

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

Lecture - 6b
Pairwise Alignment II

(Refer Slide Time: 00:17)

Example

1: AATCTATA
 2: AAGATA

Left

AATCTATA
 AAGATA

Middle

AATCTATA
 AAGATA

Right

match score = 1, if seq1(i) = seq2(i)
 match score = 0, if seq1(i) ≠ seq2(i)

The alignment scores are 4,1,3

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

Identity matrix

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Here I show the example. So, the same sequence I give AATCTATA, this is second one AAGATA. If I want to align these 2 sequences without introducing any gaps how many ways we can align 3 ways.

Student: 3 ways.

Because I do not want to introduce any gaps. So, the first sequence; I put here, the second sequence here I put the left side at the left mode side and here I put it in the middle and here I put the right side. Now we see which method is the good which alignment is the best. So, we use some scores, for example, if we take match score one or the if this sequence 1 and sequence 2 are same and the residues are the same then we give match score one.

We put the match score 0 if they don't have any match if there is any change if the residues are the same we give score one and if the residues are different we give score 0.

So, in this case, if you take the first alignments how many matches 1, 2, 3, 4 matches. So, the score will be;

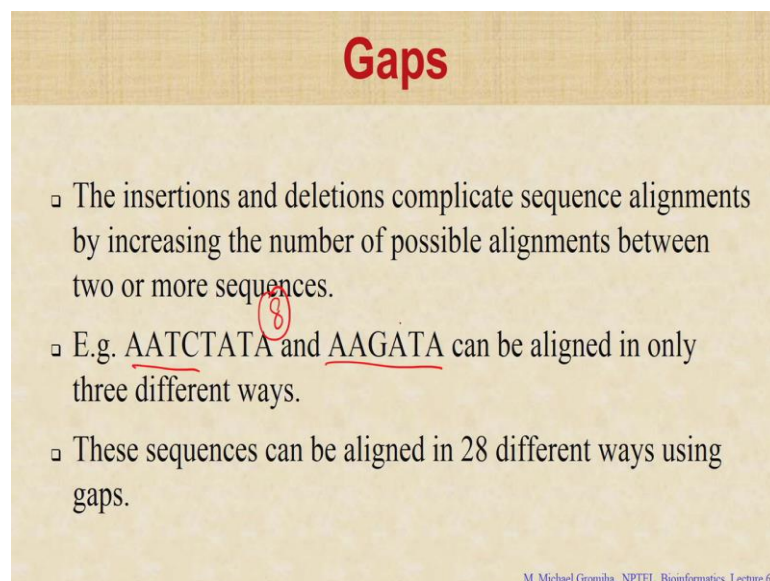
Student: 4.

4 because all the mismatch we put 0. So, 4 plus this 0 that is equal to 4 if we take the second example what is a score?

Student: 1.

1 because only one match and because the third alignment 3 right 1, 2, 3; matched with 3; let us say; we have the identity matrix, we use ATCG, see if it is same we put 1 and if it is different we put 0 that is fine.

(Refer Slide Time: 01:57)



Gaps

- The insertions and deletions complicate sequence alignments by increasing the number of possible alignments between two or more sequences.
- E.g. AATCTATA and AAGATA can be aligned in only three different ways.
- These sequences can be aligned in 28 different ways using gaps.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Now, if you introduce gap why do you introduce gaps.

Student: To better alignment.

To better alignment right, if you there is some changes right. So, we need the gaps. So, in this case, if you do the gaps, how many different ways you can align these 2 sequences the length is 1, 2, 3, 4, 5, 6, 7, 8 right. So, here you have a length as 3 6. So, how many different ways you can align these 2 sequences if you introduce gaps there are various ways.

So, there are 28 different ways you can align these 2 different sequences.

(Refer Slide Time: 02:37)

Gaps

$\begin{array}{c} \text{AATCTATA} \\ \text{AAG-AT-A} \end{array}$	$\begin{array}{c} \text{AATCTATA} \\ \text{AA-G-ATA} \\ 5-2+0=3 \end{array}$	$\begin{array}{c} \text{AATCTATA} \\ \text{AAG--ATA} \\ 5-2+0=3 \end{array}$
---	--	--

An alignment that includes gaps, an additional term, the **gap penalty** must be included in the scoring function.

Gap penalty = -1, if seq1(i) = "-" or seq2(i) = "-"

match score = 1, if seq1(i) = seq2(i)

^{mis} mismatch score = 0, if seq1(i) ≠ seq2(i)

The scores are 1, 3, 3

M. Michael Grunha, NPTEL, Bioinformatics, Lecture 6

So, show an example this is sequence 1, here sequence 2, this is sequence 3, here I put 2 gaps, here 1 gap is here, 1 gap is here, here also the again same here right and here you put the gap here, here you put the gap here, here these 2 gaps are similar together.

So, here we have to give a penalty now the previous alignment we gave the match score and we gave the mismatch score in this case here we have the gap. So, if we need to introduce gap penalty. So, here I put minus 1 if there is a dash either in the first sequence or in the second sequence then the match score is one this same as before see if this is equal to sequence 1 equal to sequence 2.

So, if there is a mismatch you can put the mismatch score. So, you consider mismatch score if sequence 1 is not equal to sequence 2, if you do this what is the score for the first alignments. So, this is 1, 2, 3, right for the alignment match score is 3.

Student: (Refer Time: 03:32).

This is minus 1, this is minus 2 and others are 0. So, this will be one. So, if we take this alignment. So, what is the matching score; 1, 2, 3, 4, 5.

Student: Minus 2.

Minus 2.

Student: 3.

plus 0 this is equal to 3. So, for this one.

Student: Again 1 3.

1, 2, 3, 4, 5; 5 minus 2 plus 0 this is equal to 3, right. So, if you the earlier alignment we got the score here without gapping gap. So, we get the score of 4 1 3, now we align with the gaps. So, if we change these score we get 1 3 3.

(Refer Slide Time: 04:25)

Sequence alignment

Position	1	2	3	4	5	6	7	8	9	10	
Seq A:	V	E	I	T	G	E	I	S	T		
Alignment 1		P	R	E	-	T	E	R	I	T	
Score:	0	-1	1	-1	1	0	0	0	-1	1	
Total:	1										
Seq A:	V	E	I	T	G	E	I	S	T		
Alignment 2		P	R	E	T	-	E	R	I	T	
Score:	0	0	0	1	-1	1	0	0	1		
Total:	2										
Seq A:	-	V	E	I	T	G	E	-	I	S	T
Alignment 3	P	R	E	-	T	-	E	R	I	-	T
Score:	-1	0	1	-1	1	-1	1	-1	1	-1	1
Total:	0										

Handwritten notes: For Alignment 1, a red arrow points to the score 1, and a calculation $4 - 3 + 0 = 1$ is written. For Alignment 2, a calculation $3 - 1 = 2$ is written. To the left of the alignments, 'Seq A' is written next to 'VEITGEIST' and 'Seq B' next to 'PRETERIT'.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Now, next question is if you have this gaps and if you have this mismatch whether we need to give same weight or different weight before that I give another example.

So, here we have the protein sequence. So, here give the DNA sequence and I give another protein sequence here, this is sequence A here, sequence B there are different ways we can align these 2 sequences. So, here if I run the first one. So, I put some gaps in the sequence A another sequence B in the second alignment there is no gap in the sequence A, but some sequence in B and the third one we will introduce gaps.

So, we take one example say if you take the first alignment. So, this is same T and T are same I and I are same this T and T are same. So, this will be 1, 2, 3, 4 right and here this is a gap. So, minus 1 this is a gap minus 1 and here again a gap this minus 1. So, 4 minus 3 that is equal to plus 0 that is equal to 1. In the second one, so, we did not give gaps.

So, in this case, here this is 1, 2, 3 minus 1 this is equal to 2. So, if you compare the alignment between one and 2 what is the difference? gaps. So, the difference is a gap because of the gaps we give the penalty. So, this is why alignment 2 is better than alignment one. So, if we try to minimize the gaps and maximize the score right this we can do in several different ways.

So, now another issue is we introduce gaps now if you see this alignment. So, we introduce gaps, but there is a difference between the introduce in the gaps if you see the first alignment how many gaps we introduced 2; 2 are the same positions or different positions how many different 2 different times we introduced gaps this one time we introduce here the second time we introduce here, here we introduce one time here and second time here and third one we introduce gap together.

See if you look into these evolutionary rates or look into this is the origin of different organisms. So, getting continuous gap is more closer than introducing gaps at the different places or the otherwise if your insertions or deletions which happened at different places are less probable than having this insertion-deletions or together. So, if we wanted to take into this account; what you have to do.

Student: (Refer Time: 07:01).

We have to give some penalties for the origination between the gap and we have to give a penalty for each gap, right. So, we will do that. So, we have 2 sequences, arbitrary sequences one is 12 residues, another nine residues. So, we have a shortage of 3 residues. So, we have introduced 3 gaps.

(Refer Slide Time: 07:14)

Origination and length penalties

- One method to further distinguish between alignment is to differentiate between the alignments that contain many isolated gaps and those that contain fewer, but longer, sequence of gaps. E.g.
- Consider two arbitrary sequences of lengths 12 and 9. Any alignment will have a shortage of 3 gaps in the shorter sequence.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Either these 3 gaps can be together or we can separate the gaps here and there. So, if we have the homologous sequences and some residues are missing at the N terminal or the C terminal, in this case, we can keep these 3 residues right either insertion in one sequence or the deletion in the other sequence right at one phase. So, in this case, we can cover to other sequences.

(Refer Slide Time: 07:52)

Origination and length penalties

Gap penalty is divided into:

- Origination penalty (for starting new series of gaps on one of the sequences being aligned)
- Length penalty (number of sequential missing characters)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

In this case, we give penalty based on the gaps where we introduce. So, given an example right. So, one is the origination penalty. So, that depends upon the gaps, we

introduced here for each series of gaps where we introduce in a sequence we give a penalty that is called origination penalty. And the second one is length penalty there is a number of missing characters when the sequence 1 and the sequence 2 when we align each other how many places where this is not how many places you have the gaps ok.

(Refer Slide Time: 08:16)

Origination and length penalties – contd..

A AATCTATA	AATCT A TATA	A A TCTATA
AAG A T A	AA-G-ATA	AAG--ATA

$3 + 0 + (2 \times -2) + (2 \times -1) = 3 - 4 - 2$

E.g. Origination penalty: -2; length penalty: -1, match score: +1 and mismatch score: 0

Case 1: $2 \times -2 + 2 \times -1 + 3 \times 1 + 3 \times 0 = -4 - 2 + 3 + 0 = -3$

Case 2: -1 $5 - 4 - 2 = -1$

Case 3: $1 \times -2 + 2 \times -1 + 5 \times 1 + 1 \times 0 = -2 - 2 + 5 + 0 = +1$ $5 - 2 - 2 = 1$

Case 2 and case 3 are different; they were same in previous scoring method

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, now we gave the 3 sequences the same 3 sequences here we have the gaps at 2 different places and here also we have the gaps at 2 different places and here the gaps are at the same place. So, if we take the origination penalty of minus 2 because we are introducing right, we do not want to introduce because nature does not want to select these insertion-deletions. So, we give the penalty of minus 2.

Length penalty of minus 1 and the match score of one and the mismatch score is 0, if you do this, what is the score for the first alignments; what is a match score.

Student: 3.

3 mismatch score is 0 anyway and what is how many originating penalty.

Student: 2, 2 and 2 minus 2.

2 times right one, we originate here one we originate 2 here rights. So, plus 2 into minus 2 right and plus how many length penalties.

Student: 2.

2 times multiply it by.

Student: Minus 1.

Minus 1. So, this will get 3 minus 4.

Student: Minus 2.

Minus 2 right minus 2. So, minus 6 3 minus 6 that is equal to;

Student: Minus 3.

Minus 3 right, fine. Now, we take the second one. So, here what is the score 5.

Student: Plus 3.

Minus 4 minus 2; fine, this equal to minus 1 if we take the last 2 1. So, here 5 minus 2 minus 2 correct this is equal to 1. So, if you compare these 2 alignments here, we introduce only once here if it is twice right as per the selection by nature, right we can this will prefer this alignment done this one because we insert 2 times here. So, here we get the better score than the other ones here we use the very simple one putting the value of one and minus 1 for the penalty right minus 2 for the origination penalty and 0 for the mismatch, but is it reliable.

(Refer Slide Time: 10:34)

Scoring matrices

- In previous alignment, for non-gap positions scores are given as 1 for match and 0 for mismatch.
- These scores can be further refined based on the substitutions.
- For example Alanine is replaced with Valine or Lysine or Lysine ^{ive}
- Once the alignment score for each possible pair of nucleotides/amino acid residues are determined, the resulting scoring matrix is used to score each non-gap position in the alignment.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, if you give just one and 0 is it fine for the alignment, because 1 and 0. So, for that, we can change for example, if you in the example case of nucleic acids rights say the followings are DNA.

So, how many different bases 4 different bases. So, that the current this is scenario whatever the changes we get the same score, but that is not reliable right sometimes we change purine by purine sometimes we change pyrimidine by pyrimidine sometimes we change purine to pyrimidine and pyrimidine to purine likewise in the amino acids different types of changes. For example, if you replace alanine by valine. So, there is an effort because alanine is a hydrophobic amino acid valine is hydrophobic.

What is the difference between these two?

Student: Sidechain.

Side chain this is bulkier than alanine. Alanine has only one CH₂ group, valine has 3. So, we have the bulkier group, but if you replace alanine by lysine there is a positive charge. So, here this is hydrophobic this will change the environment. So, it may not be good to have the same score if you replace the alanine with the valine or a lysine right say we need to consider the effect of the mutations.

So, in this case, we have to give the scoring in a different way right how to do that, but the blast.

(Refer Slide Time: 11:56)

Scoring matrices

- For nucleotides scoring is simple:
- In BLAST, same nucleotides are given a score of +5 and different ones have -4
- Case 2: matching nucleotides: mild reward (+1)
- Transitions (purine to purine, A or G)/ pyrimidine to pyrimidine (C or T): mild penalty (-1)
- Transversions (purine to pyrimidine or vice versa); severe penalty (-5)

In the blasts sometimes they use the same nucleotides is score 5 and the difference here 4 or you can give the mild reward plus 1 or we can give the score of minus 1 in the case of transitions; what are transitions; what is called transition.

Student: Purine to purine.

Purine to purine; like A or G right and pyrimidine to pyrimidine C or C right they change vice versa, then there is another called transversions so; that means, purine by pyrimidine and the pyrimidine to purine why they give the penalty of minus 5 and minus 1.

Student: (Refer Time: 12:27).

Right because what is the definition of the purines.

Student: To (Refer Time: 12:32).

Right, but the pyrimidine they have how many rings.

Student: Purine 2 rings, pyrimidine 1.

Purine 2 rings, pyrimidine 1 right (Refer Time: 12:39) here you are changing 2 rings to 2 rings or 1 ring to 1 ring that is fine this way they put the penalty of minus 1 and the other way if you change the other way around. So, 1 ring with the 2 rings or 2 rings to the 1 rings. So, it may create either the crowded situation or that is totally free this way they do not avoid the situations. So, they put minus 5 ok.

(Refer Slide Time: 12:59)

Scoring matrices..contd

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

BLAST matrix

	A	T	C	G
A	1	-5	-5	-1
T	-5	1	-1	-5
C	-5	-1	1	-5
G	-1	-5	-5	1

Transition-Transversion matrix

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Now, these are matrices.

So, here instead of 4 nucleotides right A T C G. So, earlier we use 1 1 1 1. Now we change its right this is a; if it is a score, match score we put one right if it is purine to purine or pyrimidine to pyrimidine we give minus 1 right if it is another way around purine to pyrimidine or pyrimidine to purine. So, we give minus 5 if you make this score what will happen in the alignment.

Student: mostly that alignment will be preferred where purine to purine.

Purine right. So, either they try to match or if you want to mutate, they try to make the similar type of amino acid residue or nucleotides they avoid the other way around because some (Refer Time: 13:47) instances we cannot avoid in this case we use, but this case we will give less score this is a minus 5 this is for the nucleotides what will happen in the case of amino acids how to deal with the amino acids.

Student: Then the 20 cross 20 matrix.

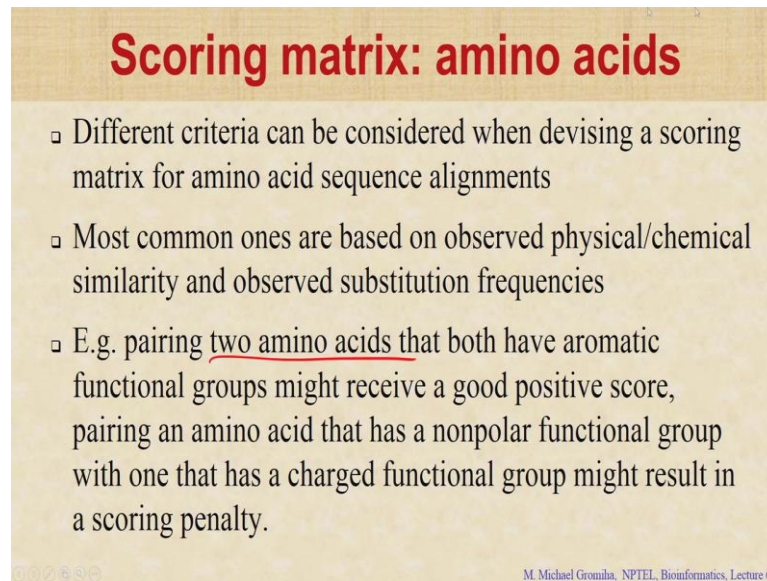
Right. So, we will have 20 different amino acid residues. So, 20 different amino acids are classified into 2 major groups for 2 different major groups.

Student: Hydrophobic.

Hydrophobic and hydrophilic and hydrophobic we have different groups like aliphatic, aromatic or Sulphur containing residues right and the hydrophilic, positive charge.

Negative charge as well as the polar right. So, now, we can see the mutations whether these 2 amino acids.

(Refer Slide Time: 14:31)



Scoring matrix: amino acids

- Different criteria can be considered when devising a scoring matrix for amino acid sequence alignments
- Most common ones are based on observed physical/chemical similarity and observed substitution frequencies
- E.g. pairing two amino acids that both have aromatic functional groups might receive a good positive score, pairing an amino acid that has a nonpolar functional group with one that has a charged functional group might result in a scoring penalty.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Right or both have the aromatic functional group. So, in this case, we can give a good positive score or we can if it is a nonpolar functional group with a charged group this is the aromatic to aromatic right that is fine aliphatic to aromatic is fine. So, if there is a non-polar group with the charged group, here we give the penalty because they alter situations right this alter the stability or alter the function like we discussed in the case of sickle-cell anemia (Refer Time: 14:59). So, what is the mutation?

Student: (Refer Time: 15:02) glutamine acid 6 to valine

Glutamine acids 6 to valine right. It causes the diseases right (Refer Time: 15:08) sickle-cell anemia. So, in this case, we need to give a penalty. So, that we can align like this then what are the different ways to have this matrix. So, we have 20 different amino acids what are the possibilities of changing a specific amino acid to other amino acids.

(Refer Slide Time: 15:28)

Scoring matrix: amino acids

- Scoring matrices have been derived based on residue hydrophobicity, charge and size
- Another option is based on genetic code: minimum number of nucleotide substitutions are necessary to convert a codon from one residue to other

4 → 20

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Either you can change the hydrophobicity like just we discussed or we can see the charge or we can see the size small amino acids to amino acid here, they do not care about the hydrophobicity, but they give it a small residues like serine to alanine or glycine to serine or they are the bulkier groups lysine to phenylalanine right. So, they give the size right you can derive the matrices right they can allow.

Then another option is the genetic code. So, how many nucleotide substitutions are necessary to convert a codon to an amino acid right. So, how many nucleotides?

Student: 3.

Right totally 4 nucleotides at ATCG, right. So, 4 bit of how many amino acids 20 right. So, in this case we discussed earlier about the genetic code some cases we have only one mutation sometimes there is 2 right. So, depending upon number of substitutions in the nucleotide they lead to the amino acid

So, we can accordingly you can change single substitution they group together 2 substitutions, they group together right likewise, they can make the genetic code to see how you can reliably you can align the protein sequences.

(Refer Slide Time: 16:33)

Scoring matrix: amino acids

- A common method for deriving scoring matrices is to observe the actual substitution rates among various amino acid residues in nature.
- If substitution between two amino acid residues is observed frequently, then positions in which these residues are aligned favorably.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

And the common method right then you can align the hydrophobicity you can align with the size you can align with the charge you can align with the number of changes in the codon, but the common method to derive this scoring matrix it's mainly the take the substitution rates actually substitution rates.

For example, I showed the hemoglobin sequences take the actual sequence and just they align and then see what is the actual rate how many times alanine is mutated to valine how many times alanine is mutated to aspartic acid. So, to take the real once and from this real cases they derive the matrices how many what is the probability of a specific residue to A; to be mutated to the residues B, right see if it is high or low in the real cases right for the different organisms.

So, based on that they derive the matrix this is called the scoring matrices and how to do that say.

(Refer Slide Time: 17:27)

Scoring matrix: amino acids

- Likewise, alignments between residues that are not observed to interchange frequently in natural evolution is penalized.
- One commonly used scoring matrix based on observed substitution rates is the point accepted mutation (PAM) matrix.
- The scores in a PAM matrix are computed by observing the substitutions that occur in alignments between similar sequences.

*100 residues → 90 matches
(10) mismatches*

*seq 1 2 AITV
seq 2 1 LAATV*

*I → A
A → I*

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

In the alignment, if the residues are aligned quite frequently they give the positive values and their alignment if they are not observed in this case we need to penalize. So, we gave the less score. So, how to do this from the alignment of different sequences of various this homologies right homology means how far they are similar right that I will discuss in the later classes.

So, the derived matrix that is called point accepted mutation matrix right, this matrix called the PAM matrix what is the PAM matrix.

Student: They will derive a like scores on the basis of how frequent the mutation occurs.

Occurs right. So, this is P A M stands for.

Student: Point accepted.

Point accepted mutation matrix right. So, you can see you can derive some PAM matrix right by the substitutions that occur in the alignments between similar sequences. So, if we have sequence 1 sequence 2 several pairs of sequences right some of them, you can see 100 percent match some of them 90 percent match some of them; 80 percent match right, they use different types of sequences with the different matches, for example, 90 percent.

So, if we take all the sequences which we have the similarity of more than 90 percent in this case if there are 10 residues nine will match one is different like we have 100 residues 90 will match and if we take the 100 residues if it is 90 percent sequence identity. So, 90 will match right 90 matches and 10 mismatches and this takes this 10 and see the rate which residue in sequence 1 is mutated to which residue in 2. See if you look into this very carefully there are 2 sequences right sequence 1 sequence 2 for example, AITV. So, here AATV; what is a substitution.

Student: I to A.

I to A; if I take this as sequence 1 and this as sequence 2 then what is the substitution a to;

Student: I.

I. So, it is changed. So, when they derive this substitution matrix say they do not care whether this from one sequence first 2 to 1. So, they take this as similar. So, for development of the PAM matrix, these where they get the diagonal matrix. So, you have one side we get the data this is second data is this mirror image of the other right that is we get the diagonal matrix.

(Refer Slide Time: 19:57)

Development of PAM matrix

1. Alignment is constructed with very high sequence identity (usually >85%).
200 → 170 residues
2. The relative mutability, m_j , 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
A → G
A → K
A → 19 mutations

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Student: Cysteine is mutated to (Refer Time: 21:28).

And how many times right; Cysteine is mutated to.

Student: (Refer Time: 21:34) Methionine.

Right, methionine right. So, so methionine is replaced to cysteine right; how far how many times that they have this mutation cysteine to methionine right now this is the specific mutations in the alignment that depends upon how many times this occurs in the sequence and also this is how many times cysteine (Refer Time: 21:55) is mutated to other residues right. So, we need to consider the all these concepts to derive the PAM matrix.

So, for example, if you take any sequence for example, if you take alanine to glycine what are the different aspects we need to consider to derive the PAM matrix frequency of alanine right first we need a frequency of alanine and then;

Student: Mutation.

Mutation frequency of alanine say this is one and the second one is how many times A is mutated. So, mutation frequency and then.

Student: (Refer Time: 22:32).

Then how this specific pairs right for example, how many what are the preferred specific mutations, for example, the mutations A to G, right we need to consider all these aspects right to derive the PAM matrix.

(Refer Slide Time: 22:49)

Development of PAM matrix..contd

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.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, now we normalize the frequency of occurrence of each amino acid and finally, take the log of each value; this will give you the PAM matrix why did you take the log.

Because otherwise (Refer Time: 22:57) there will be a large number. So, in this case, we can explain right. So, in this case on the log scale. So, you can explain the probability of each amino acid residue to be replaced with the other residues in the evolutionary rates. So, here different matrix, for example, a PAM 1. So, this means one substitution for hundred residues right this is called the PAM one matrix this is also called log-odds matrix right because the entries are based on the log of the substitution probability, then let us see how we derived the PAM matrix.

(Refer Slide Time: 23:29)

Development of PAM matrix..contd

- 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 6

I will show one example right and we will derive the matrix.

For example, if we have the 10 sequences we align the different sequence right and then see what are the different mutations and how we account these mutations to construct the PAM matrix. So, essentially if you take PAM one right this is to compare this sequence is a closely related. That means, they take highly-highly homolog sequences and if you take the PAM one hundred one thousand this mainly with the distant relationship. So, they use various levels of a sequence of homology to derive the matrix.

Normal normally you can use PAM 250, this is the usual once we for any alignment right for the generally aligning 2 sequences now we will see how we derive the PAM matrix. So, what is essentially a PAM matrix what is an expansion for PAM matrix point acceptability matrix.

So, now what is how the matrix looks like 20 by 20 matrix. So, all the 4 hundred elements are different or any anything is similar.

Student: Symmetric.

Symmetric right because of a diagonal matrix because of the reason I explained earlier. So, we get a symmetric matrix.

(Refer Slide Time: 24:35)

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, $F_{G,A} = 3$; 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)
3. ACGCTAFKL (1: I→L)
4. GCGCTGFKI (2: A→G)
5. GCGCTLFKI (2: A→L)
6. ASGCTAFKL (3: C→S)
7. ACGCTAFKL (3: G→A)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, now we have the different sequences sequence 1, 2, 3, 4, 5, 6, 7, I gave the different sequences and from different one sequence another sequence you can see the mutation say from one to two. So, the question number one A is mutated to glycine, right.

For example, if you take this one here ACG CT AF KI right this is one, next go to 2 what is the mutation A is mutated to G right. So, now, here the sequence is GCG, this is same right C TA F KI, then the third sequence we get one is. So, here the mutation is I to L right. So, here this is the change. So, we get this sequence ACG C TA F K L.

Now, the second one; we make another change a to G. So, here makes changes A to G, we get this sequence GC G GC G CTG F K I, we can see the change C to G, this A to G right, fine A to G, this is the A and here you change to G and here you can see the change to A to L with this a here and here is AL. So, here A to L and then we go to the other different mutations the third one that is C to S right from here C to S and the last one you can see G to A, here G to A.

So, you can construct a tree depending upon the substitutions in each sequence. So, I think I will discuss the development of the PAM matrix in next class. So, the first recap what are the different aspects we discussed in today's class; a first alignment what is an alignment.

Student: Comparison of 2.

Comparison of 2 sequences right. So, when you compare 2 sequences their sequence A and sequence B what are various different changes.

Student: Mutations.

Mutation; what is a mutation?

Student: Substitution.

This is a substitution. Change of 1 nucleotide; 1 amino acid by the other one right say what is an insertion.

Student: 1 or 2 nucleotide.

One or two nucleotides or amino acids, they have inserted the sequence; what are deletions?

Student: (Refer Time: 27:13) 1 or 2 amino acids.

Amino acids residues or nucleotides are deleted from the first sequence, right. So, there are different ways, then how to align sequences we will discuss about different aspects first one; we have 2 sequences of different length right, you can align without any gaps and second aspect we introduced a gap right in different places right and the third we discussed the difference between the origination and the gap penalty; how many gaps and how many times we introduced gaps right based on that we aligned, then you for a scoring we have the other various ways to score; the first one is for the if you take nucleotide; this is a how to order different ways to score.

Student: Transition.

Transitions or the transversions as well as a match score right. In the case of amino acids. So, what are the different ways to score?

Student: (Refer Time: 28:07).

Based on the hydrophobicity.

Student: (Refer Time: 28:10)

Based on charge.

Student: (Refer Time: 28:11) size.

Based on size and based on the changes in the codons.

Student: (Refer Time: 28:16).

And the actual one we can check the evolutionary rates if we align 2 sequences what is the actual changes from one sequence to the second sequence right. So, based on that we can derive PAM matrix. So, what are the various factors one has to consider for the development of PAM matrix?

Student: Frequency of amino acids.

frequency of amino acids.

Student: Mutation frequency.

frequency of mutation.

Student: specific mutations (Refer Time: 28:38).

And the probability of specific mutations right and we include some normalization factors. Finally take the log to get the log-odd matrix, right. This is why PAM matrix is also called log odd matrix, right. Then for the specific example, we will discuss in the next class.

Thanks for the kind attention.