Smith & Waterman Algorithm for Pairwise Local Alignment

Kawshik Kumar Paul(1705043)
Monirul Haque Imon(1705054)
Maisha Rahman Mim (1705060)

Department of Computer Science and Engineering
Bangladesh University of Engieering and Technology
Dhaka, Bangladesh

July 7, 2021

Lecture Outline

1. Sequence Alignment
2. Sequence Edits
3. Score Matrix
4. Dynamic Programming
5. Local Sequence Alignment
6. Smith & Waterman Algorithm
7. Example

1 Sequence Alignment

2 Sequence Edits

3 Score Matrix

4 Dynamic Programming

5 Local Sequence Alignment

6 Smith & Waterman Algorithm

7 Example

Sequence Alignment

- Why do we need to align sequence?

Why do we need to align sequence?

- Comparing DNA/protein sequences for
  - Similarity
  - Homology
- Prediction of function
- Construction of phylogeny Shotgun assembly
  - End-space-free alignment / overlap alignment
- Finding motifs
- Understanding Evolutionary Relationships

Sequence Alignment

- Procedure of comparing to (Pairwise) or more (Multiple) sequences by searching for a series of individual characters that are in the same order in the sequence.

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

Sequence Alignment

- **Definition**

  Given two strings $x = x_1 x_2 ... x_m$ and $y = y_1 y_2 ... y_n$,
  an alignment is an assignment of gaps to positions 0 ... M in x and 0 ... N in y, so as to line up each letter in one sequence with either a letter or a gap in the other sequence.

A Simple Alignment

- Let us try to align two short nucleotide sequences:
  -AATCTATA and AAGATA

- Without considering any gaps (insertions/deletions) there are 3 possible ways to align these sequences

```
A A T C T A T A      A A T C T A T A      A A T C T A T A
A A G A T A            A A G A T A            A A G A T A
```

- Which one is better?

What is a Good Alignment

A G G C T A G T T , A G C G A A G T T

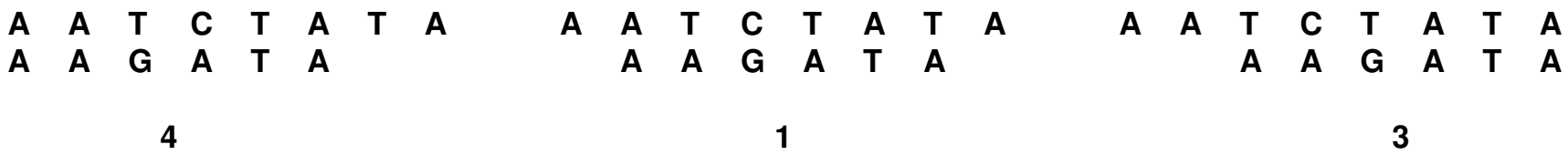
A G G C T A G T T - **A G C G A A G T T T**	**Matches = 6** **Mistatches = 3** **Gap = 1**
A G G C T A - G T T - **A G - C G A A G T T T**	**Matches = 7** **Mismatches = 1** **Gaps = 3**
A G G C - T - G T T - **A G - C G - A A G T T**	**Matches = 7** **Mismatches = 0** **Gaps = 5**

Scoring the Alignments

- We need to have a scoring mechanism to evaluate alignments
  - match score
  - mismatch score
- We can have the total score as:

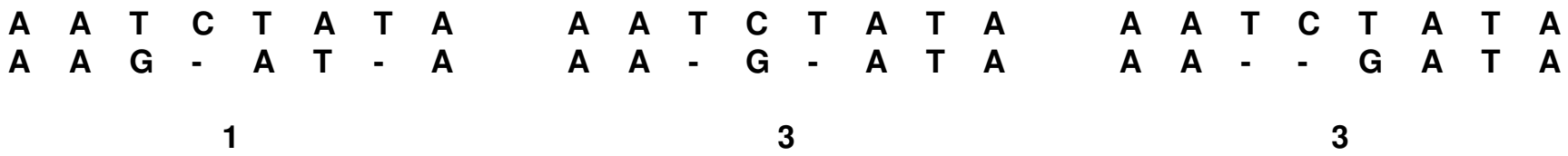
$$\sum_{i=1}^{n} \textit{match or mismatch score at position } i$$

- For the simple example, assume a match score of 1 and a mismatch score of 0:



Simple Alignment with Gaps

- Considering gapped alignments vastly increases the number of possible alignments:



- If gap penalty is -1, what will be the new scores?

1 Sequence Alignment

2 Sequence Edits

3 Score Matrix

4 Dynamic Programming

5 Local Sequence Alignment

6 Smith & Waterman Algorithm

7 Example

Sequence Edits

Lets do some sequence edits and view scores

Sequence Edits

Three types of sequence edits

1. Mutations
2. Insertions
3. Deletions

Mutations

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

Insertions

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

Deletions

A G G C C T C
A G G . C T C

Scoring Function

Match: +m

Mismatch: -s

Gap: -d

Score F = (# matches) x m - (# mismatches) x s - (# gaps) x d

1 Sequence Alignment

2 Sequence Edits

3 Score Matrix

4 Dynamic Programming

5 Local Sequence Alignment

6 Smith & Waterman Algorithm

7 Example

Score Matrix

- Assign scores to each pair of symbol
  - Higher score means more similarity.

Score Matrix

- DNA
  - Match = +1
  - Mismatch = -3
  - Gap penalty = -5
  - Gap extension penalty = -2
- Protein sequences
  - Blossum62 matrix
  - Gap open penalty = -11
  - Gap extension penalty = -1

1 Sequence Alignment

2 Sequence Edits

3 Score Matrix

4 Dynamic Programming

5 Local Sequence Alignment

6 Smith & Waterman Algorithm

7 Example

- How do we compute the best alignment?

Alignment is Additive

- Observation:
  The score of aligning $x_1 \ldots x_M$ and $y_1 \ldots y_N$ is additive

  Say that $x_1 \ldots x_i \quad x_{i+1} \ldots x_M$
  aligns to $y_1 \ldots x_j \quad x_{j+1} \ldots x_N$

  The two scores add up:

  F(x[1:M], y[1:N]) = F(x[1:i], y[1:j]) + F(x[i+1:M], y[j+1:N])

Types of Alignment

- Global
  - Strings of similar size
    - Genes with a similar structure
    - Larger regions with a preserved order (syntenic regions)
- Local
  - Finding similar regions among:
    - Dissimilar regions
    - Sequences of different lengths

Dynamic Programming

- Instead of evaluating every possible alignment, we can create a table of partial scores by breaking the alignment problem into subproblems.
- Consider two sequences CACGA and CGA
  - we have three possibilities for the first position of the alignment

First Position	Score	Remaining seqs
C C	+1	ACGA GA
– C	-1	CACGA GA
C –	-1	ACGA CGA

1 Sequence Alignment

2 Sequence Edits

3 Score Matrix

4 Dynamic Programming

5 Local Sequence Alignment

6 Smith & Waterman Algorithm

7 Example

Local Sequence Alignment

- Suppose we have a long DNA sequence (eg 4000 bp) and we want to compare it with the complete yeast genome (12.5Mbp)
- What if only a portion of our query, say 200 bp length, has strong similarity to a gene in yeast.

Local Sequence Alignment Problem

- Given two strings

$$x = x_1 \ldots\ldots x_M$$

$$y = y_1 \ldots\ldots x_N$$

Find substring y', x' whose similarity (optimal global alignment value) is maximum.

$$x = aaaacc\underline{cccggg}gtta$$

$$y = tt\underline{cccggg}aaccaacc$$

1 Sequence Alignment

2 Sequence Edits

3 Score Matrix

4 Dynamic Programming

5 Local Sequence Alignment

6 Smith & Waterman Algorithm

7 Example

Smith & Waterman Algorithm

- $F(i, j)$= optimal local similarity among suffixes $A(1 : i)$ and $B(1 : j)$
- Recurrence relation
  - $F(i, 0) = 0$
  - $F(0, j) = 0$
  - F(i,j) = max $\begin{cases} 0 \\ F(i, j-1) + s(-, B(j)) \\ F(i-1, j) + s(A(i), -) \\ F(i-1, j-1) + s(A(i), B(j)) \end{cases}$

1 Sequence Alignment

2 Sequence Edits

3 Score Matrix

4 Dynamic Programming

5 Local Sequence Alignment

6 Smith & Waterman Algorithm

7 Example

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

	–	E	Q	L	L	K	A	L	E	F	K	L
–												
K												
V												
L												
E												
F												
G												
Y												

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i,0) = 0
F(0,j) = 0

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0											
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i, j) = max [0, F(i,j-1) + s(-,Q(j)),F(i-1,j) + s(P(i),-),F(i-1,j-1) + s(P(i),Q(j)]

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0										
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i, j) = max [0, F(i,j-1) + s(-,Q(j)),F(i-1,j) + s(P(i),-),F(i-1,j-1) + s(P(i),Q(j)]

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0									
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i, j) = max [0, F(i,j-1) + s(-,Q(j)),F(i-1,j) + s(P(i),-),F(i-1,j-1) + s(P(i),Q(j)]

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0								
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i, j) = max [0, F(i,j-1) + s(-,Q(j)),F(i-1,j) + s(P(i),-),F(i-1,j-1) + s(P(i),Q(j)]

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0							
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i, j) = max [0, F(i,j-1) + s(-,Q(j)),F(i-1,j) + s(P(i),-),F(i-1,j-1) + s(P(i),Q(j)]

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4						
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i, j) = max [0, F(i,j-1) + s(-,Q(j)),F(i-1,j) + s(P(i),-),F(i-1,j-1) + s(P(i),Q(j)]

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Algorithm

F(i, j) = max [0, F(i,j-1) + s(-,Q(j)),F(i-1,j) + s(P(i),-),F(i-1,j-1) + s(P(i),Q(j)]

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: . . . . . F
P: . . . . . F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: . . . . E F
P: . . . . E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: . . . L E F
P: . . . L E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: . . - L E F
P: . . V L E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: . A - L E F
P: . - V L E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: K A - L E F
P: K - V L E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: K A - L E F
P: K - V L E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: K - A L E F
P: K V - L E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q = E Q L L K A L E F K L P = K V L E F G Y

Linear gap model

Gap= -1
Match= 4
Mismatch= -2

Alignment

Q: K A L E F
P: K V L E F

	–	E	Q	L	L	K	A	L	E	F	K	L
–	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

References

- https://ocw.metu.edu.tr/course/view.php?id=37
- https://www.cs.cmu.edu/ ckingsf/bioinfo-lectures/local.pdf

Thank you

Stay Home, Stay Safe