Bioinformatics: Sequence Alignment

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

Inspiring Excellence

Book Reference

Bioinformatics Algorithms, An Active Learning Approach , Vol 1, Chapter 5

Sequence Alignment

Biological Question

- How can we find similarity between two sequences?
- Why is it important?
- Similar Sequence → Similar Structure → Similar Function
- The purpose of sequence alignment is to line up all residues in the inputted sequences for maximal level of similarity, in the sense of their functional or evolutionary relationship.

ATGCATGC	A**TGCATGC**-	A**TGC**-**TTA**-
TGCATGCA	-**TGCATGC**A	-**TGCA**TT**A**A

Pairwise Sequence Similarity:
`https://www.ebi.ac.uk/jdispatcher/psa`

Pairwise Sequence Alignment

```
#=======================================
#
# Aligned_sequences: 2
# 1: test1
# 2: test2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 105
# Identity:      96/105 (91.4%)
# Similarity:    99/105 (94.3%)
# Gaps:           0/105 ( 0.0%)
# Score: 522.0
#
#
#=======================================
```

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

```
test1              1 MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYT     50
                       |||||||||||:.||:|||||||||||||||||||||||||||||.|:|||
test2              1 MGDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAAGFSYT     50

test1             51 AANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLK    100
                       .|||||||.||||||||||||||||||||||||.|||||.||||||||||
test2             51 DANKNKGITWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLK    100

test1            101 KATNE    105
                     |||||
test2            101 KATNE    105
```

BRAC UNIVERSITY
Inspiring Excellence

Sequence Alignment Problem

- Input Data:
  - Two sequences $S1$ and $S2$
- Parameter (s)
  - A scoring function f for
    - Substitutions
    - Gaps
- Output
  - The optimal alignment of $S1$ and $S2$ that has the maximal score.

$$\arg\max_{align}(f(align(S1, S2)))$$

- Enumerating all possible alignments is not feasible.
- A residue can either align to another residue or to a gap.
- We can use **dynamic programming**.

Formulation

- Align two sequences, x and y
- $F(i,j)$ is the score of the best alignment between $x_{1\ldots i}$ and $y_{1\ldots j}$
- $s(A,B)$ is the score for substituting A with B
- d is the (linear) gap penalty

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) & \textbf{\textit{x}}_{\textbf{i}} \textbf{ aligned to } \textbf{\textit{y}}_{\textbf{j}} \\ F(i-1, j) + d & \textbf{\textit{x}}_{\textbf{i}} \textbf{ aligned to } \textbf{\textit{a gap}} \\ F(i, j-1) + d & \textbf{\textit{y}}_{\textbf{j}} \textbf{ aligned to } \textbf{\textit{a gap}} \end{cases}$$

Dynamic Programming

- Break the problem into smaller sub-problems.
- Solve these sub-problems optimallyrecursively.
- Use these optimal solutions to construct anoptimal solution for the original problem.

Input Sequence 1: AAG

Input Sequence 2: AGC

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

For simplicity, let's set (i.e. linear gap penalty)

gap OPEN (d) = gap EXTEND (e) = -5

```
GAC-AT
C-ACAT
```

(-7) + (-5) + (-7) + (-5) + 2 + 2= -20

Sequence Alignment: Example

		A	A	G
A				
G				
C				

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0			
A				
G				
C				

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

Sequence Alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

$$
\begin{tikzcd}
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] \\
& s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j)
\end{tikzcd}
$$

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0	-5	-10	-15
A	-5			
G	-10			
C	-15			

Sequence Alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

$$\begin{tikzcd}
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] \\
& s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j)
\end{tikzcd}$$

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

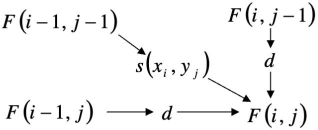
		A	A	G
	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-3	-1
C	-15	-8	-8	**-6**

Sequence Alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

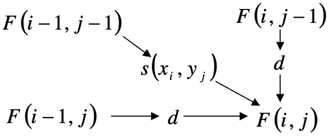
		A
	0	-5
A	-5	2



	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A
	0	-5
A	-5	2



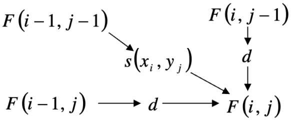
-5 + (-5) = -10

$\boxed{0 + 2 = 2}$

-5 + (-5) = -10

Sequence Alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-3	-1
C	-15	-8	-8	-6

Sequence Alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

$$
\begin{tikzcd}
F(i-1,j-1) \arrow[dr] & & & F(i,j-1) \arrow[d] \\
& s(x_i,y_j) \arrow[drr] & & d \arrow[d] \\
F(i-1,j) \arrow[r] & d \arrow[rr] & & F(i,j)
\end{tikzcd}
$$

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-3	-1
C	-15	-8	-8	-6

Sequence Alignment: Example

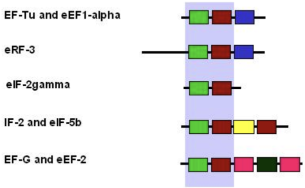
- Trace back to the upper left. Each arrow introduces one symbol at the end of each aligned sequence.

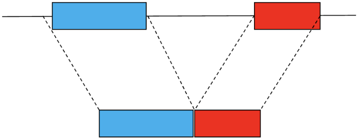
```
A A G –
– A G C

A A G –
A – G C
```

		A	A	G
	0	-5		
A		2	-3	
G				-1
C				**-6**

- Two functionally related proteins might be significantly/ largely different in their whole sequences but share similar important functional domains. The sequence fragment of the functional domain might be very conservative across different proteins in the same protein family, and determine the biological function.
- Secondly, the new discovery of introns required the DNA sequence alignment algorithms to be able to handle large deletions and interspersed conserved fragments (exons) and variable fragments (introns). Needleman-Wunsch $\rightarrow$ Smith-Waterman





Sequence Alignment: Local

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

Global alignment

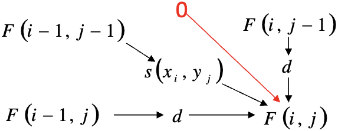
$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ \boxed{0} \end{cases}$$

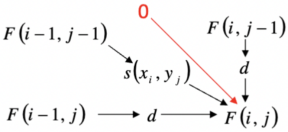
Local alignment

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ \boxed{0} \end{cases}$$



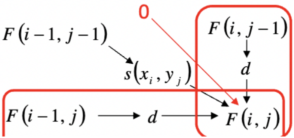
	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal local alignment of AAG and AGC.
Use a linear gap penalty of d = -5.

		A	A	G
A				
G				
C				

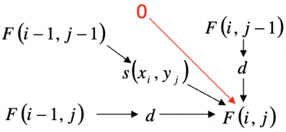
	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal local alignment of AAG and AGC.
Use a linear gap penalty of d = -5.

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

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

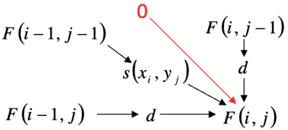


Find the optimal local alignment of AAG and AGC.
Use a linear gap penalty of d = -5.

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

Sequence Alignment: Local

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal local alignment of AAG and AGC.
Use a linear gap penalty of d = -5.

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

- Trace back begins at the highest score in the matrix and continues until you reach 0.

A G
A G

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

- And also the secondary best alignment

A
A

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

A	A	G	–
–	A	G	C

A	A	G	–
A	–	G	C

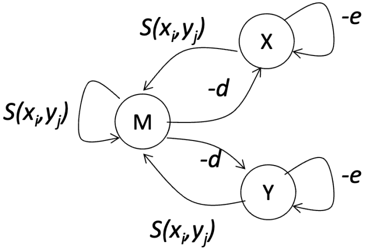
$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$

A	G
A	G

A
A

Affine Gap Penalty: Alignment as a series of state(s)



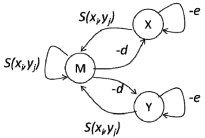
M	*Match (**not necessarily identical**)*
X	*Insert* at sequence X (delete at sequence Y)
Y	*Insert* at sequence Y (delete at sequence Y)

d	Gap open
e	Gap Extension

Affine Gap Penalty: Alignment as a series of state(s)

- M (i,j) is the score of the best alignment between $x_{1\ldots i}$ and $y_{1\ldots j}$, given $\boldsymbol{x_i}$ **aligned to** $\boldsymbol{y_j}$
- X (i,j) is the score of the best alignment between $x_{1\ldots i}$ and $y_{1\ldots j}$, given $\boldsymbol{x_i}$ **aligned to *a gap***
- Y (i,j) is the score of the best alignment between $x_{1\ldots i}$ and $y_{1\ldots j}$, given $\boldsymbol{y_j}$ **aligned to *a gap***

$$M(i,j) = \max \begin{cases} M(i-1, j-1) + s(x_i, y_j) \\ X(i-1, j-1) + s(x_i, y_j) \\ Y(i-1, j-1) + s(x_i, y_j) \end{cases}$$



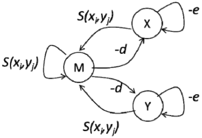
$$X(i,j) = \max \begin{cases} M(i-1, j) - d \\ X(i-1, j) - e \end{cases} \qquad Y(i,j) = \max \begin{cases} M(i, j-1) - d \\ Y(i, j-1) - e \end{cases}$$

x_i aligned to y_j

$$M(i,j) = \max \begin{cases} M(i-1,j-1) + s(x_i, y_j) & \text{after a match} \\ X(i-1,j-1) + s(x_i, y_j) & \multirow{2}{*}{after a gap} \\ Y(i-1,j-1) + s(x_i, y_j) & \end{cases}$$

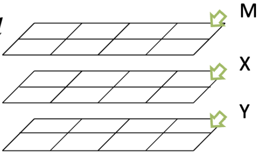
x_i aligned to *a gap*

$$X(i,j) = \max \begin{cases} M(i-1,j) - d \\ X(i-1,j) - e \end{cases}$$



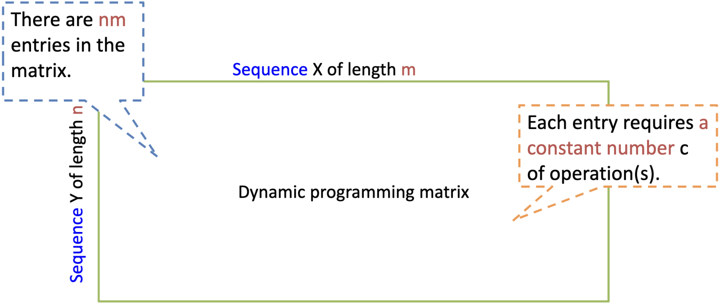
y_j aligned to *a gap*

$$Y(i,j) = \max \begin{cases} M(i,j-1) - d \\ Y(i,j-1) - e \end{cases}$$



Runtime?

- How to do the search on 550K proteins or billions of reads?
- BLAST: `https://blast.ncbi.nlm.nih.gov/Blast.cgi`



O(mn) operations needed in total.

Multiple Sequence Alignment

- If sequence similarity is weak, pairwise alignment may not identify biologically related sequences.
- simultaneous comparison of many sequences often allows us to find similarities that pairwise sequence comparison fails to reveal.
- Bioinformaticians sometimes say that while pairwise alignment whispers, multiple alignment shouts.

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA

SFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTEATIGA
-YEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
----PTMISSLEILFAAGDRLSSQDAILARRAVGSGV-Y-NAYGPTENTVLS
```

```
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA

SFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTEATIGA
-YEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
----PTMISSLEILFAAGDRLSSQDAILARRAVGSGV-Y-NAYGPTENTVLS
```

BRAC UNIVERSITY
Inspiring Excellence

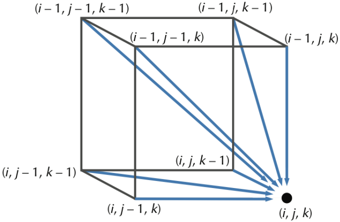
Multiple Sequence Alignment

A	T	-	G	T	T	a	T	A
A	g	C	G	a	T	C	-	A
A	T	C	G	T	-	C	T	c

0	1	2	2	3	4	5	6	7	8
0	1	2	3	4	5	6	7	7	8
0	1	2	3	4	5	5	6	7	8

$$(0,0,0) \to (1,1,1) \to (2,2,2) \to (2,3,3) \to (3,4,4) \to (4,5,5) \to (5,6,5) \to (6,7,6) \to (7,7,7) \to (8,8,8)$$

- A multiple alignment of t strings $v_1, \cdots, v_t$, also called a t-way alignment, is specified by a matrix having t rows, where the i-th row contains the symbols of v_i in order, interspersed with space symbols.



$$s_{i,j,k} = \max \begin{cases} s_{i-1,j,k} & + \; Score(v_i, -, -) \\ s_{i,j-1,k} & + \; Score(-, w_j, -) \\ s_{i,j,k-1} & + \; Score(-, -, u_k) \\ s_{i-1,j-1,k} & + \; Score(v_i, w_j, -) \\ s_{i-1,j,k-1} & + \; Score(v_i, -, u_k) \\ s_{i,j-1,k-1} & + \; Score(-, w_j, u_k) \\ s_{i-1,j-1,k-1} & + \; Score(v_i, w_j, u_k) \end{cases}$$

- In the case of t sequences of length n, the alignment graph consists of approximately n^t nodes, and each node has up to $2_t - 1$ incoming edges, yielding a total runtime of $O(n^t 2^t)$
- As t grows, the dynamic programming algorithm becomes impractical.