Υπολογιστική Βιολογία : BIO315

[4] Σύγκριση Αλληλουχιών Στοιχίσεις και Ταχείες Αναζητήσεις

Στόχοι του μαθήματος

- Να περιγράψει τη διαδικασία σύγκρισης δύο αλληλουχιών με σκοπό την εξαγωγή του βαθμού ομοιότητάς τους.
- Να περιγράψει τη βασικότερη μεθοδολογία σύγκρισης αλληλουχιών που είναι η στοίχιση.
- Να αναδείξει χαρακτηριστικά της διαδικασίας στοίχισης που σχετίζονται με:
  - Την έκταση της στοίχισης (ολική ή τοπική)
  - Την αξιολόγηση της ομοιότητας με βάση πίνακες βαθμονόμησης
- Να περιγράψει τη μεθοδολογία ταχέων αναζητήσεων (rapid searches) μέσω του συχνότερα χρησιμοποιούμενου προγράμματος βιοπληροφορικής που ονομάζεται BLAST.

Στο τέλος του μαθήματος θα πρέπει να μπορείτε:

- Να διακρίνετε μεταξύ ολικής και τοπικής στοίχισης
- Να υλοποιήσετε μια ολική στοίχιση μέσω του αλγορίθμου Needleman-Wunsch
- Να υλοποιήσετε μια τοπική στοίχιση μέσω του αλγορίθμου Smith-Waterman
- Να εντοπίσετε αλληλουχίες με ομοιότητα σε βάσεις δεδομένων με τη χρήση του BLAST

Δύο αλληλουχίες

```
H. sapiens        MTENSTSAPAAKPKRAKATLL
D. melanogaster   MSDSAVATSASPVAAPPA
```

Εικόνα 4.1: *Τμήματα της αμινοξικής αλληλουχίας της ιστόνης H1 για τον άνθρωπο (H. sapiens) και τη μύγα των φρούτων (D. melanogaster)*

Το βιολογικό πρόβλημα.
Πώς θα συγκρίνουμε τις αλληλουχίες;

Δύο αλληλουχίες μπορούν να συγκριθούν με διάφορους τρόπους.

Απλές προσεγγίσεις όπως η απόσταση Hamming είναι αποτελεσματικές για μικρές αλληλουχίες όπου η υπέρθεση της A1 με την A2 είναι προφανής

Τι συμβαίνει όμως με μεγαλύτερες αλληλουχίες όπως αυτές που αντιστοιχούν π.χ. σε γονίδια, οικογένειες γονιδίων ή και ολόκληρα χρωμοσώματα;

Σε αυτές τις περιπτώσεις η λύση είναι η “στοίχιση”.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     *****: .****:  * *:** * :*****.:* ********..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:************:***:**.: .****************     :  *.: :
```

Πώς θα κάνουμε τη στοίχιση;

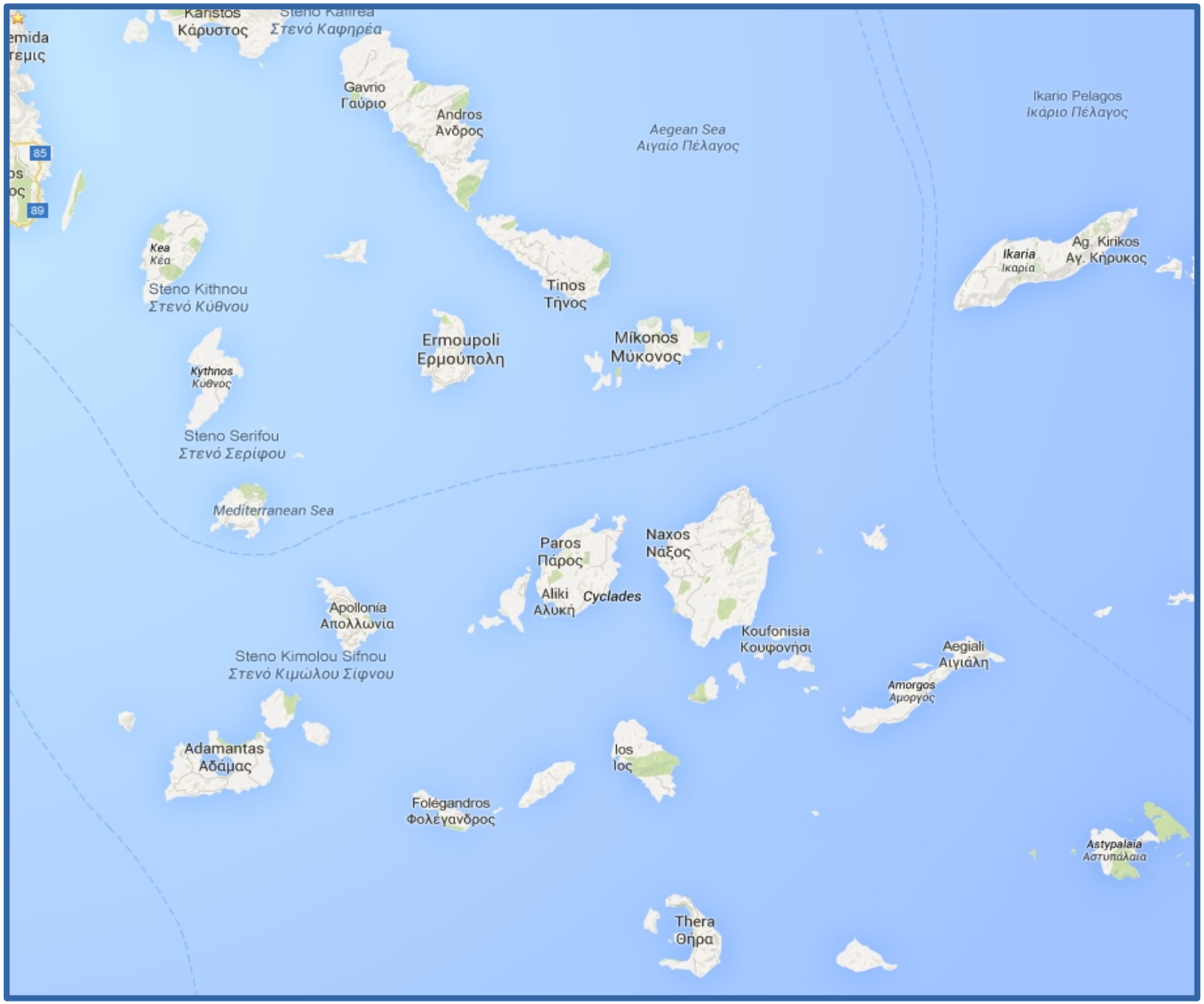
Δύο αλληλουχίες μπορούν να στοιχηθούν με πολλούς τρόπους, ο αριθμός των οποίων αυξάνεται εκρηκτικά με το μήκος τους

Αν προσπαθήσουμε να δούμε την ομοιότητά τους αυτή μπορεί να διαφέρει πολύ ανάλογα με το πώς θα τις στοιχήσουμε. Μια στοίχιση αποδίδει ομοιότητα 1/10 ενώ μια εναλλακτική μπορεί να δώσει έως και 8/12 (επιτρέποντας μετατόπιση των αλληλουχιών).

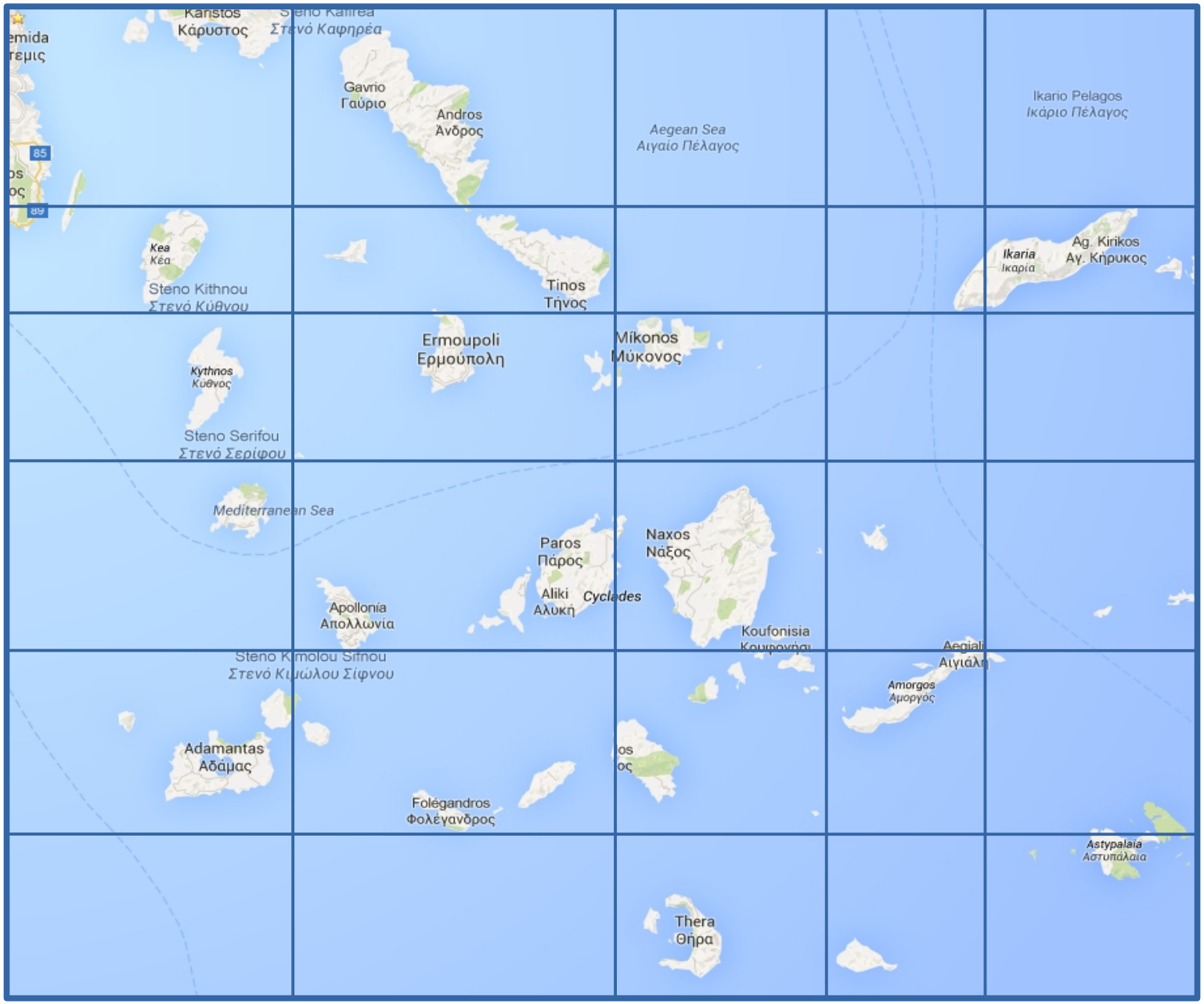
Το ερώτημα γίνεται:

Με ποιον τρόπο μπορούμε να στοιχήσουμε δύο αλληλουχίες ώστε να πετύχουμε το μέγιστο βαθμό ομοιότητας;

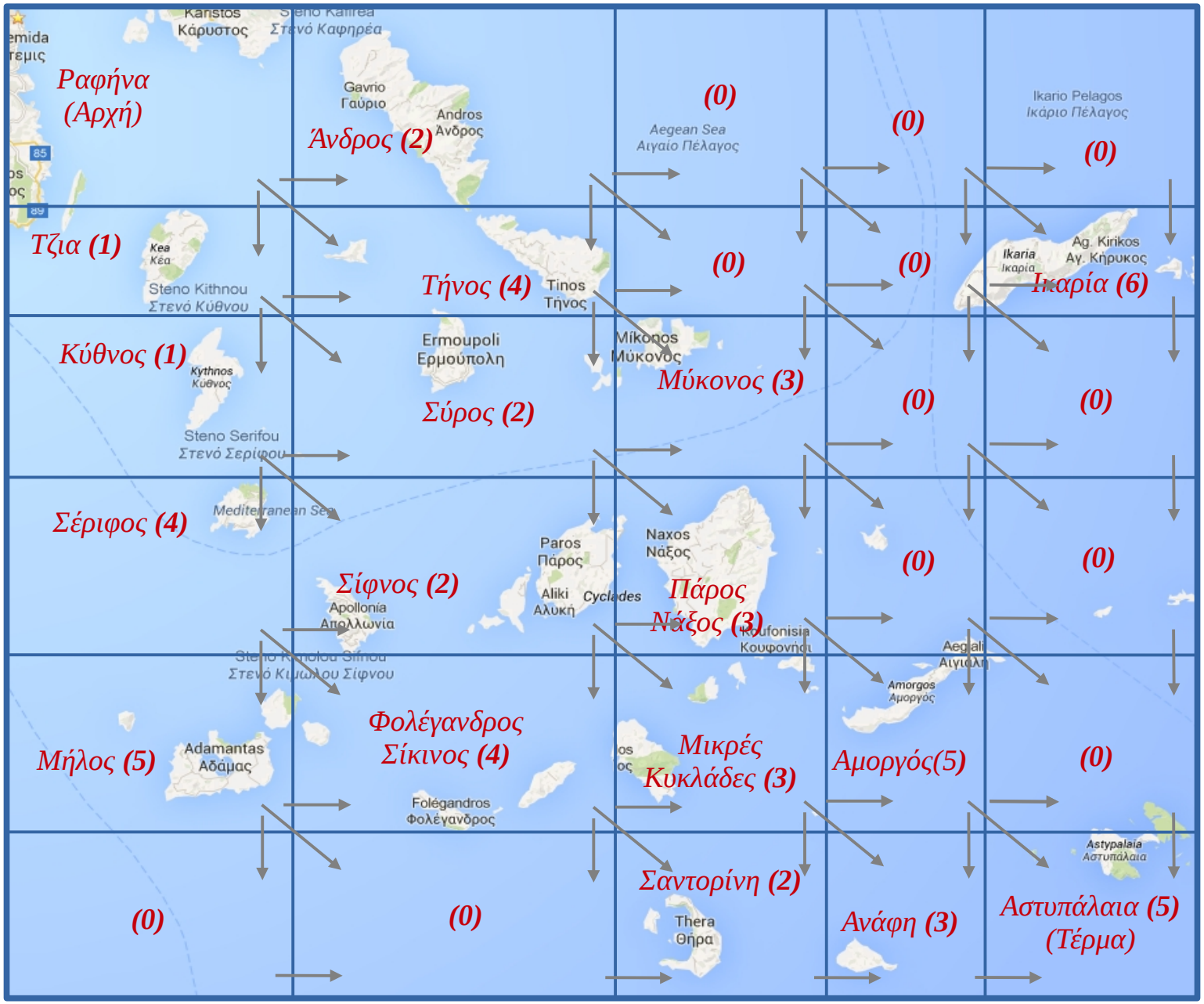
Κρουαζιέρα θα σε πάω (I)



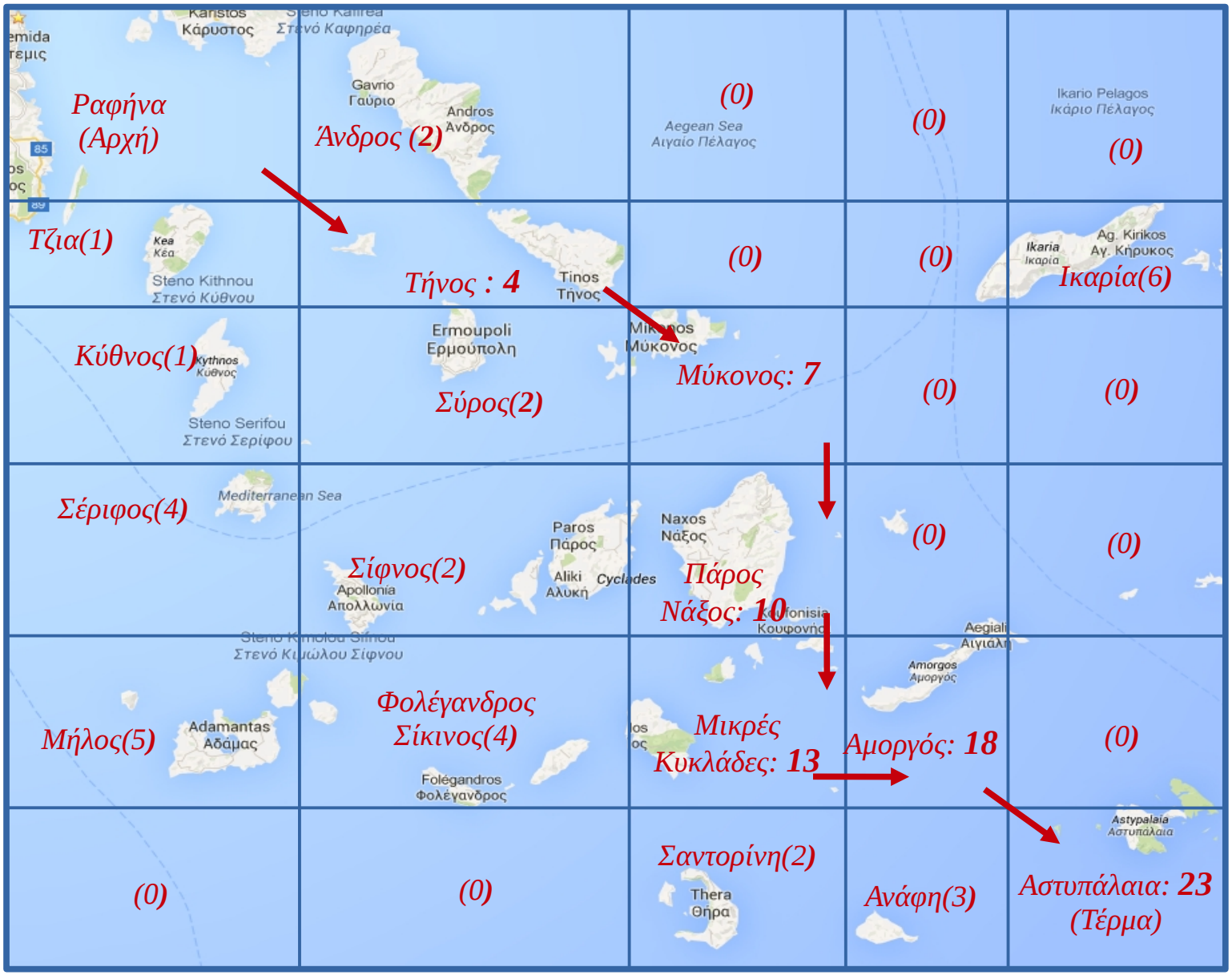
Κρουαζιέρα θα σε πάω (II)



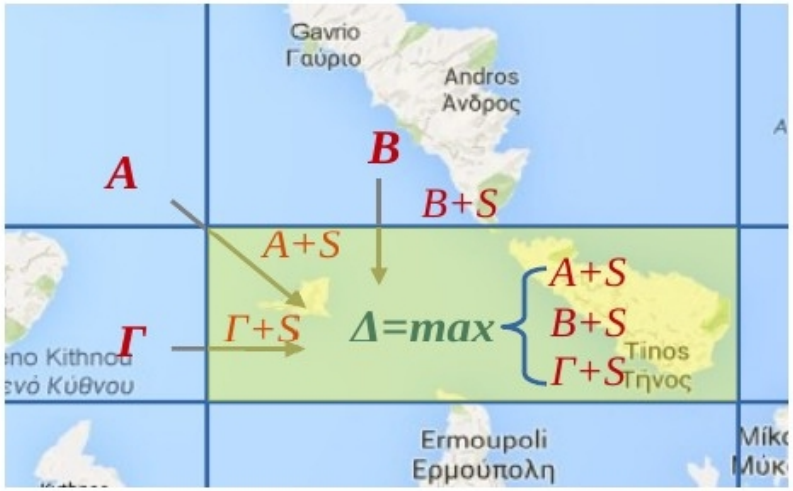
Οργανώνοντας το ταξίδι

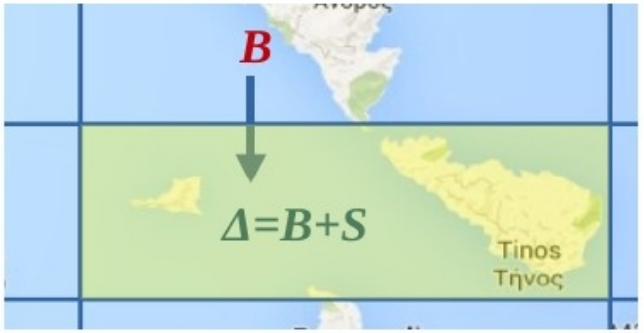


Η “άπληστη” διαδρομή. Είναι η καλύτερη;

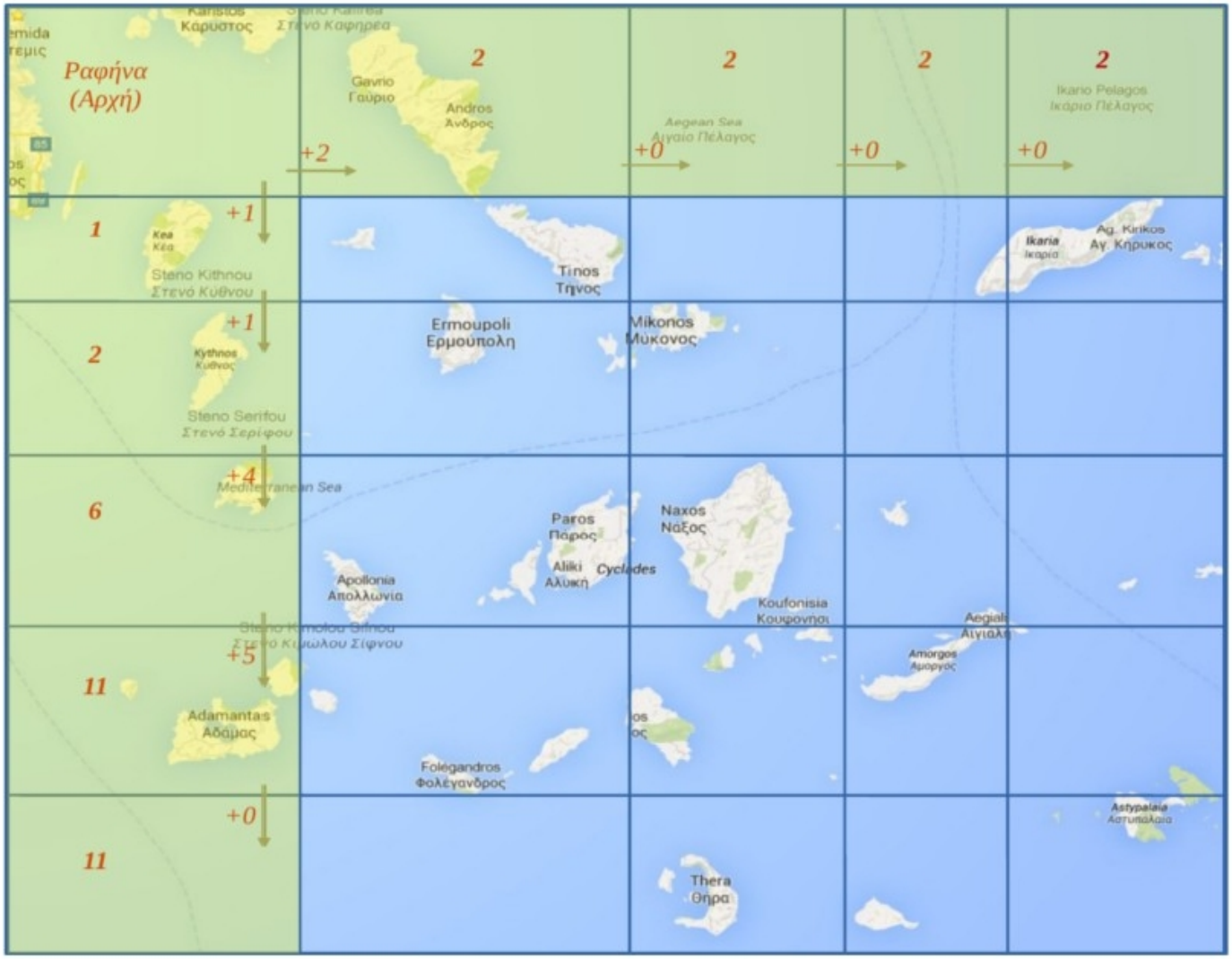


Η “εξαντλητική” προσέγγιση



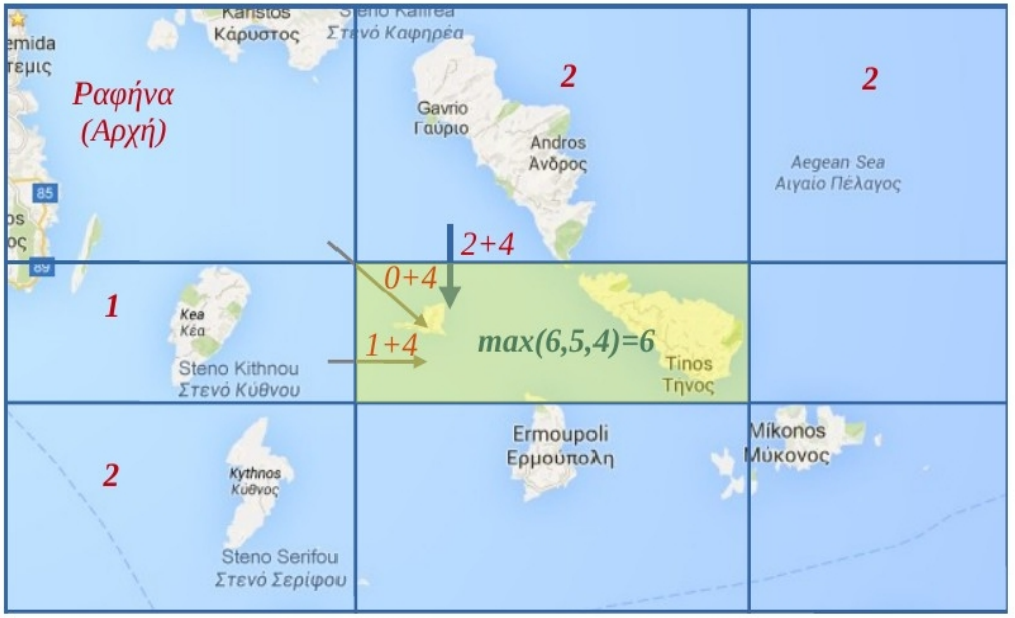


Η “εξαντλητική” προσέγγιση

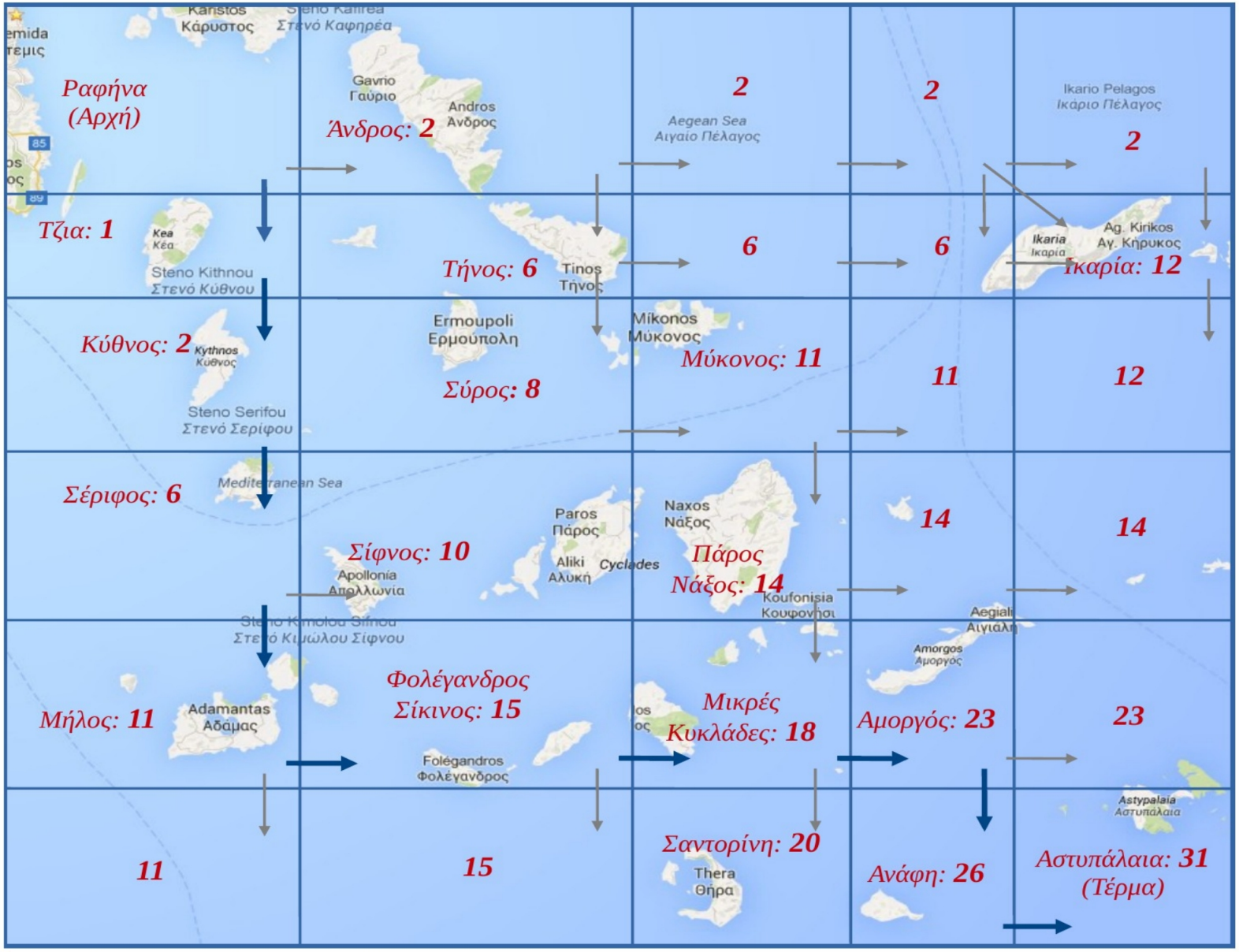


Η “εξαντλητική” προσέγγιση

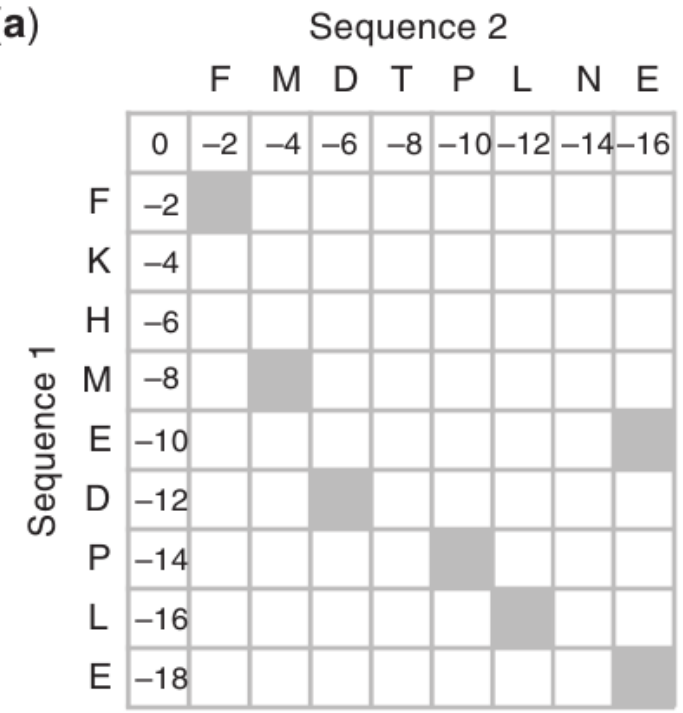
Επαναλαμβάνοντας την απλή διαδικασία επιλογής που είδαμε πιο πάνω μπορούμε να βρούμε ποια είναι η καλύτερη διάδρομη μέχρι οποιοδήποτε σημείο.



Και συνεπώς μέχρι και τον τελικό προορισμό



Ο αλγόριθμος Needlemann-Wunsch (I)

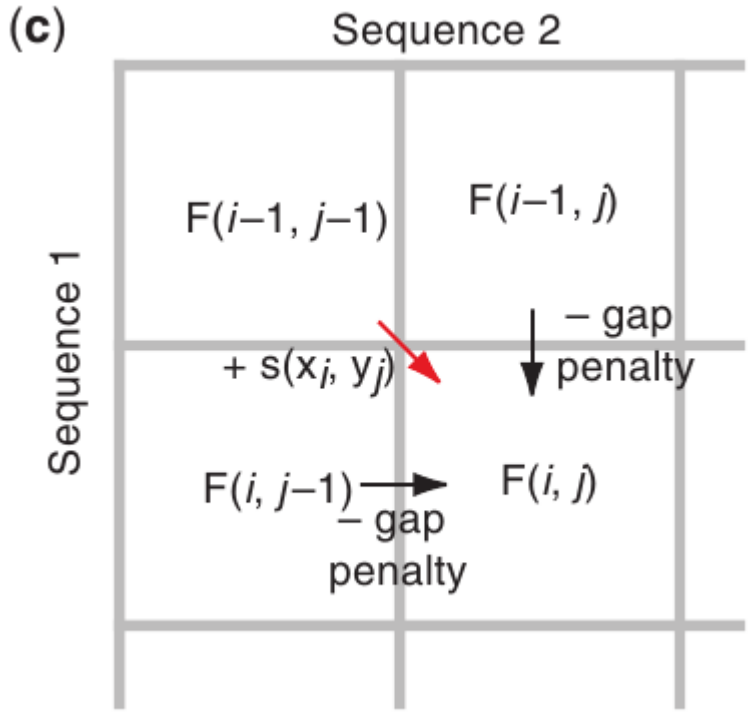


Θα χρησιμοποιήσουμε την τακτική του ταξιδιώτη.

Δημιουργούμε ένα πλέγμα με τις δύο αλληλουχίες στην πρώτη γραμμή και στήλη.

Γεμίζουμε την πρώτη γραμμή και στήλη με τις τιμές που αντιστοιχούν στην εισαγωγή “κενών”.

Ο αλγόριθμος Needlemann-Wunsch (II)



(b)

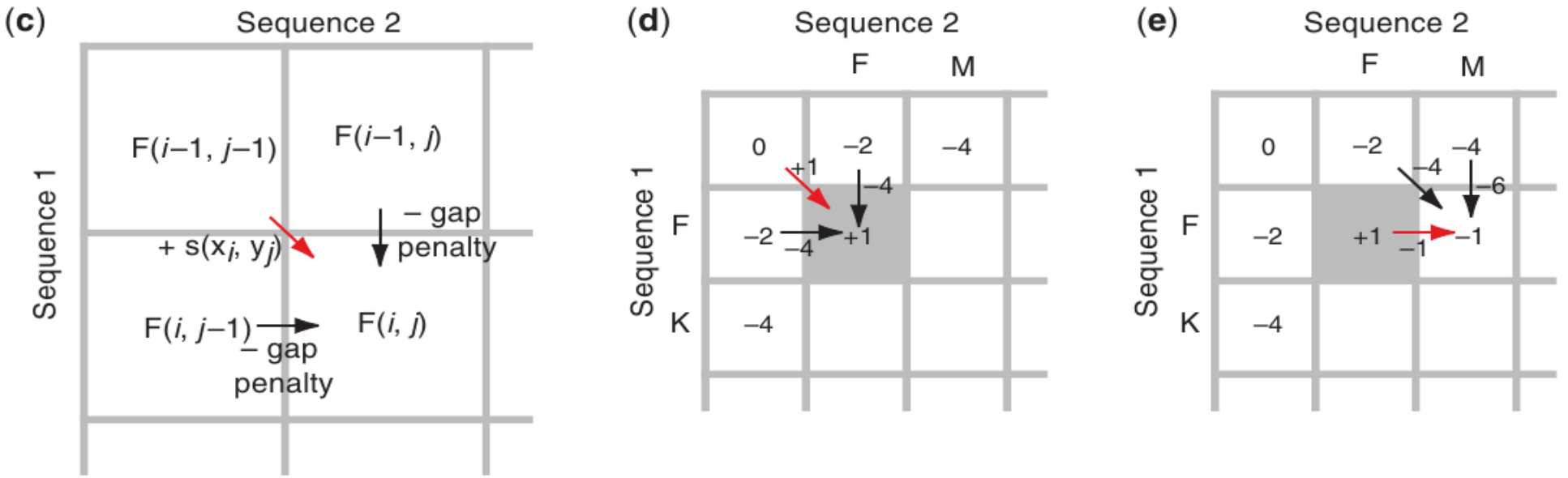
$$\text{Score} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_i) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
–2 (mismatch)
–2 (gap penalty)

Θα χρησιμοποιήσουμε ένα σχήμα βαθμονόμησης που περιέχει τιμές για την ταύτιση, την αντικατάσταση και το κενό

Για κάθε στοιχείο του πίνακα εξετάζουμε όλες τις πιθανές καταστάσεις: α) εισαγωγή κενού στη μία, β) στην άλλη αλληλουχία και γ) στοίχιση μεταξύ των 2, στην τελευταία περίπτωση αποδίδουμε το σκορ που δίνει η στοίχιση ανάλογα με τον **αν στη συγκεκριμένη θέση έχουμε ταύτιση ή όχι**.

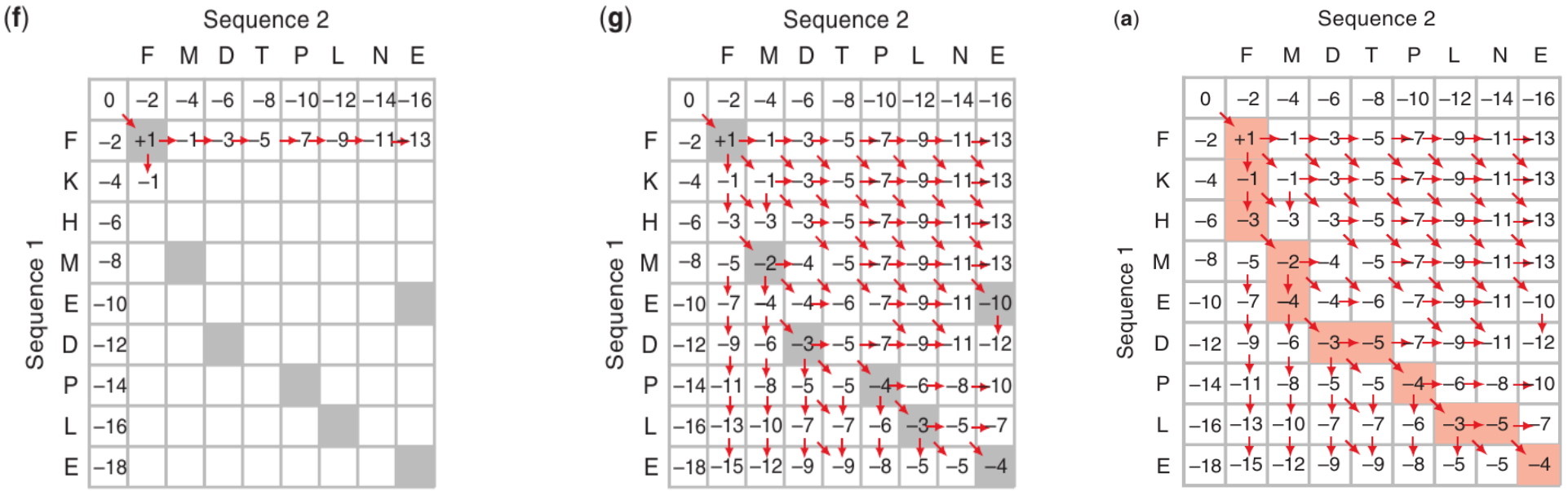
Ο αλγόριθμος Needlemann-Wunsch (III)



Από τις τρεις περιπτώσεις κρατάμε το σκορ που αθροιζόμενο στο προηγούμενο δίνει τη μέγιστη τιμή. Στη συνέχεια **σημειώνουμε το σημείο από το οποίο "φτάσαμε"** στο κάθε επόμενο (αυτό δηλαδή από το οποίο προήλθε η μέγιστη τιμη).

Συμπληρώνουμε με τον ίδιο τρόπο όλον τον πίνακα. Το τελικό σημείο (κάτω δεξιά) περιέχει **το τελικό αποτέλεσμα της στοίχισης.**

Ο αλγόριθμος Needlemann-Wunsch (IV)



Το αποτέλεσμα είναι ένας πίνακας γεμάτος τιμές και “βέλη” (pointers).

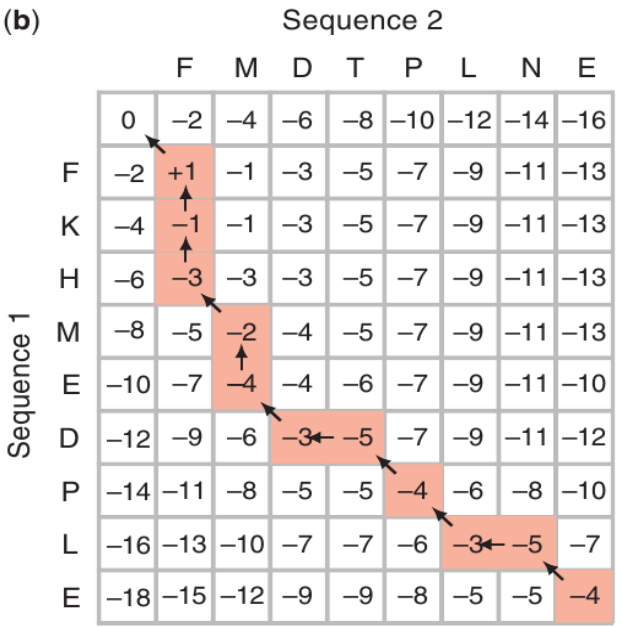
Αυτό που μας ενδιαφέρει είναι να ανατρέξουμε στη σειρά των βελών που μας έφεραν στο τελικό στοιχείο (κάτω δεξιά).

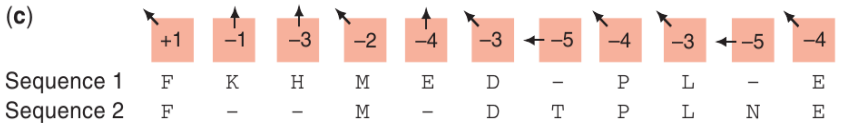
Αυτή η διαδικασία λέγεται αναδρομή (backtracking) και συνίσταται στο να ακολουθήσουμε την αντίστροφη πορεία από το τελικό σημείο προς την αρχή ακολουθώντας τα σημειωμένα “βέλη” που κάθε φορά έδωσαν τη μέγιστη τιμή

Ο αλγόριθμος Needlemann-Wunsch (V)

Η τελική στοίχιση προκύπτει από τον πίνακα, το τελικό σκορ ομοιότητας είναι αυτό που προκύπτει στο κάτω δεξιά στοιχείο του πίνακα.

Η λογική που ακολουθείται είναι αυτή του **δυναμικού προγραμματισμού** και ως συνέπεια εξασφαλίζεται πως το **αποτέλεσμα θα είναι η καλύτερη δυνατή στοίχιση** δεδομένου του σχήματος που περιλαμβάνει τον πίνακα αντικαταστάσεων και τις ποινές κενών.





Από τι εξαρτάται η στοίχιση;

- Η διαδικασία είναι ντετερμινιστική και απολύτως επαναλήψιμη
- Η τελική στοίχιση εξαρτάται μόνο από το σύστημα βαθμονόμησης

Η σημασία των τιμών ταύτισης/αντικατάστασης/κενών

A

START		M	S	D	S	A	V	A	T	S	A	S	P	V	A	A	P	P	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18
M	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16
T	-2	0	0	-1	-2	-3	-4	-5	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14
E	-3	-1	-1	-1	-2	-3	-4	-5	-5	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14
N	-4	-2	-2	-2	-2	-3	-4	-5	-6	-6	-6	-7	-8	-9	-10	-11	-12	-13	-14
S	-5	-3	-1	-2	-1	-2	-3	-4	-5	-5	-6	-5	-6	-7	-8	-9	-10	-11	-12
T	-6	-4	-2	-2	-2	-2	-3	-4	-3	-4	-5	-6	-6	-7	-8	-9	-10	-11	-12
S	-7	-5	-3	-3	-1	-2	-3	-4	-4	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-8	-6	-4	-4	-2	0	-1	-2	-3	-3	-1	-2	-3	-4	-5	-6	-7	-8	-9
P	-9	-7	-5	-5	-3	-1	-1	-2	-3	-4	-2	-2	-1	-2	-3	-4	-5	-6	-7
A	-10	-8	-6	-6	-4	-2	-2	0	-1	-2	-3	-3	-2	-2	-1	-2	-3	-4	-5
A	-11	-9	-7	-7	-5	-3	-3	-1	-1	-2	-1	-2	-3	-3	-1	0	-1	-2	-3
K	-12	-10	-8	-8	-6	-4	-4	-2	-2	-2	-2	-2	-3	-4	-2	-1	-1	-2	-3
P	-13	-11	-9	-9	-7	-5	-5	-3	-3	-3	-3	-3	-1	-2	-3	-2	0	0	-1
K	-14	-12	-10	-10	-8	-6	-6	-4	-4	-4	-4	-4	-2	-2	-3	-3	-1	-1	-1
R	-15	-13	-11	-11	-9	-7	-7	-5	-5	-5	-5	-5	-3	-3	-3	-4	-2	-2	-2
A	-16	-14	-12	-12	-10	-8	-8	-6	-6	-6	-4	-5	-4	-4	-2	-2	-3	-3	-1
K	-17	-15	-13	-13	-11	-9	-9	-7	-7	-7	-5	-5	-5	-5	-3	-3	-3	-4	-2
A	-18	-16	-14	-14	-12	-10	-10	-8	-8	-8	-6	-6	-6	-6	-4	-2	-3	-4	-3
T	-19	-17	-15	-15	-13	-11	-11	-9	-7	-8	-7	-7	-7	-7	-5	-3	-3	-4	-4
L	-20	-18	-16	-16	-14	-12	-12	-10	-8	-8	-8	-8	-8	-8	-6	-4	-4	-4	-5
L	-21	-19	-17	-17	-15	-13	-13	-11	-9	-9	-9	-9	-9	-9	-7	-5	-5	-5	-5

M - - - S D S A V A T S A S P - V A - A P P A
M T E N S T S A P A - - A K P K R A K A T L L

Ταύτιση = 1
Αντικατάσταση = -1
Κενό = -1

Ομοιότητα = 9/18 = 50%

B

START		M	S	D	S	A	V	A	T	S	A	S	P	V	A	A	P	P	A
	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	18
M	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16
T	-2	0	-1	-2	-3	-4	-5	-6	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14
E	-3	-1	-2	-3	-4	-5	-6	-7	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15
N	-4	-2	-3	-4	-5	-6	-7	-8	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16
S	-5	-3	-1	-2	-3	-4	-5	-6	-7	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14
T	-6	-4	-2	-3	-4	-5	-6	-7	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15
S	-7	-5	-3	-4	-2	-3	-4	-5	-6	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13
A	-8	-6	-4	-5	-3	-1	-2	-3	-4	-5	-3	-4	-5	-6	-7	-8	-9	-10	-11
P	-9	-7	-5	-6	-4	-2	-3	-4	-5	-6	-4	-5	-3	-4	-5	-6	-7	-8	-9
A	-10	-8	-6	-7	-5	-3	-4	-2	-3	-4	-5	-6	-4	-5	-3	-4	-5	-6	-7
A	-11	-9	-7	-8	-6	-4	-5	-3	-4	-5	-3	-4	-5	-6	-4	-2	-3	-4	-5
K	-12	-10	-8	-9	-7	-5	-6	-4	-5	-6	-4	-5	-6	-7	-5	-3	-4	-5	-6
P	-13	-11	-9	-10	-8	-6	-7	-5	-6	-7	-5	-6	-4	-5	-6	-4	-2	-3	-4
K	-14	-12	-10	-11	-9	-7	-8	-6	-7	-8	-6	-7	-5	-6	-7	-5	-3	-4	-5
R	-15	-13	-11	-12	-10	-8	-9	-7	-8	-9	-7	-8	-6	-7	-8	-6	-4	-5	-6
A	-16	-14	-12	-13	-11	-9	-10	-8	-9	-10	-8	-9	-7	-8	-6	-7	-5	-6	-4
K	-17	-15	-13	-14	-12	-10	-11	-9	-10	-11	-9	-10	-8	-9	-7	-8	-6	-7	-5
A	-18	-16	-14	-15	-13	-11	-12	-10	-11	-12	-10	-11	-9	-10	-8	-6	-7	-8	-6
T	-19	-17	-15	-16	-14	-12	-13	-11	-9	-10	-11	-12	-10	-11	-9	-7	-8	-9	-7
L	-20	-18	-16	-17	-15	-13	-14	-12	-10	-11	-12	-13	-11	-12	-10	-8	-9	-10	-8
L	-21	-19	-17	-18	-16	-14	-15	-13	-11	-12	-13	-14	-12	-13	-11	-9	-10	-11	-9

M - S D S A V A T S A S P V A A P - - - P A - - -
M T E N S - - - T S A - P - A A P K R A K A T L L

Ταύτιση=1
Αντικατάσταση=-2
Κενό= -1

Ομοιότητα=10/18=55.6%

Πίνακες αντικατάστασης

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Ζευγαρωτή Τοπική Στοίχιση

Τι συμβαίνει στις περιπτώσεις που:

- Οι δύο αλληλουχίες διαφέρουν αρκετά σε μήκος
- Μας ενδιαφέρει μόνο ένα τμήμα που έχει μεγάλη ομοιότητα

Σε περιπτώσεις σαν τις παραπάνω εφαρμόζουμε μια παραλλαγή του NW αλγορίθμου για να πετύχουμε μια **τοπική στοίχιση.** Αυτή θα αντιστοιχεί:

Σε μια στοίχιση που η μία (η μικρότερη) αλληλουχία έχει στοιχηθεί έναντι της μεγαλύτερης

Μόνο τα τμήματα με μεγάλη ομοιότητα από κάθε αλληλουχία έχουν στοιχηθεί

Ο αλγόριθμος Smith-Waterman

Ακολουθεί την ίδια λογική με τον αλγόριθμο των NW ωστόσο έχει δύο βασικές διαφορές:

Τα στοιχεία του πίνακα που δίνουν αρνητικές τιμές **μηδενίζονται**.

Η αναδρομή δεν ξεκινά από το κάτω δεξιά στοιχείο αλλά από αυτό **με τη μεγαλύτερη τιμη** (σημείο Α)

Η αναδρομή συνεχίζεται **ως το πρώτο μηδενικό στοιχείο** που συναντούμε (σημείο Β)

Η τελική στοίχιση είναι έτσι **μερική** και καλύπτει το τμήμα από το σημείο Α στο Β.

(a)

		Sequence 1													
			C	A	G	C	C	U	C	G	C	U	U	A	G
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
	U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
	G	0.0	0.0	0.0	**1.0**	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
	C	0.0	1.0	0.0	0.0	**2.0**	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
	C	0.0	1.0	0.7	0.0	1.0	**3.0**	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
Sequence 2	A	0.0	0.0	2.0	0.7	0.3	**1.7**	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
	U	0.0	0.0	0.7	1.7	0.3	1.3	**2.7**	2.3	1.0	0.7	1.7	2.0	1.0	1.0
	U	0.0	0.0	0.3	0.3	1.3	1.0	2.3	**2.3**	2.0	0.7	1.7	2.7	1.7	1.0
	G	0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	**3.3**	2.0	1.7	1.3	2.3	2.7
	A	0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
	C	0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
	G	0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
	G	0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

(b)

```
sequence 1   GCC-UCG
sequence 2   GCCAUUG
```

(c)

```
sequence 1   CAGCC-UCGCUUAG
sequence 2   AAUGCCAUUGACGG
```

Ολική και Τοπική στοίχιση κατά ζεύγη
Αλγόριθμοι NW(Needleman-Wunsch) και SW (Smith-Waterman)

NW-Alignment/BLOSUM62

```
H.sapiens          1 ----------------------MTENSTSAPAAKPKRAKATLLSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKVGENADSQ-----IKLSIKRLVTTGVLKQTKGVGASGSFRLA----KSDEPK--------------K    103
                                           :.:.::.:...|.|:|.||   ..||....|:.|:|:..|.|.|||..:|:|||.:.||    .|:|     ||..:|..|..|.|.||||.||||||:|:    |..:||              |
D.melanogaster     1 MSDSAVATSASPVAAPPATVEKKVVQKKASGSAGTKAKKASAT---PSHPPTQQMVDASIKNLKERGGSSLLAIKKYITATYK----CDAQKLAPFIKKYLKSAVVNGKLIQTKGKGASGSFKLSASAKKEKDPKAKSKVLSAEKKVQSK    143

H.sapiens        104 SVAFKKTKKEIKKVAT--------PKKASKPKKAASKAPTKKPKATPVKKA---KKKLAATPKK--AKKPKTV-----KAKPVKASKPKK--------AKPVKPKAKSSAKRAGKKK    194
                     .||.||.....||.|.        .|||...||.|....|:|.||...||.   |.|.|||..|  |.|||.|     ||||..::||||        |...|||||::|.:.
D.melanogaster   144 KVASKKIGVSSKKTAVGAADKKPKAKKAVATKKTAENKKTEKAKAKDAKKTGIIKSKPAATKAKVTAAKPKAVVAKASKAKPAVSAKPKKTVKKASVSATAKKPKAKTTAAKK----    256
```

```
# Length: 267
# Identity:       89/267 (33.3%)
# Similarity:    112/267 (41.9%)
# Gaps:           84/267 (31.5%)
# Score: 280.0
```

SW-Alignment/BLOSUM62

```
H.sapiens          3 ENSTSAPAAKPKRAKATLLSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKVGENADSQ-----IKLSIKRLVTTGVLKQTKGVGASGSFRLA----KSDEPK--------------KSVAFKKTKKEIKKVAT--------PK    121
                     :.:::.:...|.|:|.||   ..||....|:.|:|:..|.|.|||..:|:|||.:.||    .|:|     ||..:|..|..|.|.||||.||||||:|:    |..:||              |.||.||.....||.|.       .|
D.melanogaster    27 KKASGSAGTKAKKASAT---PSHPPTQQMVDASIKNLKERGGSSLLAIKKYITATYK----CDAQKLAPFIKKYLKSAVVNGKLIQTKGKGASGSFKLSASAKKEKDPKAKSKVLSAEKKVQSKKVASKKIGVSSKKTAVGAADKKPKAK    169

H.sapiens        122 KASKPKKAASKAPTKKPKATPVKKA---KKKLAATPKK--AKKPKTV-----KAKPVKASKPKK--------AKPVKPKAKSSAKR    189
                     ||...||.|....|:|.||...||.   |.|.|||..|  |.|||.|     ||||..::||||        |...|||||::|.:
D.melanogaster   170 KAVATKKTAENKKTEKAKAKDAKKTGIIKSKPAATKAKVTAAKPKAVVAKASKAKPAVSAKPKKTVKKASVSATAKKPKAKTTAAK    255
```

```
# Length: 236
# Identity:       89/236 (37.7%)
# Similarity:    111/236 (47.0%)
# Gaps:           56/236 (23.7%)
# Score: 281.0
```

Εικόνα 4.8: *α) Ολική και β) Τοπική στοίχιση των αλληλουχιών της ιστόνης H4 του ανθρώπου και της D. melanogaster. Η στοίχιση έγινε μέσω της διαδικτυακής εφαρμογής του European Bioinformatics Institute (EBI) http://www.ebi.ac.uk/Tools/psa*

Ταχείες Αναζητήσεις

Πρόκειται για προσεγγίσεις που έχουν σκοπό την γρήγορη αναζήτηση ομοιοτήτων.

Το βάρος δίνεται στην ταχύτητα, έτσι δεν μας ενδιαφέρει τόσο το score της κάθε στοίχισης ή αν είναι το μέγιστο δυνατό.

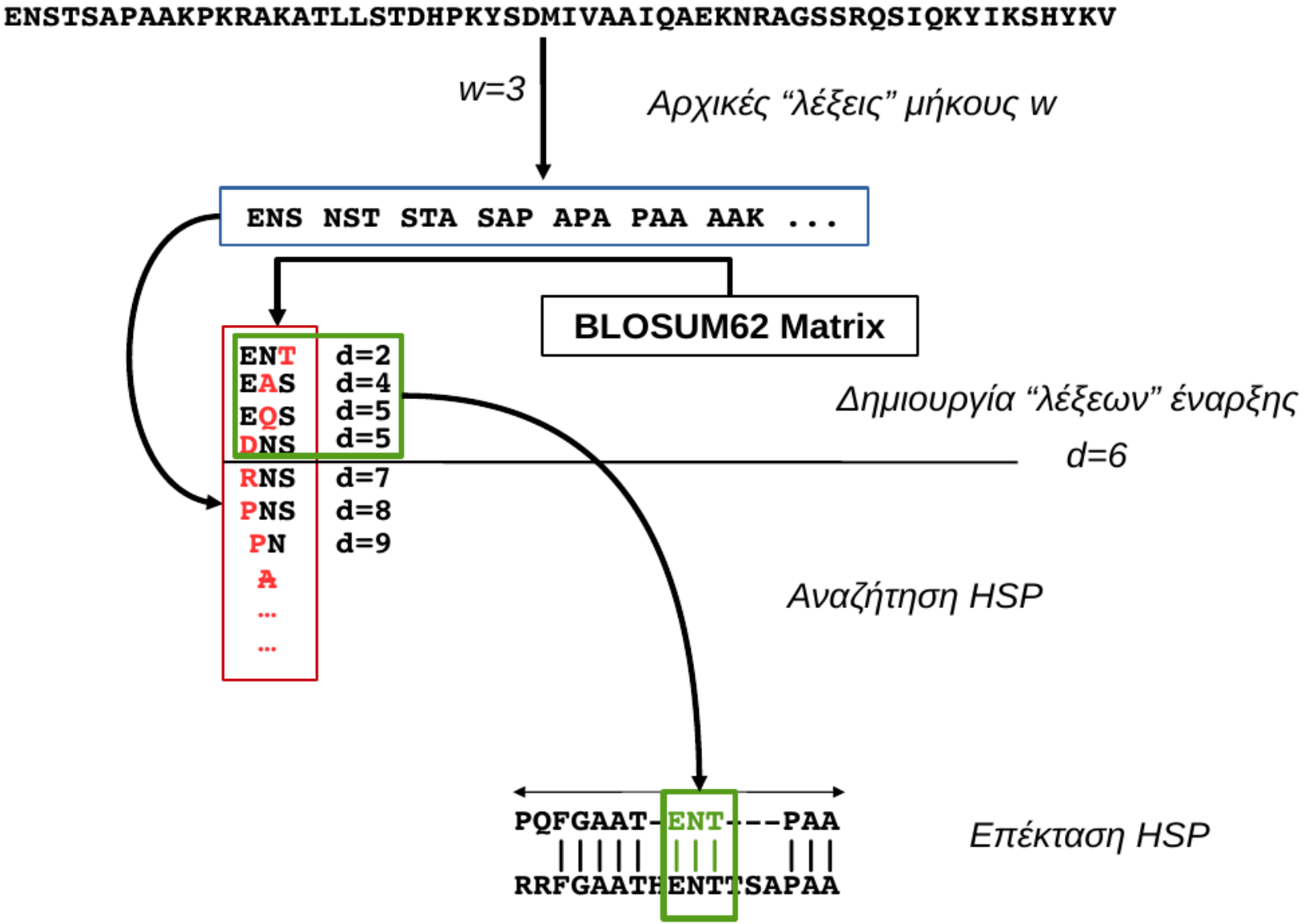
Μας ενδιαφέρει περισσότερο η ταχεία αποκομιδή παρόμοιων αλληλουχιών για περαιτέρω ανάλυση.

Με προσεκτική επιλογή παραμέτρων μπορούμε να πετύχουμε έναν ικανοποιητικό συνδυασμό ταχύτητας και ακρίβειας.

BLAST! Το πιο συχνά χρησιμοποιούμενο εργαλείο βιοπληροφορικής

- Είναι το πιο ευρέως χρησιμοποιούμενο πρόγραμμα βιοπληροφορικής.
- Χρησιμοποιείται από χιλιάδες χρήστες καθημερινά λόγω της ταχύτητας του.
- Δεδομένης μιας αλληλουχίας αποδίδει ταχύτατα τις πιο κοντινές της από πλευράς ομοιότητας μέσα από μια βάση δεδομένων

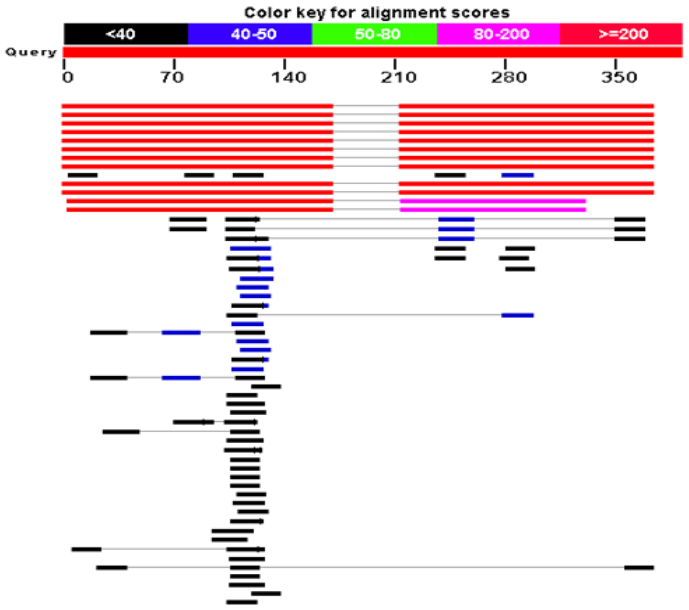
Ταχείες αναζητήσεις αλληλουχιών σε βάσεις δεδομένων. Ο αλγόριθμος του BLAST



Ο αλγόριθμος του BLAST (II)

Τα αποτελέσματα εμφανίζονται σε μορφή στοίχισης έχοντας ενσωματωμένη τη στατιστική σημασία κάθε στοίχισης και σχετικά στοιχεία σε score, μήκος στοίχισης και E-value

Το E-value είναι η εκτίμηση για το πόσο πιθανό είναι να βρεθεί το σκορ ομοιότητας σε αυτό το ύψος δεδομένων των μηκών της αλληλουχίας αναζήτησης και του συνολικού αριθμού αλληλουχιών στη βάση δεδομένων.



Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value	
ref|NP_058652.1|	hemoglobin, beta adult minor chain [Mus musculu	244	2e-65	U G
ref|NP_032246.2|	hemoglobin, beta adult major chain [Mus musculu	228	2e-60	U G
ref|XP_978992.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	226	3e-60	G
ref|NP_032247.1|	hemoglobin Y, beta-like embryonic chain [Mus mu	223	4e-59	U G
ref|NP_032245.1|	hemoglobin Z, beta-like embryonic chain [Mus mu	223	6e-59	U G
ref|XP_998314.1|	PREDICTED: similar to Hemoglobin beta-H1 sub...	203	4e-53	G
ref|XP_978924.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	187	2e-48	G
ref|XP_912634.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	161	2e-40	G
ref|XP_488069.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	154	3e-38	U G
ref|NP_032244.1|	hemoglobin alpha 1 chain [Mus musculus]	105	1e-23	U G
ref|XP_994669.1|	PREDICTED: similar to Hemoglobin alpha subun...	101	3e-22	G
ref|XP_356935.3|	PREDICTED: similar to Hemoglobin alpha subun...	100	4e-22	U G
ref|NP_034535.1|	hemoglobin X, alpha-like embryonic chain in ...	94.0	4e-20	U G
ref|NP_001029153.1|	similar to hemoglobin, theta 1 [Mus musculus	88.2	2e-18	U G
ref|NP_778165.1|	hemoglobin, theta 1 [Mus musculus]	73.9	5e-14	U G
ref|XP_978150.1|	PREDICTED: similar to hemoglobin, beta adult...	41.6	2e-04	G
ref|NP_795942.2|	5'-nucleotidase, cytosolic II-like 1 protein [M	28.9	1.5	U G

BLAST E-value

Η τιμή E (Expect-value) είναι η τιμή που περιγράφει τον αριθμό των hits (αποτελεσμάτων) που περιμένουμε να έχουμε με δεδομένων των:

S = σκόρ ομοιότητας

m = μήκος αλληλουχίας

n = μήκος αλληλουχιών στη βάση δεδομένων

με K,λ = σταθερές παραμέτρους

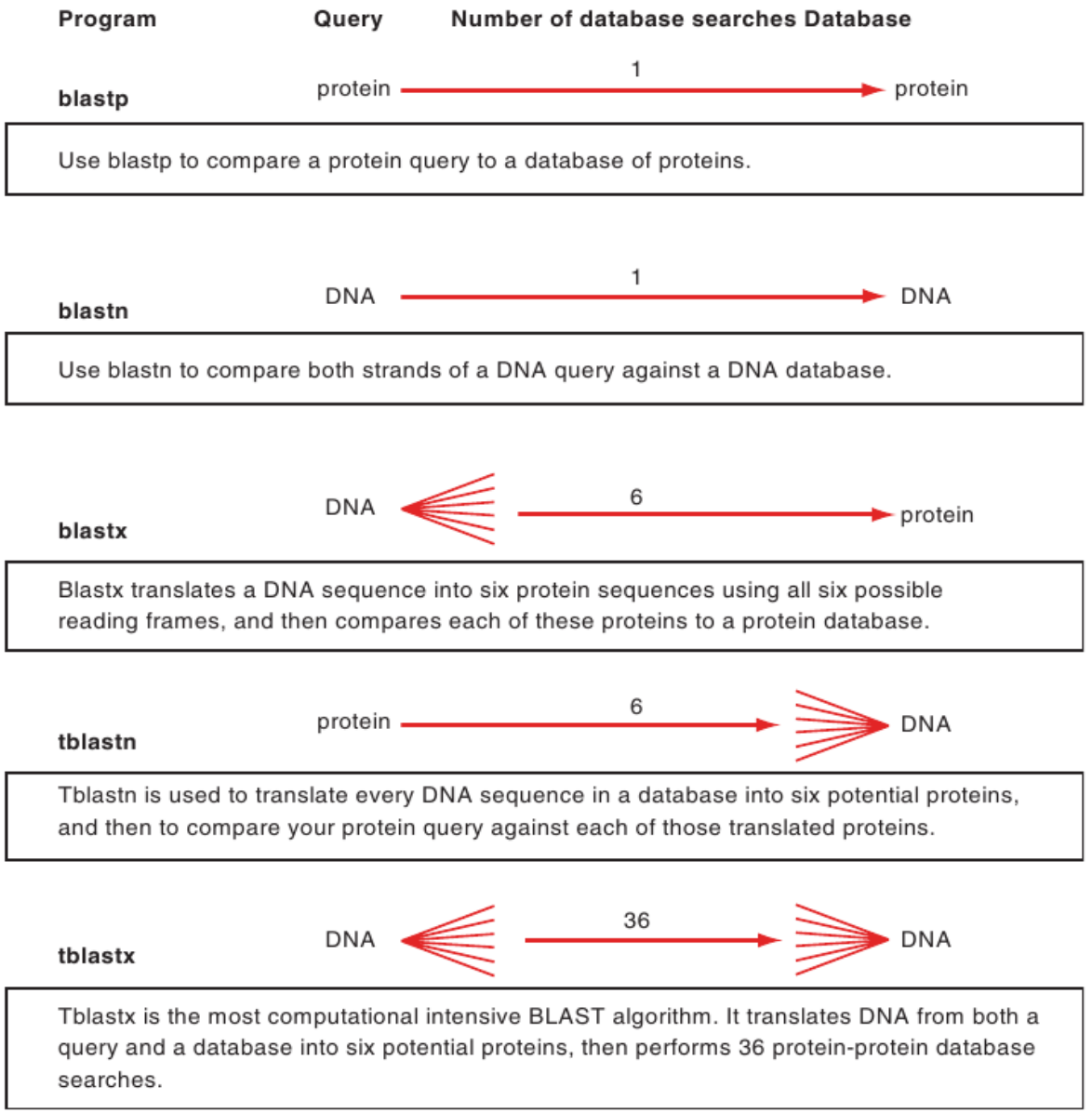
Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value	
ref|NP_058652.1|	hemoglobin, beta adult minor chain [Mus musculu	244	2e-65	U G
ref|NP_032246.2|	hemoglobin, beta adult major chain [Mus musculu	228	2e-60	U G
ref|XP_978992.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	226	3e-60	G
ref|NP_032247.1|	hemoglobin Y, beta-like embryonic chain [Mus mu	223	4e-59	U G
ref|NP_032245.1|	hemoglobin Z, beta-like embryonic chain [Mus mu	223	6e-59	U G
ref|XP_998314.1|	PREDICTED: similar to Hemoglobin beta-H1 sub...	203	4e-53	G
ref|XP_978924.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	187	2e-48	G
ref|XP_912634.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	161	2e-40	G
ref|XP_488069.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	154	3e-38	U G
ref|NP_032244.1|	hemoglobin alpha 1 chain [Mus musculus]	105	1e-23	U G
ref|XP_994669.1|	PREDICTED: similar to Hemoglobin alpha subun...	101	3e-22	G
ref|XP_356935.3|	PREDICTED: similar to Hemoglobin alpha subun...	100	4e-22	U G
ref|NP_034535.1|	hemoglobin X, alpha-like embryonic chain in ...	94.0	4e-20	U G
ref|NP_001029153.1|	similar to hemoglobin, theta 1 [Mus musculus	88.2	2e-18	U G
ref|NP_778165.1|	hemoglobin, theta 1 [Mus musculus]	73.9	5e-14	U G
ref|XP_978150.1|	PREDICTED: similar to hemoglobin, beta adult...	41.6	2e-04	G
ref|NP_795942.2|	5'-nucleotidase, cytosolic II-like 1 protein [M	28.9	1.5	U G

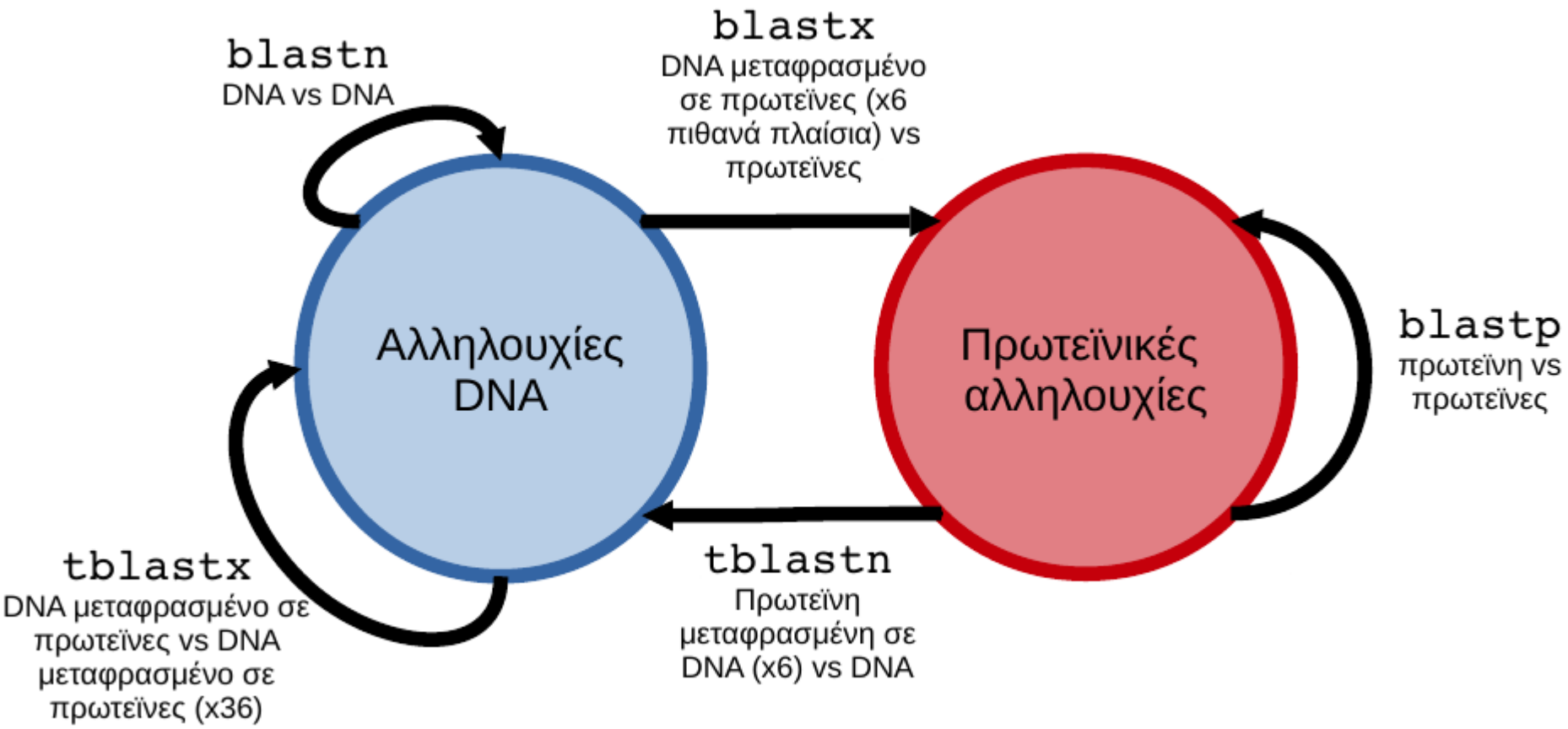
$$E = Kmne^{-(\lambda S)}$$

Παραλλαγές του BLAST

Ανάλογα με το είδος της αλληλουχίας αναζήτησης και της βάσης δεδομένων στόχου.

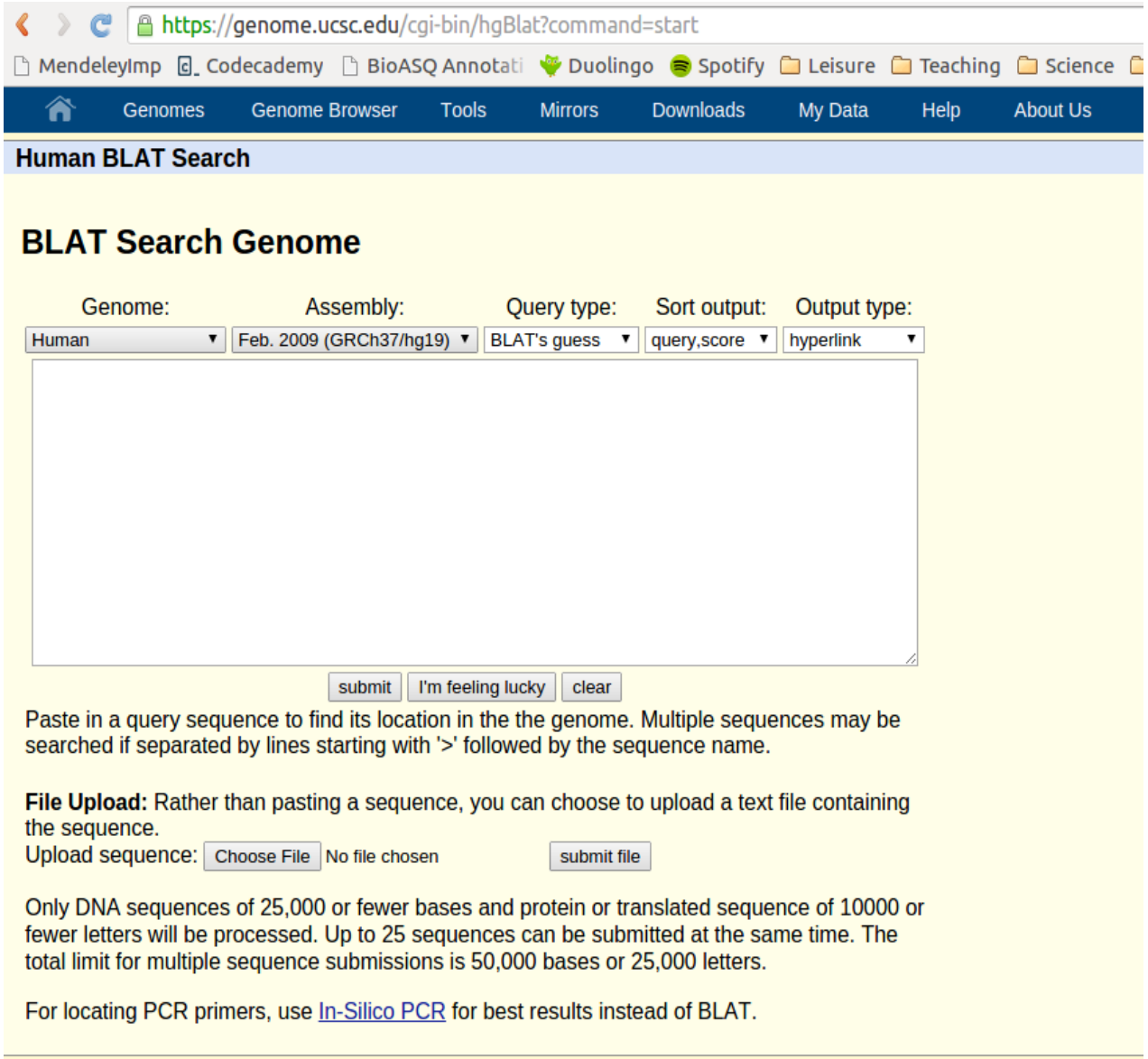


Παραλλαγές του BLAST



BLAT

- Το BLAT χρησιμοποιεί τη λογική του BLAST για την αναζήτηση αλληλουχιών μέσα σε ένα συγκεκριμένο γονιδίωμα αντί για το σύνολο των γνωστών αλληλουχιών
- Το πλεονέκτημά του είναι ότι αποθηκεύει ολόκληρο το γονιδίωμα σε μια δομή που μοιάζει με δέντρο καταλήξεων κάνοντας την αναζήτηση πολύ πιο γρήγορη.



Διαβάστε περισσότερα

Για τη στοίχιση Needleman-Wunsch

Το Κεφάλαιο 6 του An Introduction to Bioinformatics Algorithms των Pevzner & Jones αναφέρεται όχι μόνο στο πρόβλημα της στοίχισης αλλά γενικότερα στις εφαρμογές αλγορίθμων Δυναμικού Προγραμματισμού για βιολογικά προβλήματα. Το πρόβλημα του "Ταξιδιώτη στο Αιγαίο" αποτελεί παραλλαγή του αντίστοιχου του "Τουρίστα στο Μανχάταν" που περιγράφεται εκεί.

Στο Κεφάλαιο 2 του "Ανάλυση Βιολογικών Αλληλουχιών" Durbin et al (Εκδόσεις Πεδίο) περιέχεται μια πολύ αναλυτική περιγραφή του προβλήματος αλλά και της διαδικασίας.

Για τη στοίχιση Smith-Waterman

Στο ίδιο με το παραπάνω (Κεφάλαιο 2 Durbin et al)

Για τους πίνακες αντικατάστασης:

Τα δύο βασικά άρθρα για τους πίνακες PAM και BLOSUM αντίστοιχα είναι τα (Dayhoff, Schwartz, and Orcutt 1978) και (Henikoff and Henikoff 1992). Το Κεφάλαιο 3 του Bioinformatics and Functional Genomics (Pevsner 2015) συζητάει σε βάθος τόσο την ιστορία της ανάπτυξής τους όσο και τις λεπτές αλλά σημαντικές διαφορές μεταξύ των διαφορετικών πινάκων.

Για το BLAST:

Η βασική εργασία των (Altschul et al. 1990) αποτελεί μια από τις πιο συχνά αναφερόμενες εργασίες παγκοσμίως ανεξάρτητα από επιστημονικό πεδίο. Πιο αναλυτικές πληροφορίες μπορούν να βρεθούν στο Κεφάλαιο 4 του Bioinformatics and Functional Genomics (Pevsner 2015) όπου ο συγγραφέας αφιερώνει ένα ολόκληρο Κεφάλαιο στο BLAST και κυρίως στην πρακτική του εφαρμογή. Από τον ειδικευμένο σε τεχνικά εγχειρίδια εκδότη O' Reilly Press, κυκλοφορεί ένα τεχνικό βιβλίο που είναι αφιερωμένο αποκλειστικά σε αυτό (Korf, Yandell, and Bedell 2003)