Βιοπληροφορική

Ενότητα 7: Σύγκριση αλληλουχιών – Part II

Αν. καθηγητής Αγγελίδης Παντελής
e-mail: paggelidis@uowm.gr
ΕΕΔΙΠ Μπέλλου Σοφία
e-mail: sbellou@uowm.gr

Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών

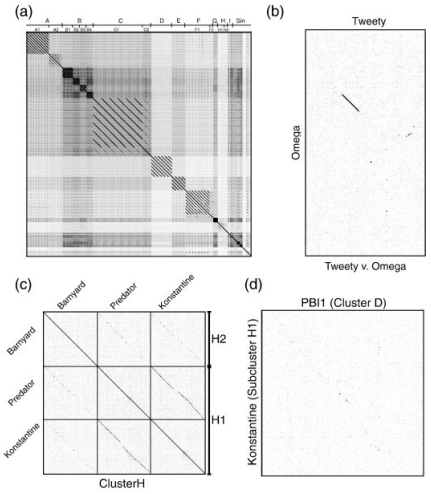
Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ψηφιακά Μαθήματα στο Πανεπιστήμιο Δυτικής Μακεδονίας**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σύγκριση αλληλουχιών – Part II



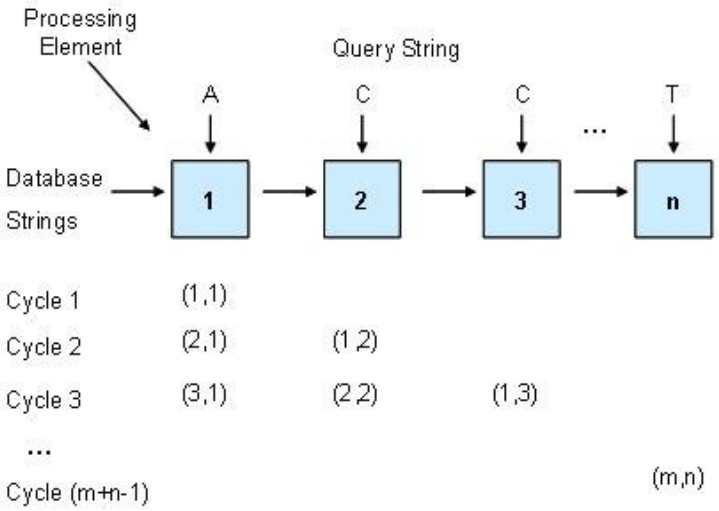
		A	T	T	C	G	T	A	C	T	T	A	G	T
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13
C	-1	-1	-2	-3	-2	-4	-6	-7	-8	-9	-10	-11	-12	-13
T	-2	-2	0	0	-2	-3	-3	-5	-7	-7	-7	-9	-11	-11
T	-3	-3	0	1	-1	-3	-2	-4	-6	-6	-6	-8	-10	-10
A	-4	-2	-2	-1	0	-2	-4	-1	-3	-5	-7	-5	-7	-9
G	-5	-4	-3	-3	-2	1	-1	-3	-2	-4	-6	-8	-4	-6
C	-6	-6	-5	-4	-2	-1	0	-2	-2	-3	-5	-7	-6	-5
T	-7	-7	-5	-4	-4	-3	0	-1	-3	-1	-1	-3	-5	-5
A	-8	-6	-7	-6	-5	-5	-2	1	-1	-3	-2	0	-2	-4
A	-9	-6	-7	-8	-7	-6	-4	1	0	-2	-4	0	-1	-3
T	-10	-8	-5	-5	-7	-8	-4	-1	0	1	1	-1	-1	0
C	-11	-10	-7	-6	-4	-6	-6	-3	0	-1	0	0	-2	-2
A	-12	-10	-9	-8	-6	-5	-7	-3	-2	-1	-2	1	-1	-3
G	-13	-12	-11	-10	-8	-5	-6	-5	-4	-3	-2	-1	2	0

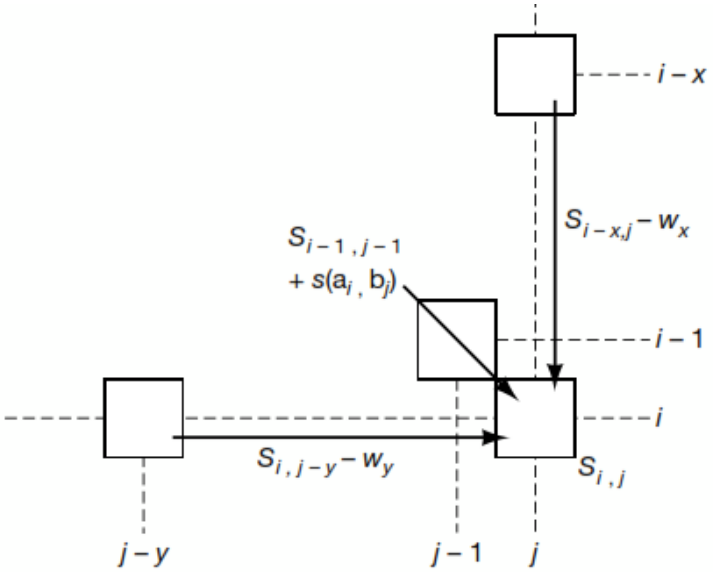
So the best alignment would be:

-- = gap
| = match

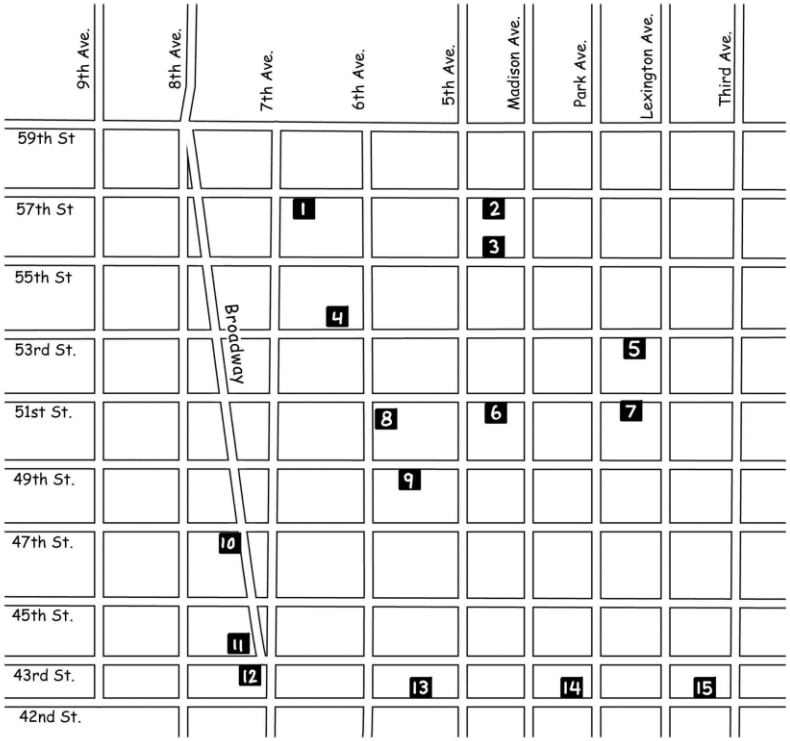
ATTCG--TACTTAGT
CTTAGCTAATCAG--

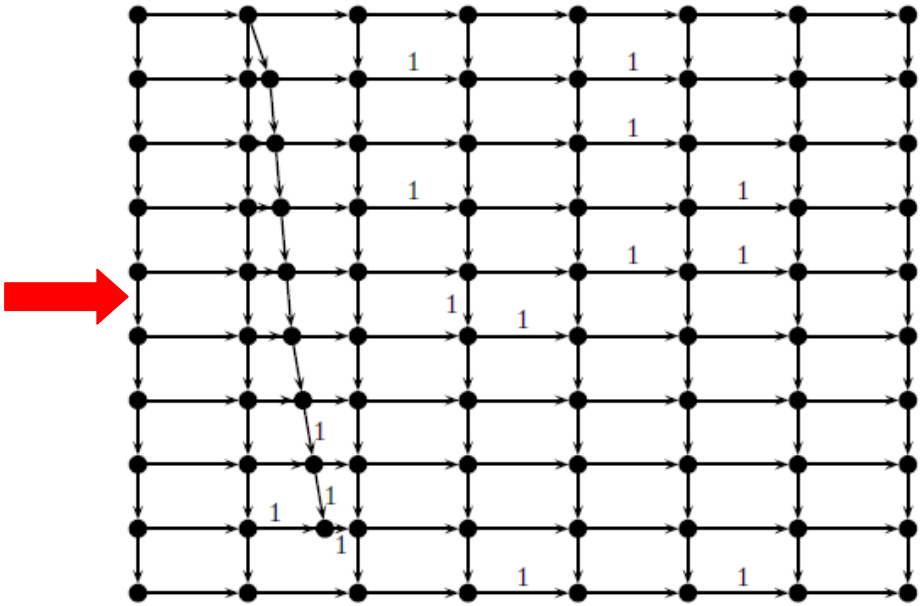
Αλγόριθμοι δυναμικού προγραμματισμού





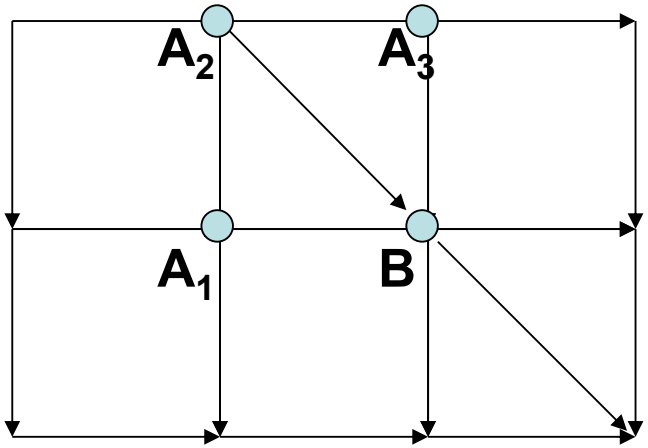
Κατευθυνόμενα ακυκλικά γραφήματα Directed acyclic graphs, DAG





Γράφημα G = (V,E), όπου V: κορυφές και E: ακμές του γραφήματος

Manhattan Is Not A Perfect Grid

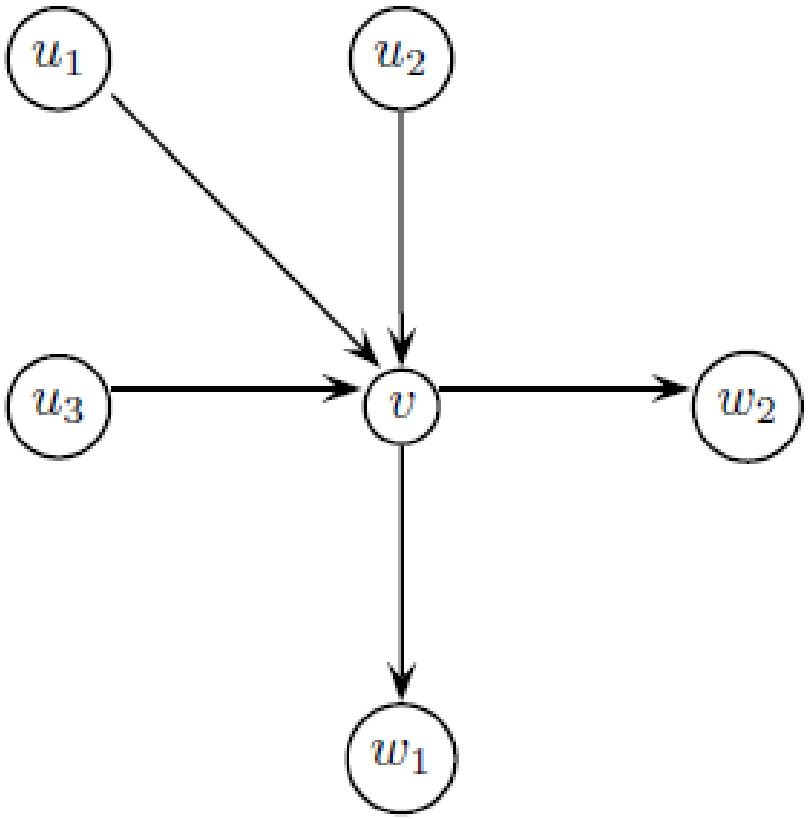


What about diagonals?

- The score at point B is given by:

$$s_B = \textbf{max of} \begin{cases} s_{A1} + \text{weight of the edge } (A_1, B) \\ s_{A2} + \text{weight of the edge } (A_2, B) \\ s_{A3} + \text{weight of the edge } (A_3, B) \end{cases}$$

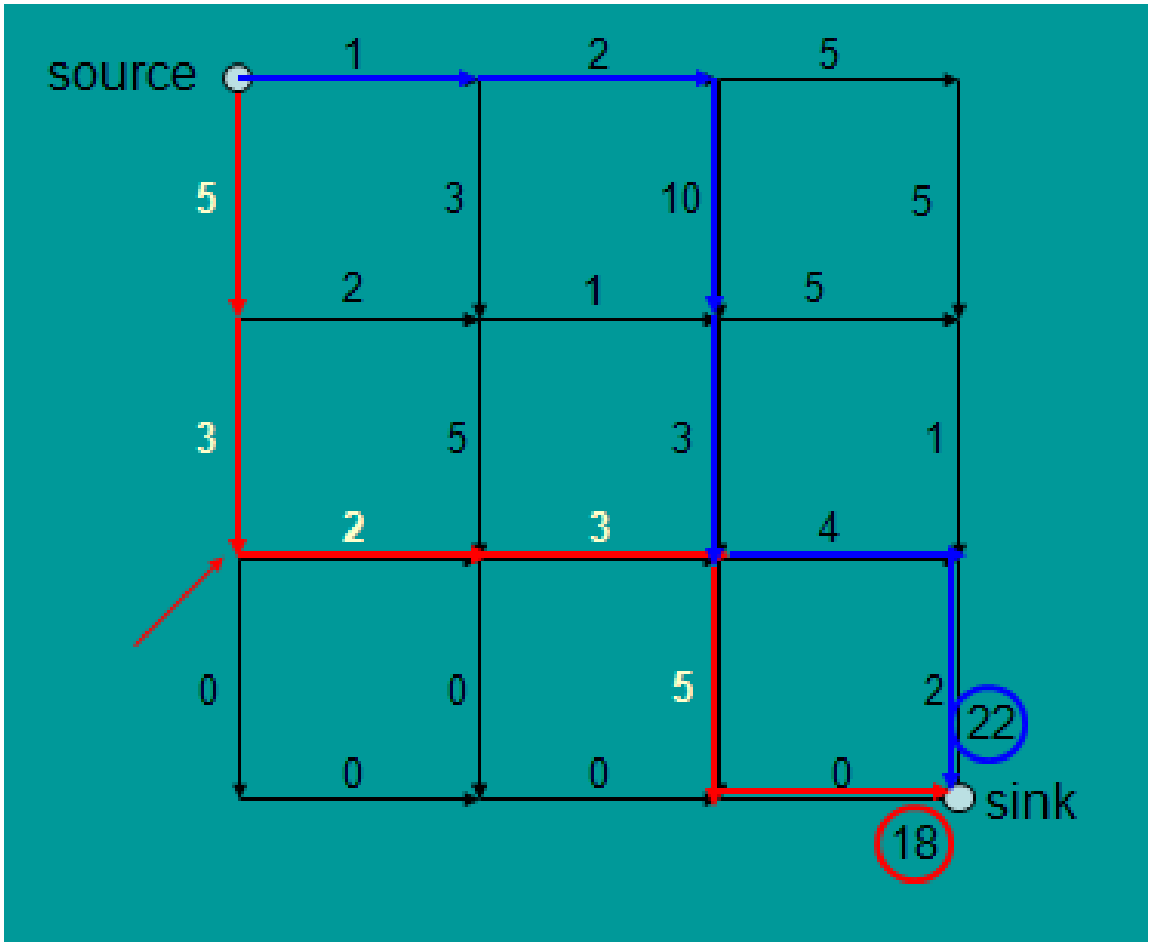
Κατευθυνόμενα ακυκλικά γραφήματα
Directed acyclic graphs, DAG



- Μία ακμή του γραφήματος G μπορεί να οριστεί σε σχέση με την κορυφή προέλευσης της u και την κορυφή προορισμού της v ως (u,v).
- **Εισερχόμενος βαθμός κορυφής:** Ο αριθμός των εισερχόμενων ακμών μιας κορυφής – πρόγονοι.
- **Εξερχόμενος βαθμός κορυφής:** Ο αριθμός των εξερχόμενων ακμών μιας κορυφής – απόγονοι.
- u: πρόγονος (predecessor) της κορυφής v αν (u,v) ε E.

- **Γράφημα G = (V,E),** όπου V: κορυφές και E: ακμές του γραφήματος.

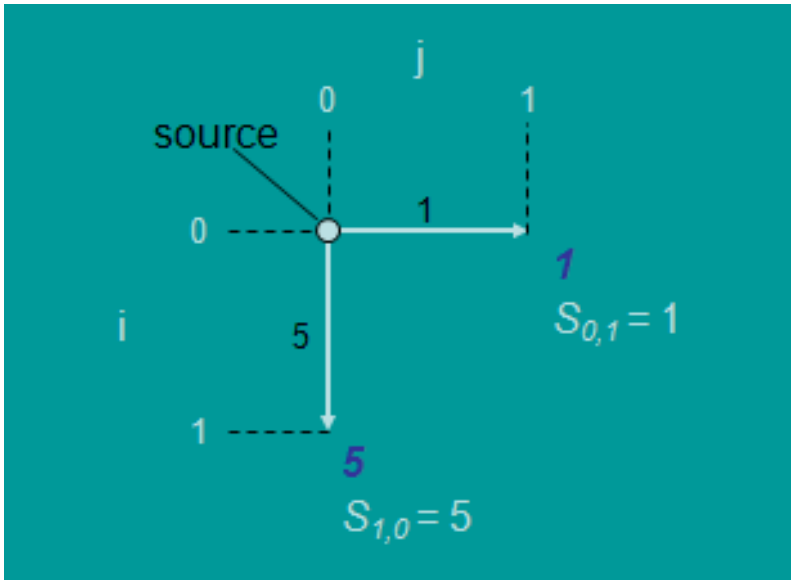
MTP (Manhattan Tourist Problem): Greedy Algorithm Is Not Optimal



MTP: Simple Recursive Program

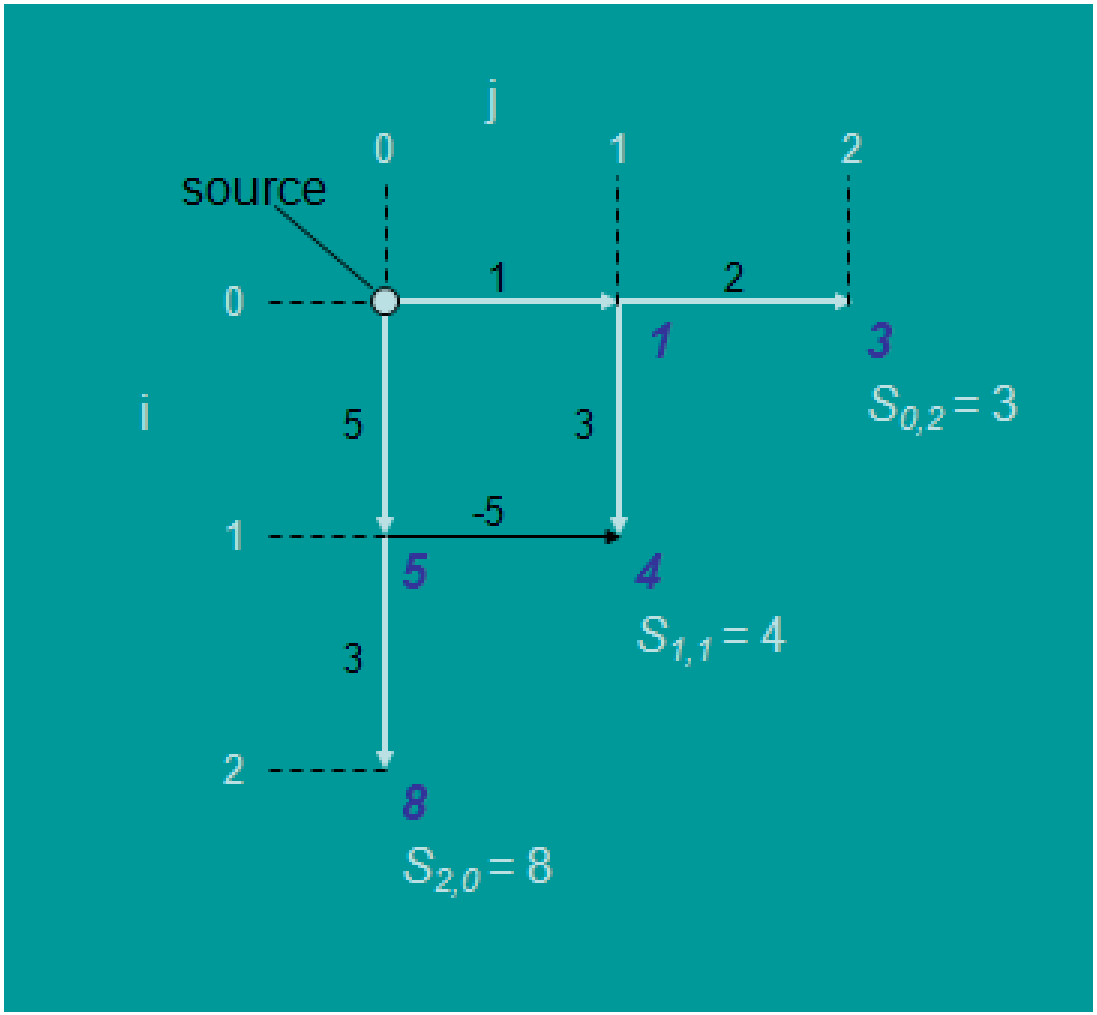
```
MT(n,m)
    if n=0  or m=0
        return MT(n,m)
    x ← MT(n-1,m)+
    length of the edge from (n- 1,m)
  to (n,m)
    y ← MT(n,m-1)+
    length of the edge from (n,m-1)
  to (n,m)
    return max{x,y}
```

MTP: Dynamic Programming

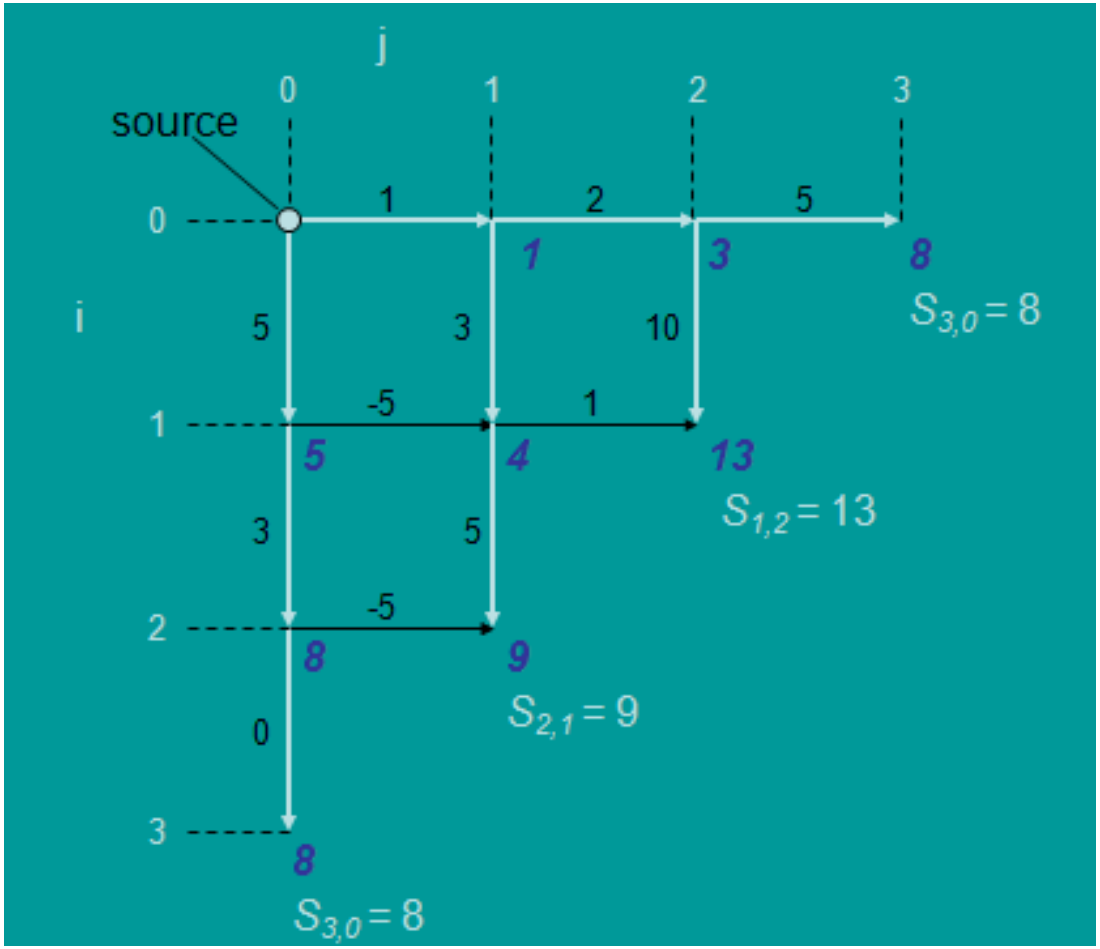


- Calculate optimal path score for each vertex in the graph.
- Each vertex's score is the maximum of the prior vertices score plus the weight of the respective edge in between.

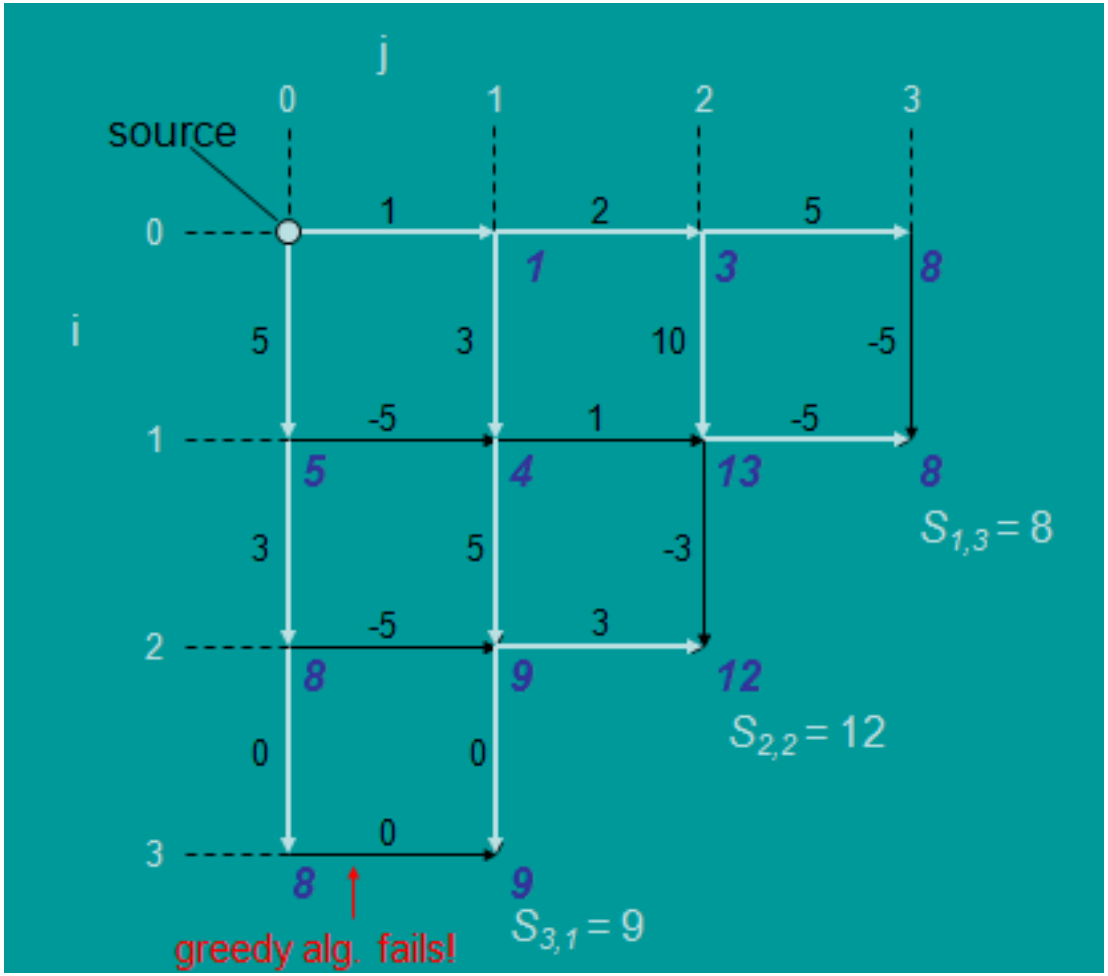
MTP: Dynamic Programming (cont'd) (1/5)



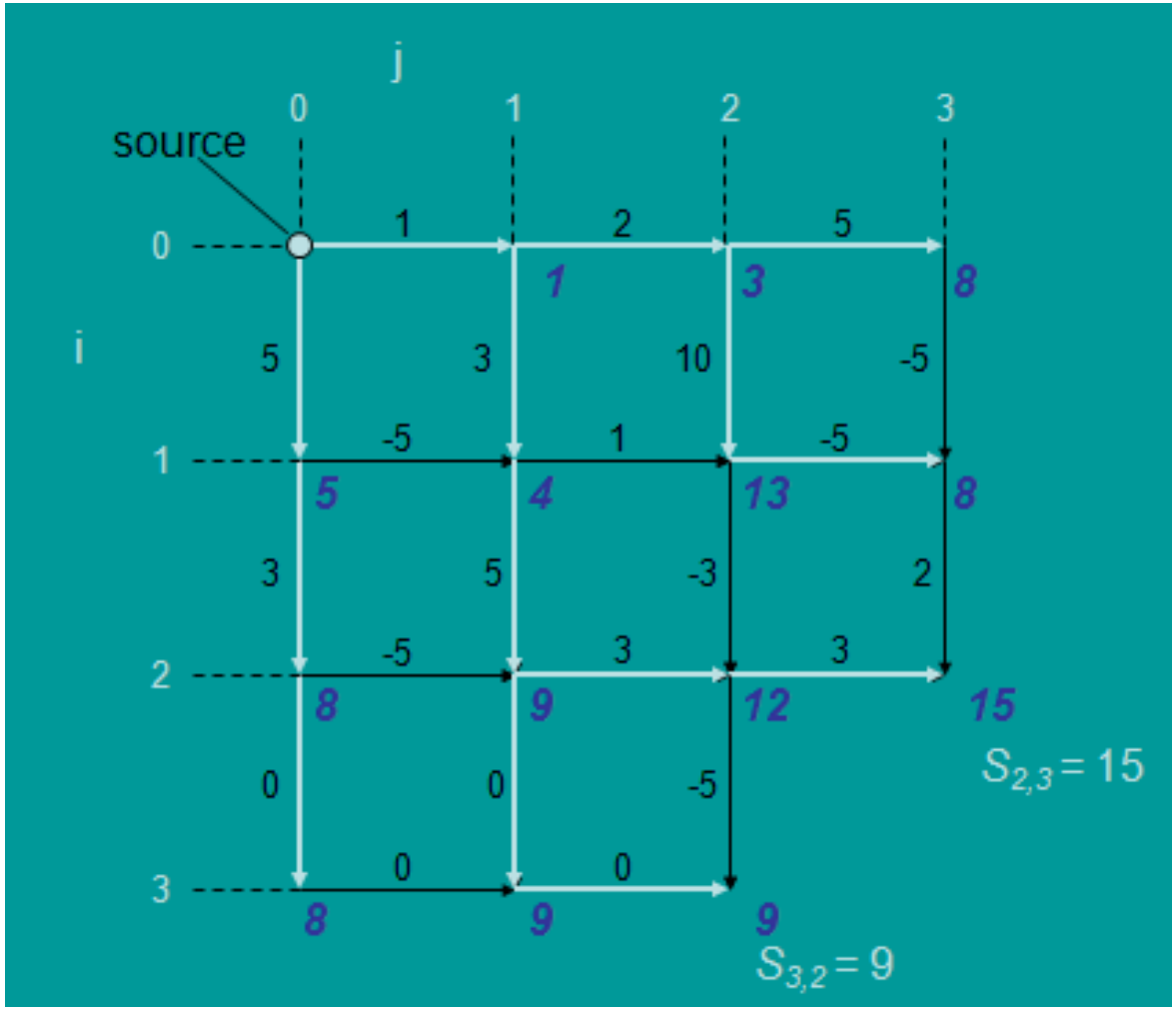
MTP: Dynamic Programming (cont'd) (2/5)



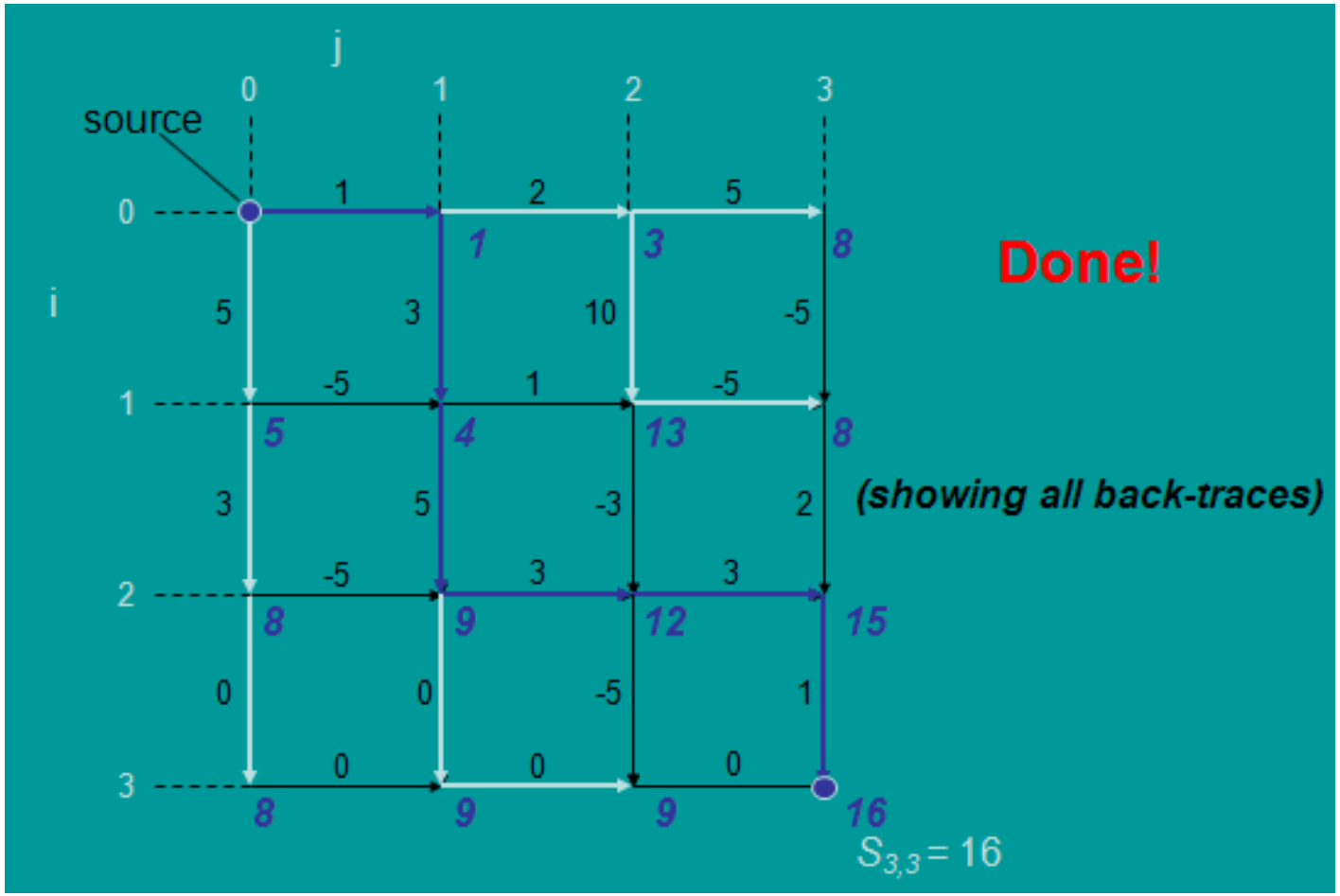
MTP: Dynamic Programming (cont'd) (3/5)



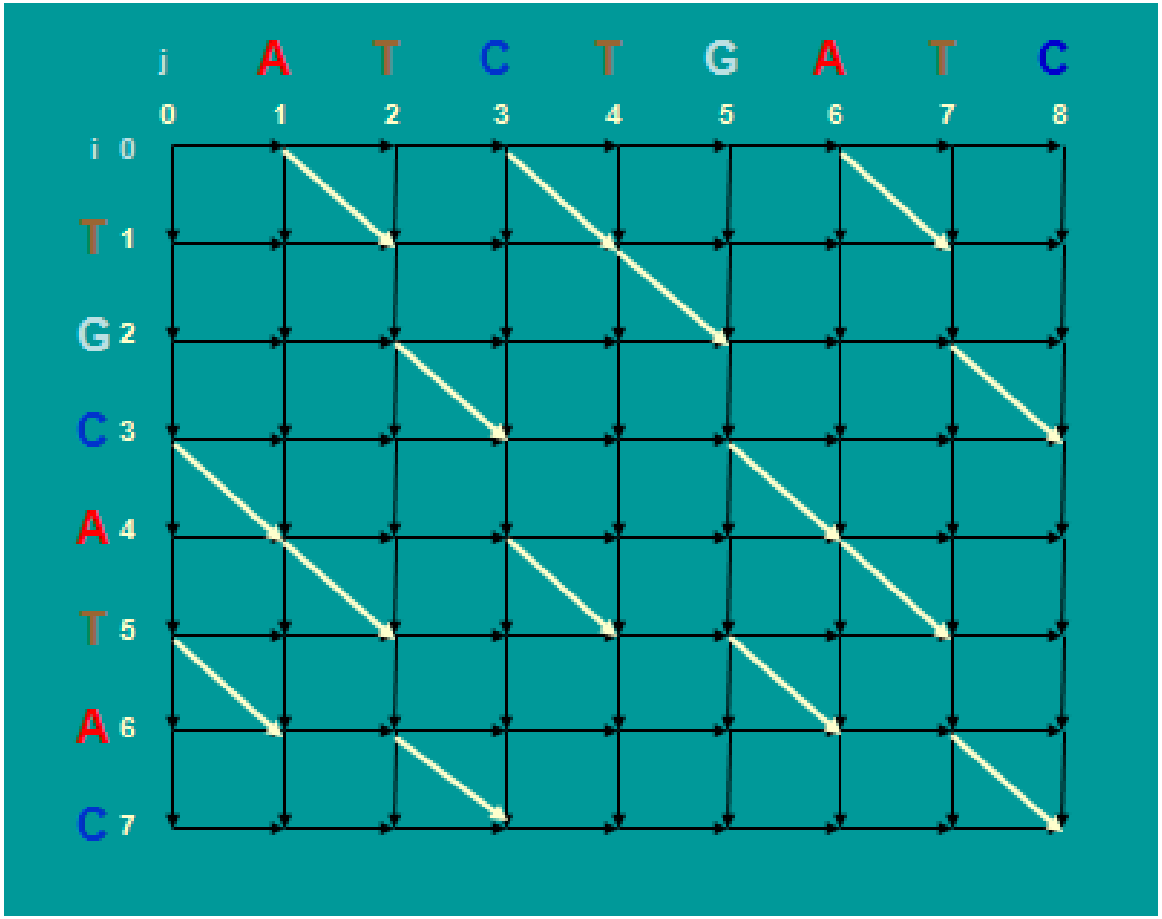
MTP: Dynamic Programming (cont'd) (4/5)



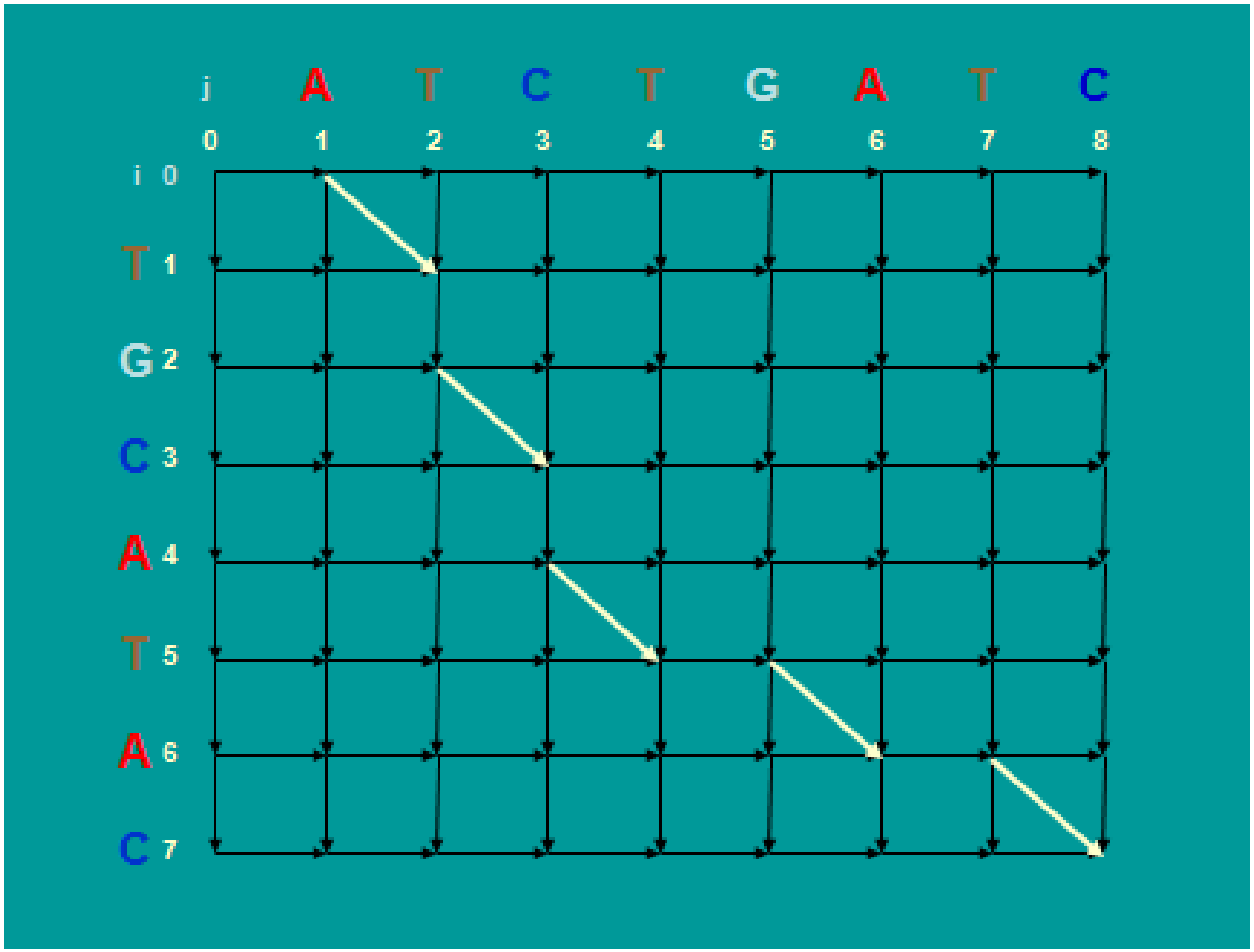
MTP: Dynamic Programming (cont'd) (5/5)



Edit Graph for LCS Problem (1/3)

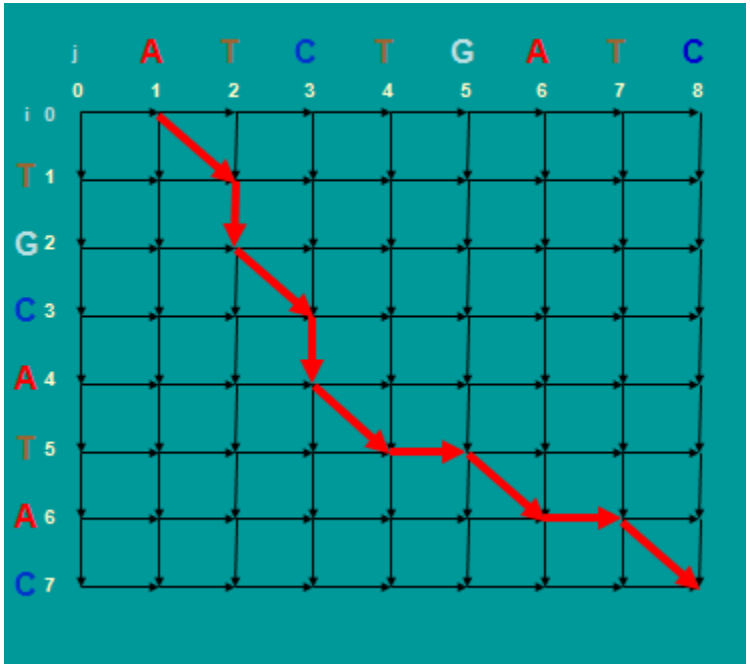


Edit Graph for LCS Problem (2/3)



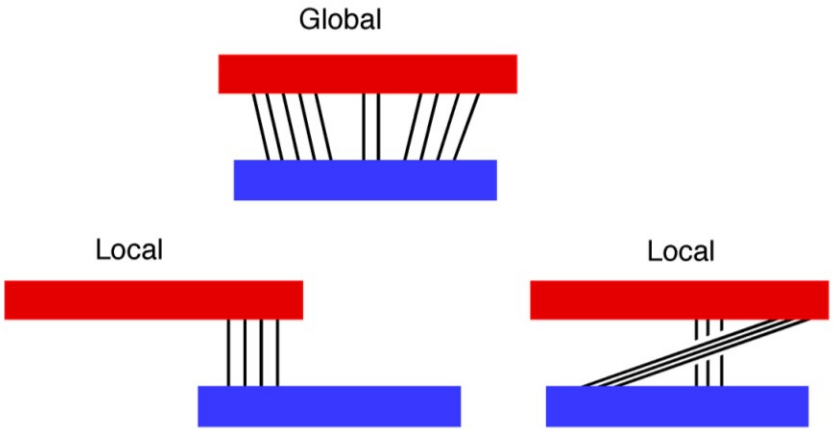
Edit Graph for LCS Problem (3/3)

- Every path is a common subsequence.
- Every diagonal edge adds an extra element to common subsequence.
- LCS Problem: Find a path with maximum number of diagonal edges.

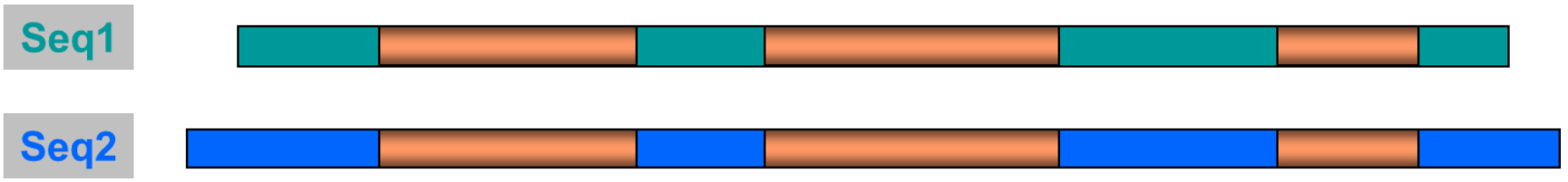


Τοπική ή ολική στοίχιση;;

- Ολική στοίχιση όταν:
  - Δύο αλληλουχίες προέρχονται από τον ίδιο πρόγονο.
  - Όταν δύο αλληλουχίες έχουν περίπου το ίδιο μήκος.
- Τοπική στοίχιση όταν:
  - Δύο αλληλουχίες έχουν μία κοινή περιοχή, η οποία αποτελεί μέλος του συνολικού τους μήκους.
  - Δύο αλληλουχίες έχουν διαφορετικό μήκος.
  - Μία αλληλουχία αποτελεί μέρος της άλλης αλληλουχίας.



Global Alignment – Ολική στοίχιση



Αλγόριθμος Needleman-Wunsch

Ολική στοίχιση – Αλγόριθμος Needleman-Wunsch (1/3)

- **Ολική στοίχιση** δύο αλληλουχιών. Προσπάθειες να συγκρίνουμε **όλα τα κατάλοιπα των δύο αλληλουχιών**
- Δύο αλληλουχίες:
  - $\mathbf{x_1, x_2 \ldots x_n}$
  - $\mathbf{y_1, y_2 \ldots y_n}$
- Κατασκευάζεται ο πίνακας S**(i, j), 0 ≤ i ≤n, 0 ≤ j ≤ m**
- Διατρέχουμε τον πίνακα από πάνω αριστερά προς κάτω δεξιά τοποθετώντας τη βαθμολογία που προκύπτει από:

Ολική στοίχιση – Αλγόριθμος Needleman-Wunsch (2/3)

1. Να στοιχηθεί το x_i με το y_j ⟶
2. Να στοιχηθεί το xi με το κενό ⟶
3. Να στοιχηθεί το y_j με το κενό ⟶

MAX

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

όπου

- **$s(x_i,y_j)$:** η βαθμολογία για τη στοίχιση των καταλοίπων x_i με y_j
- **d:** ποινή για το κενό

Για την πρώτη γραμμή: F(i,0) = -id

Για την πρώτη στήλη: F(0,j) = -jd

Χρήση των αλγορίθμων δυναμικού προγραμματισμού για σύγκριση αλληλουχιών (1/2)

1.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap				
b2	2 gaps				
b3	3 gaps				
b4	4 gaps				

2a.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11			
b2	2 gaps				
b3	3 gaps				
b4	4 gaps				

2b.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11			
b2	2 gaps	s12			
b3	3 gaps				
b4	4 gaps				

2c.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21		
b2	2 gaps	s12			
b3	3 gaps				
b4	4 gaps				

Χρήση των αλγορίθμων δυναμικού προγραμματισμού για σύγκριση αλληλουχιών (2/2)

2d.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21		
b2	2 gaps	s12	s22		
b3	3 gaps				
b4	4 gaps				

3. Part of trace back matrix

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21	s31	s41
b2	2 gaps	s12	s22	s32	s42
b3	3 gaps	s13	s23	s33	s43
b4	4 gaps	s14	s24	s34	s44

4. Trace back matrix

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21 B	s31	s41
b2	2 gaps	s12	s22	s32	s42
b3	3 gaps	s13	s23 A	s33	s43
b4	4 gaps	s14	s24	s34	s44

Alignment A:	a1	a2	a3	a4	
	b1	b2	b3	b4	
Alignment B:	a1	a2	a3	a4	–
	b1	–	b2	b3	b4

Ολική στοίχιση – Αλγόριθμος Needleman-Wunsch (3/3)

1. Να στοιχηθεί το x_i με το y_j ⟶
2. Να στοιχηθεί το x_i με το κενό ⟶ **MAX**
3. Να στοιχηθεί το y_j με το κενό ⟶

$$\begin{cases} S(i,j) = S(i-1,j-1) + w(x_i, y_j) \\ S(i,j) = S(i-1,j) - d \\ S(i,j) = S(i,j-1) - d \end{cases}$$

όπου

- **$w(x_i,y_i)$:** η βαθμολογία για τη στοίχιση των καταλοίπων x_i με y_j
- **d:** ποινή για το κενό

Για την πρώτη γραμμή: F(i,0) = -id

Για την πρώτη στήλη: F(0,j) = -jd

	gap	1	2	…	j-1	j	
gap	0	-d	-2d	…	-(j-1)d	-jd	
1	-d						
2	-2d						
…							
i-1	-(j-1)d				$s_{i-1,j-1}$	$s_{i-1,j}$	
i	-id				$s_{i,j-1}$	s_{ij}	

Αλγόριθμοι δυναμικού προγραμματισμού – Παράδειγμα

1η αλληλουχία: ASRFALFF; M = 8

2η αλληλουχία: ASIRVVFALF; N = 10

Initialized Matrix – Αλγόριθμος Needleman-Wunsch (1/2)

M+2 σειρές, N+2 στήλες

	GAP	A	S	R	F	A	L	F	F
GAP									
A									
S									
I									
R									
V									
V									
F									
A									
L									
F									

Βαθμοί
Όμοιο: +2
Ανόμοιο: 0
Κενό: -1

Για την πρώτη γραμμή: S(i,0) = -id
Για την πρώτη στήλη: S(0,j) = -jd

Ποινή κενού (d): -1

Initialized Matrix – Αλγόριθμος Needleman-Wunsch (2/2)

	GAP	A	S	R	F	A	L	F	F
GAP	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1								
S	-2								
I	-3								
R	-4								
V	-5								
V	-6								
F	-7								
A	-8								
L	-9								
F	-10								

Βαθμοί

Όμοιο: +2

Ανόμοιο: 0

Κενό: -1

Για την πρώτη γραμμή: S(i,0) = -id

Για την πρώτη στήλη: S(0,j) = -jd

Matrix fill – Αλγόριθμος Needleman-Wunsch (1/4)

$$\textbf{MAX}\begin{cases} S(i,j) = S(i-1,j-1) + v(x_i, y_j) \\ S(i,j) = S(i-1,j) - d \\ S(i,j) = S(i,j-1) - d \end{cases}$$

S(1,1) = MAX [$S_{0,0}$+2, $S_{1,0}$-1, $S_{0,1}$-1] = MAX[0+2, -1-1, -1-1] = MAX[2, -2, -2] = **2**

Όμοιο κατάλοιπο: +2

Ανόμοιο κατάλοιπο: 0

Κενό: -1

	GAP	A	S	R	F	A	L	F	F
GAP	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	2							
S	-2								
I	-3								
R	-4								
V	-5								
V	-6								
F	-7								
A	-8								
L	-9								
F	-10								

Matrix fill – Αλγόριθμος Needleman-Wunsch (2/4)

$$\text{MAX}\begin{cases} S(i,j) = S(i-1,j-1) + v(x_i, y_j) \\ S(i,j) = S(i-1,j) - ! \\ S(i,j) = S(i,j-1) - ! \end{cases}$$

S(1,2) = MAX [$S_{0,1}$+2, $S_{0,2}$-1, $S_{1,1}$-1] = MAX[-1+2, -2-1, 2-1] = MAX[1, -3, -1] = **1**

Όμοιο κατάλοιπο: +2
Ανόμοιο κατάλοιπο: 0
Κενό: -1

	GAP	A	S	R	F	A	L	F	F
GAP	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	2	1						
S	-2								
I	-3								
R	-4								
V	-5								
V	-6								
F	-7								
A	-8								
L	-9								
F	-10								

Matrix fill – Αλγόριθμος Needleman-Wunsch (3/4)

$$\text{MAX}\begin{cases} S(i,j) = S(i-1,j-1) + v(x_i, y_j) \\ S(i,j) = S(i-1,j) - d \\ S(i,j) = S(i,j-1) - d \end{cases}$$

S(2,1) = MAX [$S_{1,0}$+0, $S_{1,1}$-1, $S_{2,0}$-1] = MAX[-1+0, 2-1, -2-1] = MAX[-1, 1, -3] = **1**

Όμοιο κατάλοιπο: +2

Ανόμοιο κατάλοιπο: 0

Κενό: -1

	GAP	A	S	R	F	A	L	F	F
GAP	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	2	1						
S	-2	1							
I	-3								
R	-4								
V	-5								
V	-6								
F	-7								
A	-8								
L	-9								
F	-10								

Matrix fill – Αλγόριθμος Needleman-Wunsch (4/4)

$$\textbf{MAX}\begin{cases} S(i,j) = S(i-1,j-1) + v(x_i, y_j) \\ S(i,j) = S(i-1,j) - g \\ S(i,j) = S(i,j-1) - g \end{cases}$$

S(2,2) = MAX [$S_{1,1}$+2, $S_{1,2}$-1, $S_{2,1}$-1] = MAX[+2+2, 1-1, 1-1] = MAX[+4, 0, 0] = **4**

Όμοιο κατάλοιπο: +2
Ανόμοιο κατάλοιπο: 0
Κενό: -1

	GAP	A	S	R	F	A	L	F	F
GAP	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	2	1						
S	-2	1	4						
I	-3								
R	-4								
V	-5								
V	-6								
F	-7								
A	-8								
L	-9								
F	-10								

Matrix traceback – Αλγόριθμος Needleman-Wunsch

$$\text{MAX}\begin{cases} S(i,j) = S(i-1,j-1) + v(x_i, y_j) \\ S(i,j) = S(i-1,j) - ! \\ S(i,j) = S(i,j-1) - ! \end{cases}$$

S(2,2) = MAX [$S_{1,1}$+2, $S_{1,2}$-1, $S_{2,1}$-1] = MAX[+2+2, 1-1, 1-1] = MAX[+4, 0, 0] = **4**

Όμοιο κατάλοιπο: +2
Ανόμοιο κατάλοιπο: 0
Κενό: -1

	GAP	A	S	R	F	A	L	F	F
GAP	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	2	1	0	-1	-2	-3	-4	-5
S	-2	1	4	3	2	1	0	-1	-2
I	-3	0	3	4	3	2	1	0	-1
R	-4	-1	2	5	4	3	2	1	0
V	-5	-2	1	4	5	4	3	2	1
V	-6	-3	0	3	4	5	4	3	2
F	-7	-4	-1	2	5	4	5	6	5
A	-8	-5	-2	1	4	7	6	5	6
L	-9	-6	-3	0	3	6	9	8	7
F	-10	-7	-4	-1	2	5	8	11	10

Έλεγχος της βαθμολογίας

	GAP	A	S	R	F	A	L	F	F
GAP	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	2	1	0	-1	-2	-3	-4	-5
S	-2	1	4	3	2	1	0	-1	-2
I	-3	0	3	4	3	2	1	0	-1
R	-4	-1	2	5	4	3	2	1	0
V	-5	-2	1	4	5	4	3	2	1
V	-6	-3	0	3	4	5	4	3	2
F	-7	-4	-1	2	5	4	5	6	5
A	-8	-5	-2	1	4	7	6	5	6
L	-9	-6	-3	0	3	6	9	8	7
F	-10	-7	-4	-1	2	5	8	11	10

: κενό στην κάθετη

: κενό στην οριζόντια

Όμοιο κατάλοιπο: +2

Ανόμοιο κατάλοιπο: 0

Κενό: -1

A S - R - - F A L F F

| | | | | | | |

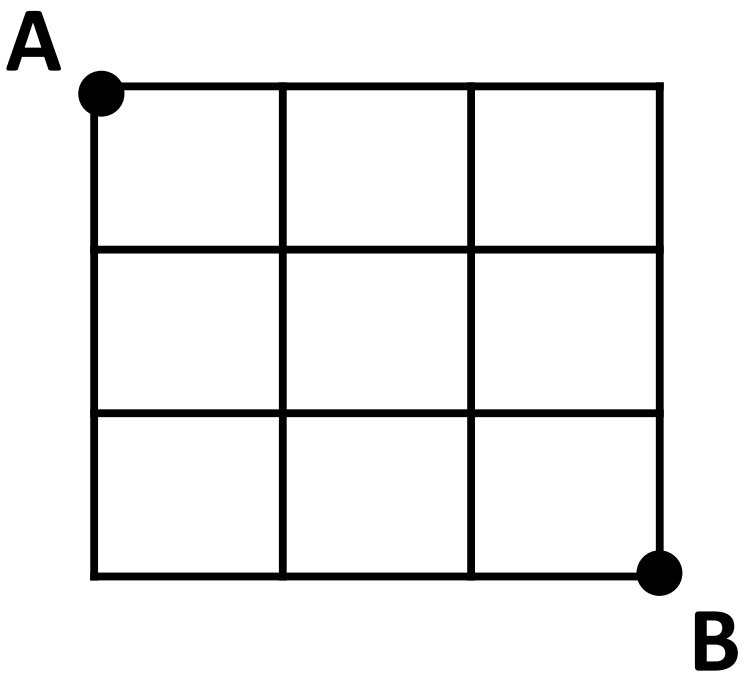
A S I R V V F A L F -

Έλεγχος της βαθμολογίας:

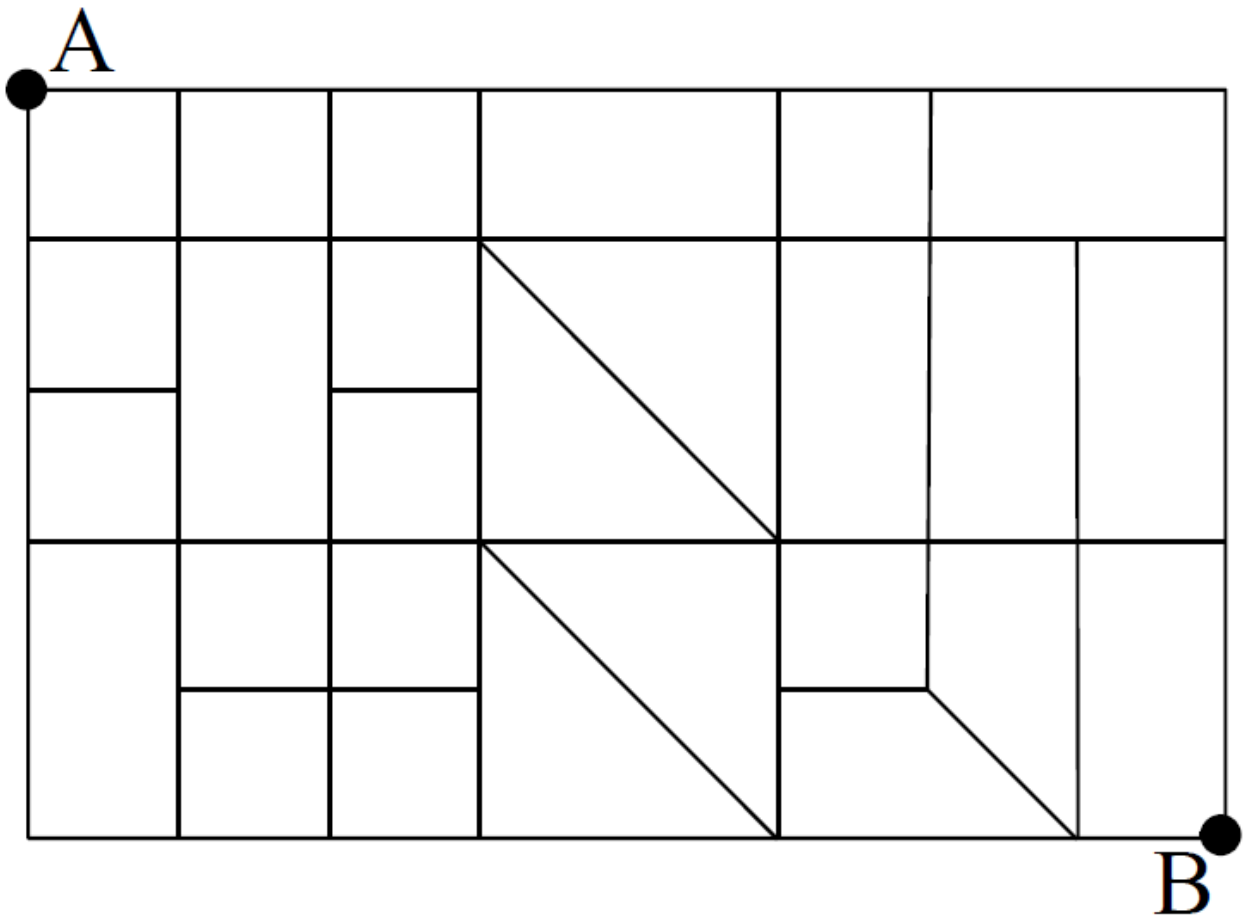
(7x2) – (4x1) = 14-4 = 10

Και ένα διαφορετικό πρόβλημα…

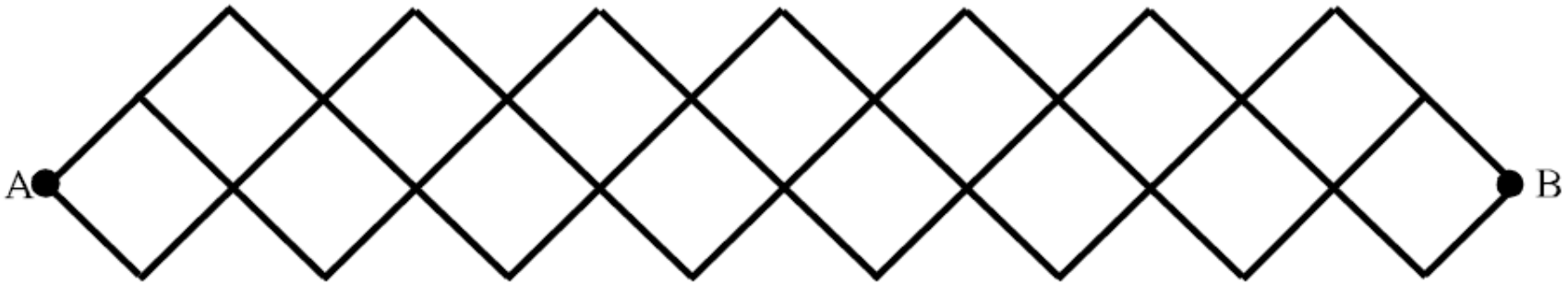
- Πόσα διαφορετικά μονοπάτια υπάρχουν που συνδέουν το Α με το Β;
- **Προσοχή:** Επιτρέπεται η κίνηση ανατολικά και κάτω. Δεν επιτρέπεται η κίνηση δυτικά και πάνω.



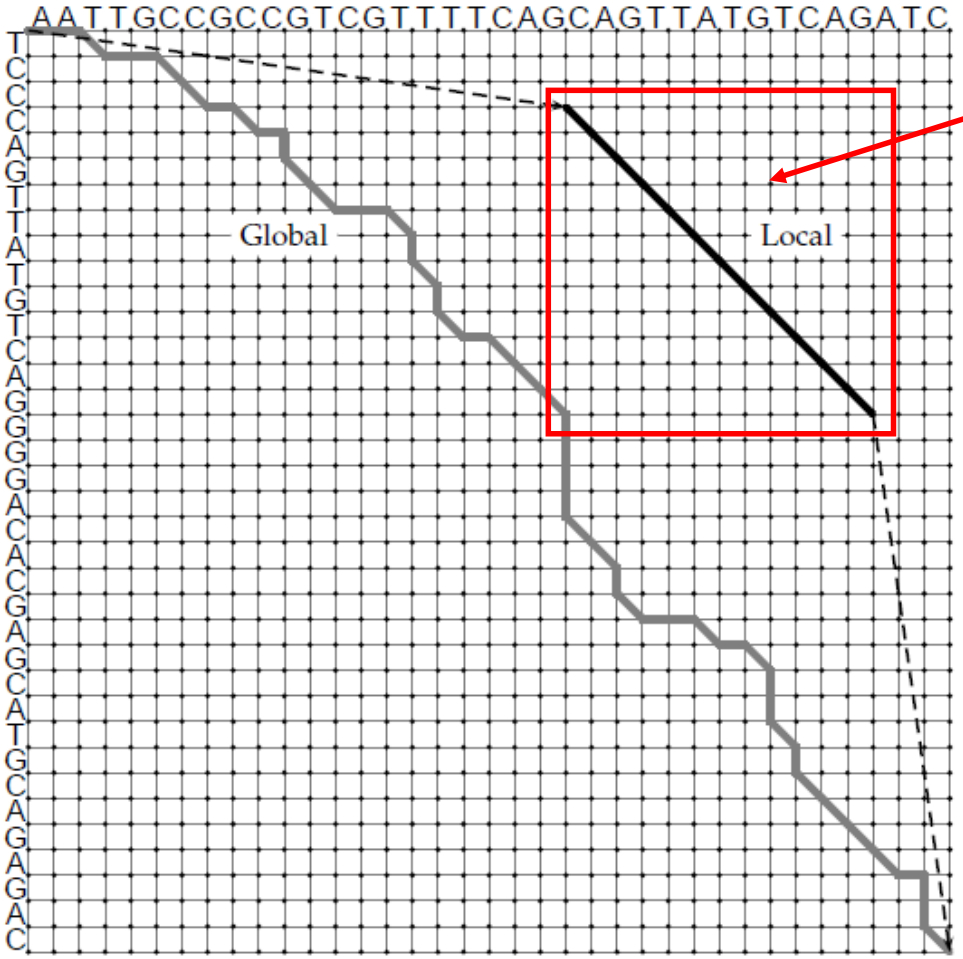
Δοκιμάστε το παρακάτω



Ή το παρακάτω



Τοπική vs. Ολική στοίχιση

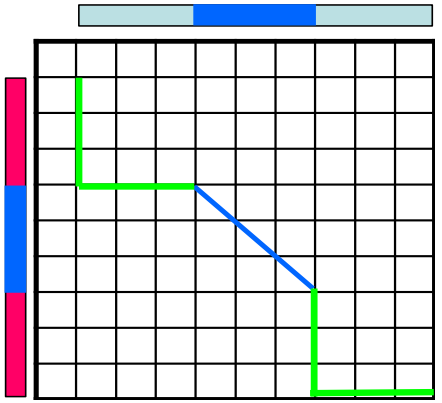


Compute a “mini” Global Alignment to get Local.

Τοπική στοίχιση αλληλουχιών

$$S = \frac{1}{3} A \cdot \delta - \frac{2}{3} A \cdot \sigma - \frac{2}{3} A \cdot \sigma = \frac{1}{3} A(\delta - 4\sigma$$

όπου δ: ταίριασμα και σ: ποινή εισαγωγής κενού



Τοπική στοίχιση τυχαίων αλληλουχιών (1/2)

$$S' = \frac{1}{4} A \cdot \delta - \frac{3}{4} A \cdot \sigma = \frac{1}{4} A(\delta - 3\sigma)$$

όπου δ: ταίριασμα και σ: ποινή εισαγωγής κενού

Τοπική στοίχιση τυχαίων αλληλουχιών (2/2)

$$S = \frac{1}{3} A(\delta - |\sigma$$

$$S' = \frac{1}{4} A(\delta - |\sigma$$

Α: μήκος αλληλουχίας
δ: ταίριασμα
σ: ποινή κενού

- **Παράδειγμα:**

Για αλληλουχία Α=30 καταλοίπων, με βαθμό ταιριάσματος, δ=3 και ποινή κενού σ=-2

S=-50 και S'=-22.5

Για αλληλουχία Α=90 καταλοίπων, με βαθμό ταιριάσματος, δ=3 και ποινή κενού σ=-2

S=-150 και S'=-67.5

Τοπική vs. Ολική στοίχιση (1/2)

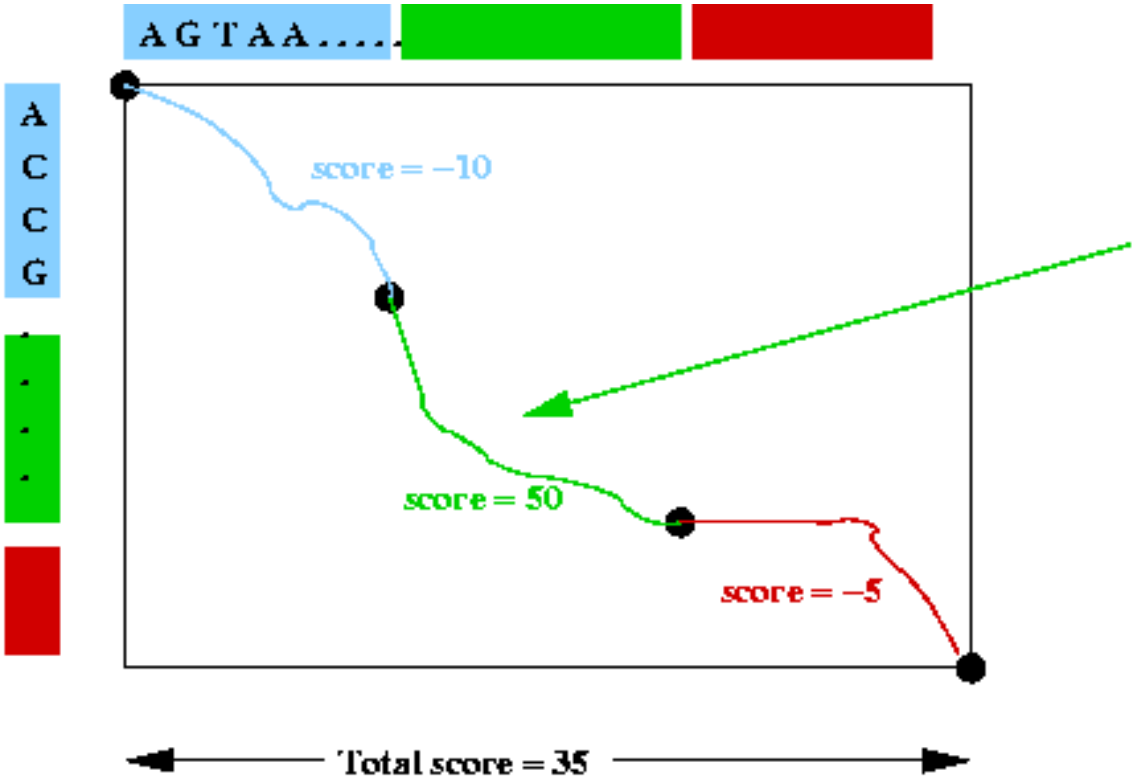
- **Ολική στοίχιση (Global Alignment)**

```
--T—-CC-C-AGT—-TATGT-CAGGGGACACG—A-GCATGCAGA-GAC
  |   || |   ||   | | | |||     || | | |   | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG—T-CAGAT--C
```

- **Τοπική στοίχιση (Local Alignment) — better alignment to find conserved segment**

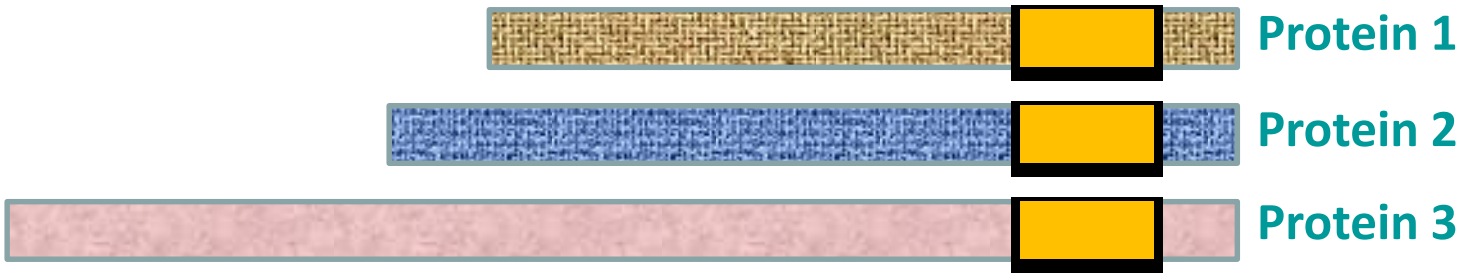
```
                    TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC
                       ||||||||||||
AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC
```

Τοπική vs. Ολική στοίχιση - Βαθμολογία



Τοπική στοίχιση μπορεί να έχει μεγαλύτερη βαθμολογία από την ολική στοίχιση

Local Alignment – Τοπική στοίχιση



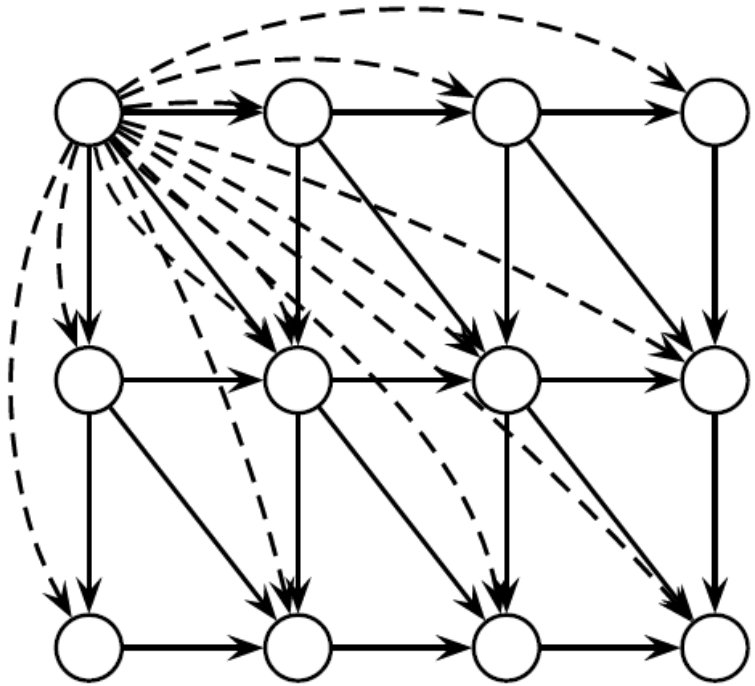
Αλγόριθμος Smith-Waterman

Τοπική στοίχιση – Χρήση σε βιολογικά ερωτήματα

- Όταν υπάρχουν βιολογικά σημαντικές ομοιότητες σε ορισμένα κομμάτια αλληλουχιών (DNA, πρωτεΐνες) σκοπός είναι η μέγιστη βαθμολογία στοίχισης s(a1 …ai, b1…bj) για όλες τις υποσυμβολοσειρές s(a1 …ai, b1…bj) των Α και Β αντίστοιχα.

- **ΠΡΟΒΛΗΜΑ ΤΟΠΙΚΗΣ ΣΤΟΙΧΙΣΗΣ: Η στοίχιση δεν επεκτείνεται σε όλο το μήκος των συμβολοσειρών.**

Τοπική vs. Ολική στοίχιση (2/2)



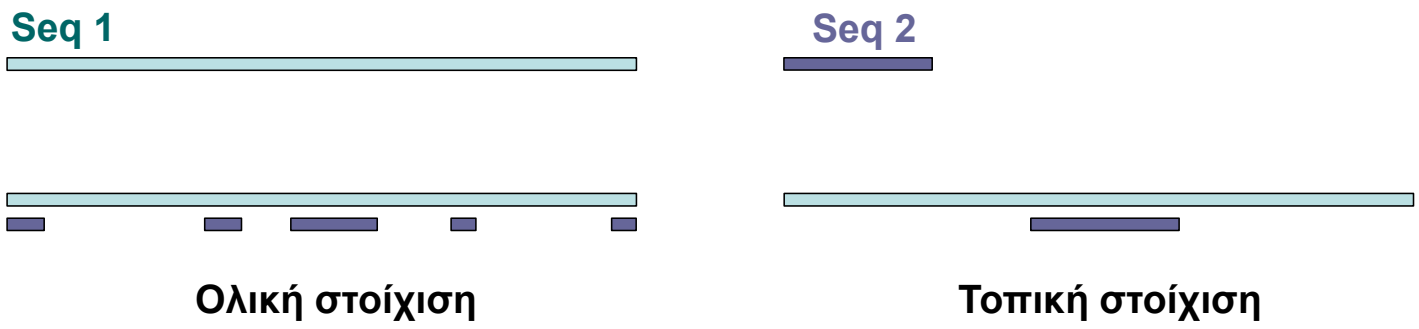
- **Προηγούμενο πρόβλημα (ολική στοίχιση):** Μεγαλύτερη διαδρομή μεταξύ (0,0) και (n,m).
- **Νέο πρόβλημα (τοπική στοίχιση):** Μεγαλύτερη διαδρομή μεταξύ (*i,j*) και (*i′, j′*).
- **<u>Διαφορετικά:</u>** Εύρεση των μεγαλύτερων διαδρομών από την κορυφή προέλευσης (0,0) προς κάθε άλλη κορυφή με την προσθήκη ακμών με συντελεστή στάθμισης 0 στο γράφημα μετασχηματισμού.
- Οι κορυφή προέλευσης (0,0) αποτελεί πρόγονο κάθε κορυφής του γραφήματος και παρέχουν δωρεάν μεταφορά από την κορυφή προέλευσης σε οποιαδήποτε άλλη κορυφή (i,j).

Τοπική στοίχιση αλληλουχιών - Παράδειγμα

- **Δεδομένα:**
  - Μία αλληλουχία DNA, 4000bp.
  - Ολόκληρο το ανθρώπινο γονιδίωμα.
- **Σκοπός:** Να συγκρίνω τις δύο αλληλουχίες και να συμπεράνω αν η αλληλουχία των 4000bp ανήκει στο ανθρώπινο γονιδίωμα.
- **Πρόβλημα:**
  - Τι θα συμβεί αν μόνο ένα μικρό μέρος, π.χ. 200bp, της αλληλουχίας παρουσιάζει αυξημένη ομοιότητα με το ανθρώπινο γονιδίωμα;
  - Μπορούμε να βρούμε αυτή την ομοιότητα των 200bp χρησιμοποιώντας αλγόριθμο ολικής στοίχισης.

Τοπική στοίχιση – Αλγόριθμος Smith-Waterman (1/3)

- Προτάθηκε το 1981 από τους Temple Smith και Mike Waterman.
- Γιατί διαλέγουμε τοπική στοίχιση:
  - Αναδεικνύει συντηρημένες περιοχές μεταξύ δύο αλληλουχιών.
  - Συγκρίνει δύο αλληλουχίες διαφορετικού μήκους.
  - Συγκρίνει δύο αλληλουχίες που είναι μερικώς όμοιες.
  - Συγκρίνει δύο αλληλουχίες από τις οποίες η μία είναι μέρος της άλλης.



Τοπική στοίχιση – Αλγόριθμος Smith-Waterman (2/3)

- Αποτελεί τροποποίηση του αλγόριθμου ολικής στοίχισης Needleman – Wunsch.
- Το βέλτιστο μονοπάτι δεν βρίσκεται αναγκαστικά στις άκρες του γραφήματος.
- Το βέλτιστο μονοπάτι μπορεί να βρίσκεται και στο εσωτερικό του.

Τοπική στοίχιση – Αλγόριθμος Smith-Waterman (3/3)

$$S(i,j) = \max\left\{\begin{array}{c} 0 \\ S(i-1,j-1) + w(x_i, y_j) \\ S(i-1,j) - d \\ S(i,j-1) - d \end{array}\right\}$$

- Τροποποιήσεις του αλγόριθμου **Needleman-Wunsch:**
  - Αρνητική βαθμολογία για ταίριασμα ανόμοιων καταλοίπων.
  - Όταν μία βαθμολογία είναι αρνητική, αντικαθίσταται με το 0.
  - **Matrix traceback:** Ξεκινά από τη μεγαλύτερη βαθμολογία και όχι από το στοιχείο (n,m).

Initialization Step – Αλγόριθμος Smith-Waterman (1/2)

- 1η αλληλουχία: ASRFALFF, m=8.
- 2η αλληλουχία: SFAL, n=4.
  - $w(x_i, y_j) = +2$ αν $x_i = y_j$ (βαθμός όμοιου καταλοίπου).
  - $w(x_i, y_j) = -1$ αν $x_i \neq y_j$ (βαθμός διαφορετικού καταλοίπου).
  - d = -1 (ποινή κενού).

ΒΑΘΜΟΛΟΓΙΑ

Όμοιο κατάλοιπο: +2

Ανόμοιο κατάλοιπο: -1

Κενό: -1

Initialization Step – Αλγόριθμος Smith-Waterman (2/2)

- 1η αλληλουχία: ASRFALFF, m=8
- 2η αλληλουχία: SFAL, n=4

- **1η σειρά:** όλα τα στοιχεία 0
- **1η στήλη:** όλα τα στοιχεία 0

$$S(i,j) = \max \begin{Bmatrix} 0 \\ S(i-1,j-1) + w(x_i, y_j) \\ S(i-1,j) - d \\ S(i,j-1) - d \end{Bmatrix}$$

	GAP	A	S	R	F	A	L	F	F
GAP	0	0	0	0	0	0	0	0	0
S	0								
F	0								
A	0								
L	0								

Matrix fill – Αλγόριθμος Smith-Waterman (1/3)

- S(1,1) = MAX[0, S(0,0)-1, S(0,1)-1, F(1,0)-1 = MAX[0,-1,-1,-1)=0

$$S(i,j) = \max \begin{Bmatrix} 0 \\ S(i-1,j-1) + w(x_i, y_j) \\ S(i-1,j) - d \\ S(i,j-1) - d \end{Bmatrix}$$

	GAP	A	S	R	F	A	L	F	F
GAP	0	0	0	0	0	0	0	0	0
S	0	**0**							
F	0								
A	0								
L	0								

- **Όμοιο κατάλοιπο: +2**
- **Ανόμοιο κατάλοιπο: -1**
- **Κενό: -1**

Matrix fill – Αλγόριθμος Smith-Waterman (2/3)

- S(1,2) = MAX[0, S(0,1)+2, S(0,2)-1, S(1,1)-1 = MAX[0,+2,-1,-1] = +2
- S(1,3) = MAX[0, S(0,2)-1, S(0,3)-1, S(1,2)-1 = MAX[0,+1,-1,-1] = +1
- S(1,4) = MAX[0, S(0,3)-1, S(0,4)-1, S(1,3)-1 = MAX[0,+1,-1,-1] = 0

$$S(i,j) = \max \begin{Bmatrix} 0 \\ S(i-1,j-1) + w(x_i, y_j) \\ S(i-1,j) - d \\ S(i,j-1) - d \end{Bmatrix}$$

	GAP	A	S	R	F	A	L	F	F
GAP	0	0	0	0	0	0	0	0	0
S	0	**0**	**+2**	**+1**	**0**				
F	0								
A	0								
L	0								

- **Όμοιο κατάλοιπο: +2**
- **Ανόμοιο κατάλοιπο: -1**
- **Κενό: -1**

Matrix fill – Αλγόριθμος Smith-Waterman (3/3)

	GAP	A	S	R	F	A	L	F	F
GAP	0	0	0	0	0	0	0	0	0
S	0	0	+2	+1	0	0	0	0	0
F	0	0	1	1	3	2	1	2	2
A	0	2	1	0	2	5	4	3	2
L	0	1	1	0	0	4	7	6	5

- **Όμοιο κατάλοιπο: +2**
- **Ανόμοιο κατάλοιπο: -1**
- **Κενό: -1**

Traceback – Αλγόριθμος Smith-Waterman (1/2)

	GAP	A	S	R	F	A	L	F	F
GAP	0	0	0	0	0	0	0	0	0
S	0	0	2	1	0	0	0	0	0
F	0	0	1	1	3	2	1	2	2
A	0	2	1	0	2	5	4	3	2
L	0	1	1	0	0	4	7	6	5

Traceback – Αλγόριθμος Smith-Waterman (2/2)

	GAP	A	S	R	F	A	L	F	F
GAP	0	0	0	0	0	0	0	0	0
S	0	0	2	1	0	0	0	0	0
F	0	0	1	1	3	2	1	2	2
A	0	2	1	0	2	5	4	3	2
L	0	1	1	0	0	4	7	6	5

→ : κενό στην κάθετη

↓ : κενό στην οριζόντια

- **Όμοιο κατάλοιπο: +2**
- **Ανόμοιο κατάλοιπο: -1**
- **Κενό: -1**

A S R F A L F F

| | | |

S - F A L

Έλεγχος της βαθμολογίας:

4x2-1x1 = 8-1 = 7

Ένας αλγόριθμος ολικής στοίχισης δίνει ολική στοίχιση και ένα τοπικής στοίχισης δίνει τοπική στοίχιση

- Η επιλογή του κατάλληλου αλγορίθμου δεν αρκεί.
- Το σύστημα βαθμονόμησης (match, mismatch, gap penalties) παίζει καθοριστικό ρόλο στο τελικό αποτέλεσμα (ολική ή τοπική).
- **Ολική στοίχιση:** Η συγκρινόμενες περιοχές έχουν μεγάλο μήκος και καλύπτουν το μεγαλύτερο μέρος των αλληλουχιών. Επίσης, εμφανίζονται πολλά κενά με σκοπό τη βέλτιστη στοίχιση.
- **Τοπική στοίχιση:** Μικρότερη από την ολική στοίχιση και δεν περιλαμβάνει πολλά κενά.
- **Υποκειμενική παρατήρηση;;;**

Τι συμβαίνει όταν στοιχίζονται τυχαίες αλληλουχίες με συνεχώς αυξανόμενο μήκος;

- **Σύστημα βαθμονόμησης για ολική στοίχιση:** Σε κάθε ταίριασμα δίνει θετική βαθμολογία και οι ποινές για λάθος στοίχιση και εισχώρηση κενών είναι χαμηλή. Επιτρέπεται η «προσπέραση» ανόμοιων περιοχών.
- **Σύστημα βαθμονόμησης για τοπική στοίχιση:** Επιλέγεται **αρνητική βαθμολογία για λάθος στοίχιση** και **υψηλή ποινή για τα κενά** με σκοπό να μην πραγματοποιηθεί στοίχιση σε περιοχές που δεν ταιριάζουν σημαντικά.
- **ΑΠΟΤΕΛΕΣΜΑ:** Στην ολική στοίχιση η βαθμολογία αυξάνεται αναλογικά με το μήκος των αλληλουχιών. Στην τοπική στοίχιση η βαθμολογία αυξάνεται αναλογικά με τη λογαριθμική τιμή του μήκους των αλληλουχιών.
- **ΑΥΤΗ Η ΔΙΑΦΟΡΕΤΙΚΗ ΣΥΜΠΕΡΙΦΟΡΑ ΤΗΣ ΒΑΘΜΟΛΟΓΙΑΣ ΣΤΟΙΧΙΣΗΣ ΤΥΧΑΙΩΝ ΑΛΛΗΛΟΥΧΙΩΝ ΔΙΑΚΡΙΝΕΙ ΤΗΝ ΟΛΙΚΗ ΑΠΟ ΤΗΝ ΤΟΠΙΚΗ ΣΤΟΙΧΙΣΗ.**

Το αντίστροφο ερώτημα: Είναι πάντα το σύστημα βαθμονόμησης αρκετό

- ΟΛΙΚΗ ΣΤΟΙΧΙΣΗ:
  - Επιτρέπονται αρνητικές βαθμολογίες.
  - Η κάθε βαθμολογία δεν μπορεί να αυτονομηθεί από την προηγούμενη.
- ΤΟΠΙΚΗ ΣΤΟΙΧΙΣΗ
  - Δεν επιτρέπονται αρνητικές βαθμολογίες.
  - Η στοίχιση ξεκινάει από τη μέγιστη βαθμολογία, επομένως, δεν ακολουθείται αναγκαστικά το μονοπάτι που διατρέχει ολόκληρο τον πίνακα.

Ολική ή τοπική στοίχιση – Παράδειγμα (1/5)

Ολική στοίχιση

		M	N	A	L	S	D	R	T
	GAP	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6	-6	-10					
G	-16	-6							
S	-20	-10							
D	-24								
R	-28								
T	-32								
T	-36								
E	-40								
T	-44								

PAM250

Gap penalty=-12-4(x-1)

M→M: 6
M→N: -2
G→M: -3
S→M: -2
M→A: -1

Ολική ή τοπική στοίχιση – Παράδειγμα (2/5)

Ολική στοίχιση

		M	N	A	L	S	D	R	T
	GAP	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6	-6	-10	-14	-18	-22	-26	-30
G	-16	-6	6	-5	-10	-13	-17	-22	-26
S	-20	-10	-5	7	-5	-8	-13	-17	-21
D	-24	-14	-8	-5	3	-5	-4	-14	-17
R	-28	-18	-14	-9	-8	3	-6	2	10
T	-32	-22	-18	-13	-11	-7	3	-7	5
T	-36	-26	-22	-17	-15	-10	-7	2	-4
E	-40	-30	-25	-21	-20	-15	-7	-8	2
T	-44	-34	-30	-24	-23	-19	-15	-8	-5

PAM250

Gap penalty=-12-4(x-1)

M→M: 6

M→N: -2

G→M: -3

M	-	N	A	L	S	D	R	T
M	G	S	D	R	T	T	E	T

6	-12	1	0	-3	1	0	-1	3	-5

M	N	A	-	L	S	D	R	T
M	G	S	D	R	T	T	E	T

6	0	1	-12	-3	1	0	-1	3	-5

Ολική ή τοπική στοίχιση – Παράδειγμα (3/5)

Τοπική στοίχιση

		M	N	A	L	S	D	R	T
	GAP	0	0	0	0	0	0	0	0
M	-12	6	0	0					
G	-16	0							
S	-20	0							
D	-24								
R	-28								
T	-32								
T	-36								
E	-40								
T	-44								

PAM250

Gap penalty=-12-4(x-1)

M→M: 6
M→N: -2
G→M: -3
S→M: -2
M→A: -1

Ολική ή τοπική στοίχιση – Παράδειγμα (4/5)

Τοπική στοίχιση

		M	N	A	L	S	D	R	T
	GAP	0	0	0	0	0	0	0	0
M	0	6	0	0	4	0	0	0	0
G	0	0	6	1	0	5	1	0	0
S	0	0	1	7	0	2	5	1	1
D	0	0	2	1	3	0	6	4	1
R	0	0	0	0	0	3	0	12	3
T	0	0	0	1	0	1	3	0	15
T	0	0	0	1	0	1	1	2	3
E	0	0	1	0	0	0	4	0	2
T	0	0	0	2	0	1	0	3	3

PAM250

Gap penalty=-12-4(x-1)

Ολική ή τοπική στοίχιση – Παράδειγμα (5/5)

Τοπική στοίχιση

		M	N	A	L	S	D	R	T
	GAP	0	0	0	0	0	0	0	0
M	0	6	0	0	4	0	0	0	0
G	0	0	6	1	0	5	1	0	0
S	0	0	1	7	0	**2**	5	1	1
D	0	0	2	1	3	0	**6**	4	1
R	0	0	0	0	0	3	0	**12**	3
T	0	0	0	1	0	1	3	0	**15**
T	0	0	0	1	0	1	1	2	3
E	0	0	1	0	0	0	4	0	2
T	0	0	0	2	0	1	0	3	3

PAM250
Gap penalty=-12-4(x-1)

S	**D**	**R**	**T**
S	**D**	**R**	**T**

2	4	6	3	15

Τέλος Ενότητας

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημείωμα Αναφοράς

- Copyright Πανεπιστήμιο Δυτικής Μακεδονίας, Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών, Αγγελίδης Παντελής. «**Βιοπληροφορική**». Έκδοση: 1.0. Κοζάνη 2015. Διαθέσιμο από τη δικτυακή διεύθυνση:

  https: //eclass.uowm.gr/courses/ICTE102/

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.