. So, you compare this values what is Q and NQ and N0 L going L L to L it is 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)

BLAST
Basic Local Alignment Search Tool

Step 2:

Use all of the 2-letter words in query sequence to scan against database sequence and mark those with score ≥ 8

Note:
Marked points can be on the diagonal and off-diagonal

Identify Word Matches

		Query Sequence							
		1	2	3	4	5	6	7	
Query Sequence		Q	L	N	F	S	A	G	W
Database Sequence	1	N							
	2	L	*						
	3	N		*					
	4	Y			*				
	5	T							
	6	P							*
	W								

NF
NY
6+3=9

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, we use this matrix and you make any cutoff score 8 this is this can be adjustable. So, you can put an example, 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 get score of 9 NF NY what is NN.

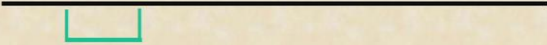
Student: 6.

NN is 6 FY Y F is 3 3. So, $6 + 3 = 9$. 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.

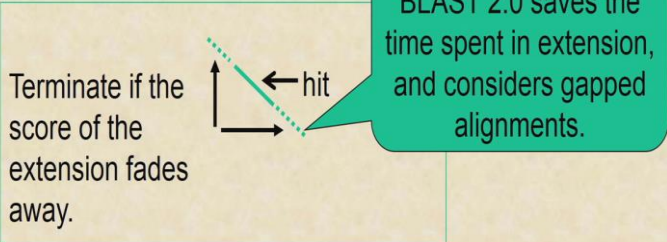
(Refer Slide Time: 20:35)

BLAST

Step2: Scan sequence b for hits.



Step 3: Extend hits.



Terminate if the score of the extension fades away.

hit

BLAST 2.0 saves the time spent in extension, and considers gapped alignments.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

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