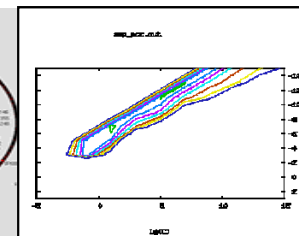
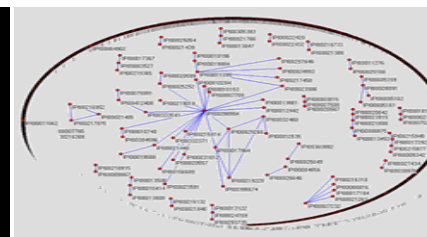
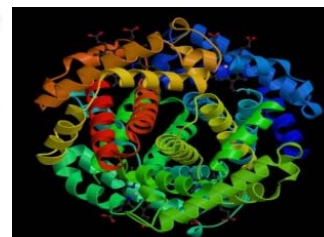
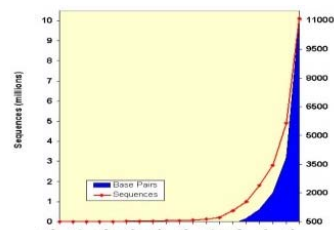


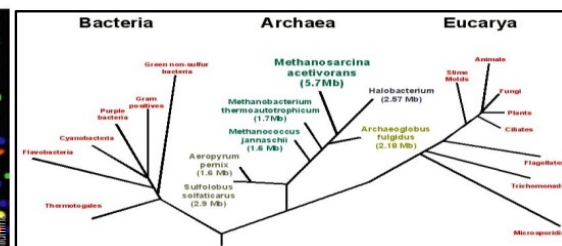
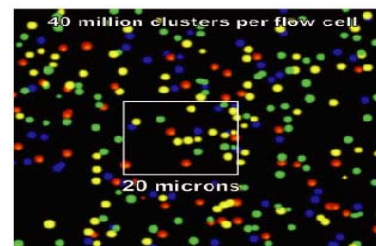
北京大学生物信息学中心 高歌
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University





TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACC
CCCTAACCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAA

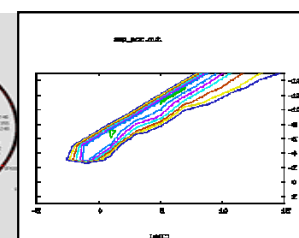
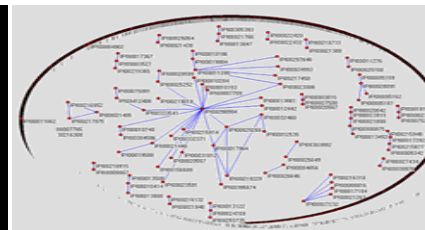
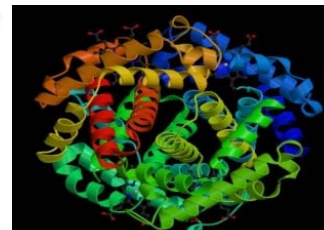
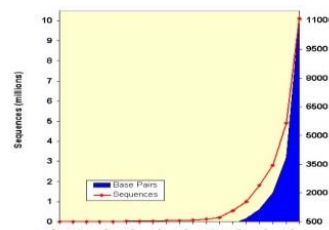
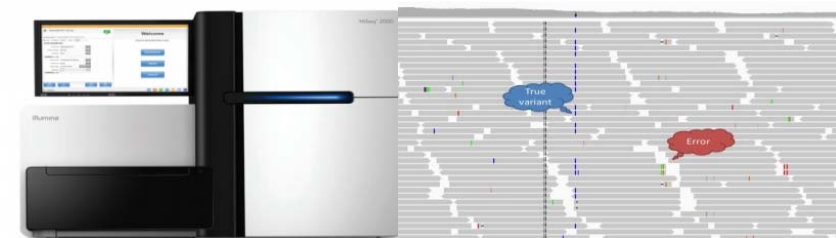


S1: Alignment with Affine Gap Penalty

北京大学生物信息学中心 高歌

Ge Gao, Ph.D.

Center for Bioinformatics, Peking University



Affine gap penalty: **opening** a gap receives a score of **d**; **extending** a gap receives a score of **e**.

$$\text{Penalty} = d + (n-1)^* e$$

```
#=====
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 292.5
#
#=====

HBA_HUMAN      1 MV-LSPADKTNVKAAWGKVGAGHAGEYGAELERMFLSFPTTKTYFPHF-D      48
               || :|.:|:|.|.||| | :..|.|.|||.|:..:|.|:..:|..| |
HBB_HUMAN      1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD      48

HBA_HUMAN     49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR      93
               || .|:..:|.|.||| |..|.:.:..:|.|:|:.....:|.|:|..|.|.
HBB_HUMAN     49 LSTPDVAVMGNPVKVKAHGKKVLGAFSDGLAHLNCLKGTFTATLSELHCDKLH      98

HBA_HUMAN     94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR      142
               |||.|:|:|.|.:.|.:.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
HBB_HUMAN     99 VDPENFRLLGNLVLCVLAHFFGKEFTPPVQAAYQKVVAGVANALAHKYH      147
```

Global alignment (Needleman-Wunsch)

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

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

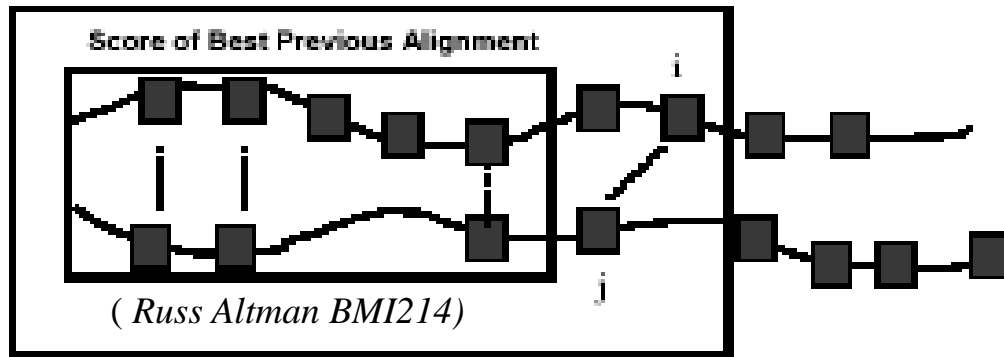
Local alignment (Smith-Waterman)

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

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

Alignment as (a series of) state(s)

New Best Alignment = Previous Best + Local Best



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

L S P -

- T P E



X M M Y

M *Match (not necessarily identical)*

X *Insert at sequence X
(delete at sequence Y)*

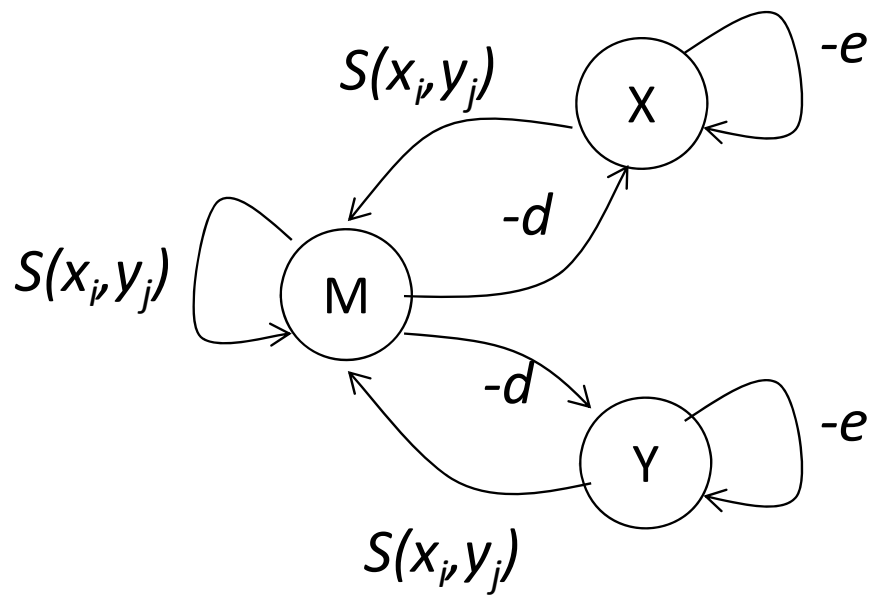
Y *Insert at sequence Y
(delete at sequence X)*

A A G -

A - G C



M X M Y

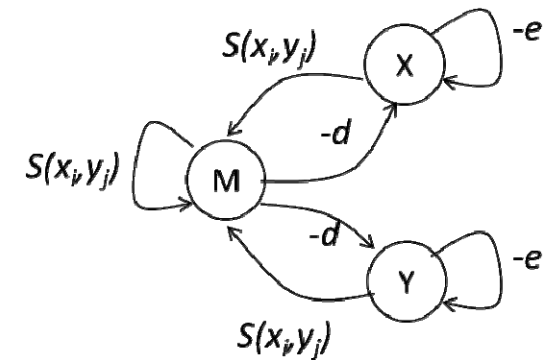


M	Match (<u>not necessarily identical</u>)
X	Insert at sequence X (delete at sequence Y)
Y	Insert at sequence Y (delete at sequence X)

d	Gap open
e	Gap Extension

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

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

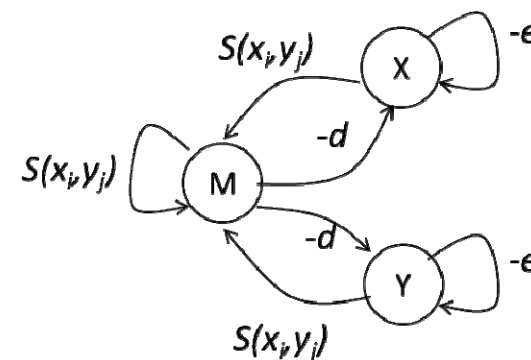


$$X(i, j) = \max \begin{cases} M(i-1, j) - d \\ X(i-1, j) - e \end{cases}$$

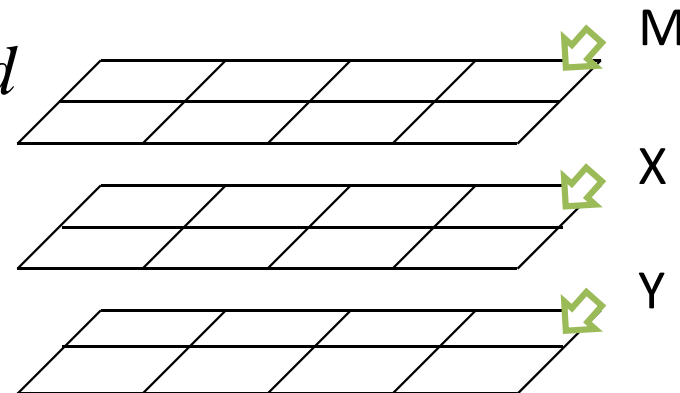
$$Y(i, j) = \max \begin{cases} M(i, j-1) - d \\ Y(i, j-1) - e \end{cases}$$

x_i aligned to y_j $M(i, j) = \max \left\{ \begin{array}{l} M(i-1, j-1) + s(x_i, y_j) \quad \text{after a match} \\ X(i-1, j-1) + s(x_i, y_j) \\ Y(i-1, j-1) + s(x_i, y_j) \end{array} \right\}$ after a gap

x_i aligned to a gap $X(i, j) = \max \left\{ \begin{array}{l} M(i-1, j) - d \\ X(i-1, j) - e \end{array} \right\}$

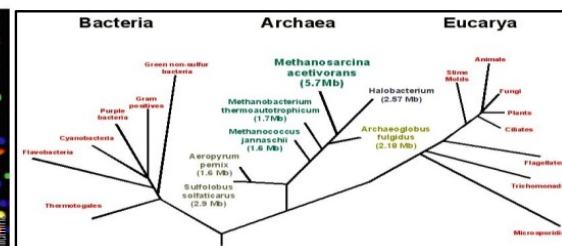
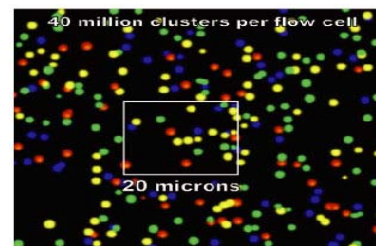


y_j aligned to a gap $Y(i, j) = \max \left\{ \begin{array}{l} M(i, j-1) - d \\ Y(i, j-1) - e \end{array} \right\}$





TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACC
CCCTAACCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAA

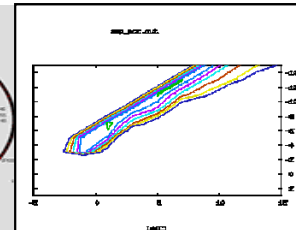
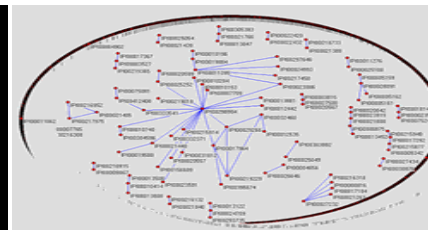
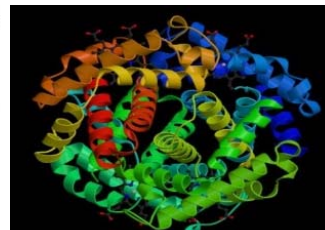
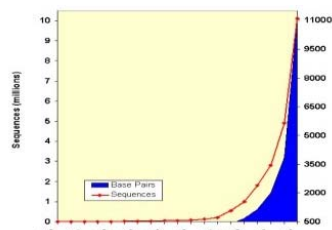
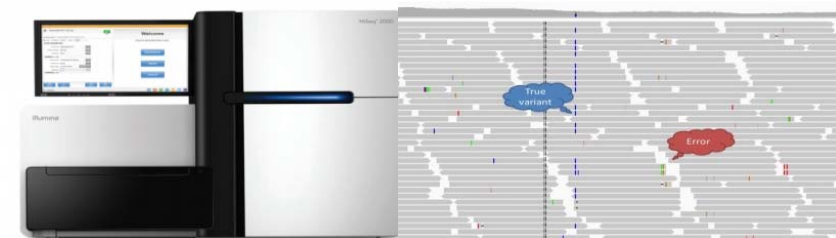


S2: Time Complexity for Needleman-Wunsch Algorithm

北京大学生物信息学中心 高歌

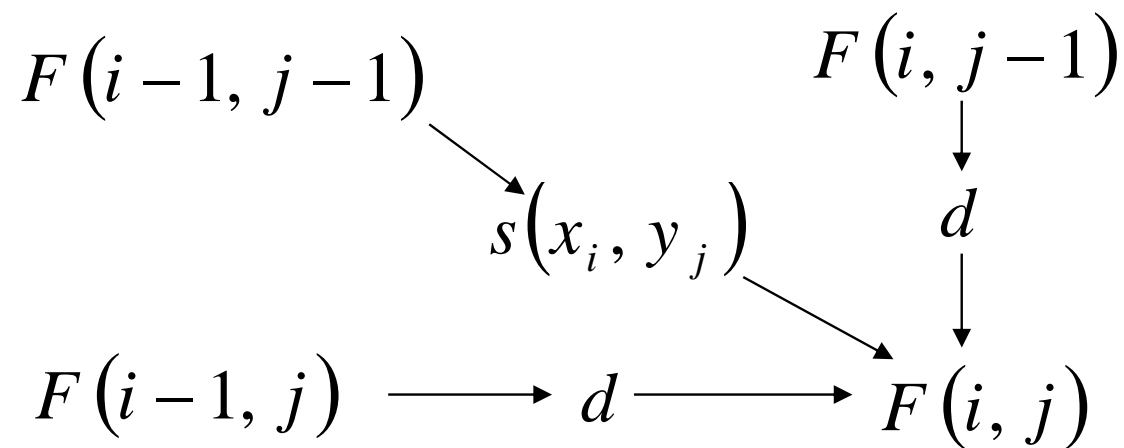
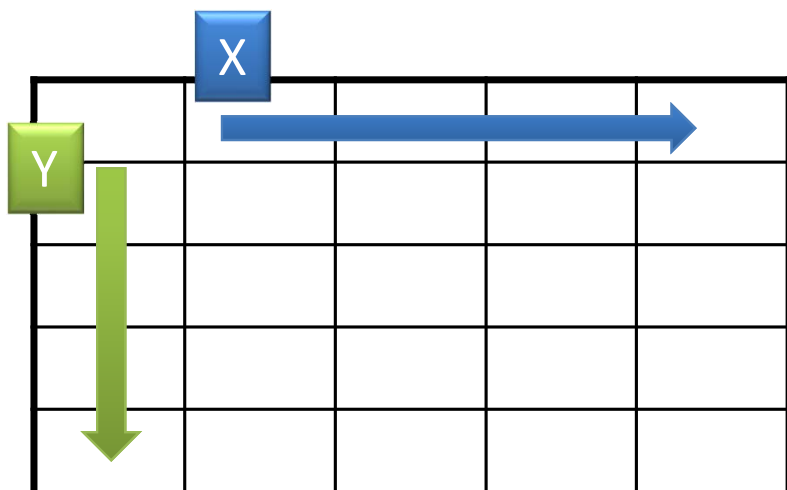
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University



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

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) & \mathbf{x_i \text{ aligned to } y_j} \\ F(i-1, j) + d & \mathbf{x_i \text{ aligned to a gap}} \\ F(i, j-1) + d & \mathbf{y_j \text{ aligned to a gap}} \end{cases}$$



There are nm entries in the matrix.

Sequence X of length m

Sequence Y of length n

Dynamic programming matrix

Each entry requires a constant number c of operation(s).

$O(mn)$ operations needed in total.

Summary Questions

- Could you write the formula for local alignment with Affine Gap Penalty? Explain your result.
- Do you think the Time Complexity for Smith-Waterman Algorithm would be different from of Needleman-Wunsch? Why?

生物信息学：导论与方法

Bioinformatics: Introduction and Methods



<https://www.coursera.org/course/pkubioinfo>