双序列比对-局部比对算法

高建召

致谢

此PPT内容参考了
杨建益老师的PPT。
https://yanglab.nankai.edu.cn/

双序列比对的算法

- Dot Matrix，点阵法
- 动态规划算法：
  - Global: Needleman-Wunsch
  - Local: Smith-Waterman
- Word or *k*-tuple算法：FASTA，BLAST

动态规划算法：局部比对

Local alignment: Smith-Waterman

Smith-Waterman algorithm

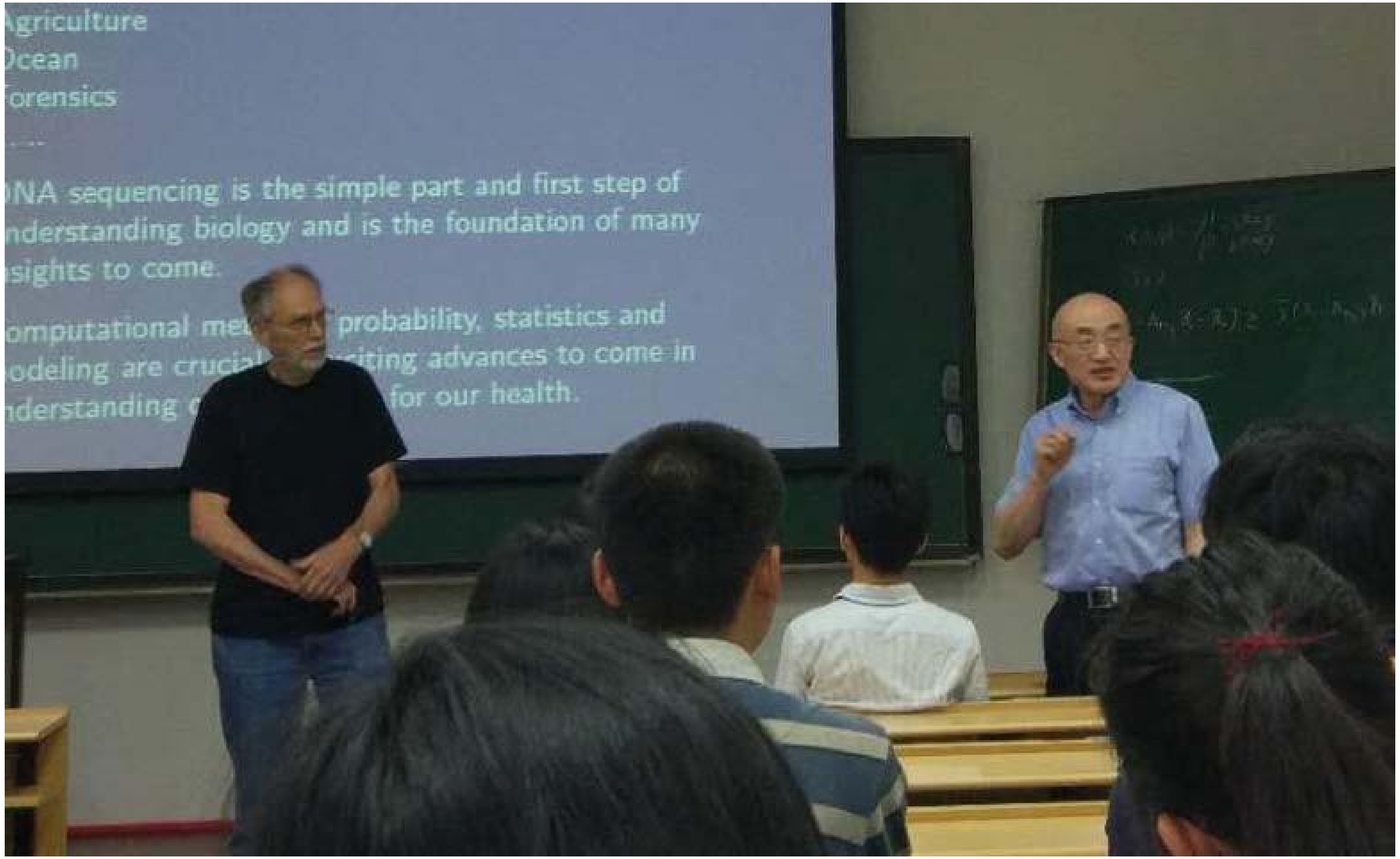
- For generating optimal local alignment?

T F Smith & M S Waterman, Identification of common molecular subsequences. J Mol Biol (1981) 147, 195-197.

Temple F. Smith (1939-)
Boston University

Michael S. Waterman (1942-)
University of Southern California (USC)

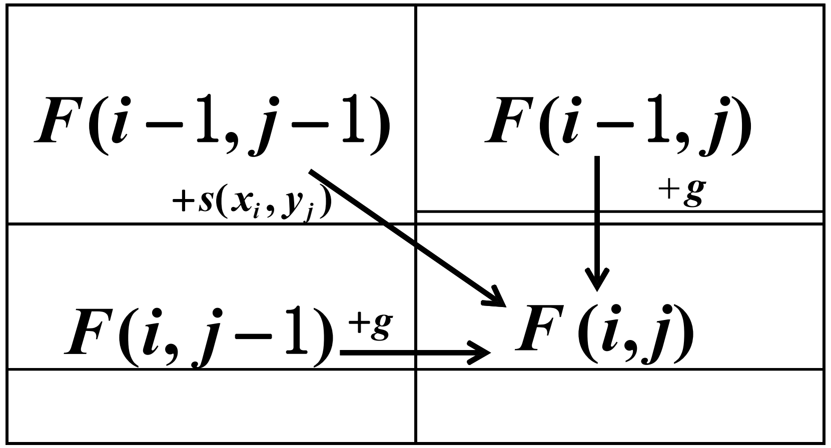
Michael S. Waterman (1942-)


Agriculture
Ocean
Forensics
DNA sequencing is the simple part and first step of
nderstanding biology and is the foundation of many
nsights to come.
omputational me probability, statistics and
odeling are crucia iting advances to come in
nderstanding for our health.

Two differences between SW and NW

- 1. Non-negative scores

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



Two differences between SW and NW

- 2. Traceback starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered.

Example

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

		A	A	G	A
T					
T					
A					
A					
G					

Example

$$s(x_i, y_j) = \begin{cases} 1, & \textbf{if } x_i = y_j \\ -1, & \textbf{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

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

Initialization

Example

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

		A	A	G	A
	0	0	0	0	0
T	0	0			
T	0	0			
A	0	↖ 1			
A	0	↖ 1			
G	0	0			

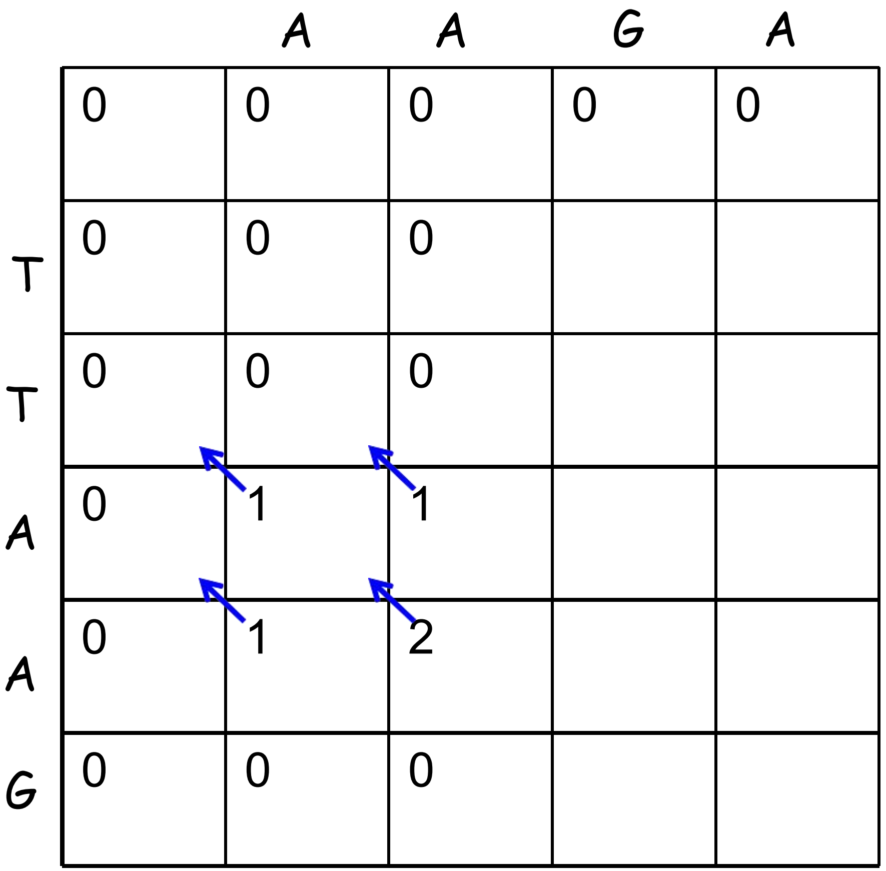
Filling….

Example

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening



		A	A	G	A
	0	0	0	0	0
T	0	0	0		
T	0	0	0		
A	0	1	1		
A	0	1	2		
G	0	0	0		

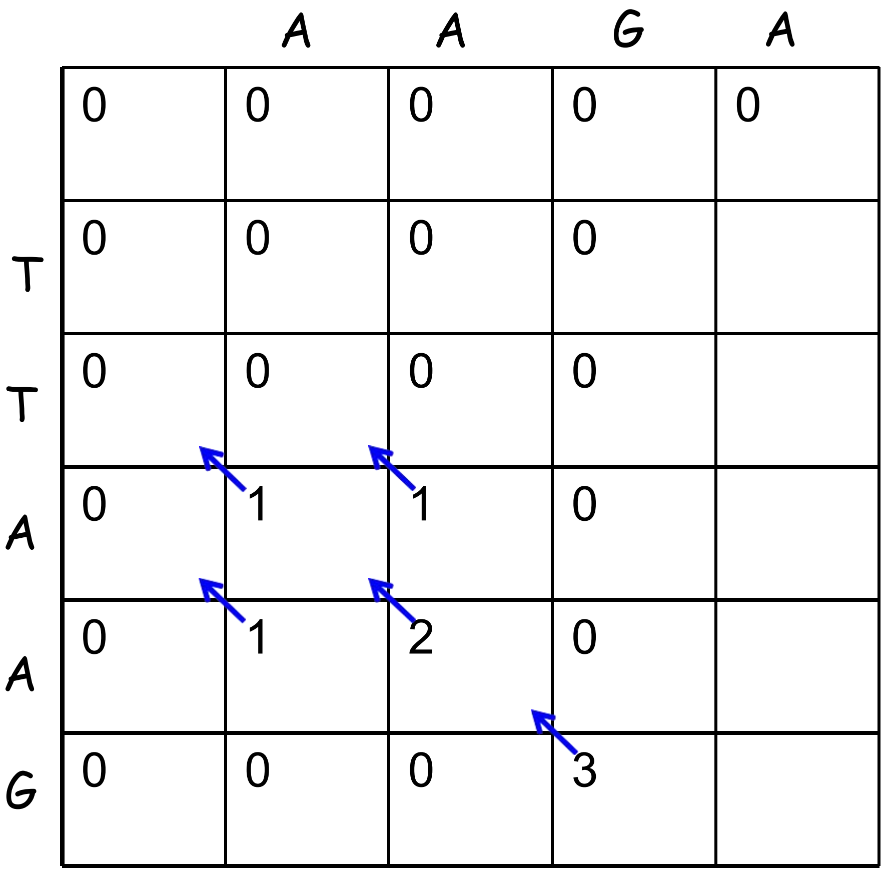
Filling….

Example

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening



		A	A	G	A
	0	0	0	0	0
T	0	0	0	0	
T	0	0	0	0	
A	0	↖1	↖1	0	
A	0	↖1	↖2	0	
G	0	0	0	↖3	

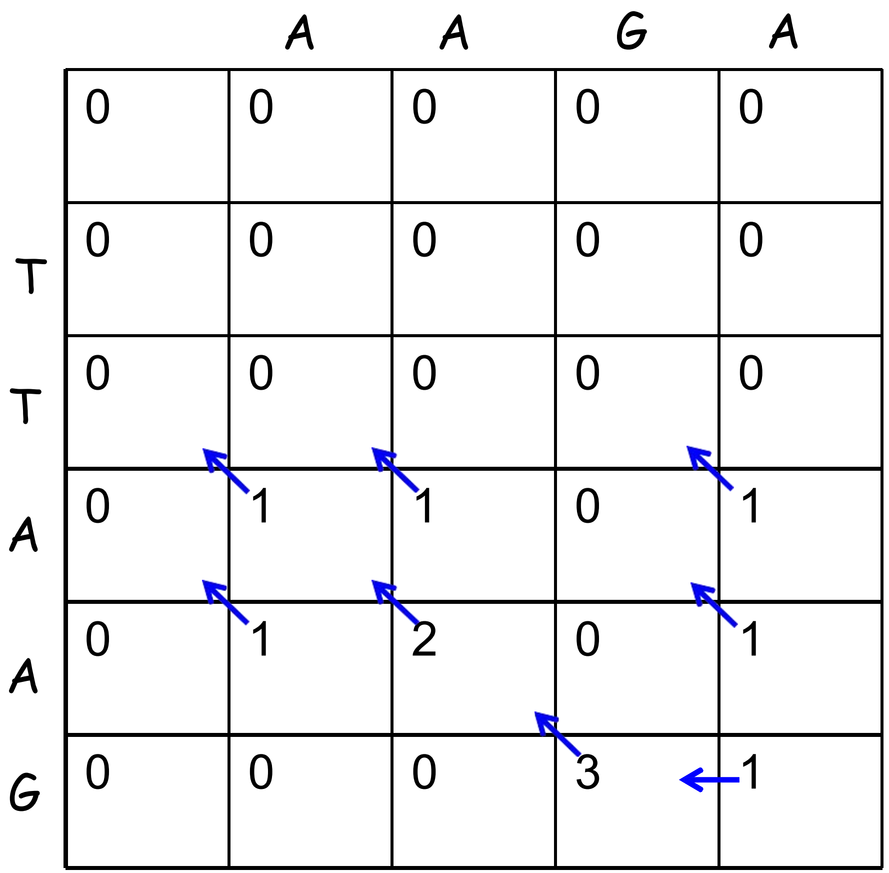
Filling....

Example

$$s(x_i, y_j) = \begin{cases} 1, & \textbf{if } x_i = y_j \\ -1, & \textbf{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

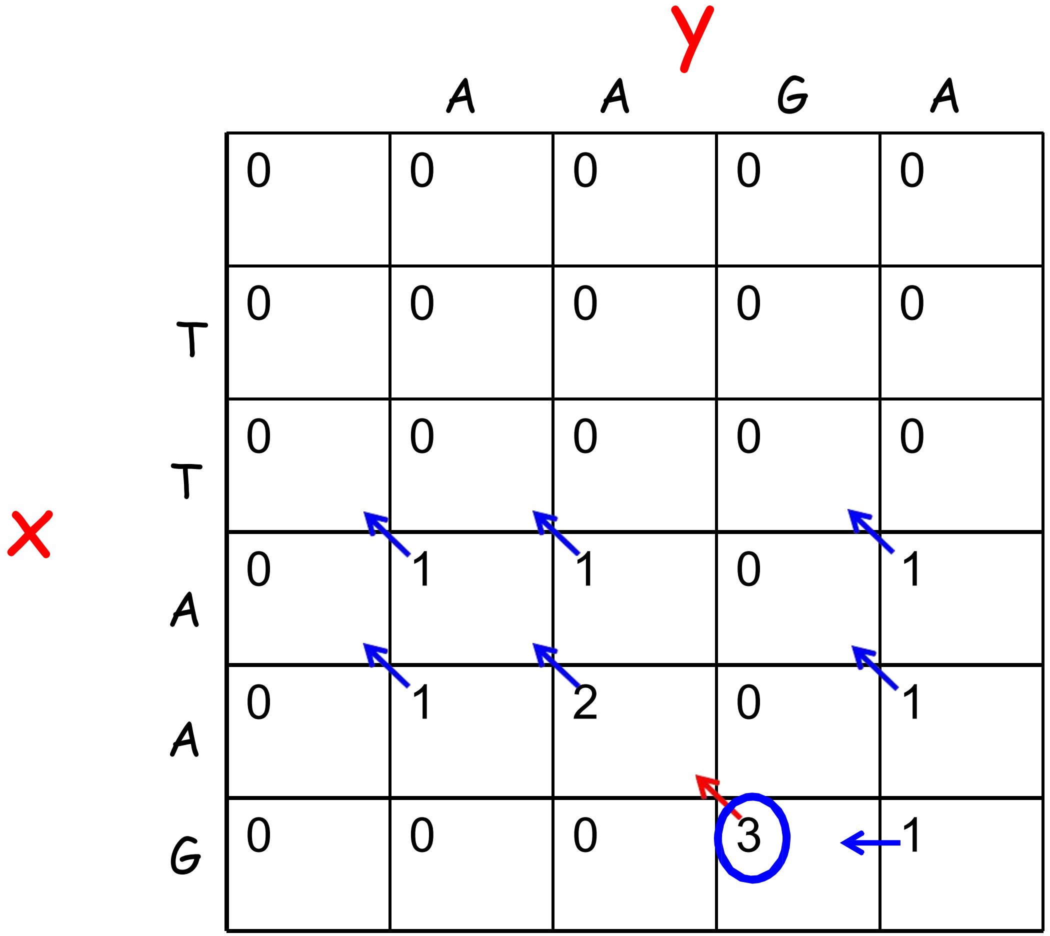


		A	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
T	0	0	0	0	0
A	0	1	1	0	1
A	0	1	2	0	1
G	0	0	0	3	1

Filling done

Example

x:G
y:G



		A	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
T	0	0	0	0	0
A	0	↖1	↖1	0	↖1
A	0	↖1	↖2	0	↖1
G	0	0	0	↖3	←1

Traceback…

Example

x : AG

y : AG

y

x

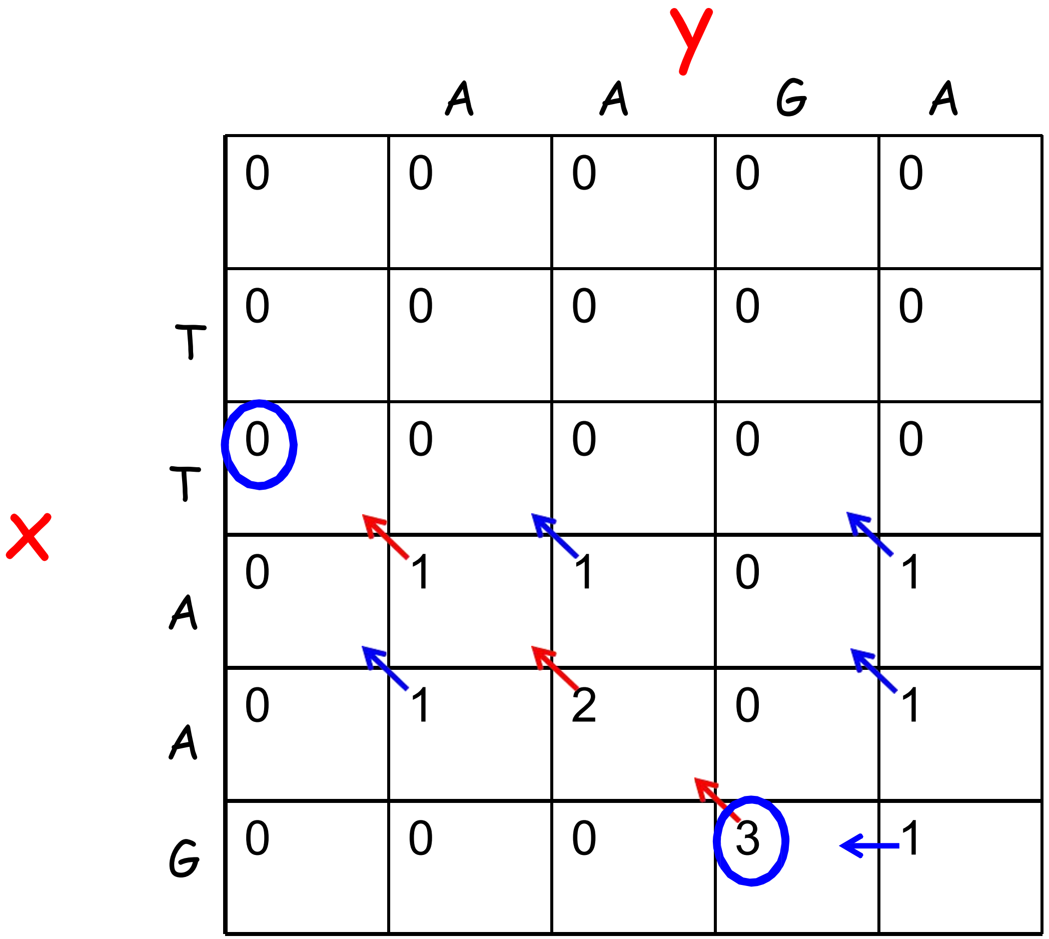
		A	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
T	0	0	0	0	0
A	0	↖1	↖1	0	↖1
A	0	↖1	↖2	0	↖1
G	0	0	0	↖3	←1

Traceback…

Example

x : AAG

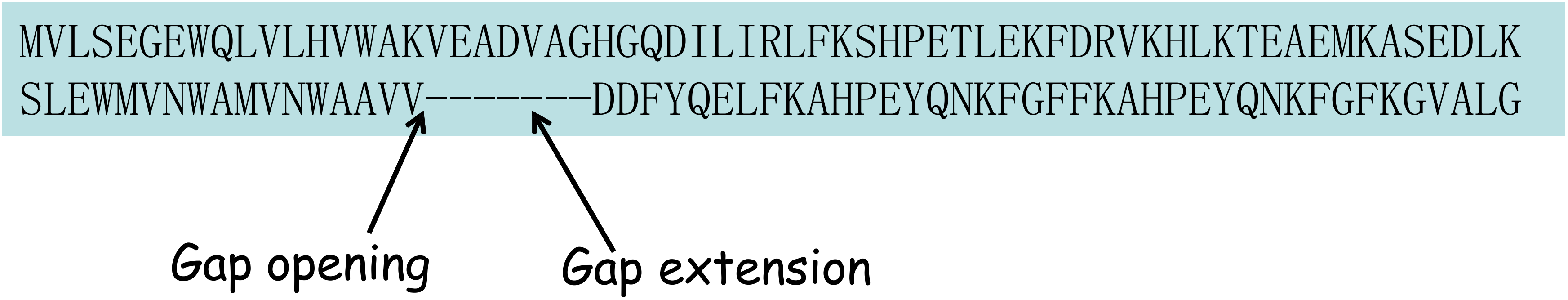
y : AAG



		A	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
T	0	0	0	0	0
A	0	↖1	↖1	0	↖1
A	0	↖1	↖2	0	↖1
G	0	0	0	↖3	←1

Traceback done

Affine Gap penalty



```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLK
SLEWMVNWAMVNWAAVV-------DDFYQELFKAHPEYQNKFGFFKAHPEYQNKFGFKGVALG
```

Gap opening

Gap extension

- Gap penalty: $w(k) = a + b \times (k-1) \quad (k \geq 1;\ a, b < 0)$
  - k: length of continous gaps
  - a: gap opening penalty
  - b: gap entension penalty

- Linear gap penalty if a=b
- Affine gap penalty if a!=b

DP for affine gap penalty case

- O. Gotoh. An improved algorithm for matching biological sequences. Journal of Molecular Biology 162 705-708 1982.

Time complexity O(mn)

带仿射罚分的全局比对

NW-DP with affine gap penalty

- Need 3 matrices instead of 1

$M(i,j)$ best score given that $x[i]$ is aligned to $y[j]$

$I_x(i,j)$ best score given that $x[i]$ is aligned to a *gap*

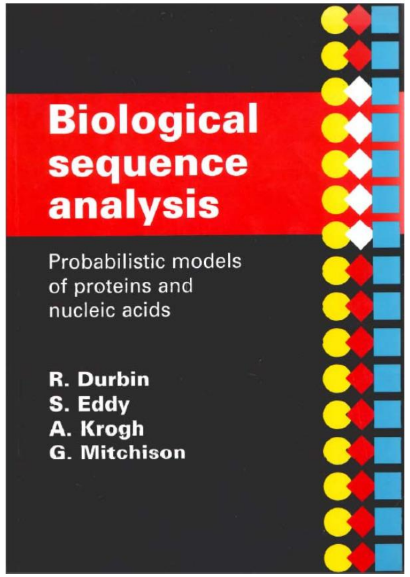
$I_y(i,j)$ best score given that $y[j]$ is aligned to a *gap*

NW-DP with affine gap penalty

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

$$I_x(i,j)=\max\begin{cases}M(i-1,j)+a\\ I_x(i-1,j)+b\end{cases}$$

$$I_y(i,j)=\max\begin{cases}M(i,j-1)+a\\ I_y(i,j-1)+b\end{cases}$$



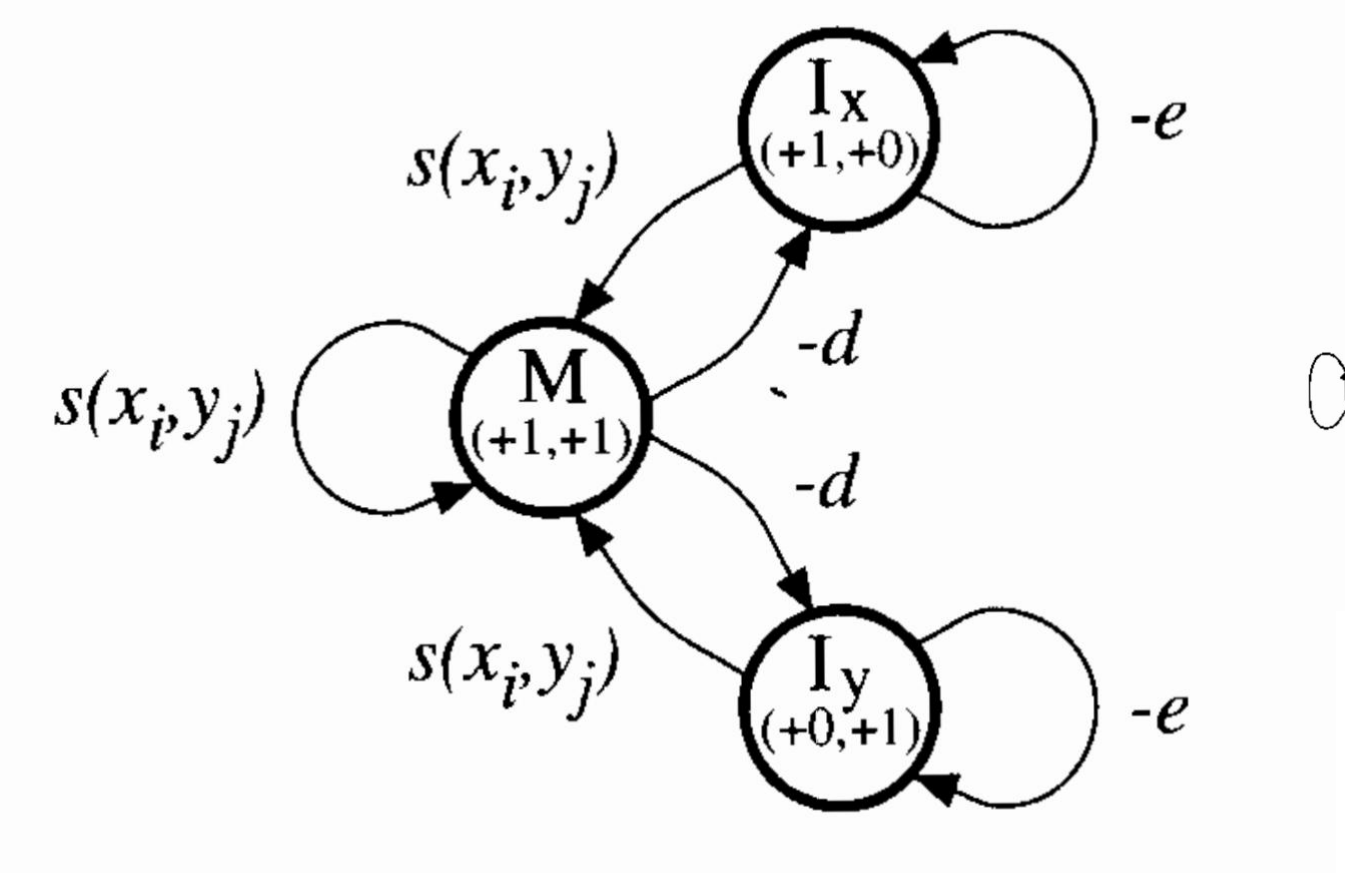
Biological sequence analysis

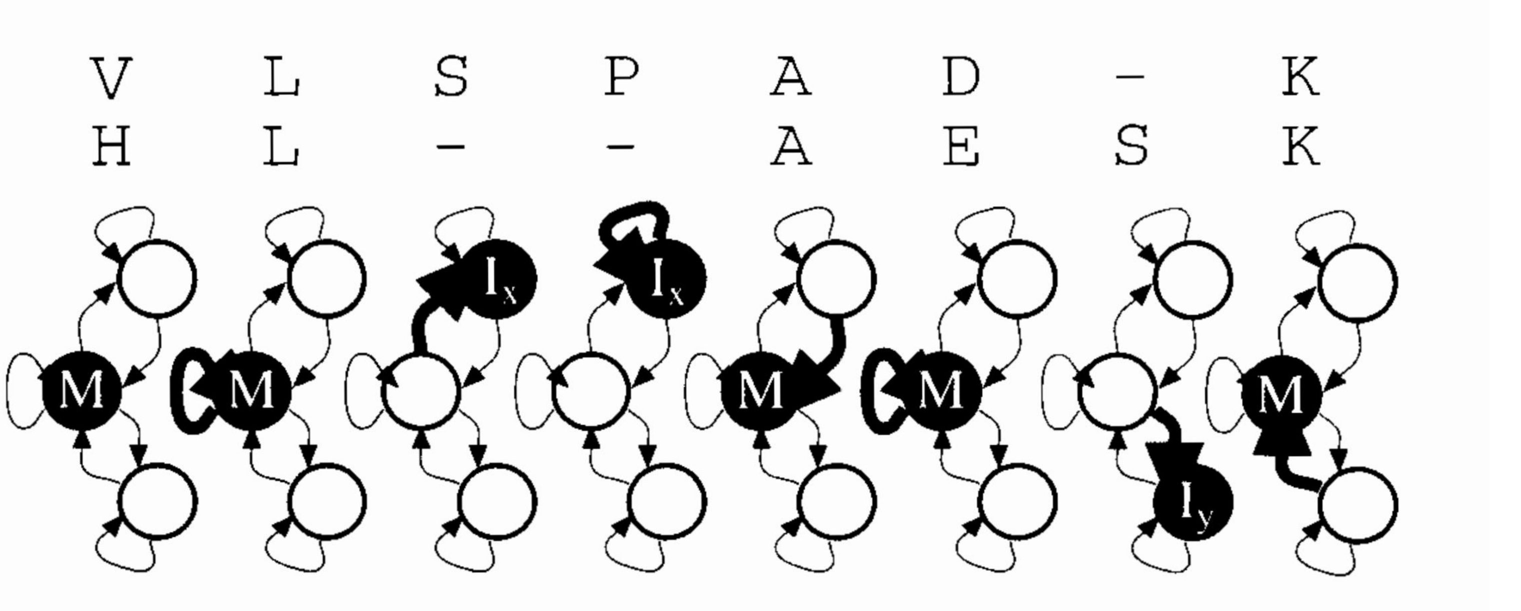
Probabilistic models of proteins and nucleic acids

R. Durbin
S. Eddy
A. Krogh
G. Mitchison

P29

NW-DP with affine gap penalty





每一个矩阵值表现为一个状态；
箭头表示状态间的转移。
一个比对就是一条由状态节点连接起来的路径。

NW-DP with affine gap penalty

- Initialization

$$\begin{cases} M(0,0) = 0; \\ M(i,0) = -\infty, M(0,j) = -\infty \quad (i, j \neq 0) \end{cases}$$

$$\begin{cases} I_x(i,0) = a + b \times i, \quad (0 \leq i \leq m) \\ I_x(0,j) = -\infty, \quad (0 < j \leq n) \end{cases}$$

$$\begin{cases} I_y(0,j) = a + b \times j, \quad (0 \leq j \leq n) \\ I_y(i,0) = -\infty, \quad (0 < i \leq m) \end{cases}$$

NW-DP with affine gap penalty

- Traceback
  - ➢ Start at the largest of $M(m,n)$, $I_x(m,n)$, $I_y(m,n)$
  - ➢ Stop at any of $M(0,0)$, $I_x(0,0)$, $I_y(0,0)$

Example

$$s(x_i, y_j) = \begin{cases} 1, & \textbf{if } x_i = y_j \\ -1, & \textbf{otherwise} \end{cases}$$

a=-3,b=-1

		A	C	A	C	T
A						
A						
T						

Example

$$s(x_i, y_j) = \begin{cases} 1, & \textbf{if } x_i = y_j \\ -1, & \textbf{otherwise} \end{cases}$$

a=-3,b=-1

Filling…

$$\begin{cases} M(0,0) = 0; \\ M(i,0) = -\infty, M(0,j) = -\infty \quad (i, j \neq 0) \end{cases}$$

$$\begin{cases} I_x(i,0) = a + b \times i, \quad (0 \le i \le m) \\ I_x(0,j) = -\infty, \quad (0 < j \le n) \end{cases}$$

$$\begin{cases} I_y(0,j) = a + b \times j, \quad (0 \le j \le n) \\ I_y(i,0) = -\infty, \quad (0 < i \le m) \end{cases}$$

		A	C	A	C	T
	M:0	-∞	-∞	-∞	-∞	-∞
	I_x:-3	-∞	-∞	-∞	-∞	-∞
	I_y:-3 ←	-4 ←	-5 ←	-6 ←	-7 ←	-8
	-∞					
A	-4					
	-∞					
	-∞					
A	-5					
	-∞					
	-∞					
T	-6					
	-∞					

Example

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

a=-3,b=-1

Filling…

$$\begin{cases} M(0,0) = 0; \\ M(i,0) = -\infty, M(0,j) = -\infty \quad (i, j \neq 0) \end{cases}$$

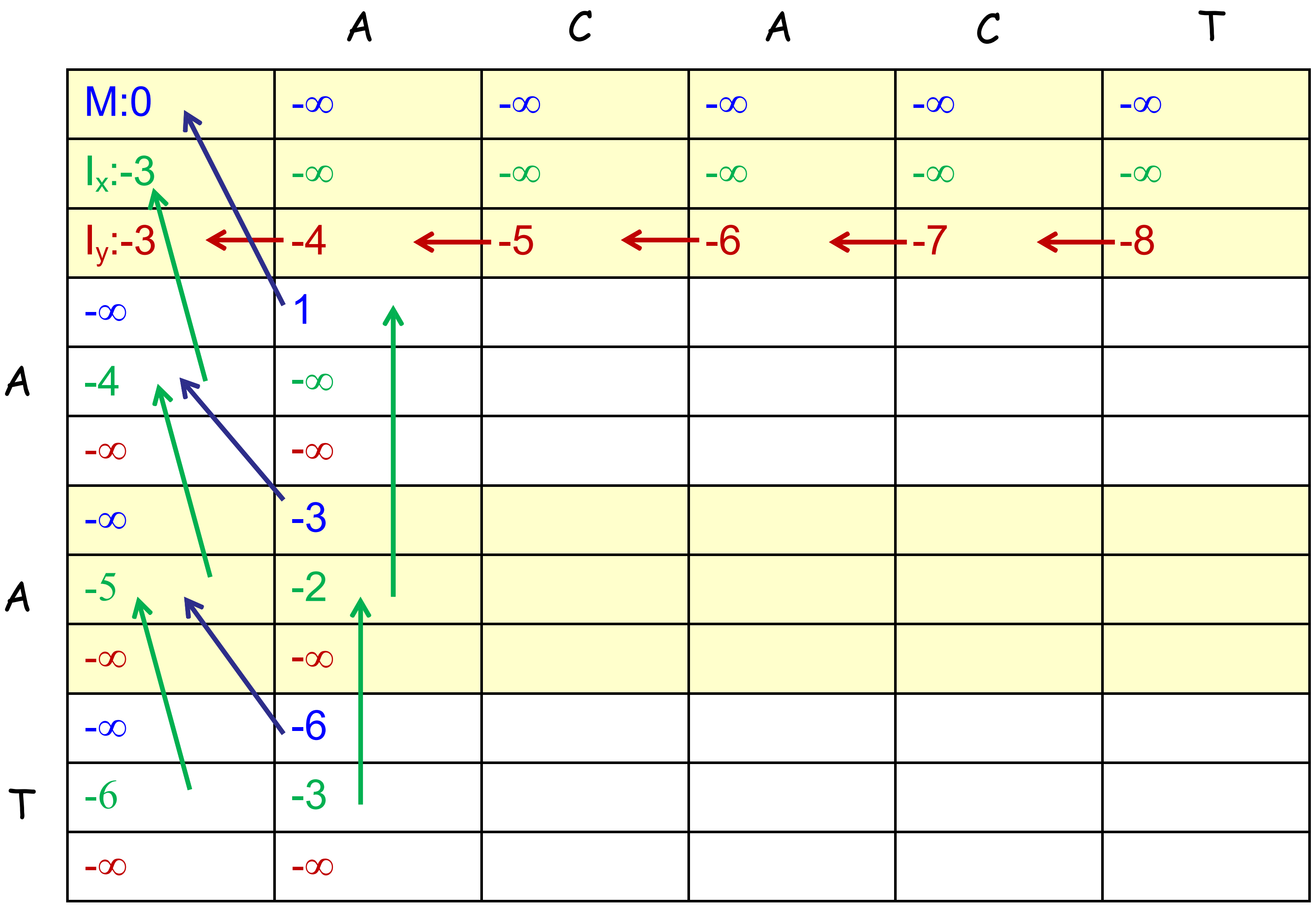
$$\begin{cases} I_x(i,0) = a + b \times i, & (0 \le i \le m) \\ I_x(0,j) = -\infty, & (0 < j \le n) \end{cases}$$

$$\begin{cases} I_y(0,j) = a + b \times j, & (0 \le j \le n) \\ I_y(i,0) = -\infty, & (0 < i \le m) \end{cases}$$

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

$$I_x(i,j) = \max \begin{cases} M(i-1, j) + a \\ I_x(i-1, j) + b \end{cases}$$

$$I_y(i,j) = \max \begin{cases} M(i, j-1) + a \\ I_y(i, j-1) + b \end{cases}$$



		A	C	A	C	T
	M:0	-∞	-∞	-∞	-∞	-∞
	Ix:-3	-∞	-∞	-∞	-∞	-∞
	Iy:-3	-4	-5	-6	-7	-8
	-∞	1				
A	-4	-∞				
	-∞	-∞				
	-∞	-3				
A	-5	-2				
	-∞	-∞				
	-∞	-6				
T	-6	-3				
	-∞	-∞				

Example

$$s(x_i, y_j) = \begin{cases} 1, & \textbf{if } x_i = y_j \\ -1, & \textbf{otherwise} \end{cases}$$

a=-3,b=-1

Filling…

$$\begin{cases} M(0,0) = 0; \\ M(i,0) = -\infty, M(0,j) = -\infty \quad (i, j \neq 0) \end{cases}$$

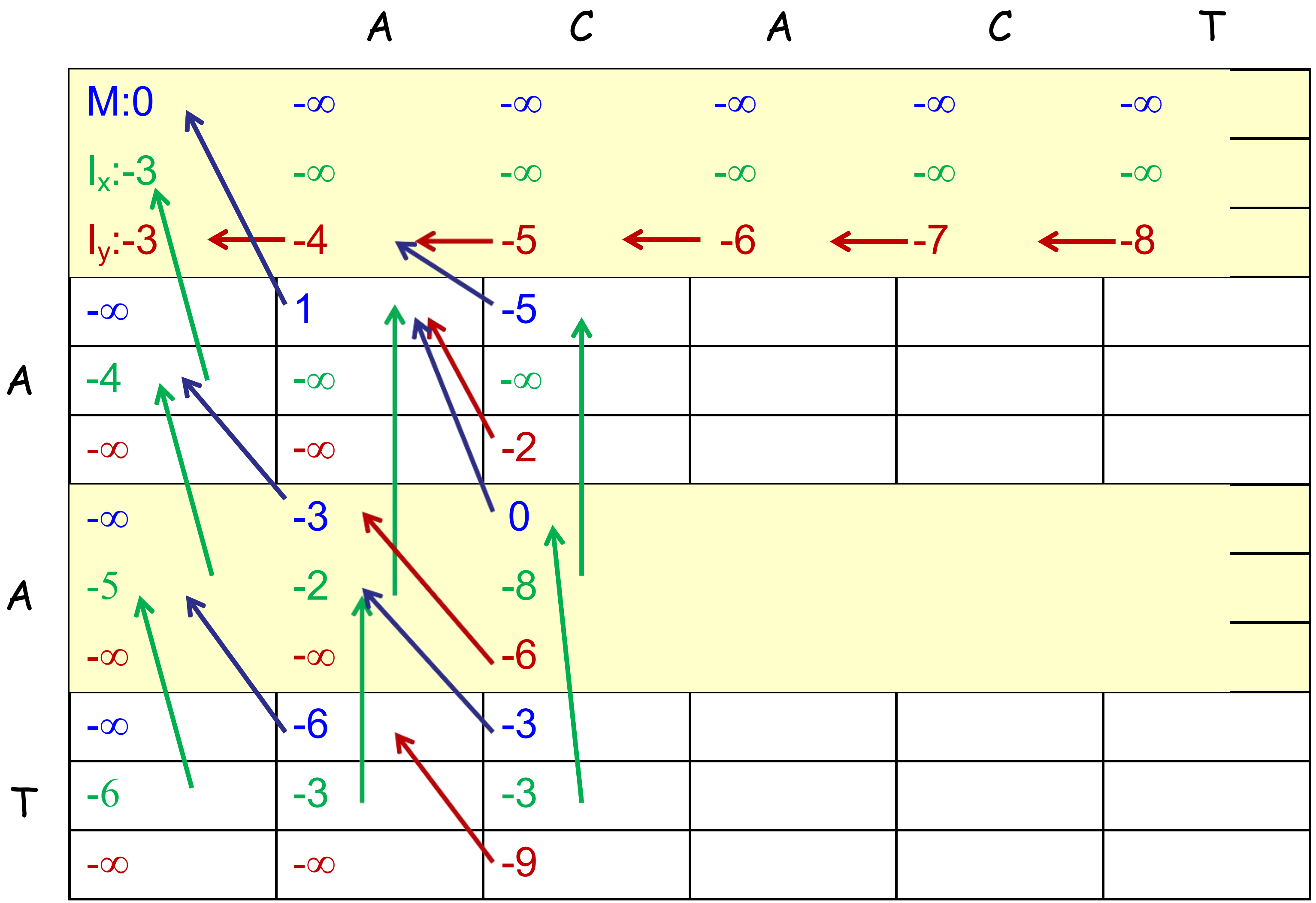
$$\begin{cases} I_x(i,0) = a + b \times i, \quad (0 \leq i \leq m) \\ I_x(0,j) = -\infty, \quad (0 < j \leq n) \end{cases}$$

$$\begin{cases} I_y(0,j) = a + b \times j, \quad (0 \leq j \leq n) \\ I_y(i,0) = -\infty, \quad (0 < i \leq m) \end{cases}$$

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

$$I_x(i,j) = \max \begin{cases} M(i-1, j) + a \\ I_x(i-1, j) + b \end{cases}$$

$$I_y(i,j) = \max \begin{cases} M(i, j-1) + a \\ I_y(i, j-1) + b \end{cases}$$



		A	C	A	C	T
	M:0 Ix:-3 Iy:-3	-∞ -∞ -4	-∞ -∞ -5	-∞ -∞ -6	-∞ -∞ -7	-∞ -∞ -8
A	-∞ -4 -∞	1 -∞ -∞	-5 -∞ -2			
A	-∞ -5 -∞	-3 -2 -∞	0 -8 -6			
T	-∞ -6 -∞	-6 -3 -∞	-3 -3 -9			

Example

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

a=-3,b=-1

Filling…

$$\begin{cases} M(0,0) = 0; \\ M(i,0) = -\infty, M(0,j) = -\infty \quad (i, j \neq 0) \end{cases}$$

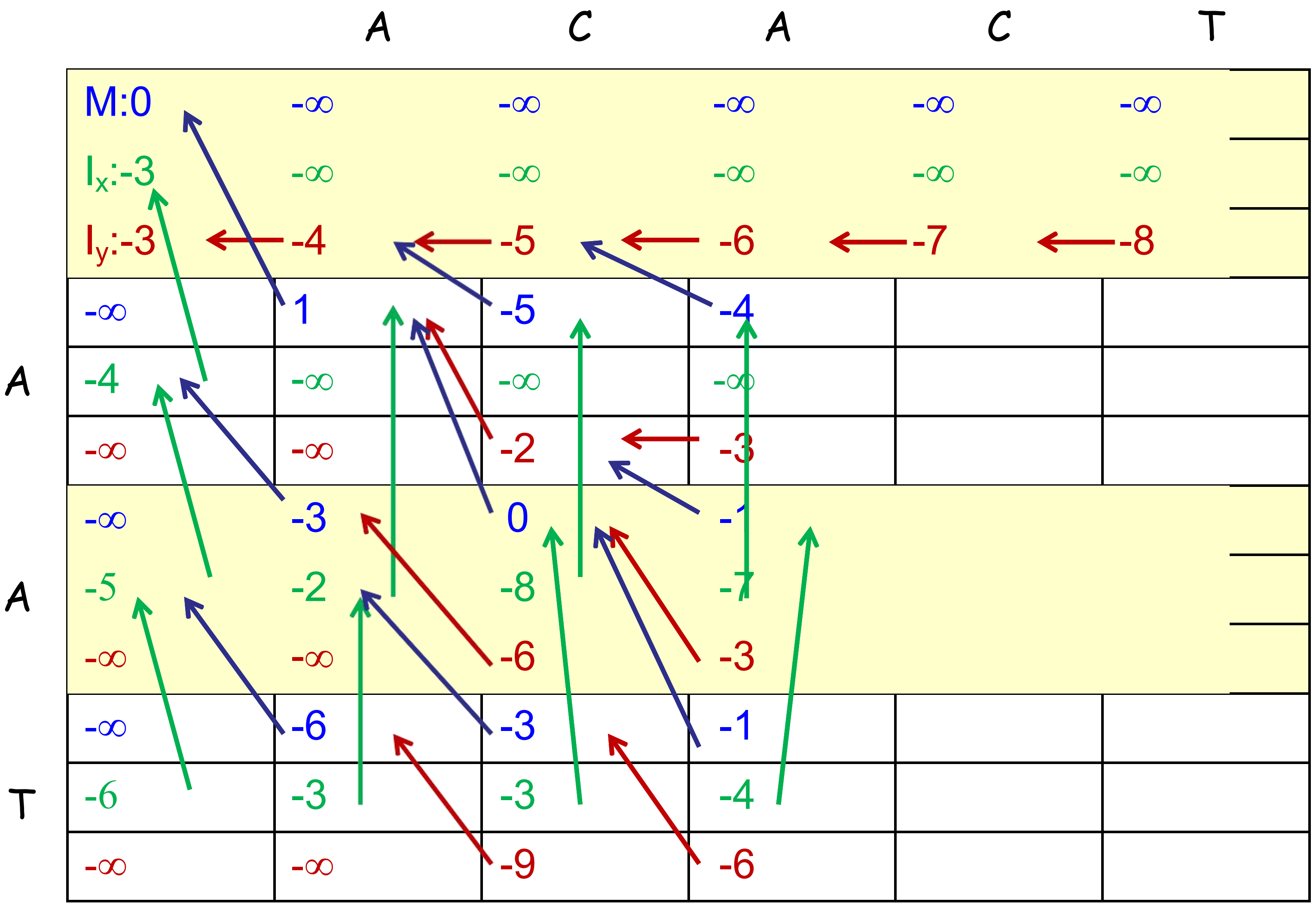
$$\begin{cases} I_x(i,0) = a + b \times i, \quad (0 \le i \le m) \\ I_x(0,j) = -\infty, \quad (0 < j \le n) \end{cases}$$

$$\begin{cases} I_y(0,j) = a + b \times j, \quad (0 \le j \le n) \\ I_y(i,0) = -\infty, \quad (0 < i \le m) \end{cases}$$

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

$$I_x(i,j) = \max \begin{cases} M(i-1, j) + a \\ I_x(i-1, j) + b \end{cases}$$

$$I_y(i,j) = \max \begin{cases} M(i, j-1) + a \\ I_y(i, j-1) + b \end{cases}$$



		A	C	A	C	T
	M:0 I_x:-3 I_y:-3	-∞ -∞ -4	-∞ -∞ -5	-∞ -∞ -6	-∞ -∞ -7	-∞ -∞ -8
A	-∞ -4 -∞	1 -∞ -∞	-5 -∞ -2	-4 -∞ -3		
A	-∞ -5 -∞	-3 -2 -∞	0 -8 -6	-1 -7 -3		
T	-∞ -6 -∞	-6 -3 -∞	-3 -3 -9	-1 -4 -6		

Example

$$s(x_i, y_j) = \begin{cases} 1, & \textbf{if } x_i = y_j \\ -1, & \textbf{otherwise} \end{cases}$$

a=-3,b=-1

Filling...

$$\begin{cases} M(0,0)=0; \\ M(i,0)=-\infty, M(0,j)=-\infty \quad (i,j\neq 0) \end{cases}$$

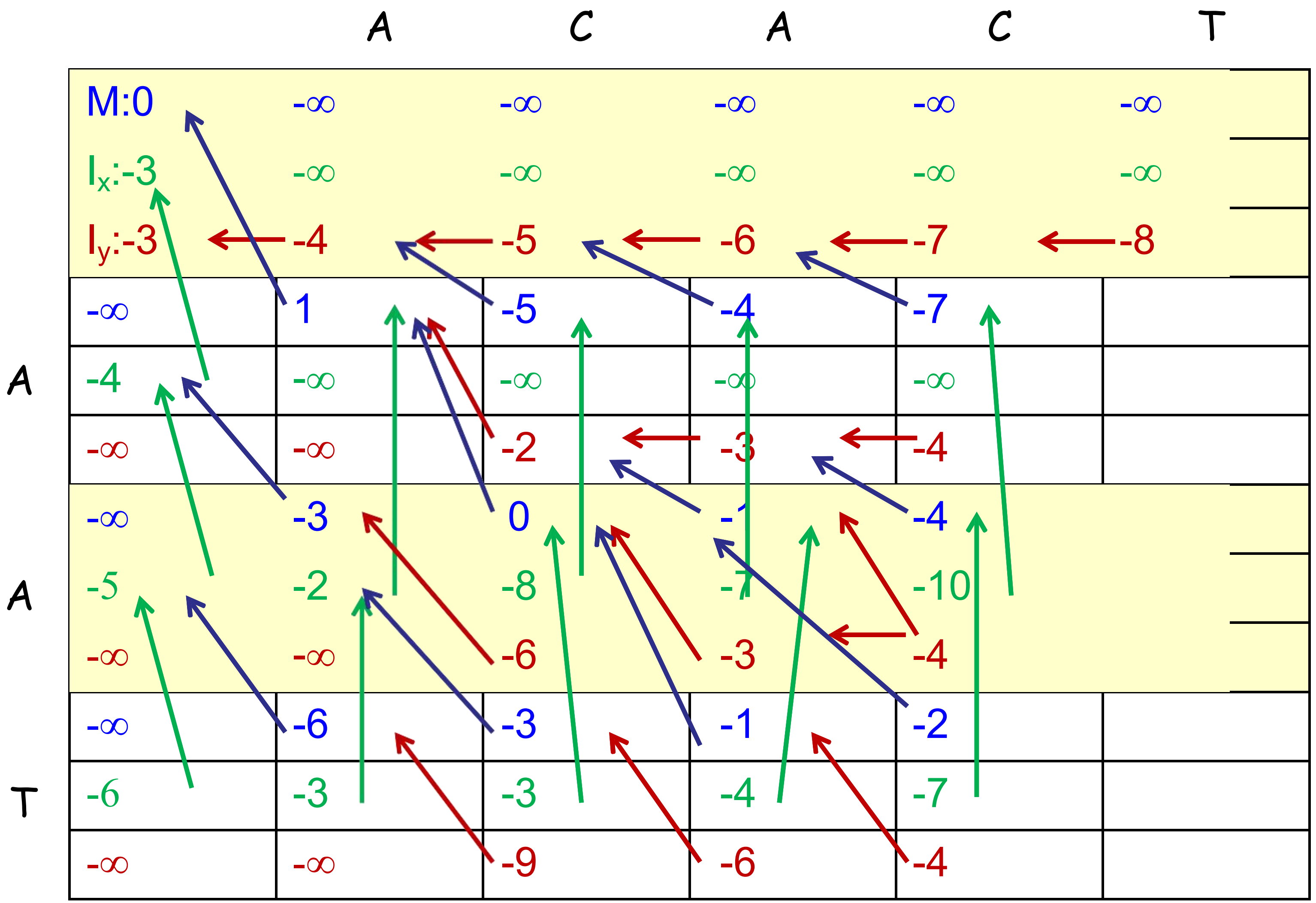
$$\begin{cases} I_x(i,0)=a+b\times i, & (0\le i\le m) \\ I_x(0,j)=-\infty, & (0<j\le n) \end{cases}$$

$$\begin{cases} I_y(0,j)=a+b\times j, & (0\le j\le n) \\ I_y(i,0)=-\infty, & (0<i\le m) \end{cases}$$

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

$$I_x(i,j)=\max\begin{cases} M(i-1,j)+a \\ I_x(i-1,j)+b \end{cases}$$

$$I_y(i,j)=\max\begin{cases} M(i,j-1)+a \\ I_y(i,j-1)+b \end{cases}$$



		A	C	A	C	T
	M:0 Ix:-3 Iy:-3	-∞ -∞ -4	-∞ -∞ -5	-∞ -∞ -6	-∞ -∞ -7	-∞ -∞ -8
A	-∞ -4 -∞	1 -∞ -∞	-5 -∞ -2	-4 -∞ -3	-7 -∞ -4	
A	-∞ -5 -∞	-3 -2 -∞	0 -8 -6	-1 -7 -3	-4 -10 -4	
T	-∞ -6 -∞	-6 -3 -∞	-3 -3 -9	-1 -4 -6	-2 -7 -4	

Example

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Filling done

a=-3,b=-1

$$\begin{cases} M(0,0) = 0; \\ M(i,0) = -\infty, M(0,j) = -\infty \quad (i, j \neq 0) \end{cases}$$

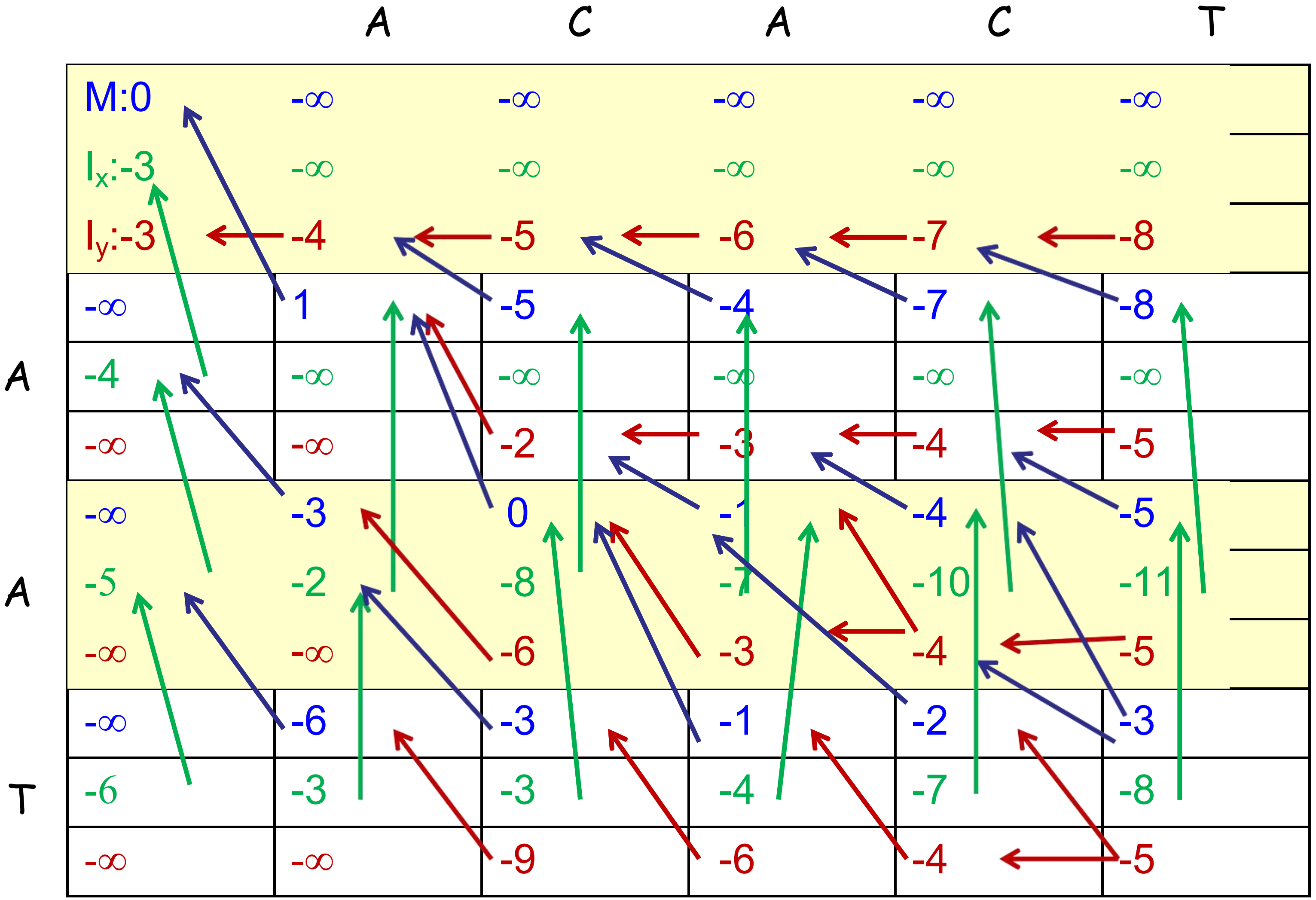
$$\begin{cases} I_x(i,0) = a + b \times i, \quad (0 \le i \le m) \\ I_x(0,j) = -\infty, \quad (0 < j \le n) \end{cases}$$

$$\begin{cases} I_y(0,j) = a + b \times j, \quad (0 \le j \le n) \\ I_y(i,0) = -\infty, \quad (0 < i \le m) \end{cases}$$

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

$$I_x(i,j) = \max \begin{cases} M(i-1, j) + a \\ I_x(i-1, j) + b \end{cases}$$

$$I_y(i,j) = \max \begin{cases} M(i, j-1) + a \\ I_y(i, j-1) + b \end{cases}$$



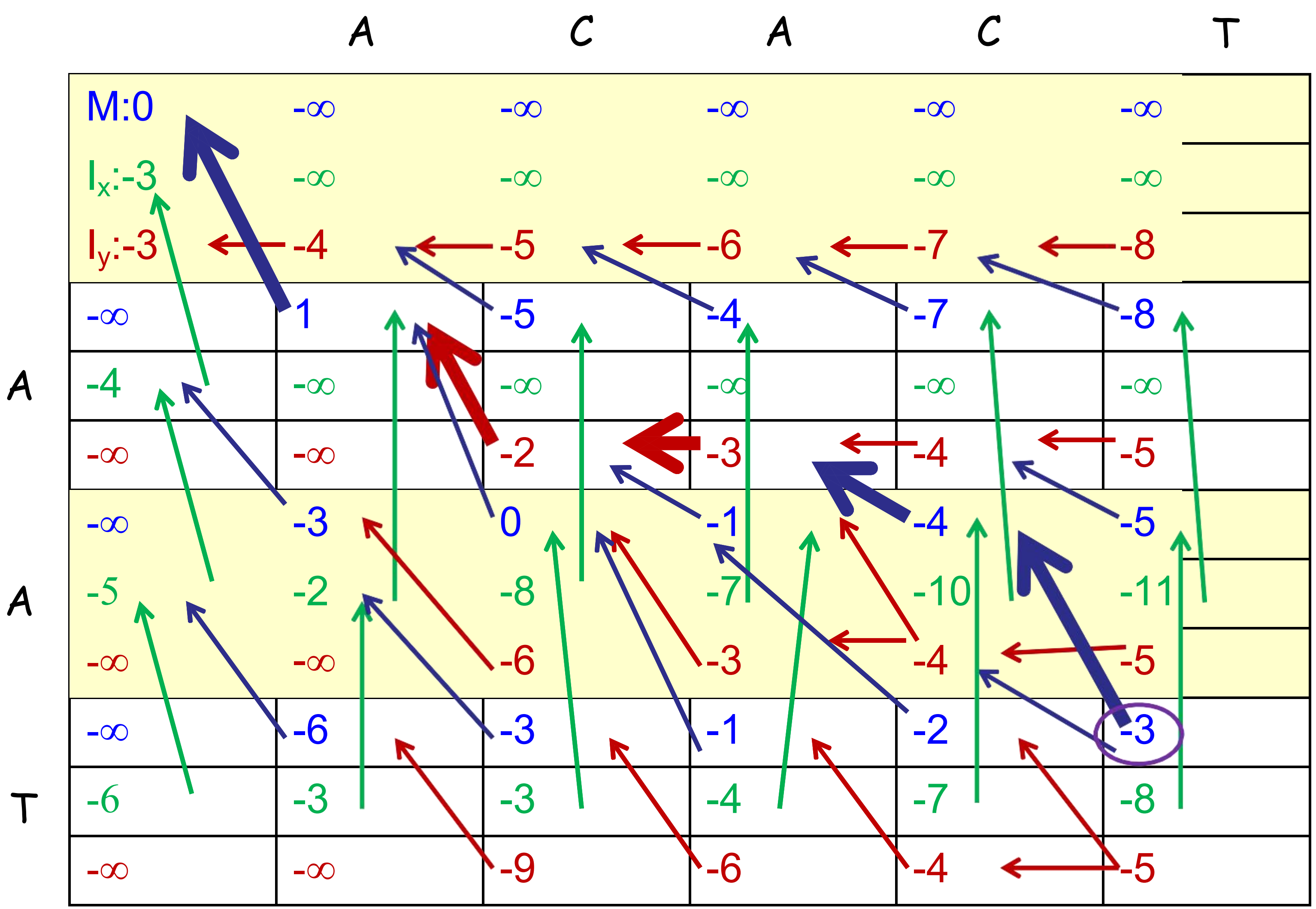
		A	C	A	C	T
	M:0	-∞	-∞	-∞	-∞	-∞
	I_x:-3	-∞	-∞	-∞	-∞	-∞
	I_y:-3	-4	-5	-6	-7	-8
	-∞	1	-5	-4	-7	-8
A	-4	-∞	-∞	-∞	-∞	-∞
	-∞	-∞	-2	-3	-4	-5
	-∞	-3	0	-1	-4	-5
A	-5	-2	-8	-7	-10	-11
	-∞	-∞	-6	-3	-4	-5
	-∞	-6	-3	-1	-2	-3
T	-6	-3	-3	-4	-7	-8
	-∞	-∞	-9	-6	-4	-5

Example

x:A--AT

y:ACACT

Traceback



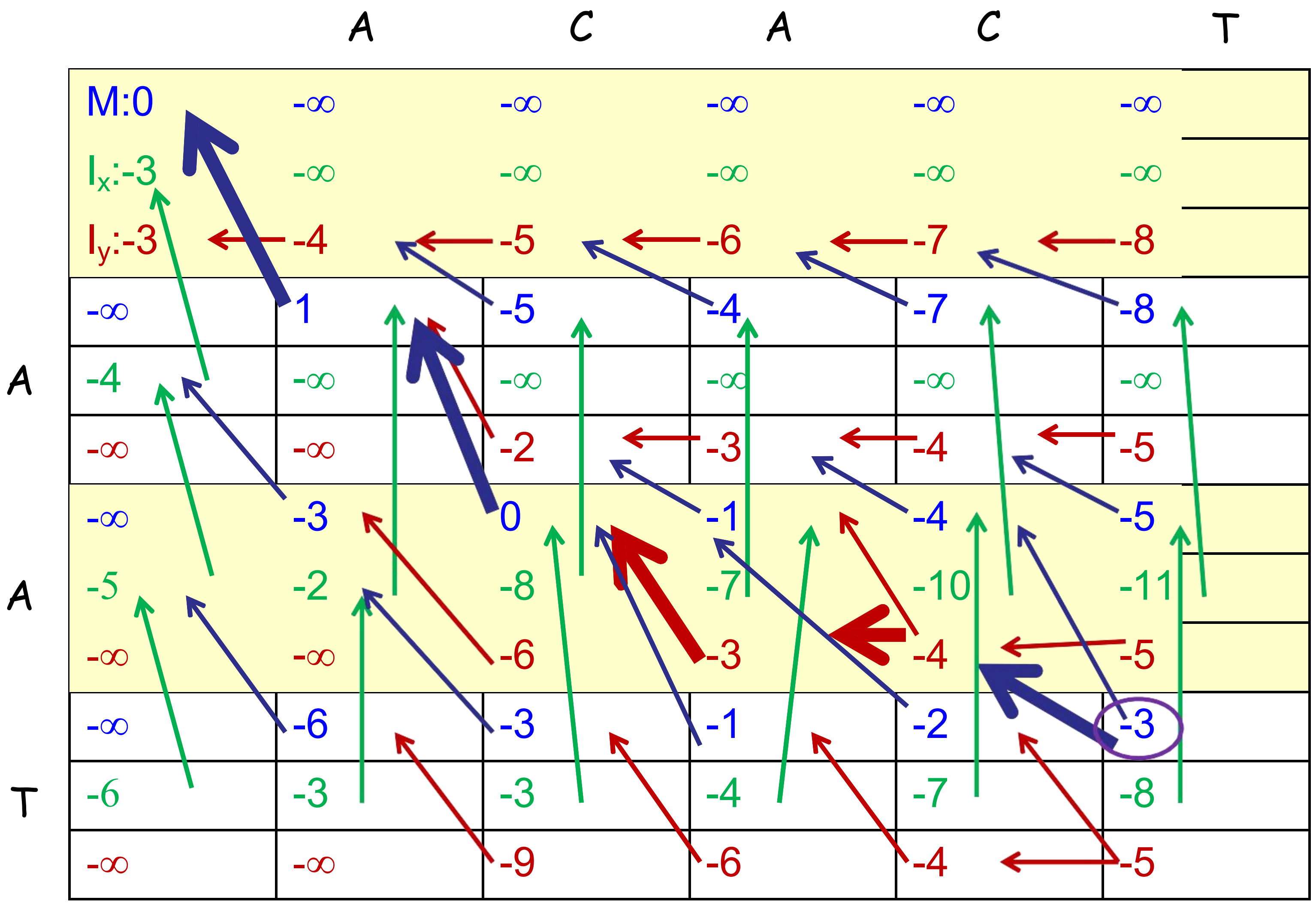
		A	C	A	C	T
	M:0 I_x:-3 I_y:-3	-∞ -∞ -4	-∞ -∞ -5	-∞ -∞ -6	-∞ -∞ -7	-∞ -∞ -8
A	-∞ -4 -∞	1 -∞ -∞	-5 -∞ -2	-4 -∞ -3	-7 -∞ -4	-8 -∞ -5
A	-∞ -5 -∞	-3 -2 -∞	0 -8 -6	-1 -7 -3	-4 -10 -4	-5 -11 -5
T	-∞ -6 -∞	-6 -3 -∞	-3 -3 -9	-1 -4 -6	-2 -7 -4	-3 -8 -5

Example

Another alignment

x:AA--T
y:ACACT

Traceback



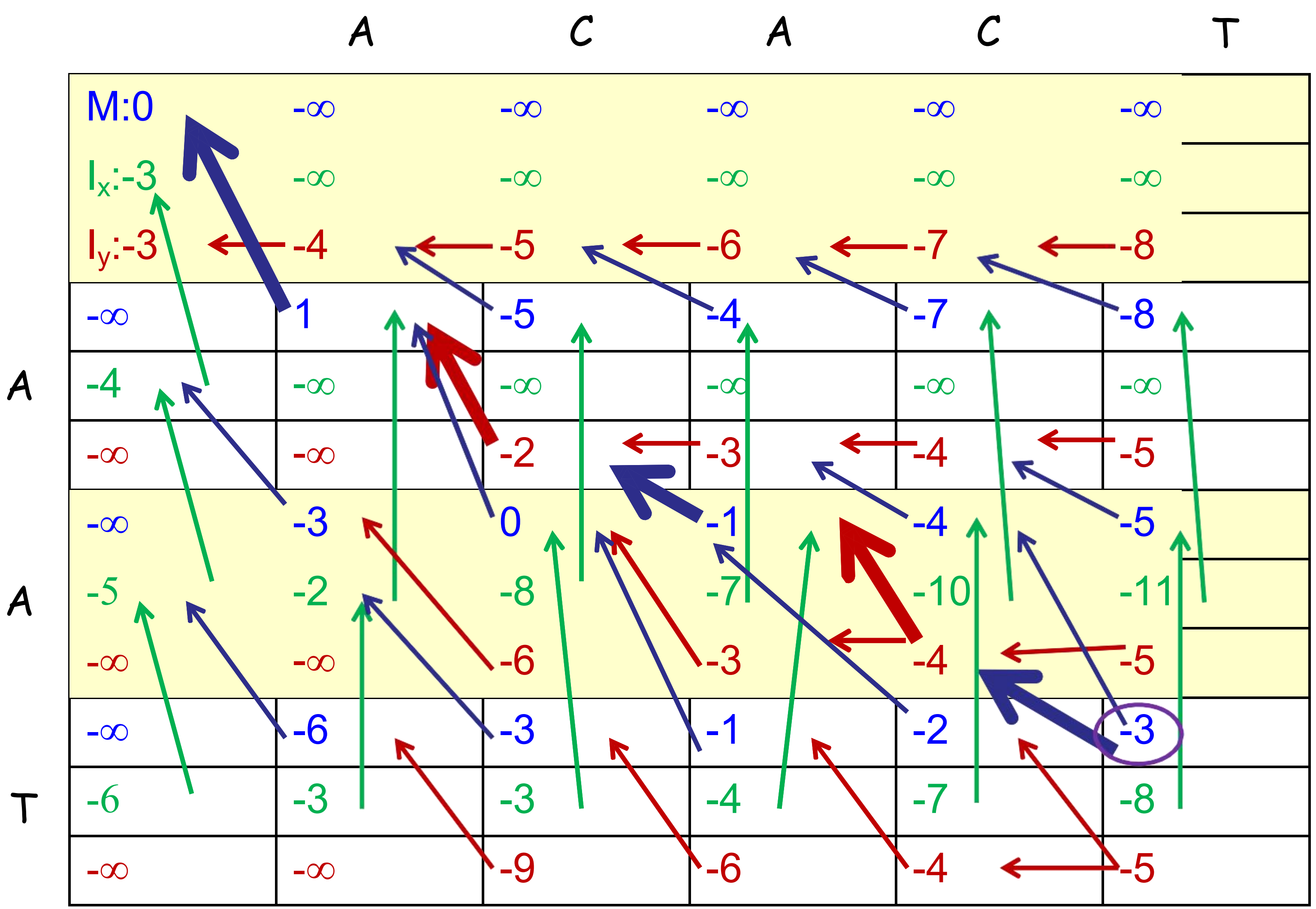
		A	C	A	C	T
	M:0 I_x:-3 I_y:-3	-∞ -∞ -4	-∞ -∞ -5	-∞ -∞ -6	-∞ -∞ -7	-∞ -∞ -8
A	-∞ -4 -∞	1 -∞ -∞	-5 -∞ -2	-4 -∞ -3	-7 -∞ -4	-8 -∞ -5
A	-∞ -5 -∞	-3 -2 -∞	0 -8 -6	-1 -7 -3	-4 -10 -4	-5 -11 -5
T	-∞ -6 -∞	-6 -3 -∞	-3 -3 -9	-1 -4 -6	-2 -7 -4	-3 -8 -5

Example

Another alignment

x:A–A–T
y:ACACT

Traceback



		A	C	A	C	T
	M:0 I_x:-3 I_y:-3	-∞ -∞ -4	-∞ -∞ -5	-∞ -∞ -6	-∞ -∞ -7	-∞ -∞ -8
A	-∞ -4 -∞	1 -∞ -∞	-5 -∞ -2	-4 -∞ -3	-7 -∞ -4	-8 -∞ -5
A	-∞ -5 -∞	-3 -2 -∞	0 -8 -6	-1 -7 -3	-4 -10 -4	-5 -11 -5
T	-∞ -6 -∞	-6 -3 -∞	-3 -3 -9	-1 -4 -6	-2 -7 -4	-3 -8 -5

Example

Another alignment

$$s(x_i, y_j) = \begin{cases} 1, & \textbf{if } x_i = y_j \\ -1, & \textbf{otherwise} \end{cases}$$

a=-3,b=-1

```
x:A--AT
y:ACACT
```

```
x:AA--T
y:ACACT
```

```
x:A-A-T
y:ACACT
```

带仿射罚分的局部比对

SW-DP with affine gap penalty

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

$$I_x(i,j)=\max\begin{cases}M(i-1,j)+a\\ I_x(i-1,j)+b\end{cases}$$

$$I_y(i,j)=\max\begin{cases}M(i,j-1)+a\\ I_y(i,j-1)+b\end{cases}$$

SW-DP with affine gap penalty

- Initialization

$$M(i,0) = M(0,j) = 0$$

$$I_x(i,0) = I_x(0,j) = -\infty$$

$$I_y(i,0) = I_y(0,j) = -\infty$$

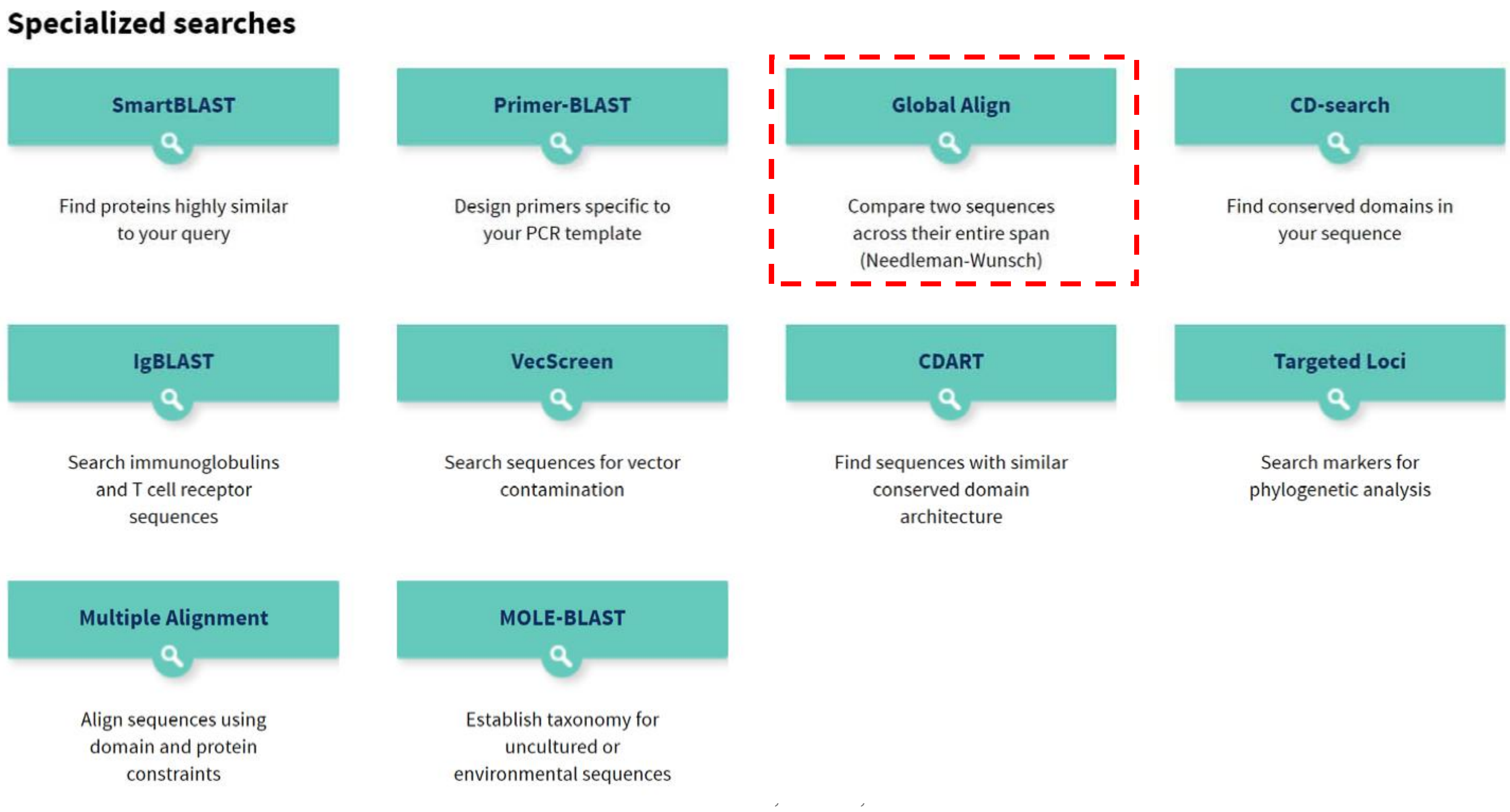
SW-DP with affine gap penalty

- Traceback
    - ➢ Start at the largest of M(i,j)
    - ➢ Stop at M(i,j)=0

软件、服务器操作

全局比对：Needleman-Wunsch算法

- ❒ https://blast.ncbi.nlm.nih.gov/Blast.cgi



Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

IgBLAST
Search immunoglobulins and T cell receptor sequences

VecScreen
Search sequences for vector contamination

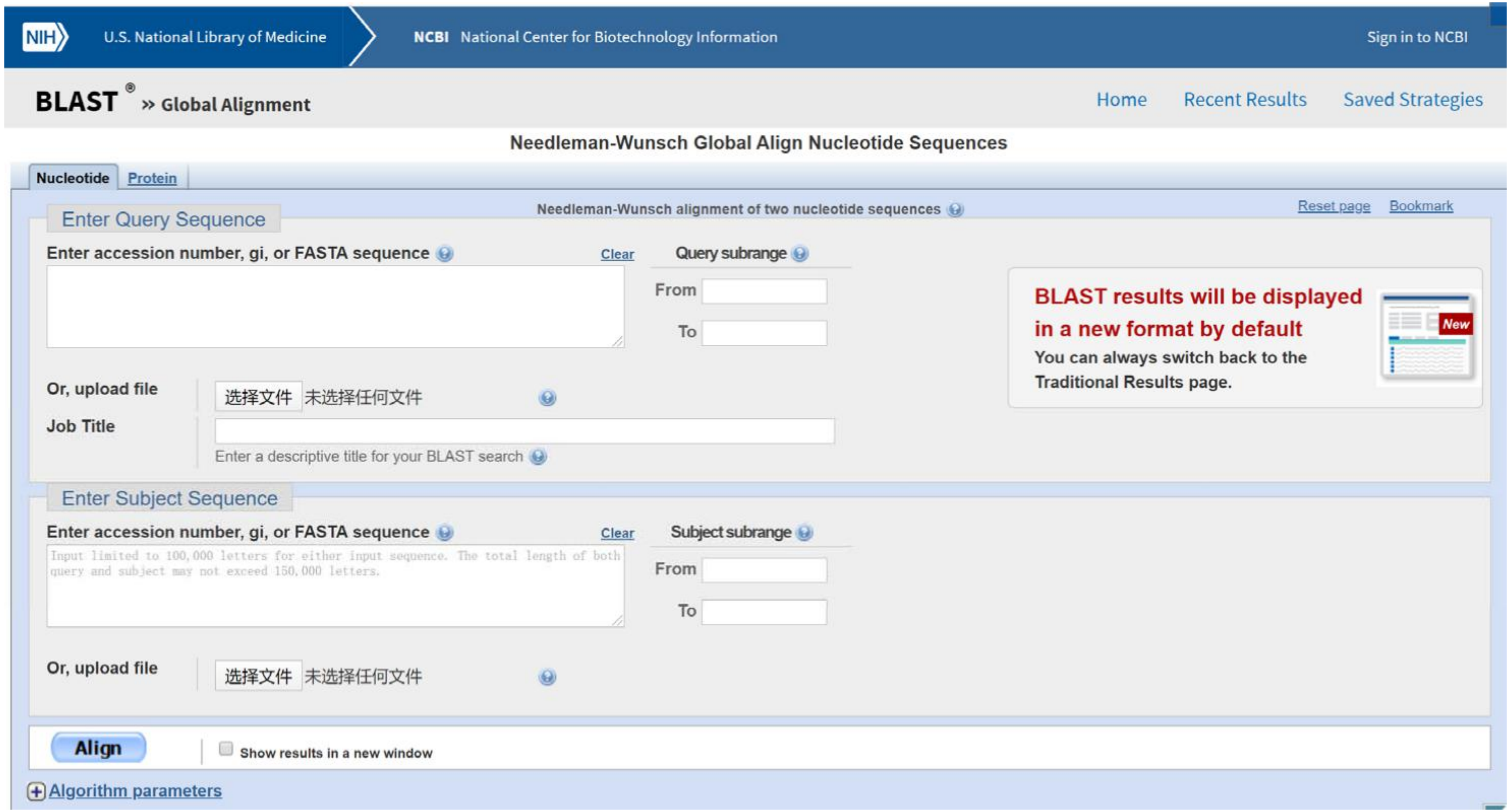
CDART
Find sequences with similar conserved domain architecture

Targeted Loci
Search markers for phylogenetic analysis

Multiple Alignment
Align sequences using domain and protein constraints

MOLE-BLAST
Establish taxonomy for uncultured or environmental sequences

Global Alignment



Needleman-Wunsch Global Align Nucleotide Sequences

Needleman-Wunsch alignment of two nucleotide sequences

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Query subrange

From

To

Or, upload file 未选择任何文件

Job Title

Enter a descriptive title for your BLAST search

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence

Input limited to 100,000 letters for either input sequence. The total length of both query and subject may not exceed 150,000 letters.

Subject subrange

From

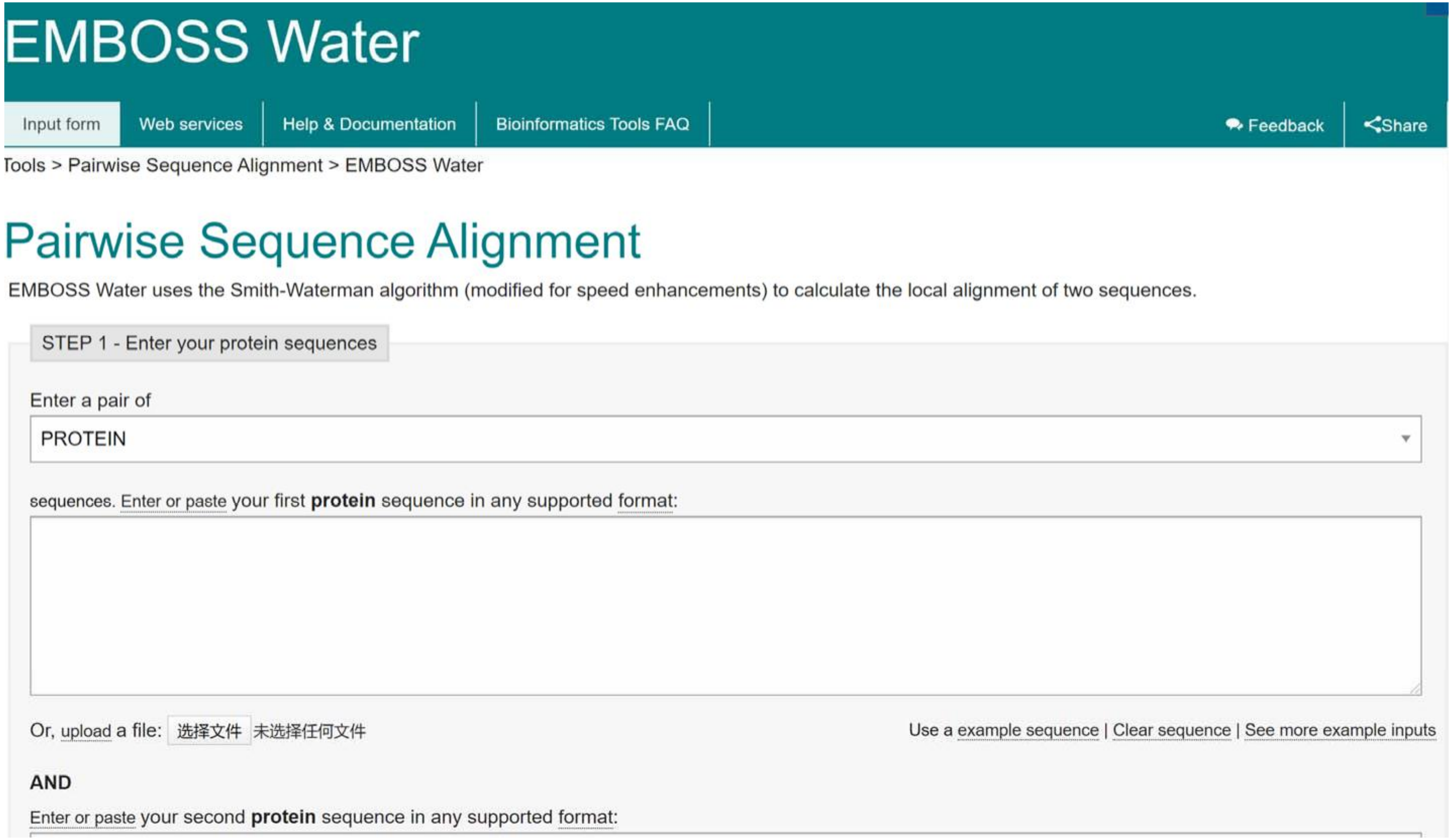
To

Or, upload file 未选择任何文件

□ Show results in a new window

局部比对：Smith-Waterman算法

- https://www.ebi.ac.uk/Tools/psa/emboss_water/



EMBOSS Water

Tools > Pairwise Sequence Alignment > EMBOSS Water

Pairwise Sequence Alignment

EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

STEP 1 - Enter your protein sequences

Enter a pair of

PROTEIN

sequences. Enter or paste your first **protein** sequence in any supported format:

Or, upload a file: 选择文件 未选择任何文件

Use a example sequence | Clear sequence | See more example inputs

AND

Enter or paste your second **protein** sequence in any supported format:

作业

Homework 2

1.Please find the best global alignment of following two sequences

- AGTTGC
- CAGA

Score matrix

	A	T	G	C
A	2	1	-1	-1
T	1	2	-1	-1
G	-1	-1	2	1
C	-1	-1	1	2

Gap penalty: open=extension=-2

2.Please find the best global alignment of following two sequences

- AGTTGC
- CAGA

Score matrix

	A	T	G	C
A	2	1	-1	-1
T	1	2	-1	-1
G	-1	-1	2	1
C	-1	-1	1	2

Gap penalty: open=-2, extension=-1

Your answers to 1,2 should include the following

1. Alignment matrix
2. Trace back path
3. Alignment result

in a similar format as the examples given in the class.

编程题目：

3*. Write a program to align any two protein sequences with BLOSUM62 matrix, available at:
http://yanglab.nankai.edu.cn/teaching/bioinformatics/BLOSUM62.txt

Gap opening=-11 Gap extension=-1

To examine whether your program is correct, you can compare your result with the program at
http://zhanglab.ccmb.med.umich.edu/NW-align/