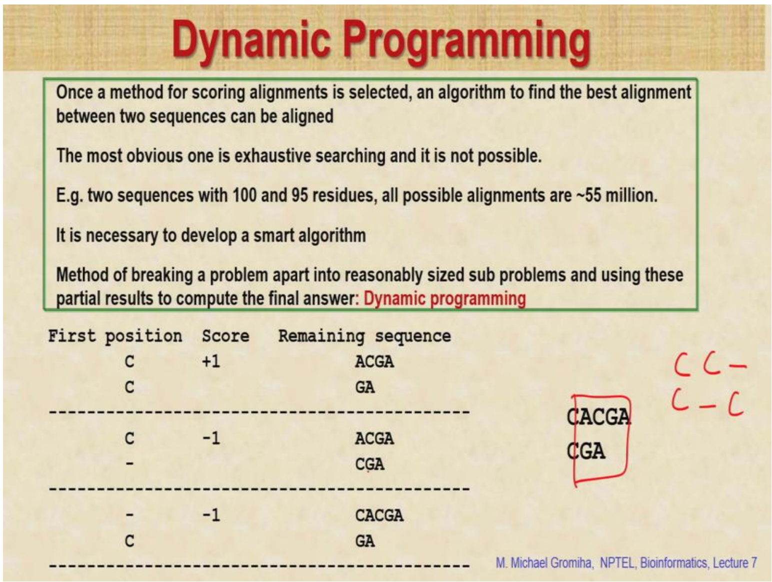
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 7b
Sequence Alignment II

(Refer Slide Time: 00:17)



Dynamic Programming

Once a method for scoring alignments is selected, an algorithm to find the best alignment between two sequences can be aligned

The most obvious one is exhaustive searching and it is not possible.

E.g. two sequences with 100 and 95 residues, all possible alignments are ~55 million.

It is necessary to develop a smart algorithm

Method of breaking a problem apart into reasonably sized sub problems and using these partial results to compute the final answer: Dynamic programming

First position	Score	Remaining sequence
C	+1	ACGA
C		GA
C	-1	ACGA
-		CGA
-	-1	CACGA
C		GA

CACGA
CGA

C C –
C – C

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Now the problem is if you have the large sequences. For example, if we have 100 sequences and 95 residue; 95 residues, right, 2 sequences; one with the 100 residues, another 95 residues and if you use the all possible alignments, you can get about 55 million possibilities. So, how to get the best alignment? So, the most obvious one is exhaustive searching, but that is not possible because of a lot of combinations.

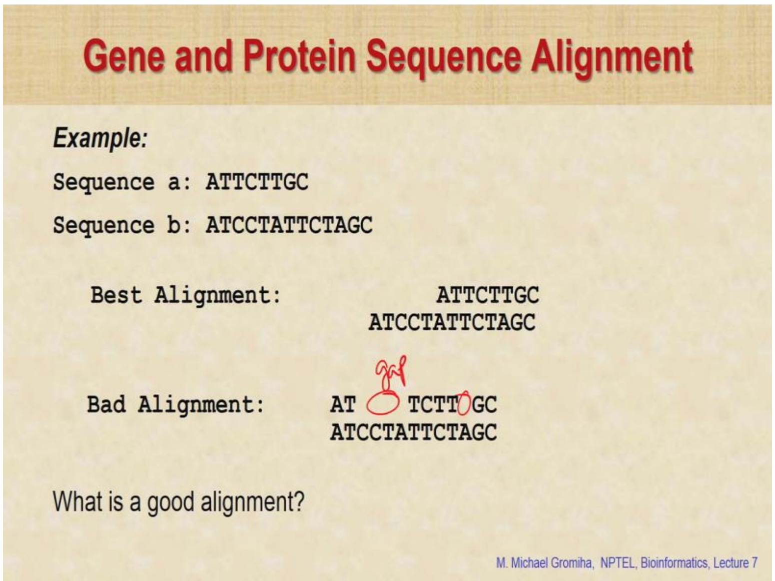
So, in this case, we need to develop a smart algorithm; different ways to get this alignment. So, one of this aspect is the dynamic programming; how it works; in this method, it breaks the problem into a reasonably smaller once and do the analysis and finally, combine together to get the final answer. For example, if I give the sequence CACGA and CGA, first we start with the first one; take the first one C and C, there are 3 different ways you can align the first one; what are 3 different ways; you can make C and C together.

Student: Second position.

Or you can put C and the gap here and you can put a gap here and C here, we do C and C together, then this is aligned. So, you get this score of +1 if this C is taken out, then this is the remaining one and the second aspect if you put the gap in the second one and the first one is already aligned. So, remaining is CGA and the second one we did not do anything. So, CGA is as it is third option if you put the gap in the first one right. So, we have; we did not use any nucleotide in the first sequence. So, it is as it is second one includes you use the c; so here residue GA. So, you can continue as it is and then appropriately score.

And finally you can add the whole scores and then see what will be the probable alignment. So, in this case there are various routes for a particular sequences pair of sequences.

(Refer Slide Time: 02:13)



Gene and Protein Sequence Alignment

Example:

Sequence a: ATTCTTGC

Sequence b: ATCCTATTCTAGC

```
Best Alignment:          ATTCTTGC
                    ATCCTATTCTAGC

                      gap
Bad Alignment:     AT    TCTT GC
                   ATCCTATTCTAGC
```

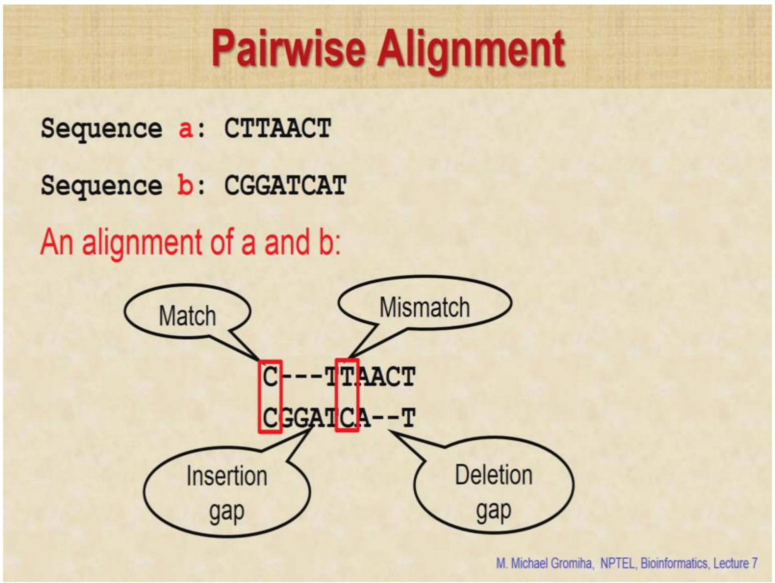
What is a good alignment?

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

For example if you see a sequence a, this is a sequence a 3 6 8; 3 6 8 nucleotides and the sequence b is a longer one; now different ways to align, for example, if you align this way or this way; here I put together everything on the right side second one; I give gaps in between here is some gap and here is a gap. So, which one is a best alignment this one or this one first if you looked into the alignment.

You can see because here everything is at the right side and some of them are align properly. So, we tell this is good alignment, second also if you see they probably put some cases which are whatever they align they put these residues and other places they put the gaps. So, which one is the good alignment. So, which one we need to use.

(Refer Slide Time: 02:57)



Pairwise Alignment

Sequence a: CTTAACT

Sequence b: CGGATCAT

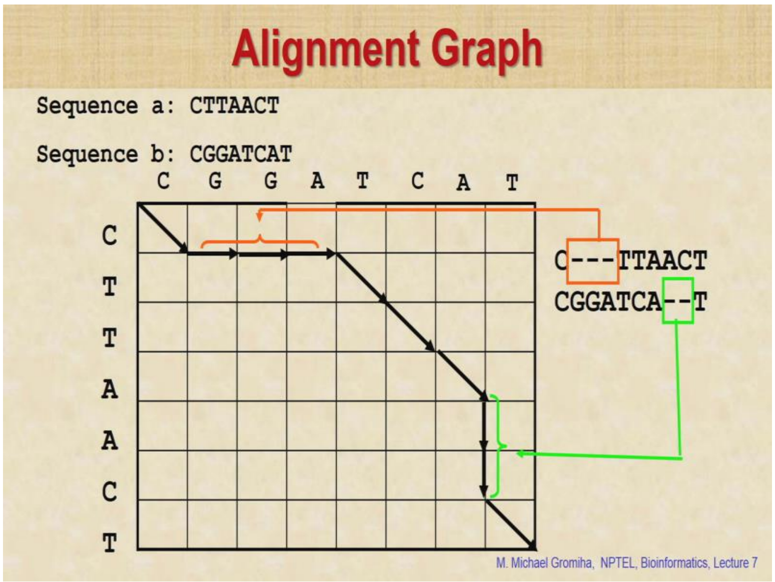
An alignment of a and b:

```
C---TTAACT
CGGATCA--T
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, different ways for example, is the 2 sequences when you align the sequence here this is a match because both are the same as I have discussed earlier, this is a mismatch, because the both are different here we inserted the gaps here. Because this is deleted. So, deletion gap here we inserted the nucleotides; so insertion gap.

(Refer Slide Time: 03:16)



Alignment Graph

Sequence a: CTTAACT

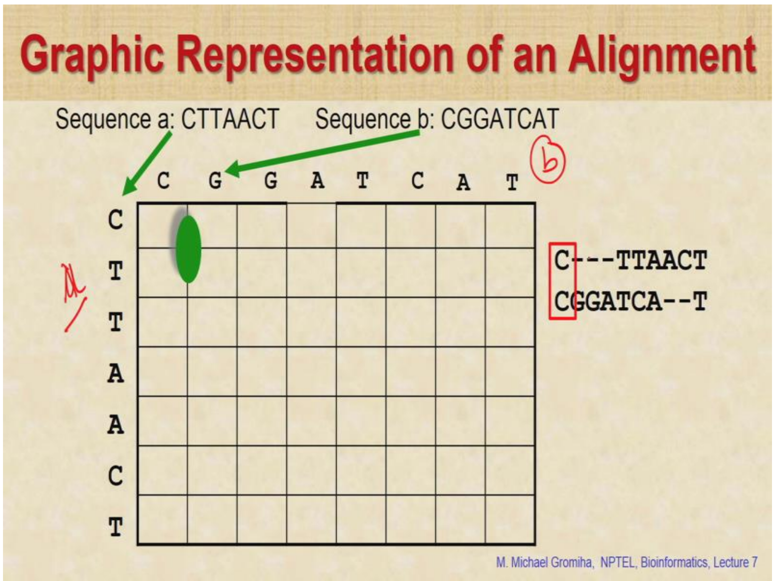
Sequence b: CGGATCAT

```
C---TTAACT
CGGATCA--T
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Now how to align; so here you put sequence a here and the sequence b here there are different ways to align. So, if you make this type of alignment first one is C. So, C and C connected and 3 gaps. So, here if you go on this direction and then again they align the 3 residues 1, 2, 3 right 1, 2, 3 and then again 2 gaps here 2 gaps and finally, they are end aligned there in the corner.

(Refer Slide Time: 03:44)



Graphic Representation of an Alignment

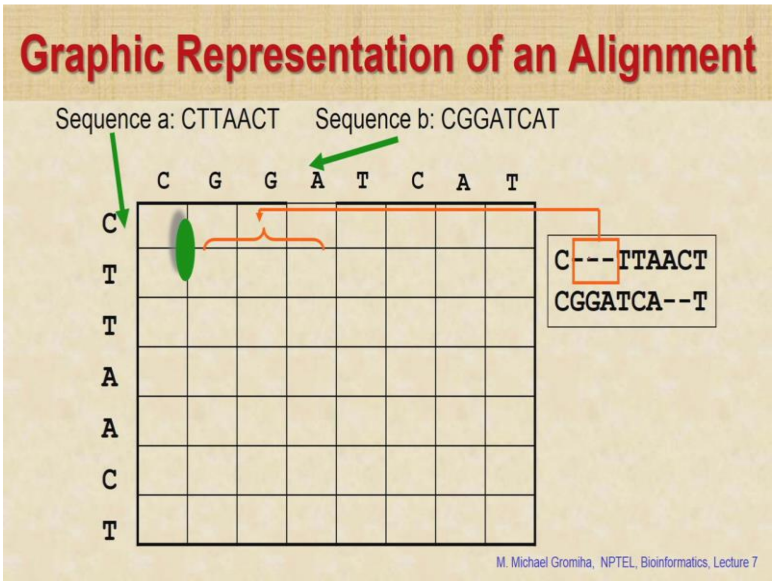
Sequence a: CTTAACT Sequence b: CGGATCAT

```
C---TTAACT
CGGATCA--T
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, take the first one this is the sequence a; this is sequence a. So, here this is sequence b; first you C and C we aligned. So, C and C we put it here and they put 3 gaps.

(Refer Slide Time: 03:57)



Graphic Representation of an Alignment

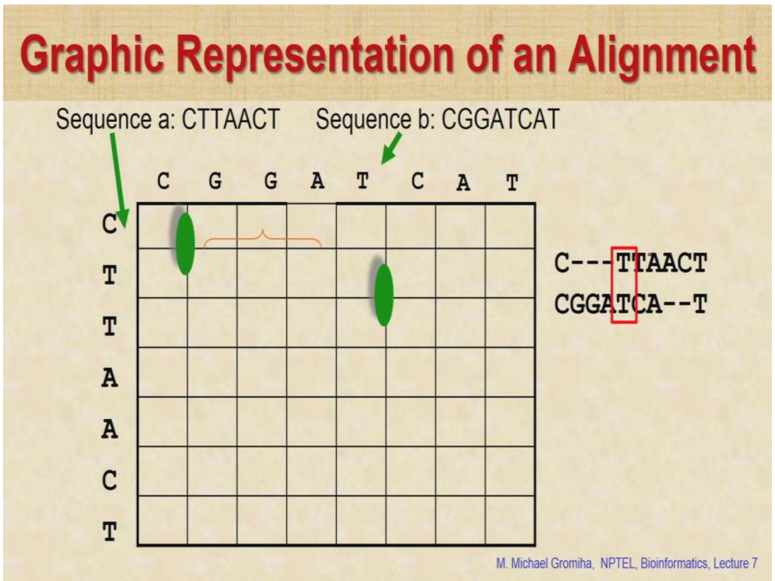
Sequence a: CTTAACT Sequence b: CGGATCAT

```
C---TTAACT
CGGATCA--T
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, in this case, we moved 3 characters 1, 2, 3; then we have 3 to 4 3 letters 1 2 3 they are aligned.

(Refer Slide Time: 04:06)



Graphic Representation of an Alignment

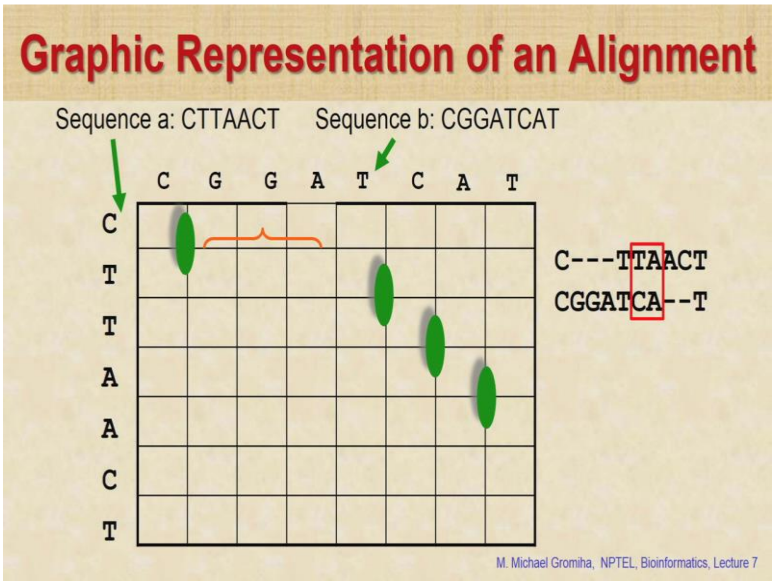
Sequence a: CTTAACT Sequence b: CGGATCAT

C---TTAACT
CGGATCA--T

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, there are corners 1, 2, 3, right these 2 are here.

(Refer Slide Time: 04:07)



Graphic Representation of an Alignment

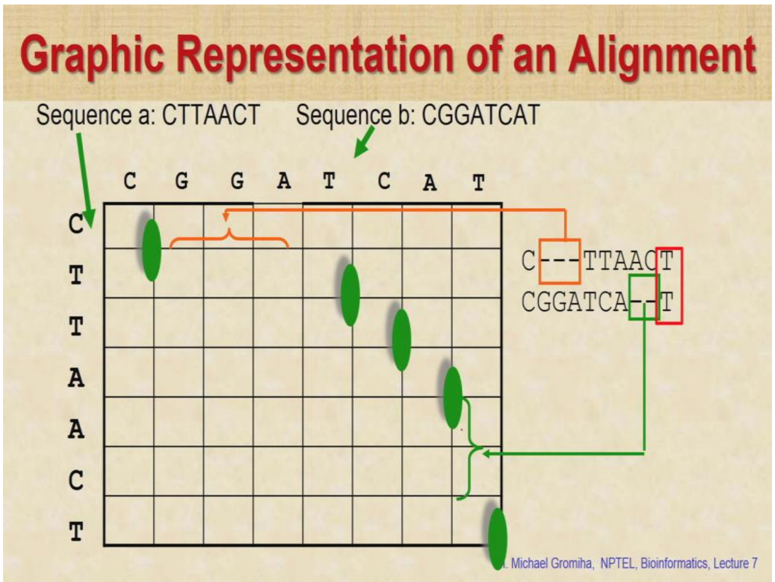
Sequence a: CTTAACT Sequence b: CGGATCAT

C---TTAACT
CGGATCA--T

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

And we have the 2 gaps.

(Refer Slide Time: 04:14)



Graphic Representation of an Alignment

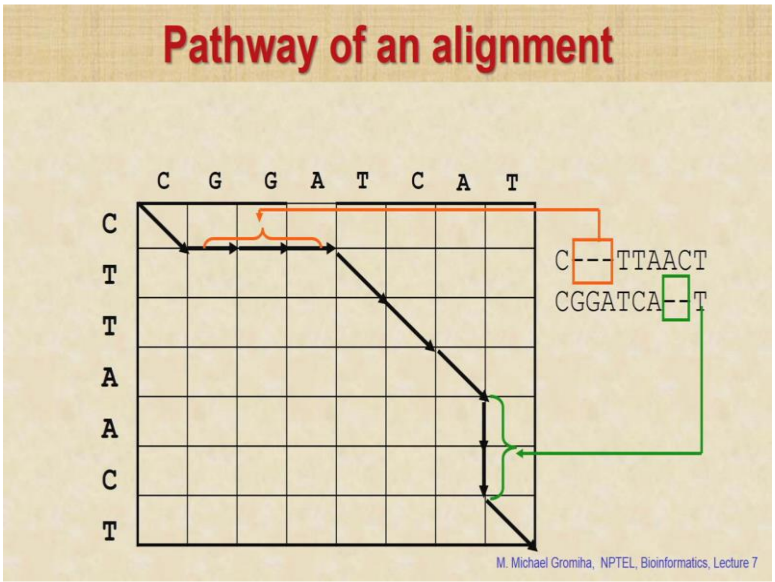
Sequence a: CTTAACT Sequence b: CGGATCAT

C---TTAACT
CGGATCA--T

. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, we go the down vertically down 2 gaps and the last one is here ok.

(Refer Slide Time: 04:18)



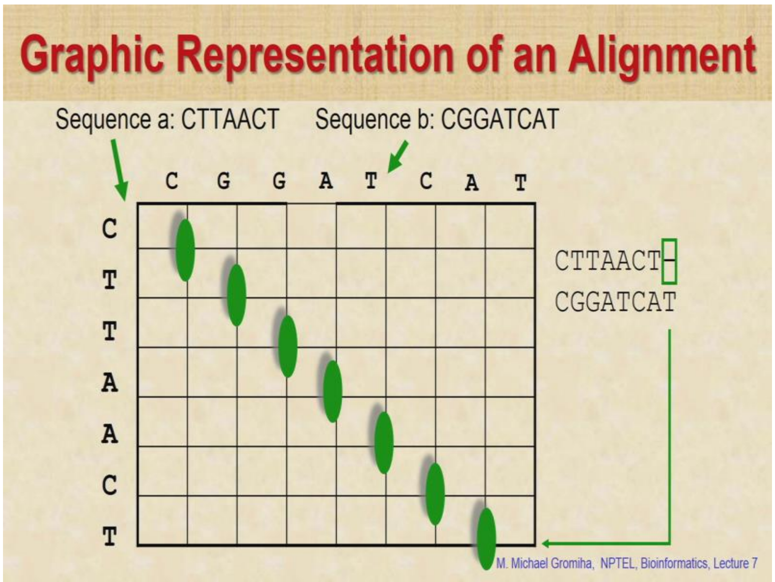
Pathway of an alignment

C---TTAACT
CGGATCA--T

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, now if you have this one; this is the path.

(Refer Slide Time: 04:21)

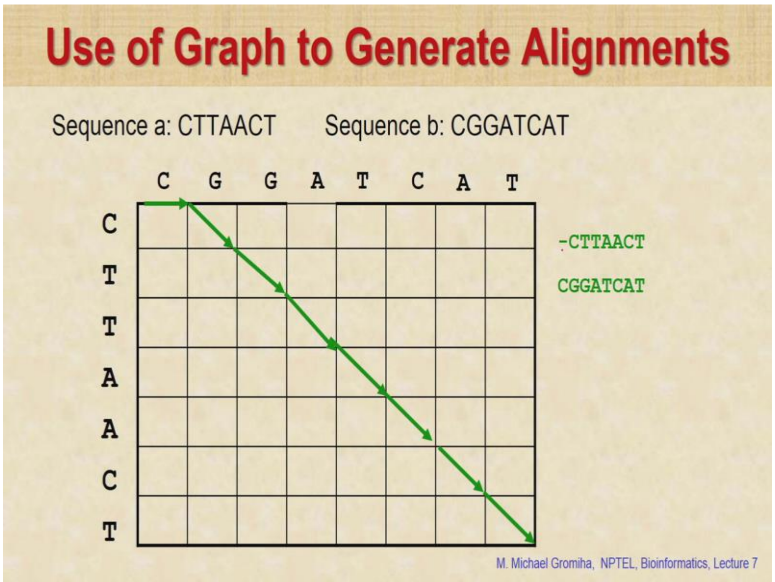


Graphic Representation of an Alignment

Sequence a: CTTAACT Sequence b: CGGATCAT

Now, if you align the sequence in a different way here we do not care about the matching or mismatching score just we try to align. Now, the last one is a gap. So, here everything we aligned. And then finally you put a gap here right here this is a gap.

(Refer Slide Time: 04:35)

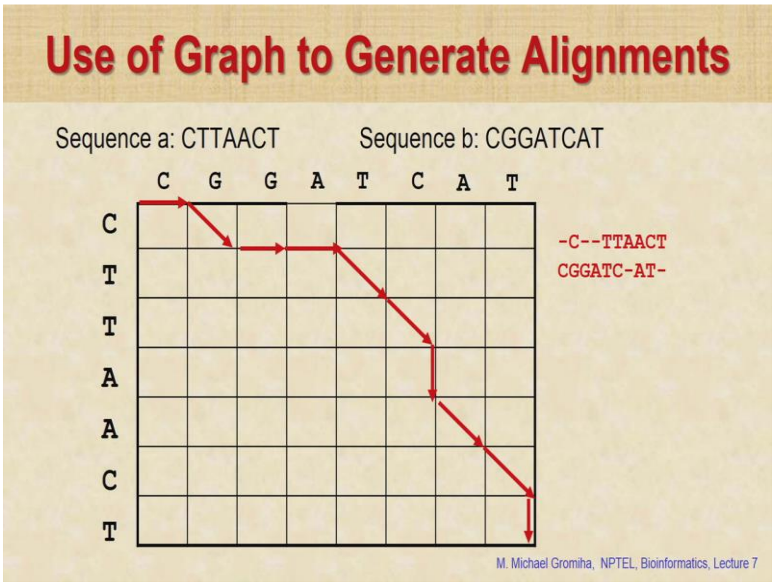


Use of Graph to Generate Alignments

Sequence a: CTTAACT Sequence b: CGGATCAT

So, now this one if you put a gap with the first one; so this is a gap and all others are aligned.

(Refer Slide Time: 04:41)



Use of Graph to Generate Alignments

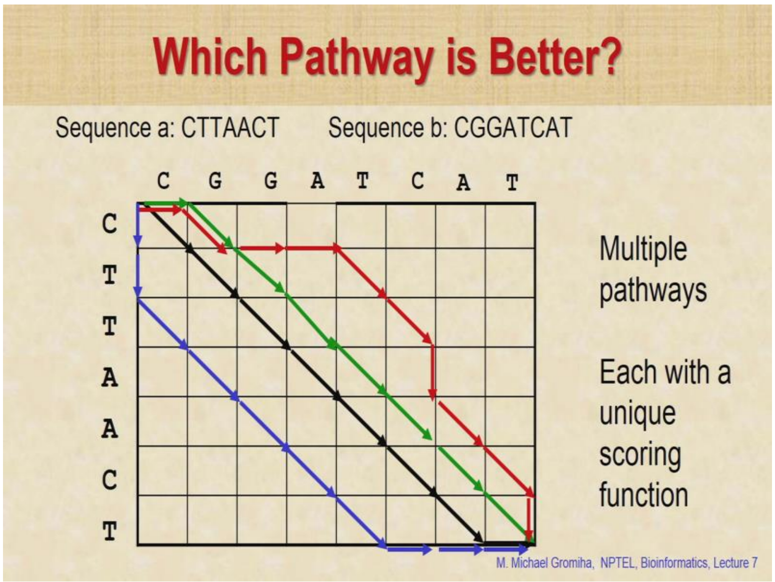
Sequence a: CTTAACT Sequence b: CGGATCAT

```
-C--TTAACT
CGGATC-AT-
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, this is another example, if you see one gap and here this is aligned C and G are aligned. So, then again 2 gaps and here again the 3 aligned and then 1 gap right and 2 are aligned and 1 gap.

(Refer Slide Time: 04:53)



Which Pathway is Better?

Sequence a: CTTAACT Sequence b: CGGATCAT

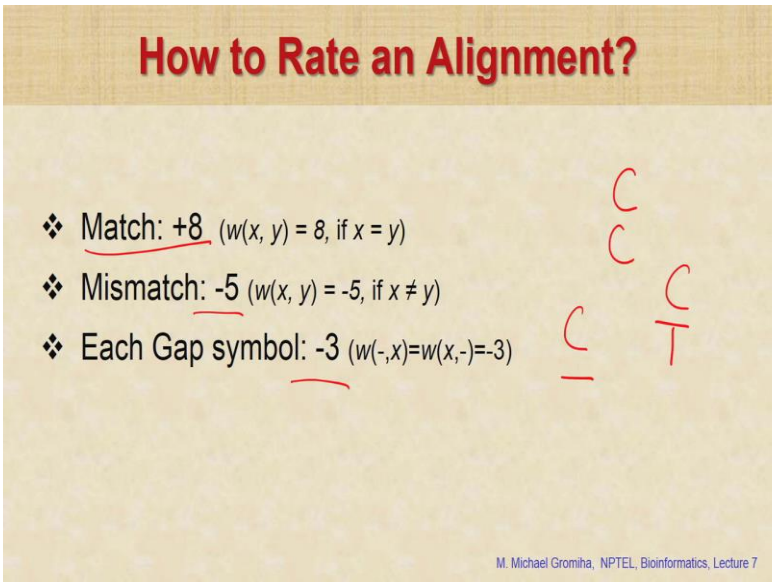
Multiple pathways

Each with a unique scoring function

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, now there are multiple ways multiple pathways to align the sequences these are the various ways now the question is which one you need to choose which one is the best one. So, in this case we give a scoring function right for example.

(Refer Slide Time: 05:07)



How to Rate an Alignment?

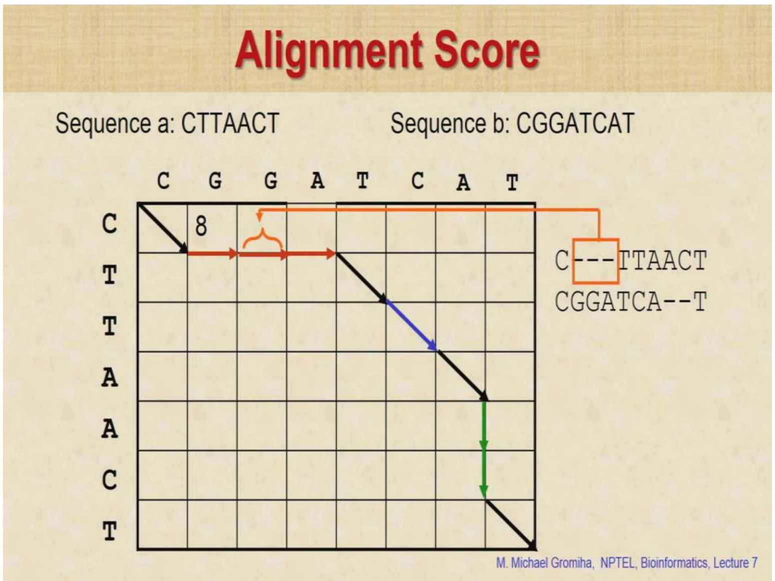
- ❖ Match: +8 (*w*(*x*, *y*) = 8, if *x* = *y*)
- ❖ Mismatch: -5 (*w*(*x*, *y*) = -5, if *x* ≠ *y*)
- ❖ Each Gap symbol: -3 (*w*(-,*x*)=*w*(*x*,-)=-3)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

If you give a scoring function if it is a match then we give a score of +8 and if it is a mismatch we give -5 and gap symbol we will give -3; for example, if it is the match means weight of x, y equal to 8 if x equal to y right if this is a C and then the C. So, we can see this match score is equal to 8 mismatch score if x is not equal to y for example one is the C another one is T.

So, we can give this score -5 the gaps symbol we give -3. For example, if it is a C and the gap gives -3, this is one example in the reality if you see the gap symbol should get more penalty than the mismatch score.

(Refer Slide Time: 05:44)



Alignment Score

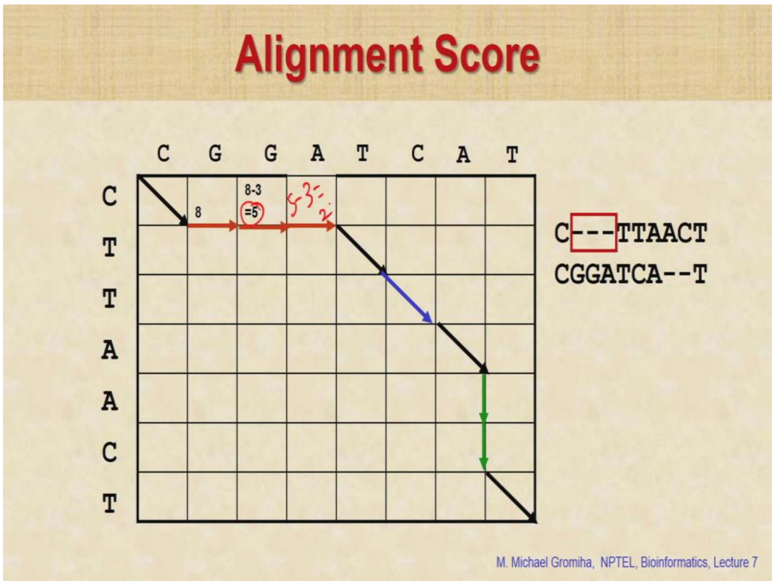
Sequence a: CTTAACT Sequence b: CGGATCAT

```
C---TTAACT
CGGATCA--T
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, we do like this first, if you see C and C match. So, score will be 8 and the second one we introduce a gap. So, what will be next one; 8 -3 = 5 because 8 here.

(Refer Slide Time: 05:54)



Alignment Score

```
C---TTAACT
CGGATCA--T
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

And because the gap -3 that is equal to 5; so again we introduce a gap. So, this number will be 5 minus.

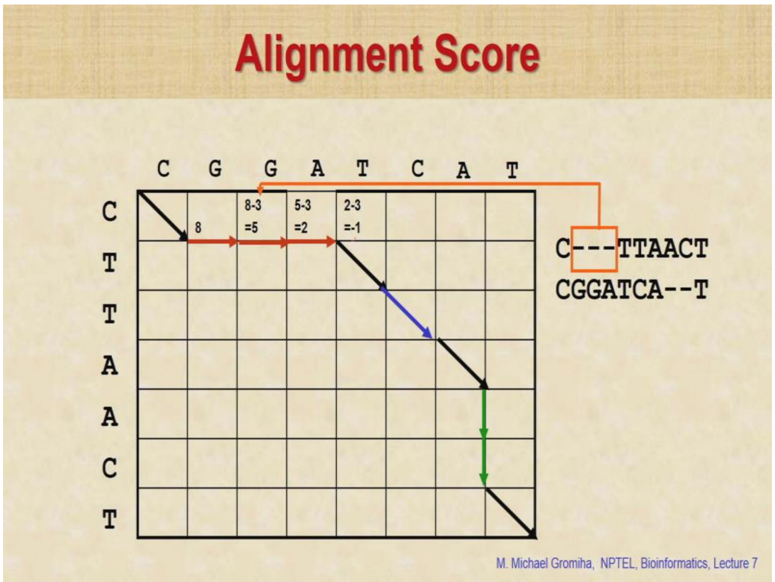
Student: 3.

3 equal to?

Student: 2.

2, fine.

(Refer Slide Time: 06:04)



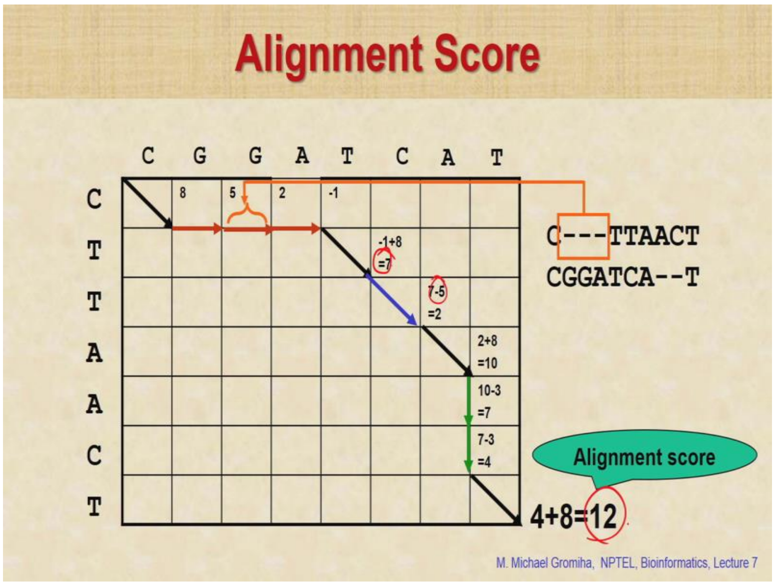
Alignment Score

C---TTAACT
CGGATCA--T

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, next again a gap -1; likewise then we add this score T plus T 3 plus score -1 plus 8 equal to 7.

(Refer Slide Time: 06:10)



Alignment Score

C---TTAACT
CGGATCA--T

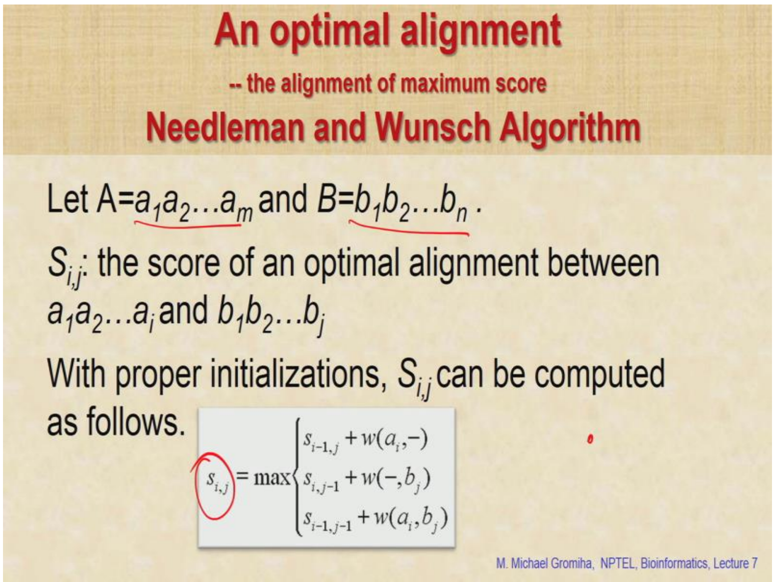
Alignment score

4+8=12

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

And then go with this one; this is a mismatch T and A. So, -5 this is equal to 2 finally, we get the score that is equal to 12.

(Refer Slide Time: 06:23)



An optimal alignment

-- the alignment of maximum score

Needleman and Wunsch Algorithm

Let A=$a_1a_2\ldots a_m$ and B=$b_1b_2\ldots b_n$.

$S_{i,j}$: the score of an optimal alignment between $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$

With proper initializations, $S_{i,j}$ can be computed as follows.

$$s_{i,j} = \max\begin{cases} s_{i-1,j} + w(a_i,-) \\ s_{i,j-1} + w(-,b_j) \\ s_{i-1,j-1} + w(a_i,b_j) \end{cases}$$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, we want to see there are various numbers you get depending upon this one alignment. So, we got different ways or pathways. So, different pathways will give you different numbers. So, to combine everything Wunsch Needleman, they brought up an algorithm a smart way to select via which way we have to align. So, for example, if first sequence is a_1, a_2 a_m and the second is b_1, b_2, b_n.

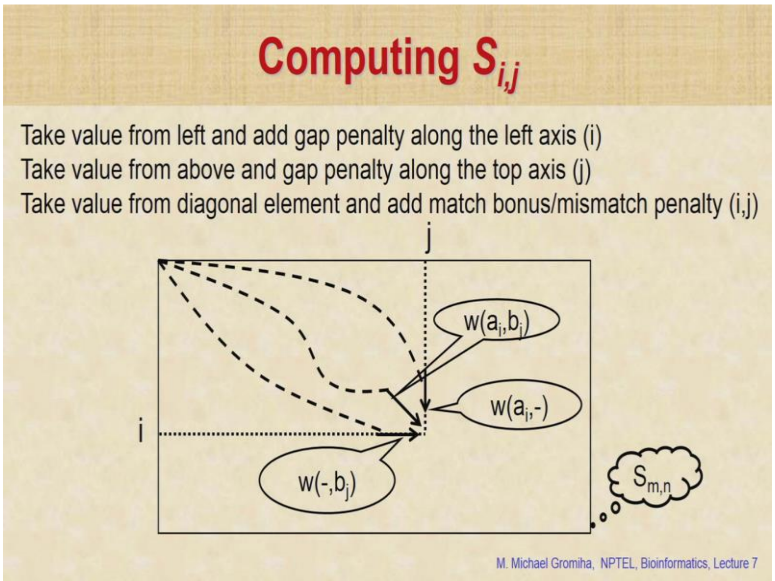
So, we can align different ways you can align. So, they give proper initialization and they give a score that is $S_{i,j}$, if you have one nucleotide and if you want to make the alignment compared with the previous ones either we can give gap for insertion or for deletion or you can compare these 2 nucleotides giving the match score or mismatch score. So, he use these values, right and take the maximum first ways whether it is insertion or deletion right or this is a matching score or mismatching score you wait depending upon a and b if it is a match we give the match score, if it is mismatch we give mismatch score, otherwise we give the gap score.

We give the maximum values because why you give the maximum values.

Student: To maximise the score (Refer Time: 07:41).

Because we want to have the best alignment; so you want to maximise the score. So, all the 3 conditions he took the maxima of all 3 options.

(Refer Slide Time: 07:50)



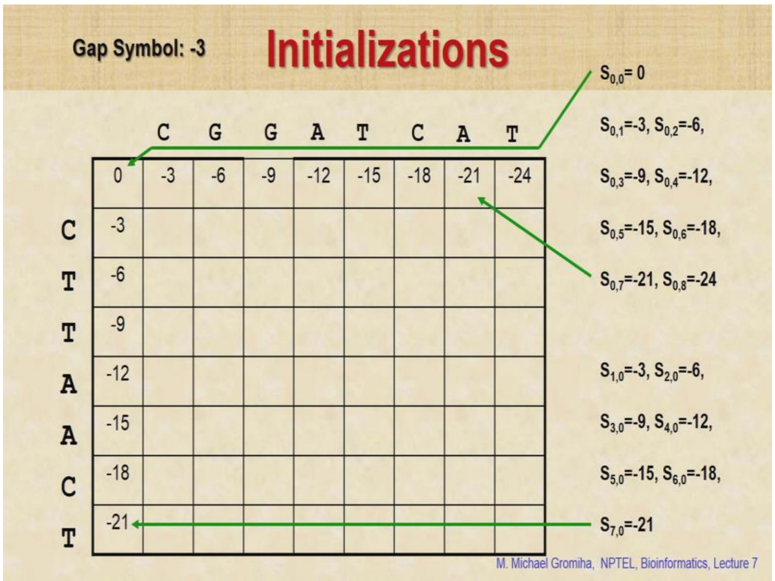
Computing $S_{i,j}$

Take value from left and add gap penalty along the left axis (i)
Take value from above and gap penalty along the top axis (j)
Take value from diagonal element and add match bonus/mismatch penalty (i,j)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, here we take the value from left and add a gap penalty along left axis or take the value from the above and the gap penalty along the y axis or take the diagonal and take the weight whether this is match or mismatch.

(Refer Slide Time: 08:06)



Gap Symbol: -3

Initializations

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3								
T	-6								
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

$S_{0,0}$= 0

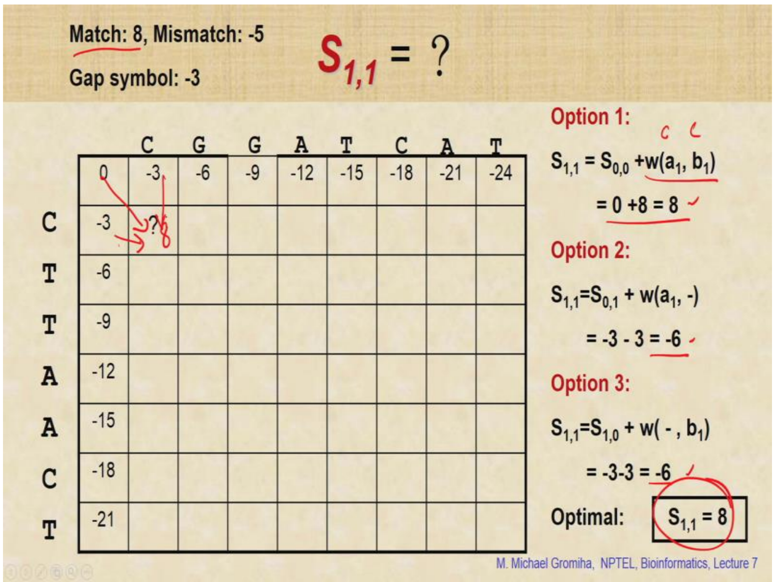
$S_{0,1}$=-3, $S_{0,2}$=-6,
$S_{0,3}$=-9, $S_{0,4}$=-12,
$S_{0,5}$=-15, $S_{0,6}$=-18,
$S_{0,7}$=-21, $S_{0,8}$=-24

$S_{1,0}$=-3, $S_{2,0}$=-6,
$S_{3,0}$=-9, $S_{4,0}$=-12,
$S_{5,0}$=-15, $S_{6,0}$=-18,
$S_{7,0}$=-21

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, how to do that? So, first we give a gap symbol of -3, then it is a, we initialize first. So, we give -3, if you completely gap, then we gives this -24 up to the both ways. Now first case is first we have to align.

(Refer Slide Time: 08:23)



Match: 8, Mismatch: -5

Gap symbol: -3

$S_{1,1}$ = ?

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	?8							
T	-6								
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

c c

$S_{1,1} = S_{0,0} + w(a_1, b_1)$

$= 0 + 8 = 8$ ✓

Option 2:

$S_{1,1} = S_{0,1} + w(a_1, -)$

$= -3 - 3 = -6$ ✓

Option 3:

$S_{1,1} = S_{1,0} + w(-, b_1)$

$= -3 - 3 = -6$ ✓

Optimal: $\boxed{S_{1,1} = 8}$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, which one is the first number. So, 3 options for this one what are 3 options.

Student: 0 0 and (Refer Time: 08:29).

One, we can come from here other one come from here or from left to right, here we introduce gap, here we introduce gap and here we introduce the alignment either match or mismatch. So, option one S 0 0 ok from here weight of a_1, b_1. So, here if you see a; what is a_1.

Student: C (Refer Time: 08:51).

C; what it b_1.

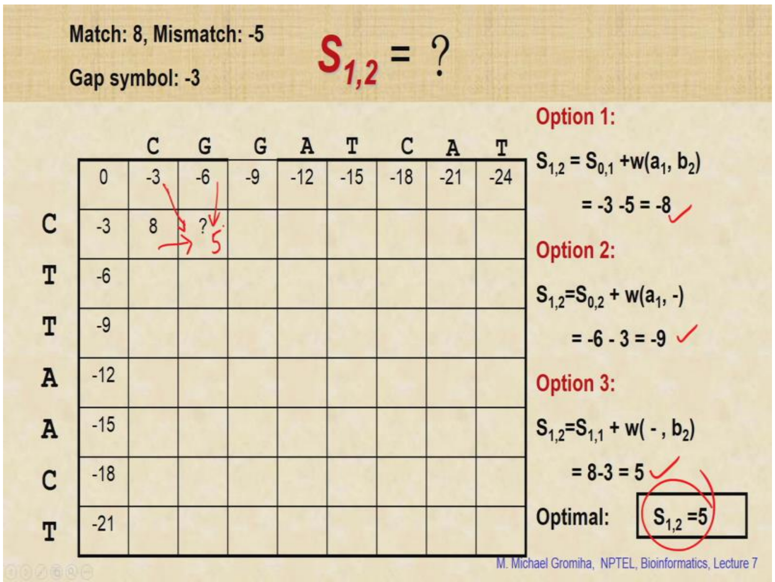
Student: C.

C. So, match; so match score equal to 8. So, we put 0 + 8 = 8, then option 2 we go from $S_{1,1}$, that is equal to $S_{0,1}$ plus weight of a_1, gap. So, do this is -3 added to -3. So, put another -3. So, this will be -6 and option 3 we go there is 1,0 and with the gap in the first term and the b_{11}. So, if you see this one this is gap -3. So, again it is -3 this will -6 now compare this 3 one is 8 one is -6 one is -6 which one is maximum.

Student: 8.

8. So, put the maximum value of 8 here.

(Refer Slide Time: 09:40)



Match: 8, Mismatch: -5

Gap symbol: -3

$S_{1,2}$ = ?

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	? 5						
T	-6								
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

$S_{1,2} = S_{0,1} + w(a_1, b_2)$

$= -3 -5 = -8$ ✓

Option 2:

$S_{1,2}=S_{0,2} + w(a_1, -)$

$= -6 - 3 = -9$ ✓

Option 3:

$S_{1,2}=S_{1,1} + w(- , b_2)$

$= 8-3 = 5$ ✓

Optimal: $\boxed{S_{1,2} =5}$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

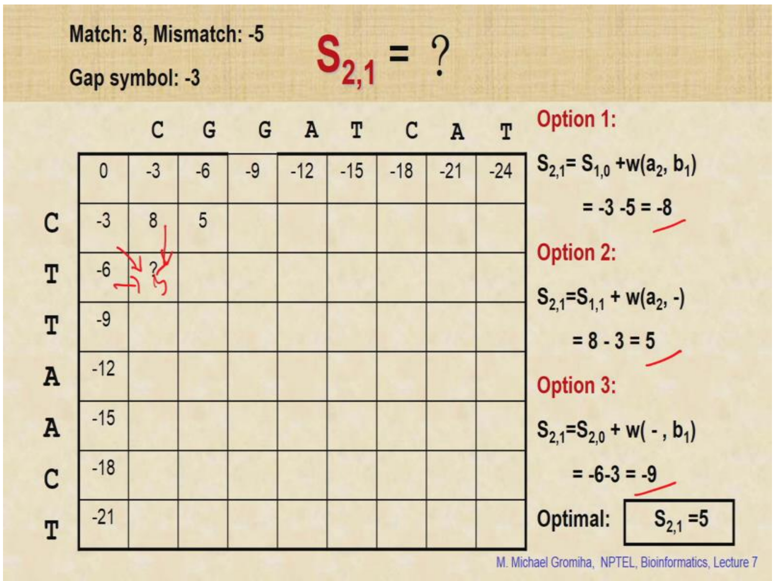
So, what is value here what are the possibilities the numbers from here what is the number -8; this way.

Student: -9.

-9 and this way.

Student: 5.

(Refer Slide Time: 09:56)



Match: 8, Mismatch: -5

Gap symbol: -3

$S_{2,1}$ = ?

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5						
T	-6	? 5							
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

$S_{2,1}= S_{1,0} + w(a_2, b_1)$

$= -3 -5 = -8$

Option 2:

$S_{2,1}=S_{1,1} + w(a_2, -)$

$= 8 - 3 = 5$

Option 3:

$S_{2,1}=S_{2,0} + w(- , b_1)$

$= -6-3 = -9$

Optimal: $\boxed{S_{2,1} =5}$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

This is 5. So, 5 is the maximum. So, you put 5 here. So, now, it is 5; what is the number here, these are the options.

Student: -8.

-8.

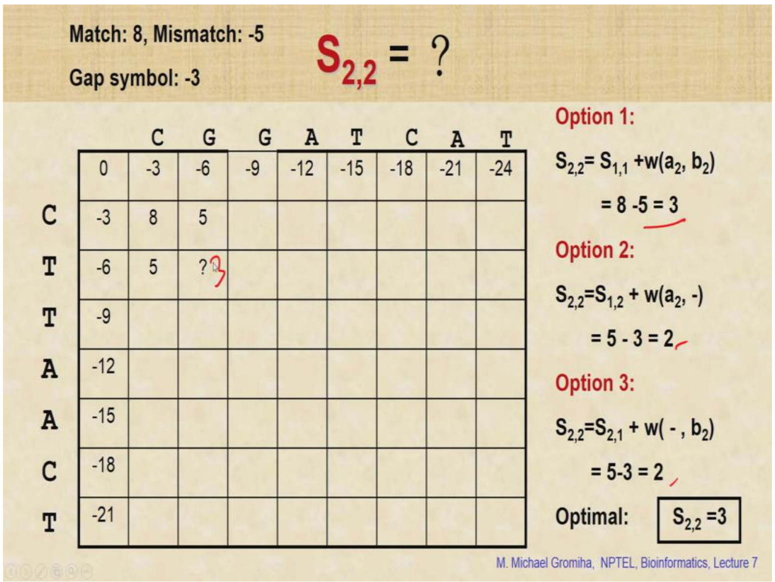
Student: (Refer Time: 10:08).

+5.

Student: 5.

And -9; so if it is maximum is 5. So, we put 5 here, fine.

(Refer Slide Time: 10:15)



Match: 8, Mismatch: -5

Gap symbol: -3

$S_{2,2}$ = ?

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5						
T	-6	5	? 3						
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

$S_{2,2} = S_{1,1} + w(a_2, b_2)$

$= 8 - 5 = 3$

Option 2:

$S_{2,2} = S_{1,2} + w(a_2, -)$

$= 5 - 3 = 2$

Option 3:

$S_{2,2} = S_{2,1} + w(-, b_2)$

$= 5 - 3 = 2$

Optimal: $\boxed{S_{2,2} = 3}$

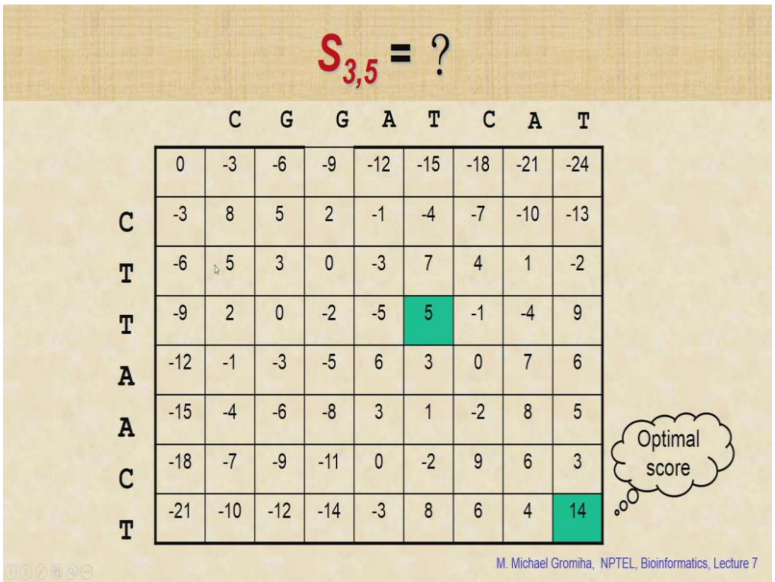
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

This one quickly 5 -2 = 3 here equal to 5 -3, 5 -3 = 2, 5 -3 = 2 and here 8 minus.

Student: 5.

5 equal to 3 because of mismatch. So, this is a maximum. So, we put 3 here.

(Refer Slide Time: 10:30)



$S_{3,5}$ = ?

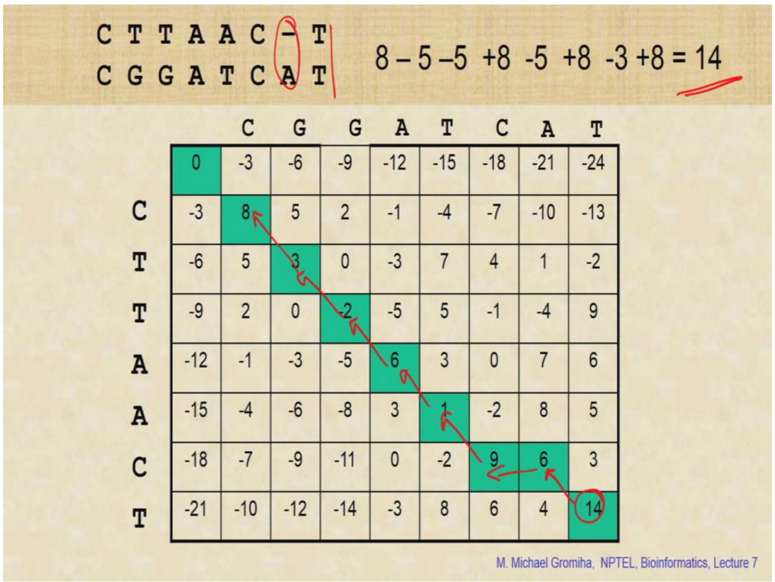
		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	5	-1	-4	9
A	-12	-1	-3	-5	6	3	0	7	6
A	-15	-4	-6	-8	3	1	-2	8	5
C	-18	-7	-9	-11	0	-2	9	6	3
T	-21	-10	-12	-14	-3	8	6	4	14

Optimal score

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Finally, we fill the metrics. So, finally we get the optimum score of 14.

(Refer Slide Time: 10:36)



C T T A A C – T
C G G A T C A T

8 – 5 –5 +8 -5 +8 -3 +8 = 14

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	5	-1	-4	9
A	-12	-1	-3	-5	6	3	0	7	6
A	-15	-4	-6	-8	3	1	-2	8	5
C	-18	-7	-9	-11	0	-2	9	6	3
T	-21	-10	-12	-14	-3	8	6	4	14

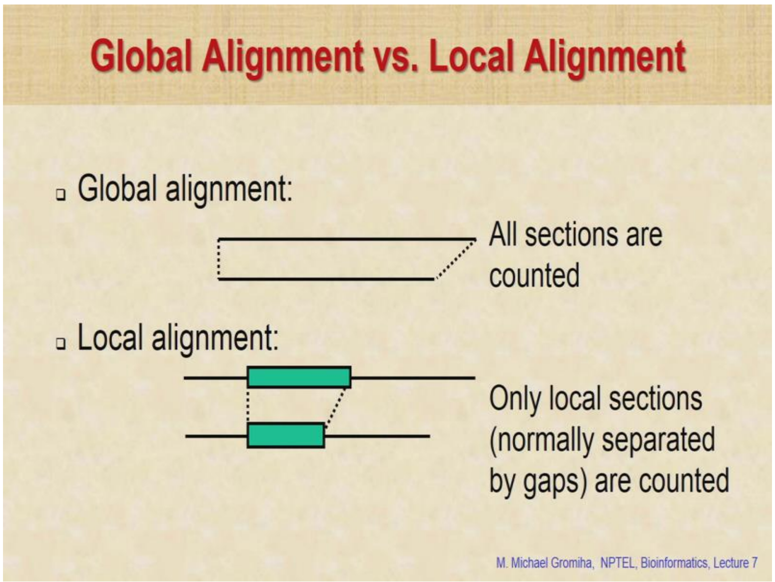
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, now the question is how to go back because we get the maximum value from the highest value you go back because there are 3 possibilities 3 directions and see which one is the best you go back and then finally, make the alignment right this is 14 is the highest one, then go back among the 3, this is the 6 and then this 9 and one 6 -2 and 3 and here we see 8.

So, now here we have 1, we introduce one gap here. So, gap is here T and T are aligned right and then if you see all others are aligned CC TG TG AA and AT CC, right. So, we aligned all the other sequences. So, you put the number. Finally, we get total of 14. Now, if you have 2 sequences we check the 2 sequences and there are 3 possibilities to fill the matrix and take the maximum number and fill the matrix and trace back to see the route.

And finally, we can make this alignment. So, here there are 2 types of alignment one alignment is the global alignment as I discussed in the previous classes. So, it considers everything all the residues are aligned.

(Refer Slide Time: 11:54)



Global Alignment vs. Local Alignment

- Global alignment:

All sections are counted

- Local alignment:

Only local sections (normally separated by gaps) are counted

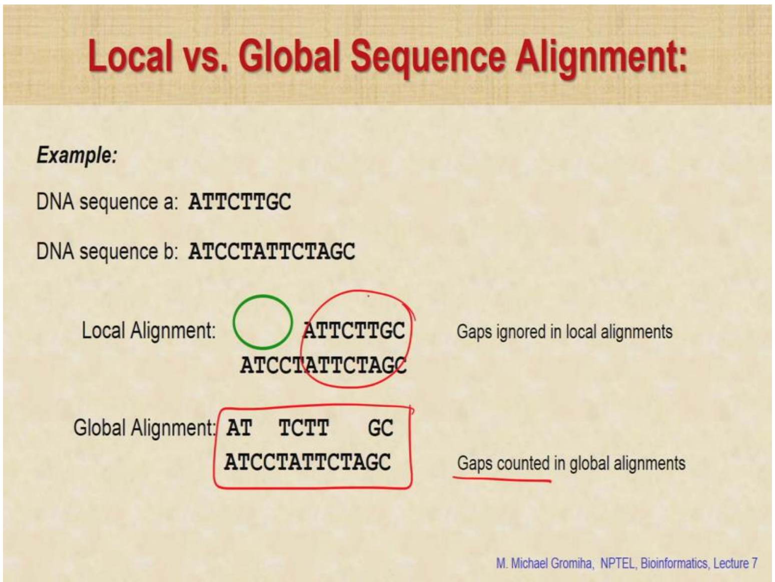
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

And the some cases; we are in only some particular regions; this is called the local alignment for why we need local alignment.

Student: For (Refer Time: 12:01).

Here, what sort repeats or some cases there are small motifs which are important for any specific functions like TATA box binding protein. So, TATA box is important for the binding. So, we want to see the small repeats or small motifs, we can get from this local alignment.

(Refer Slide Time: 12:16)



Local vs. Global Sequence Alignment:

Example:

DNA sequence a: **ATTCTTGC**

DNA sequence b: **ATCCTATTCTAGC**

Local Alignment:

```
    ATTCTTGC
ATCCTATTCTAGC
```

Gaps ignored in local alignments

Global Alignment:

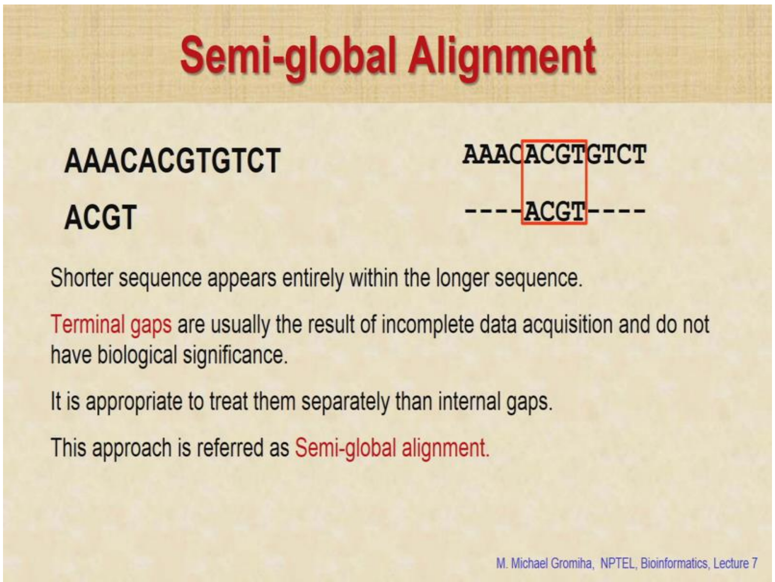
```
AT   TCTT   GC
ATCCTATTCTAGC
```

Gaps counted in global alignments

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Right for example, if this is the sequence a and sequence b some cases we considered all the sequence together. So, for the global alignment here gaps are also counted and here you can see we ignore the gaps and then see only the other regions right for local alignment.

(Refer Slide Time: 12:35)



Semi-global Alignment

AAACACGTGTCT

ACGT

```
AAACACGTGTCT
----ACGT----
```

Shorter sequence appears entirely within the longer sequence.

Terminal gaps are usually the result of incomplete data acquisition and do not have biological significance.

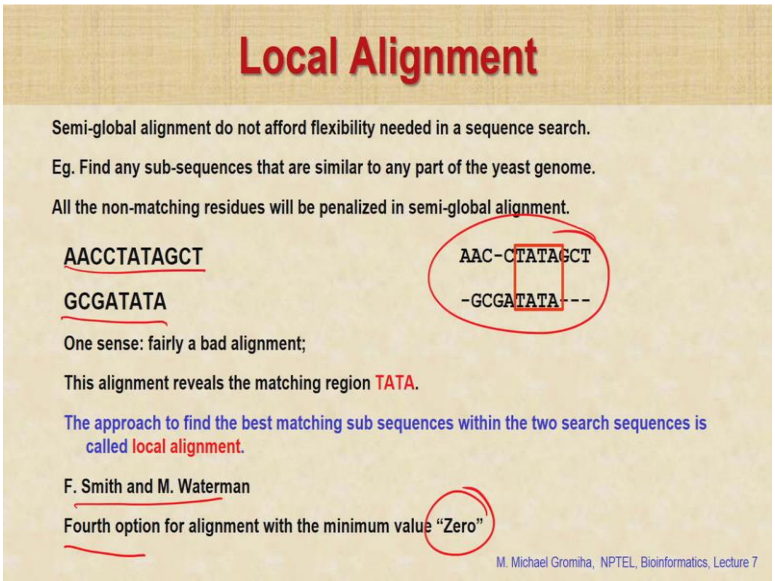
It is appropriate to treat them separately than internal gaps.

This approach is referred as Semi-global alignment.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

And also another alignment called semi global alignment this will tell you how we can align the different sequences with respect to gaps, internal or the outside; some case if you see the gaps outside these terminal gaps are usually result of incompleted data acquisition and may not have any biological significance in this case; you can remove these gaps. And if you remove the terminal gaps then we align only the sequences which is in between. So, this type of alignment; we call as semi global alignment and how to do this alignment.

(Refer Slide Time: 13:10)



Local Alignment

Semi-global alignment do not afford flexibility needed in a sequence search.

Eg. Find any sub-sequences that are similar to any part of the yeast genome.

All the non-matching residues will be penalized in semi-global alignment.

AACCTATAGCT

GCGATATA

```
AAC-CTATAGCT
-GCGATATA---
```

One sense: fairly a bad alignment;

This alignment reveals the matching region TATA.

The approach to find the best matching sub sequences within the two search sequences is called local alignment.

F. Smith and M. Waterman

Fourth option for alignment with the minimum value "Zero"

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

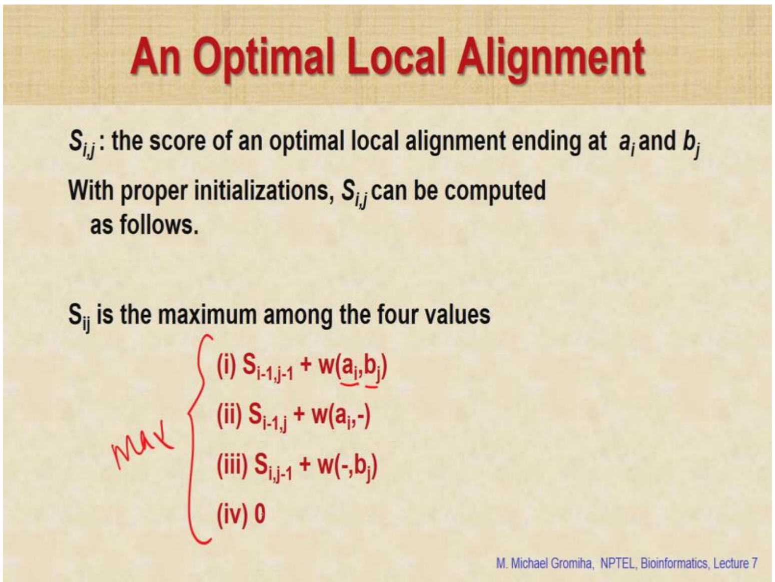
For example, if we got 2 sequences this is sequence 1 and this is sequence 2.

We can make different types of alignments for example, you can make this type of alignments; it look like a global alignment is a kind of only bad alignment because some gap are here and some gaps here and the residues are not properly aligned, but if you look at the details now it will give you the matching region called T A T A. So, then we have to extract this type of information. So, in this case, Smith and Waterman, they proposed another option; currently there are 3 options right Wunsch Needleman propose 3 options.

Student: (Refer Time: 13:47).

What are 3 options; match or mismatch deletions insertions right they take the maxima of these 3 conditions. So, they purpose and the fourth option. So, the value of 0 they do not want to have any negative values they want to have the positive values if it goes a negative just remove it because we do not have to include in the alignment. So, they included in the alignment only the positive values.

(Refer Slide Time: 14:13)



An Optimal Local Alignment

$S_{i,j}$: the score of an optimal local alignment ending at a_i and b_j

With proper initializations, $S_{i,j}$ can be computed as follows.

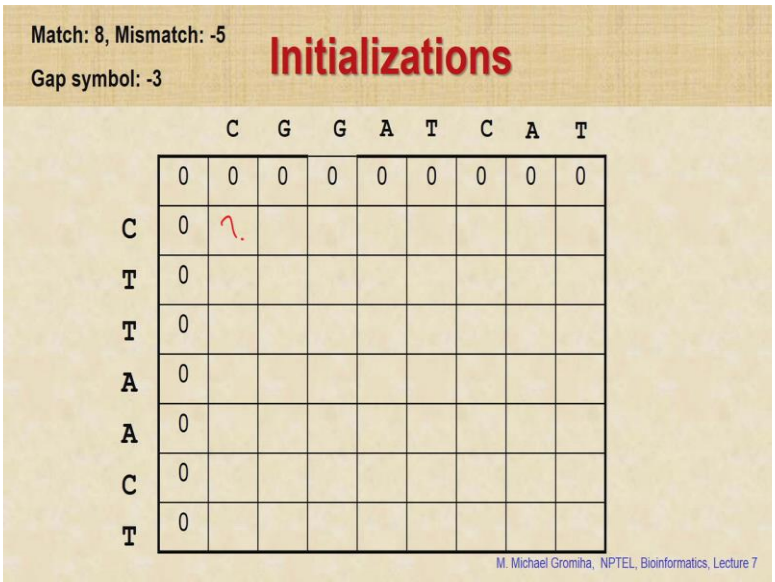
S_{ij} is the maximum among the four values

max
(i) $S_{i-1,j-1} + w(a_i,b_j)$
(ii) $S_{i-1,j} + w(a_i,-)$
(iii) $S_{i,j-1} + w(-,b_j)$
(iv) 0

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, now the condition is the maximum of these 4 options; one is Si -1 S j -1 diagonal right go through diagonal lines. So, with the a and b with respect to match and mismatch the second one is the gap third is also gap insertion and a deletion and the last option they give 0.

(Refer Slide Time: 14:38)



Match: 8, Mismatch: -5

Gap symbol: -3

Initializations

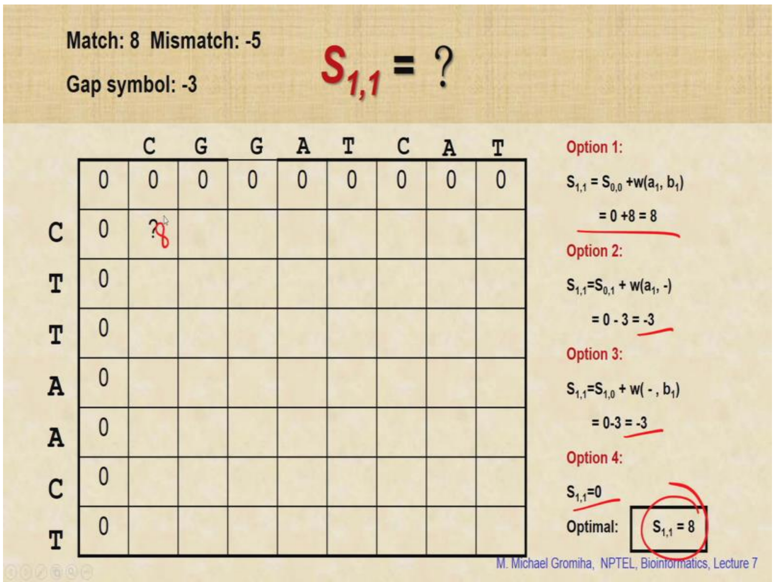
		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	7.							
T	0								
T	0								
A	0								
A	0								
C	0								
T	0								

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, if you have this same sequence I give a sequence first initialization because there is no negative values. So, put a 0, right what will come here.

Student: (Refer Time: 14:50) C and C 8 8.

(Refer Slide Time: 14:52)



Match: 8 Mismatch: -5

Gap symbol: -3

$S_{1,1}$ = ?

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	?8							
T	0								
T	0								
A	0								
A	0								
C	0								
T	0								

Option 1:

$S_{1,1} = S_{0,0} + w(a_1, b_1)$

$= 0 + 8 = 8$

Option 2:

$S_{1,1} = S_{0,1} + w(a_1, -)$

$= 0 - 3 = -3$

Option 3:

$S_{1,1} = S_{1,0} + w(-, b_1)$

$= 0 - 3 = -3$

Option 4:

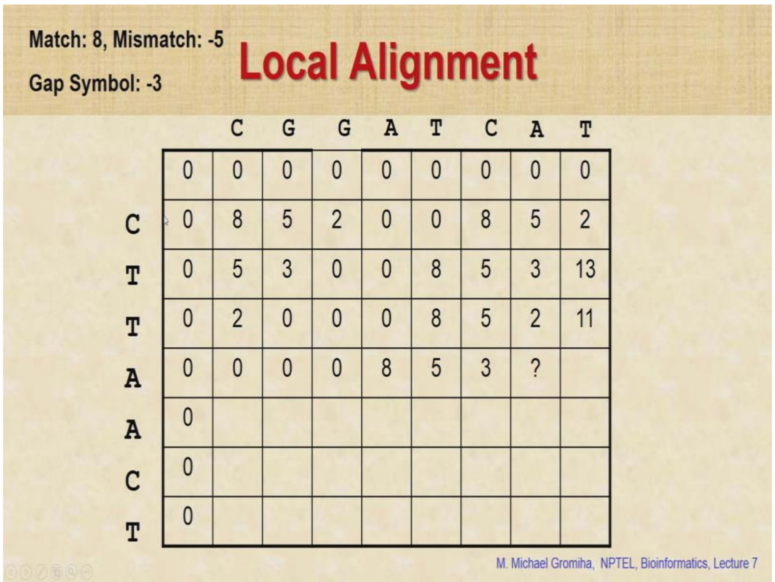
$S_{1,1} = 0$

Optimal: $\boxed{S_{1,1} = 8}$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Right this equal to 8 because C and C this is 8; this is a match and here there is a gap is -3 and the other gap is insertion -3. So, this is the option is 4 this is 0. So, the maximum of this among this 4 is 8.

(Refer Slide Time: 15:10)



Match: 8, Mismatch: -5

Gap Symbol: -3

Local Alignment

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	8	5	2	0	0	8	5	2
T	0	5	3	0	0	8	5	3	13
T	0	2	0	0	0	8	5	2	11
A	0	0	0	0	8	5	3	?	
A	0								
C	0								
T	0								

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

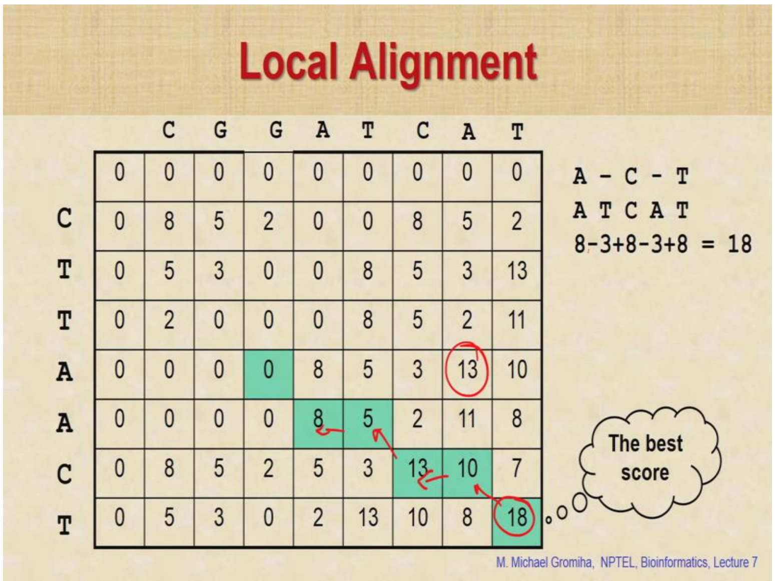
So, you put 8, right what is next one what will come what will come here.

Student: 5 + 8 is 13.

13.

Student: Yes sir.

(Refer Slide Time: 15:20)



Local Alignment

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	8	5	2	0	0	8	5	2
T	0	5	3	0	0	8	5	3	13
T	0	2	0	0	0	8	5	2	11
A	0	0	0	0	8	5	3	13	10
A	0	0	0	0	8	5	2	11	8
C	0	8	5	2	5	3	13	10	7
T	0	5	3	0	2	13	10	8	18

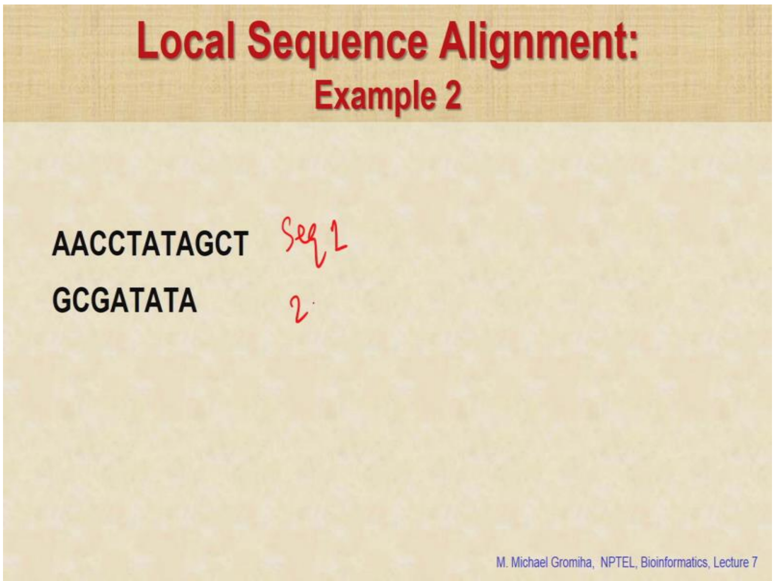
A - C - T
A T C A T
8-3+8-3+8 = 18

The best score

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Right, 13; so now, we finally completed this matrix right. So, now, the maximum score is 18. So, now, go back with this 18. So, go this 10 here; this is again among these 3 is 13. So, here this is 5, this is 8 and then 0. So, this stops here. So, make here. So, this is the 18 T and T match. So, here then we use this gap. So, A then C and C are matching right then another gap. So, A and A are matching, so it is equal to 18, right. So, I will give you another example.

(Refer Slide Time: 16:01)



Local Sequence Alignment: Example 2

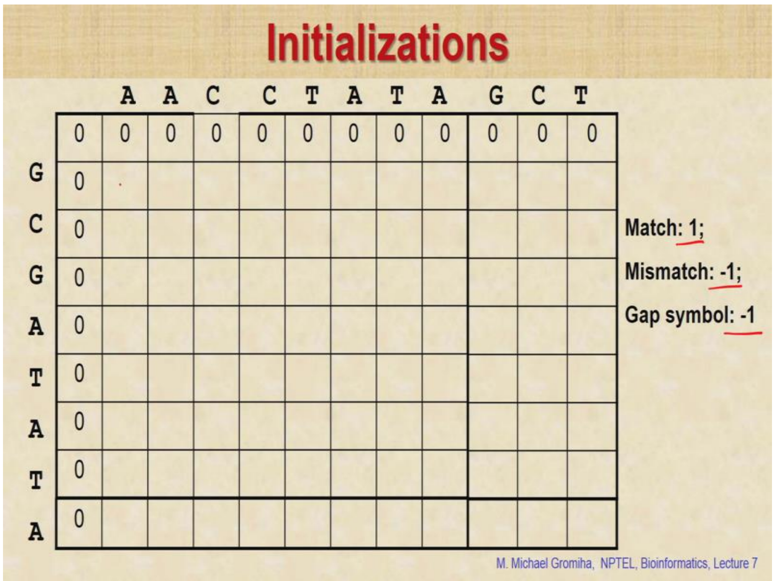
AACCTATAGCT Seq 1

GCGATATA 2

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

So, here is the 2 sequences this is sequence 1 this is 2.

(Refer Slide Time: 16:10)



Initializations

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

Match: 1;

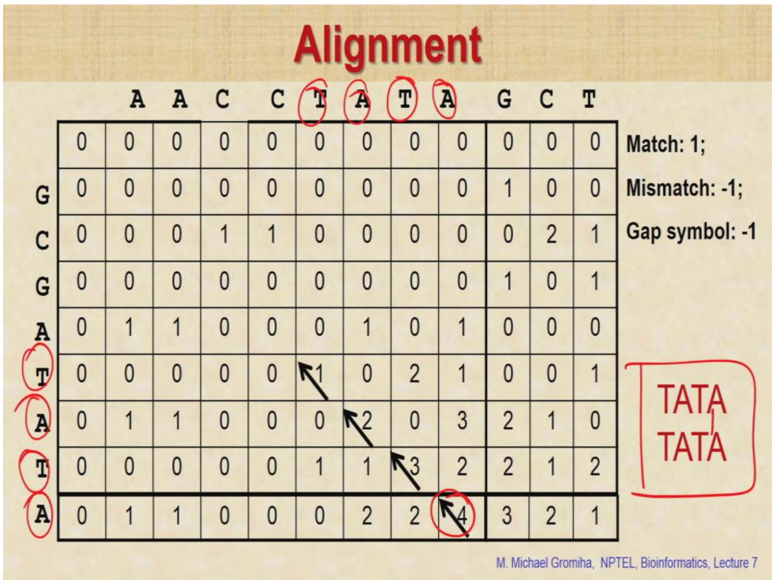
Mismatch: -1;

Gap symbol: -1

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

And how to make the alignment; so for simplicity I put the match score equal to 1 mismatch score equal to -1 and gap symbol equal to -1; so what is the value.

(Refer Slide Time: 16:26)



Alignment

		A	A	C	C	T	A	T	A	G	C	T
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	1	0	0
C	0	0	0	1	1	0	0	0	0	0	2	1
G	0	0	0	0	0	0	0	0	0	1	0	1
A	0	1	1	0	0	0	1	0	1	0	0	0
T	0	0	0	0	0	1	0	2	1	0	0	1
A	0	1	1	0	0	0	2	0	3	2	1	0
T	0	0	0	0	0	1	1	3	2	2	1	2
A	0	1	1	0	0	0	2	2	4	3	2	1

Match: 1;

Mismatch: -1;

Gap symbol: -1

TATA
TATA

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 7

Student: (Refer Time: 16:27).

0; the 0 because in this see the maxima of 4 conditions right maximum is 0, right, you don’t have a negative values right. So, it is 0. So, if you make the alignments finally, we get these numbers here equal to one this equal to 2 and this equal to 3 and this equal to 4. So, now, this is maximum right. So, now, we trace back 4, 3, 2, 1. So, finally, if you see the alignment this is A and A, A and A and

A, this A and this A right A and A aligned and the T and T this T and this T aligned and this A and this A aligned and this T and this T aligned after this 0.

So, we get this motif between these 2 sequences right this is a specific motif for the DNA binding proteins. So, we could find any specific motifs or if any patterns in any alignments; so this is the applications for this local alignment. So, there are 2 questions. So, we can use these 2 different algorithms right to align the sequences one and the sequence 2, either with the local alignment or with the global alignment. So, so far what did we discuss today?

Student: (Refer Time: 17:42) algorithms for alignment.

Yes, the algorithms for alignment; so how to if you got 2 sequences right. So, a different way is to align. So, what is dynamic programming?

Student: So, we divide alignment process in to a smaller parts.

Small parts, yes.

Student: And we will align small and small fragment we will align and (Refer Time: 18:06).

And then finally, merge together to form the complete alignment right. So, what are the 2 different algorithms we discussed today?

Student: Needleman Wunsch algorithm.

Wunsch Needleman algorithm.

Student: Smith and Waterman.

Smith and waterman algorithm Wunsch Needleman algorithm is for which type of alignments.

Student: Global.

Global alignment right what is the condition used in Wunsch Needleman algorithm.

Student: substitution is (Refer Time: 18:29) substitution score maxima (Refer Time: 18:30).

There is a maxima of.

Student: (Refer Time: 18:31) substitution.

3 different conditions; so either substitution score or the insertion or deletion in the case of local alignments.

Student: We have the fourth conditions.

Fourth condition, so what is the fourth condition.

Student: 0 value; value is the (Refer Time: 18:43).

There is 0. So, you can get the values. So, till now we use some numbers, for the match score or mismatch score or the gap we use some numbers right in the beginning of the lecture we discussed about some matrices. So, what are the matrices we discussed?

Student: PAM and BLOSUM (Refer Time: 18:58).

PAM metrics and the BLOSUM metrics; so what are the characteristics of PAM metrics?

Student: It considers mutation, sir.

It considers mutations. So, depending upon the type of mutations or based on this charge or the actually it depends on the actual mutability rate right in the in any actual cases depending on the mutability rate. So, they derive the metrics.

So, accordingly you can use the numbers available in the BLOSUM metrics or the PAM metrics to give the weightage either the match or mismatch you give the weightage then you can give the probable alignment right this is what they use in the BLAST when they implement algorithm.

So, in next class, we will discuss about the software for aligning sequences; how, we will discuss about how BLAST works. Now I will tell you how they take 2 sequences and how they align the sequences and what is the score they give how they evaluate the score between 2 different alignments. For example, if you have one sequence, if it is aligned with 3 4 different sequences which one has the highest score which one has the most probable alignment. So, they give the alignment scores they try to optimize this alignment right, we will discuss in the in next class.

Thank you for your kind attention.