BI694
Bioinformatics & Phylogenetics

Winter Semester 2017

WEEK 4

Overview

- Global alignment
- Local alignment
- Sequence similarity searches

Pairwise alignment

Global alignment: Closely related sequences which are of same length are appropriate for global alignment. Here, the alignment is carried out from beginning to end of the sequence to find the best possible alignment.

Local alignment: Sequences which are suspected to have similarity or even dissimilar sequences can be compared using local alignment. It finds the local regions with high level of similarity.

Pairwise alignment

Local Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
          |||| ||||||| |||||||||||||||||
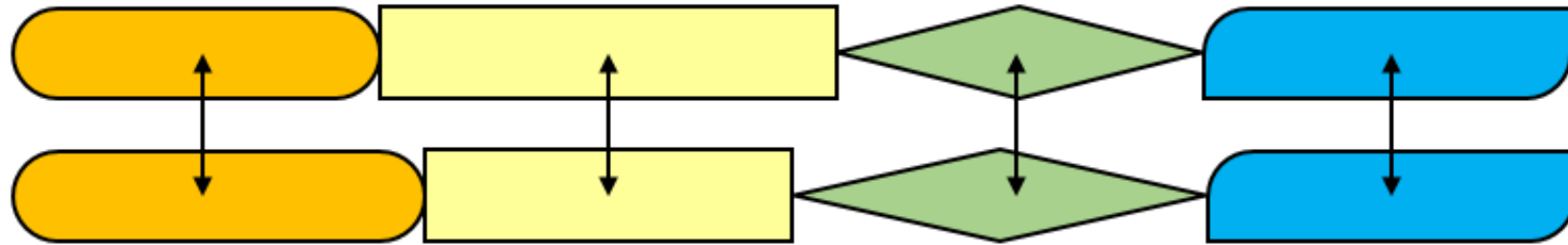Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'

Global Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
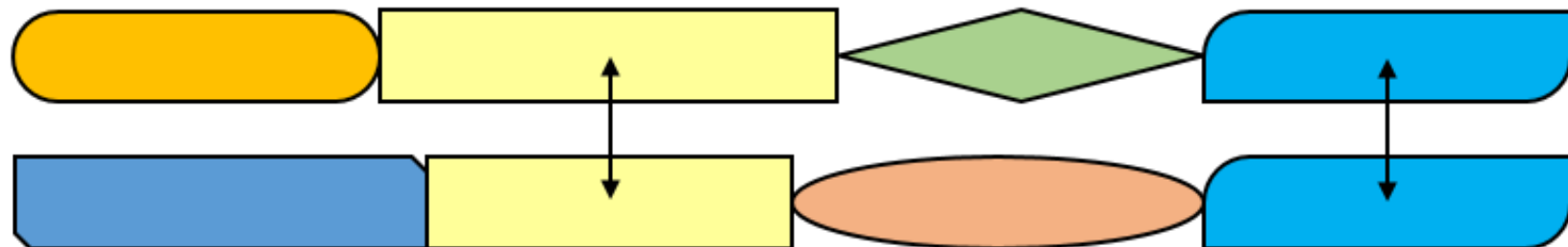 ||||||||||| ||||||| ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'

Query Sequence

Pairwise alignment

Global Alignment

Local Alignment

Pairwise alignment

These two methods of alignments are defined by different algorithms, which use scoring matrices to align the two different series of characters or patterns (sequences).

Pairwise alignment

These two methods of alignments are defined by different algorithms, which use scoring matrices to align the two different series of characters or patterns (sequences).

What is an algorithm?

Pairwise alignment

al·go·rithm

noun

noun: **algorithm**; plural noun: **algorithms**

a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

Global Alignment

- Needleman-Wunsch algorithm:
  - Published by Saul B. Needleman and Christian D. Wunsch in 1970
  - Aligns two complete sequences
  - Alignment contains all letters from both sequences
  - Comparison of closely related sequences
  - Usually used for comparing genes or proteins with similar function

Global Alignment
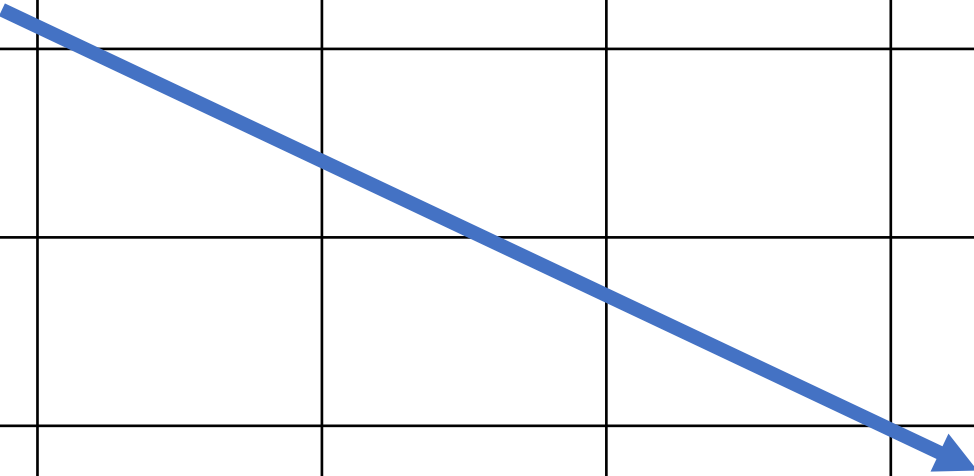
- Needleman-Wunsch algorithm:
  1. Scoring Matrix
  2. Trace Back
  3. Align Sequences

1) Scoring Matrix

	-	A	T	C	G
-	0				
T					
C					
G					

1) Scoring Matrix

	-	A	T	C	G
-	0				
T					
C					
G					

1) Scoring Matrix

	-	A	T	C	G
-	0	-2	-4	-6	-8
T	-2				
C	-4				
G	-6				

Gap = -2
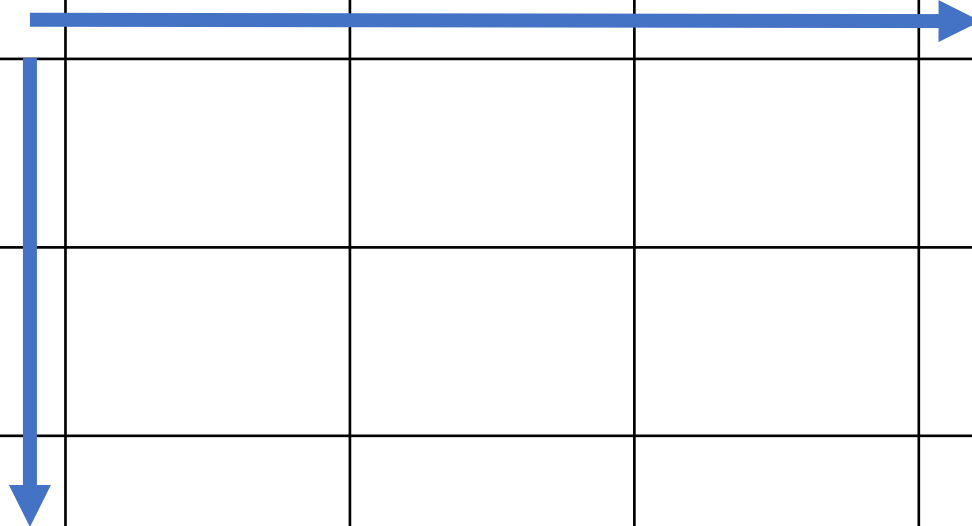
1) Scoring Matrix

	-	A	T	C	G
-	0	-2	-4	-6	-8
T	-2	-1 -4 / -4			
C	-4				
G	-6				

Gap = -2

Match = +1

Mismatch = -1

1) Scoring Matrix

	-	A	T	C	G
-	0	-2	-4	-6	-8
T	-2	-1			
C	-4				
G	-6				

Gap = -2

Match = +1

Mismatch = -1

1) Scoring Matrix

	-	A	T	C	G
-	0	-2	-4	-6	-8
T	-2	-1	-1	-3	-5
C	-4	-3	-2	0	-2
G	-6	-5	-4	-2	1

Gap = -2

Match = +1

Mismatch = -1

2) Traceback

	-	A	T	C	G
-	0	-2	-4	-6	-8
T	-2	-1	-1	-3	-5
C	-4	-3	-2	0	-2
G	-6	-5	-4	-2	1

2) Traceback

	-	A	T	C	G
-	0	-2	-4	-6	-8
T	-2	-1	-1	-3	-5
C	-4	-3	-2	0	-2
G	-6	-5	-4	-2	**1**

Find highest value

2) Traceback

	-	A	T	C	G
-	**0** ←	**-2**	-4	-6	-8
T	-2	-1	**-1**	-3	-5
C	-4	-3	-2	**0**	-2
G	-6	-5	-4	-2	**1**

Find highest value

Trace back to origin

2) Alignment

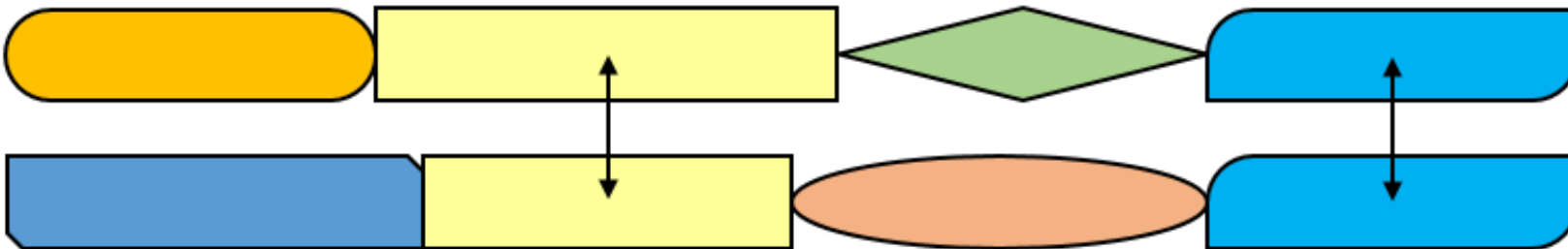|	-	A	T	C	G
-	**0**	**-2**	-4	-6	-8
T	-2	-1	**-1**	-3	-5
C	-4	-3	-2	**0**	-2
G	-6	-5	-4	-2	**1**

Diagonal = Character

Horizontal or Vertical = Gap

A T C G
- T C G

Local alignment

- Protein or gene sequences are often composed of multiple domains
- How do we find shared domains between sequences?
  - Find similar subsequences

Local Alignment

Global Alignment

- Smith-Waterman algorithm:
  - Temple F. Smith and Michael S. Waterman in 1981
  - Variant of Needleman-Wunsch algorithm
  - Origin and end of alignment differ from Needleman-Wunsch

Global Alignment

- Recall that we need:
  1. Scoring Matrix
  2. Trace Back
  3. Align Sequences

1) Scoring Matrix

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

Gap = -3

Negative values set to 0

1) Scoring Matrix

	-	T	G	A	C
-	0	0	0	0	0
G	0	0	5	2	0
G	0	0	5	3	0
A	0	0	2	10	7

Gap = -3

Match = +5

Mismatch = -2

2)Traceback

	-	T	G	A	C
-	0	0	0	0	0
G	0	0	5	2	0
G	0	0	5	3	0
A	0	0	2	10	7

Where do we start?

2)Traceback & Alignment

	-	T	**G**	**A**	C
-	0	0	0	0	0
G	0	0	5	2	0
G	0	0	**5**	3	0
A	0	0	2	**10**	7

Highest value is the origin and 0 is the end of alignment path.

A G
A G

Smith Waterman example number 2

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

Smith Waterman example number 2

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	18	14
C	0	5	1	0	0	4	5	2	9	18	14	15

Smith Waterman example number 2

Smith Waterman example number 2

Smith Waterman example number 2
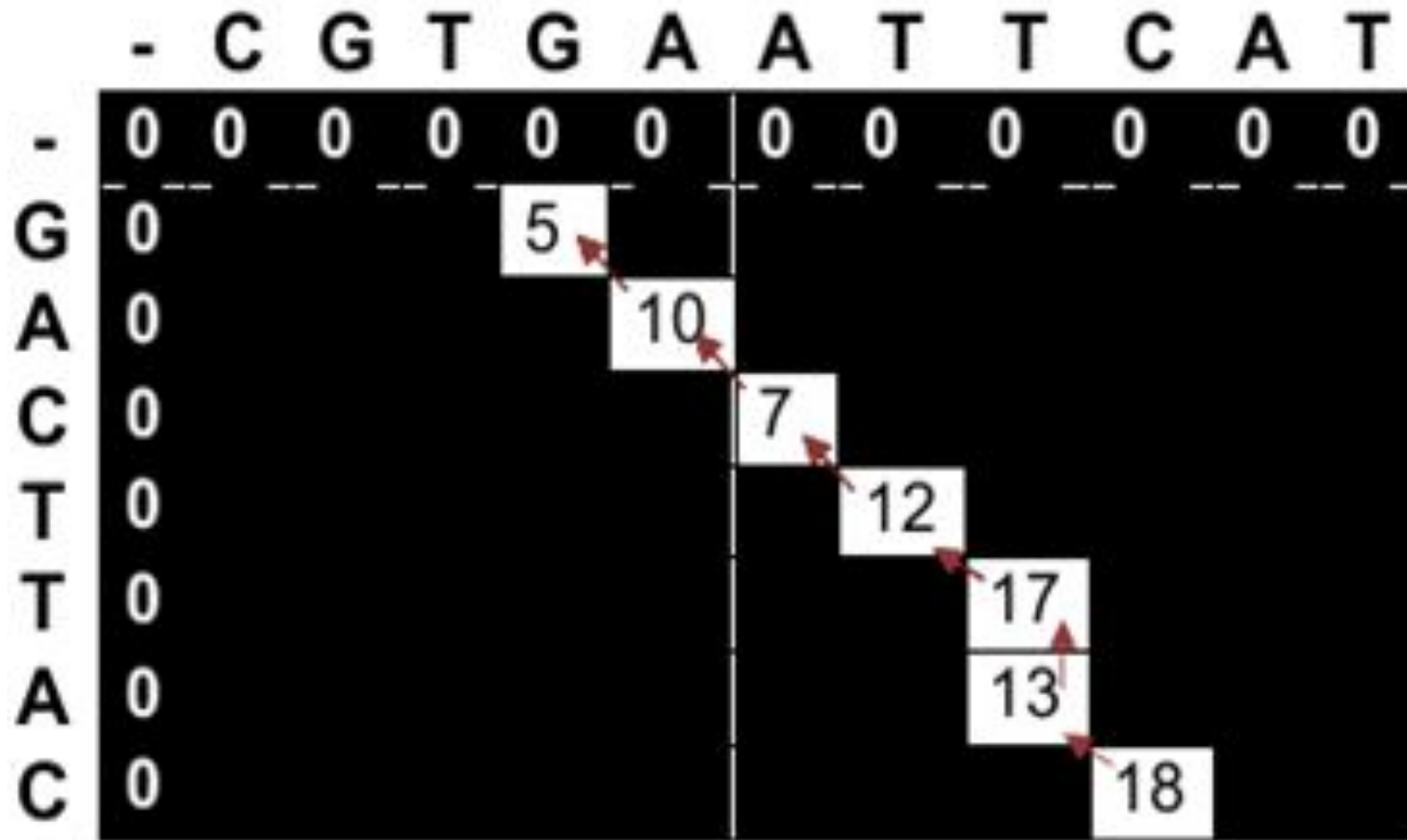
Smith Waterman example number 2

```
G A A T T C A          G A A T T - C
| |   |   |     |       | |   |   |   |
G A C T T - A          G A C T T A C


+ + - + + - +          + + - + + - +
5 5 3 5 5 4 5          5 5 3 5 5 4 5
```

Recap…

BLAST – Basic Local Alignment Search Tool

- Essential part of our bioinformatics toolkit
- Performs searches of query sequences against database
- NCBI Genbank demo…
- How do we evaluate "BLAST hits"?

Next week

- How does BLAST work
- Evaluating BLAST results
- Remote and local searches
- Output formats and parsing
- Working remotely using the secure shell

Next week

- Need to install BLAST suite from source
- Need to set up user accounts on analysis server