Σύγκριση 2 ακολουθιών

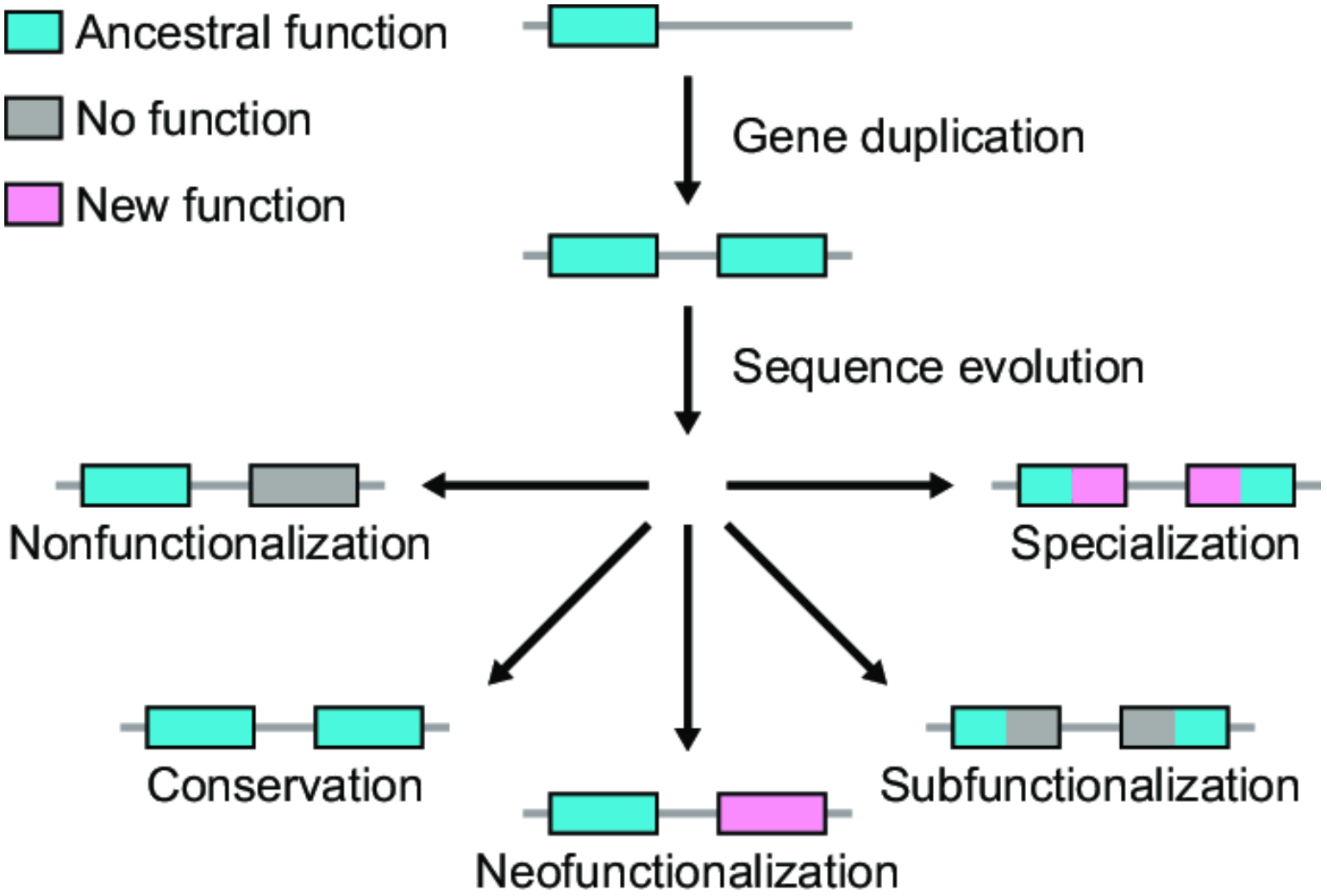
Στιγμοπίνακες (Dotplots)
Στοίχιση ακολουθιών κατά ζεύγη (Pairwise alignment)

Αναζήτηση ομόλογων ακολουθιών

Γρηγόριος Αμούτζιας
Καθηγητής Βιοπληροφορικής με έμφαση στη Μικροβιολογία
Τμήμα Βιοχημείας και Βιοτεχνολογίας Π.Θ.

Τα είδη & οι ακολουθίες των γονιδιωμάτων τους εξελίσσονται

Gene duplication

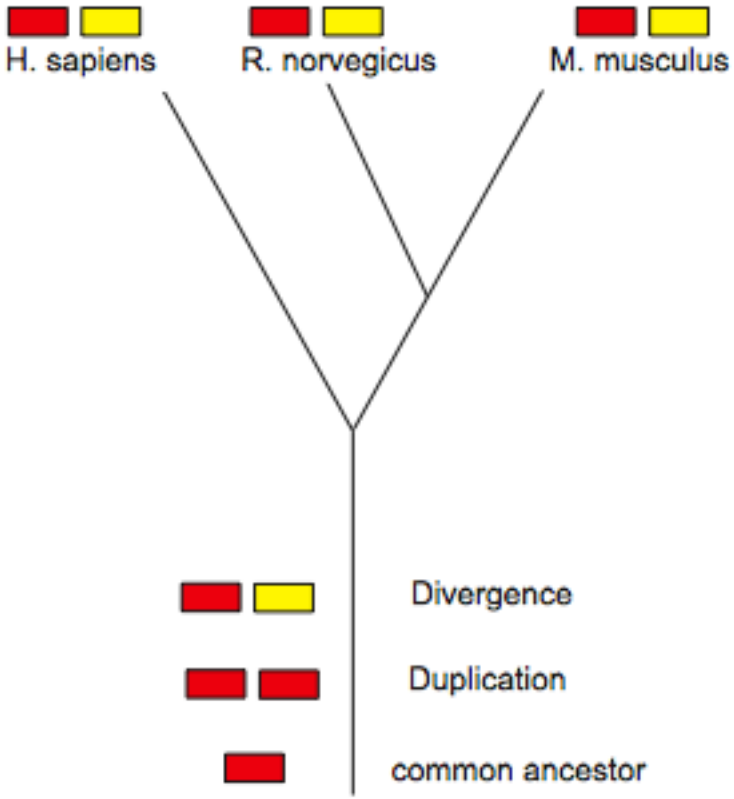


DeGiorgio and Assis, 2021. Mol Biol Evol. Mar 9;38(3):1209-1224. doi: 10.1093/molbev/msaa267.

Λίγη εξέλιξη: ομολογία

- Ομόλογα γονίδια: κοινός εξελικτικός πρόγονος. Χιμαιρικές πρωτεΐνες;
- Ορθόλογα γονίδια: προέρχονται από ειδογένεση. Ουσιαστικά, ένα γονίδιο α (μεταλλαγμένο) σε δύο διαφορετικούς οργανισμούς. Συχνά έχουν την ίδια λειτουργία
- Παράλογα γονίδια: προέρχονται από γονιδιακό διπλασιασμό. Ανήκουν στην ίδια οικογένεια
- Ξενόλογα γονίδια: από οριζόντια μεταφορά
- Παράδειγμα με Πυρηνικούς υποδοχείς

Λίγη εξέλιξη: ομολογία (ii)



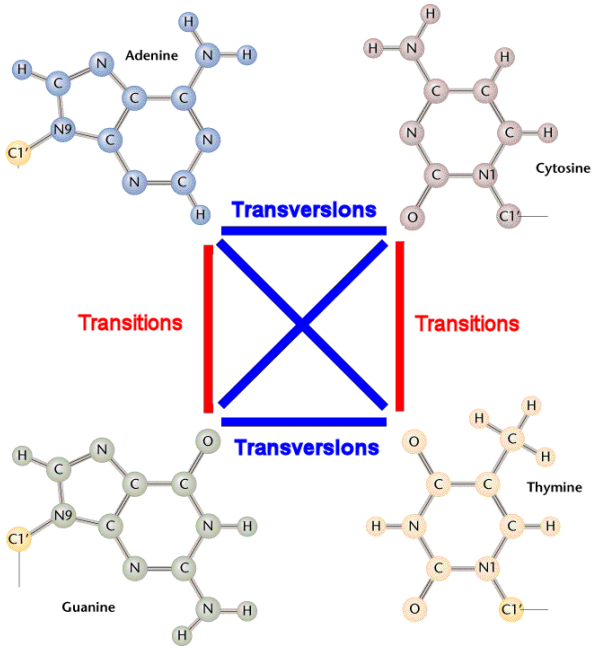
- Γονίδια του ίδιου χρώματος από διαφορετικούς οργανισμούς είναι ορθόλογα.
- Το κόκκινο και το κίτρινο από ένα οργανισμό είναι παράλογα.
- Το κόκκινο από ένα οργανισμό και το κίτρινο από ένα άλλο οργανισμό είναι έξτρα-παράλογα

Βασικότερα είδη μεταλλάξεων

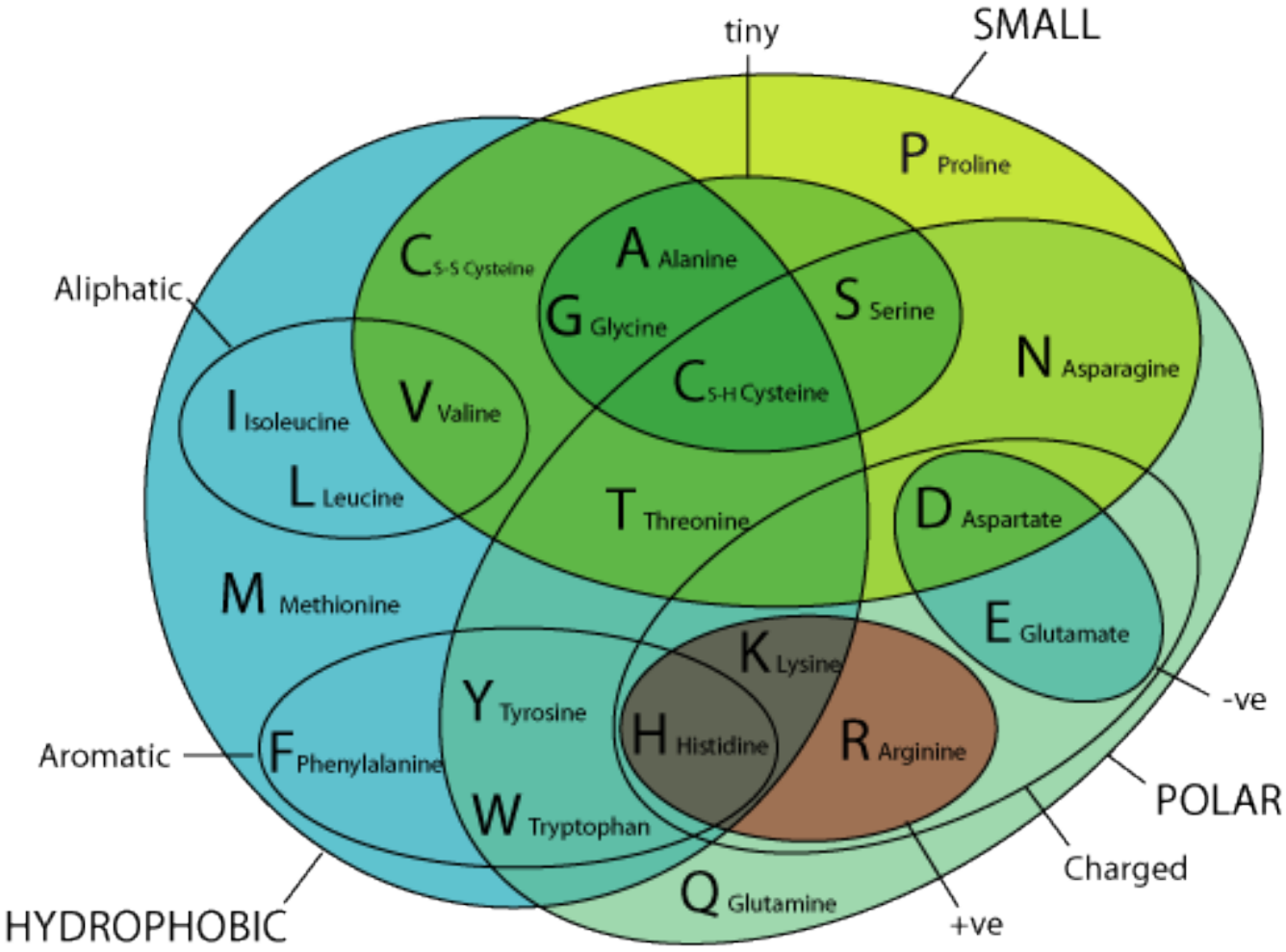
- Μεταλλάξεις σημείου (point mutations)
  - Συνώνυμες (synonymous)
  - Μη-συνώνυμες (non-synonymous)
    - Αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες
    - Αμινοξέα με διαφορετικές φυσικοχημικές ιδιότητες
  - Κωδικόνια τερματισμού

Μεταπτώσεις-μεταστροφές

- Μεταπτώσεις (Transitions)
  - Δημιουργούνται με μεγαλύτερη συχνότητα
  - Συνήθως οδηγούν σε συνώνυμες μεταλλάξεις
  - Είναι πιο συχνές στα SNPs



Κατηγοριοποίηση αμινοξέων

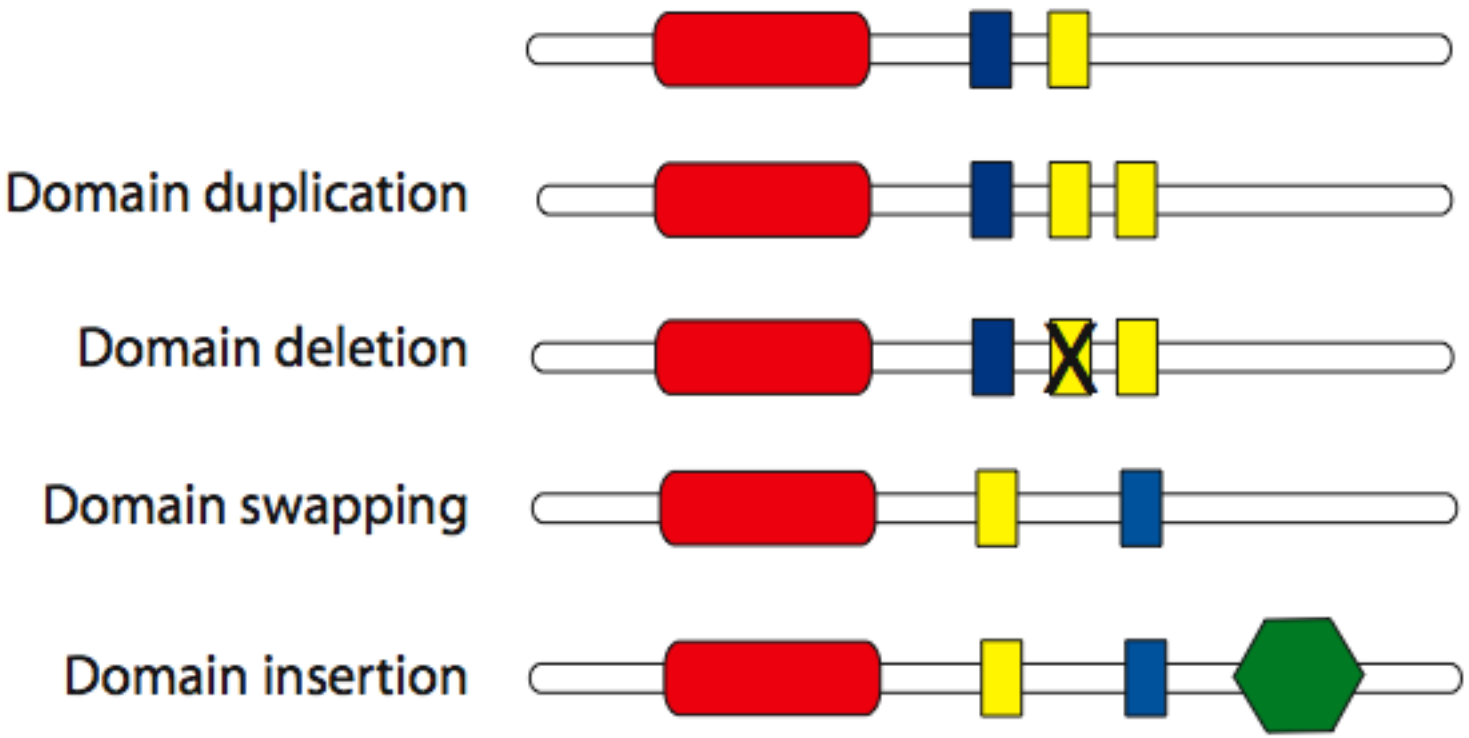


Βασικότερα είδη μεταλλάξεων

- Δομικές Αναδιατάξεις
  - Προσθήκες/απαλείψεις (insertions/deletions)
  - Αναστροφές
  - Διπλασιασμοί

Βασικότερα είδη μεταλλάξεων (ii)

- Αναδιάταξη αυτόνομων λειτουργικών περιοχών μιας πρωτεΐνης (domain rearrangements)



Όλες οι περιοχές μιας πρωτεΐνης δεν μεταλλάσονται με τον ίδιο ρυθμό

- Αυτόνομες λειτουργικές περιοχές (domains): συνήθως είναι πολύ συντηρημένες

- Περιοχές ενδογενούς δομικής αστάθειας (intrinsically disordered regions). Π.χ, ευέλικτες συνδετικές περιοχές (flexible linkers).
  - Μεταβαλλόμενο μήκος και περιεκτικότητα αμινοξέων, με παρόμοιες όμως φυσικοχημικές ιδιότητες.
  - Μεταλλάσονται γρήγορα. Το εξελικτικό σήμα μπορεί να χαθεί σύντομα
  - Συχνά δεν υπάρχει περιορισμός θέσης (π.χ φωσφορυλίωση)

Γλοβίνες

- πολύ συντηρημένη τριτοταγής δομή, λίγο συντηρημένη πρωτοταγής δομή (~10-20% ομοιότητα)

Στιγμοπίνακες (dotplots)

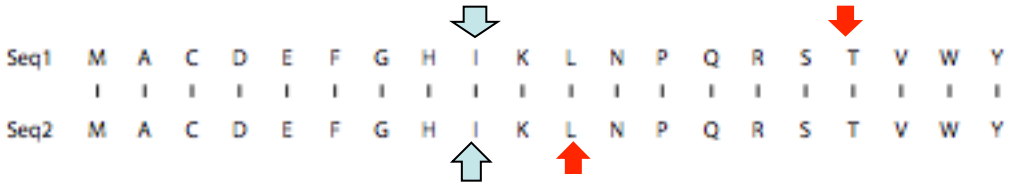
Στιγμοπίνακες (dotplots)

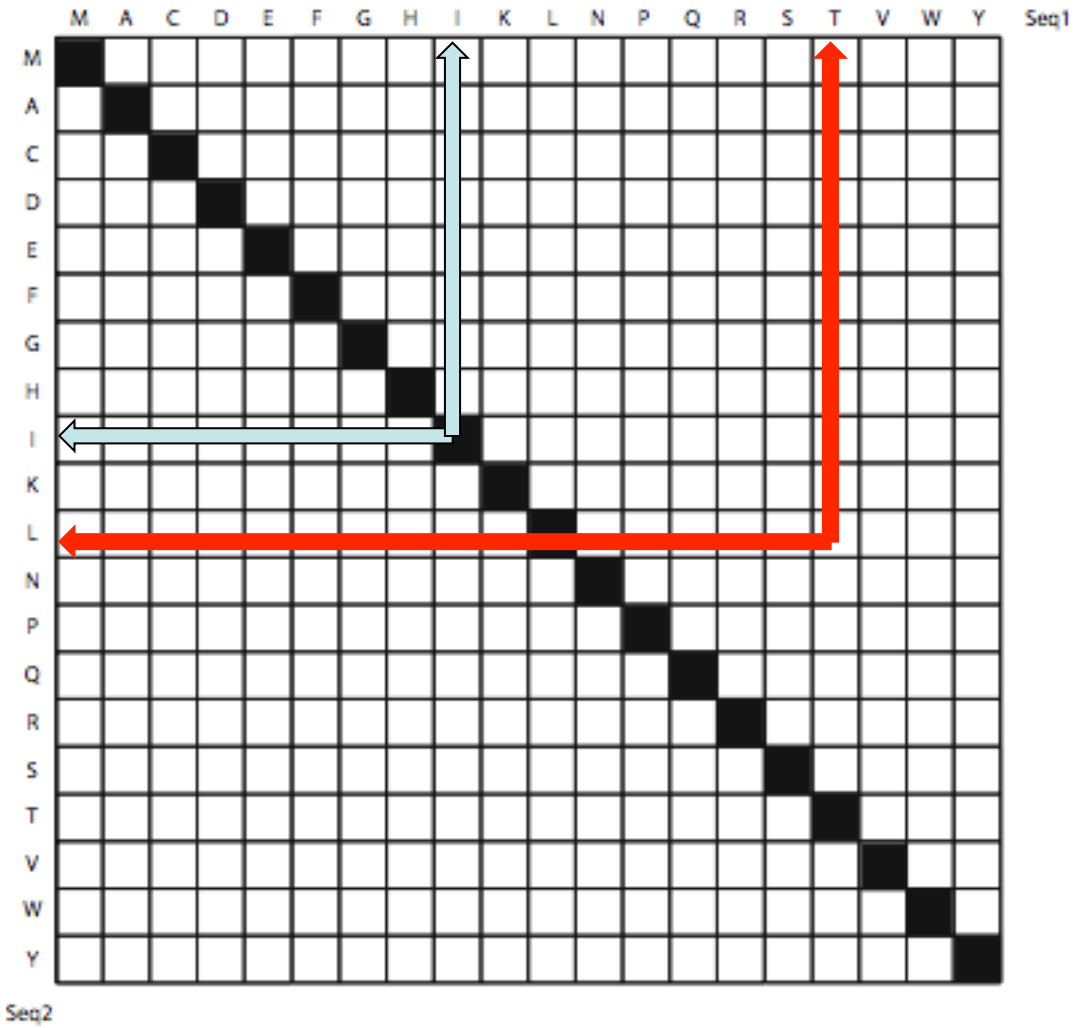
- Εισήχθησαν από τους Gibbs & McIntyre το 1970.
- Χρησιμοποιούνται για σύγκριση 2 ακολουθιών (π.χ. Πρωτεϊνών ή DNA) χωρίς να χρειάζεται να στοιχιθούν πρώτα.
- Αποκαλύπτουν
  - Προσθήκες - Απαλείψεις
  - Ευθείες ή ανεστραμμένες επαναλήψεις (π.χ χρήσιμο για RNA)
  - Περιοχές χαμηλής πολυπλοκότητας
  - Αναστροφές

- Διάφορα προγράμματα (π.χ Dotlet)

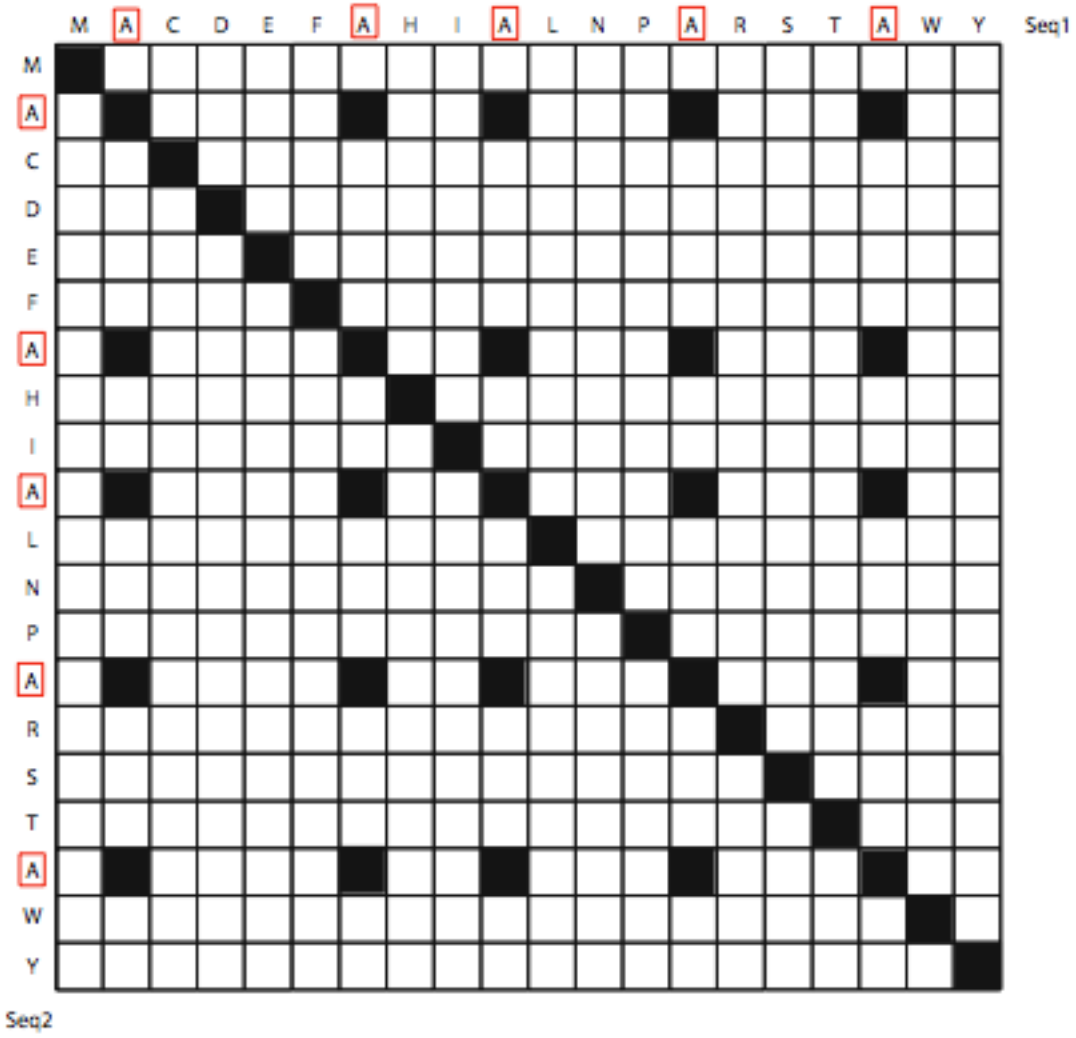
- Σε ένα βαθμό, εισέρχεται η υποκειμενικότητα στην ερμηνεία των αποτελεσμάτων.

Στιγμοπίνακες



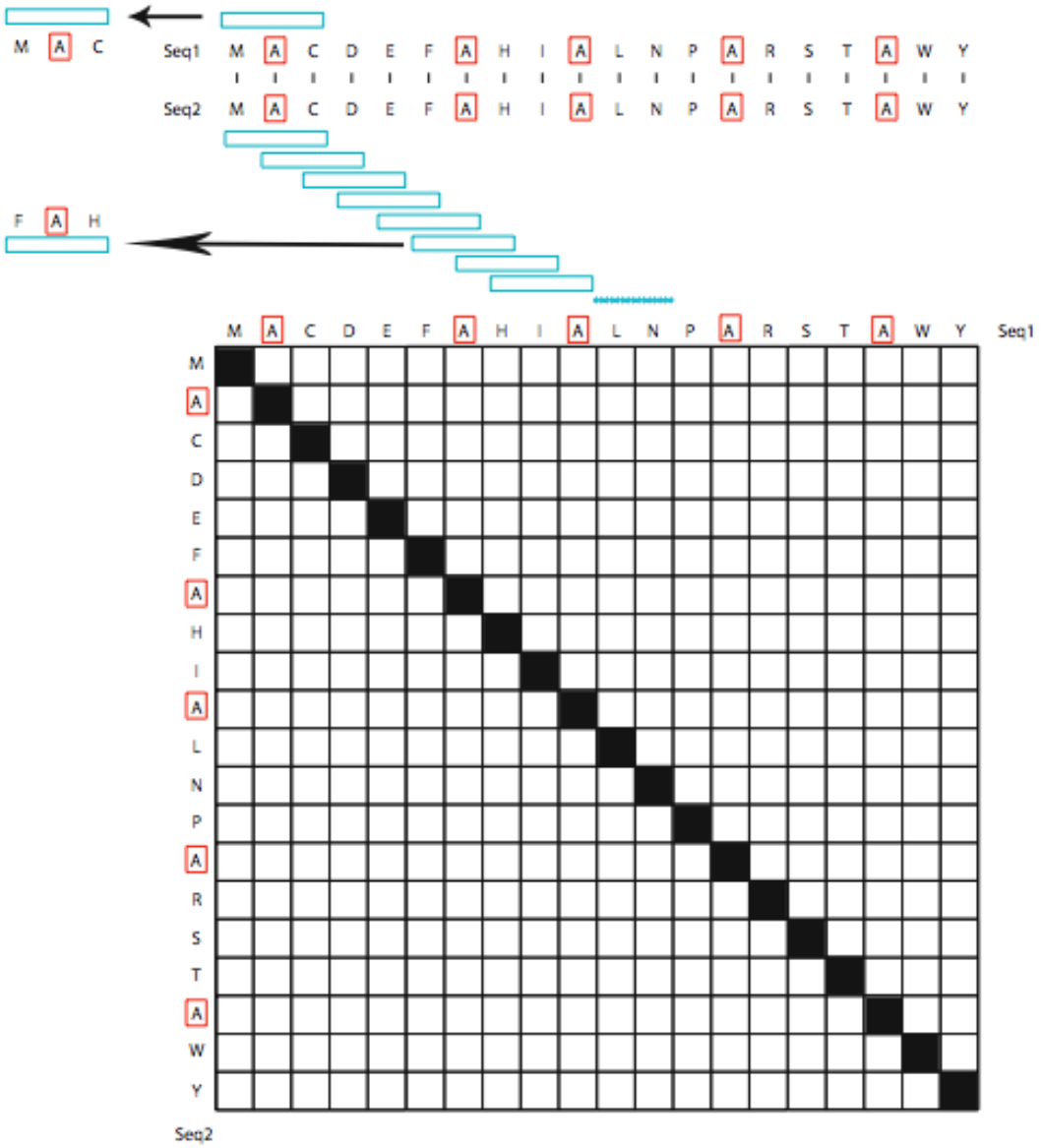


Στιγμοπίνακες - Θόρυβος

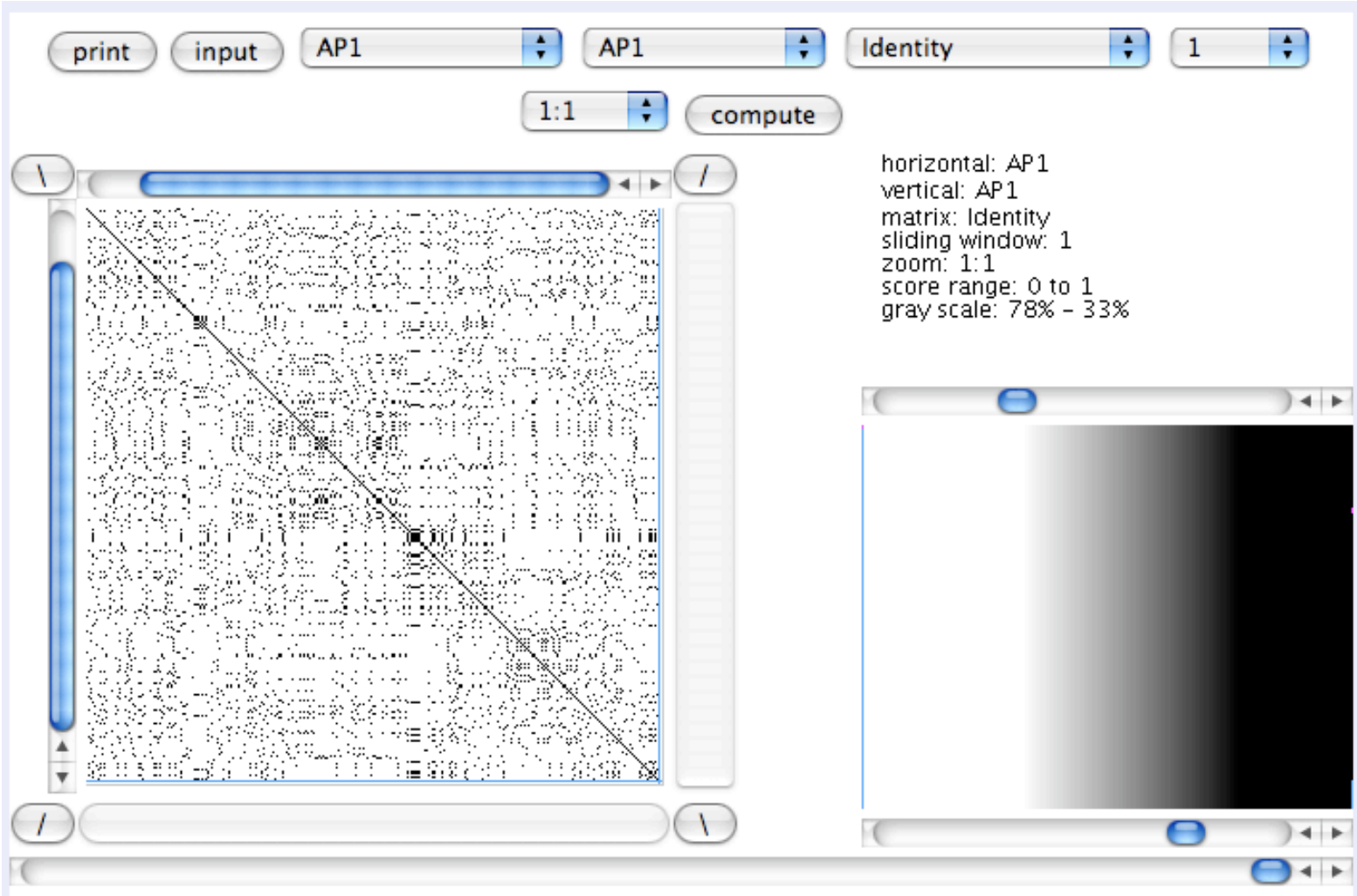


Στιγμοπίνακες -

Απαλοιφή Θορύβου με συρόμενα παράθυρα

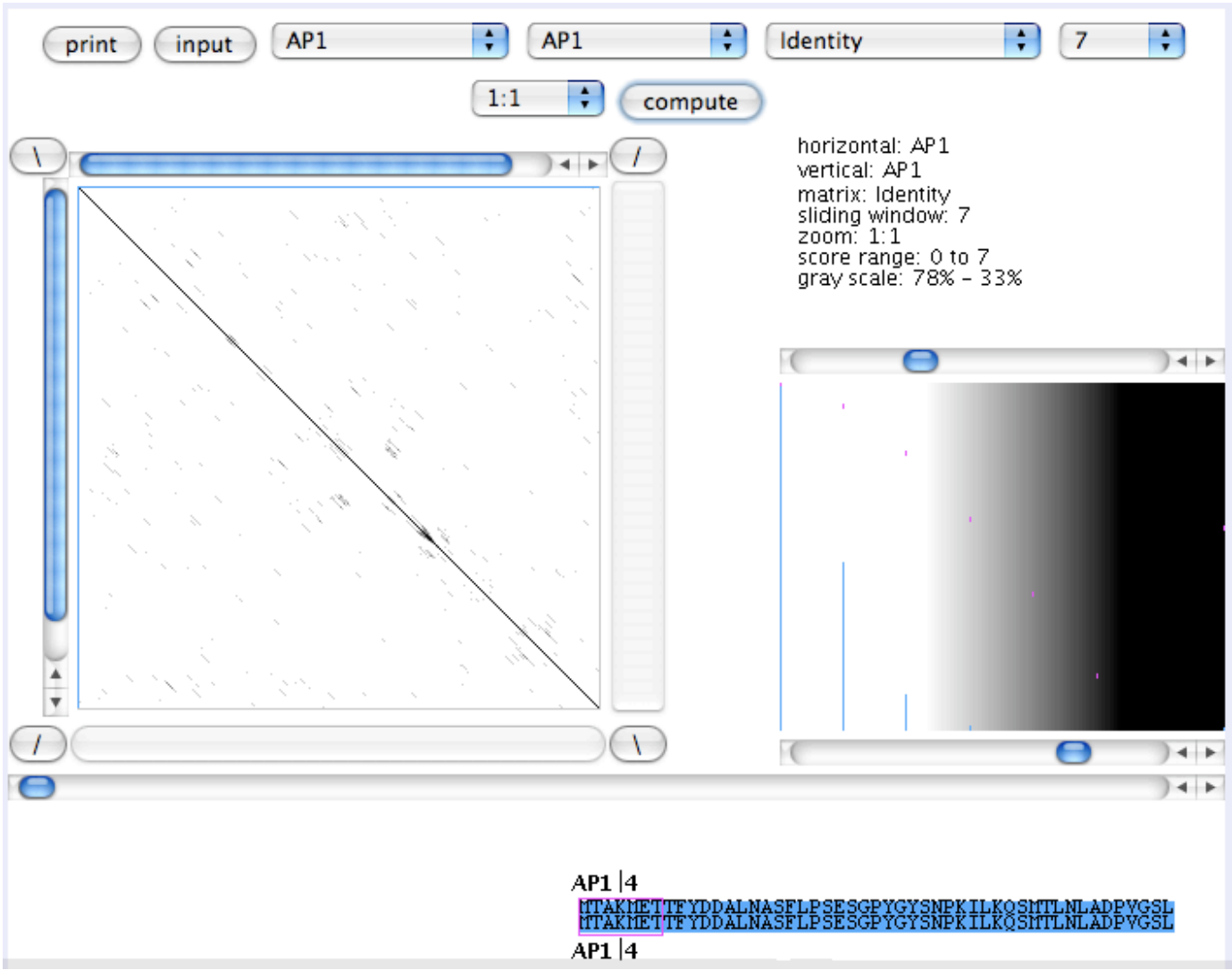


Στιγμοπίνακες



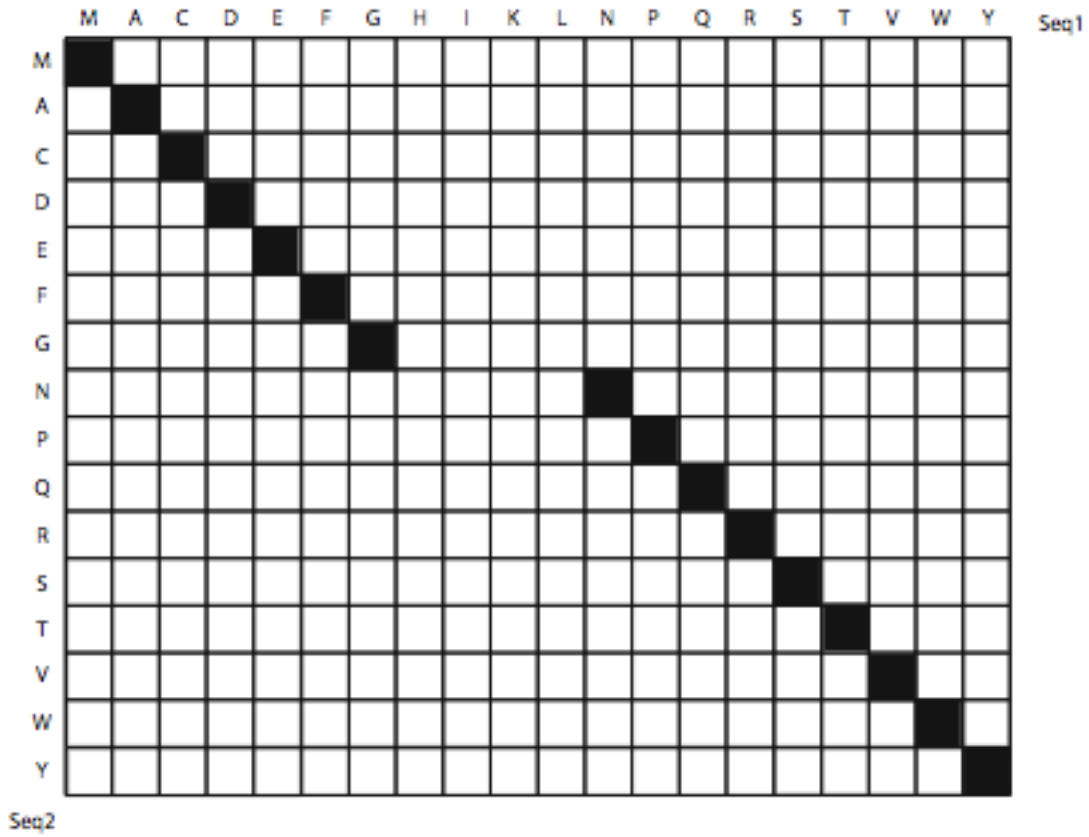
Στιγμοπίνακες

- Απαλοιφή θορύβου με συρόμενα παράθυρα
- Ο Mount προτείνει:
  - Για DNA: παράθυρο 15 χαρακτήρων με τουλάχιστον 10 αντιστοιχίσεις
  - Για πρωτεΐνες: παράθυρο 2-3 χαρακτήρων με τουλάχιστον 2 αντιστοιχίσεις



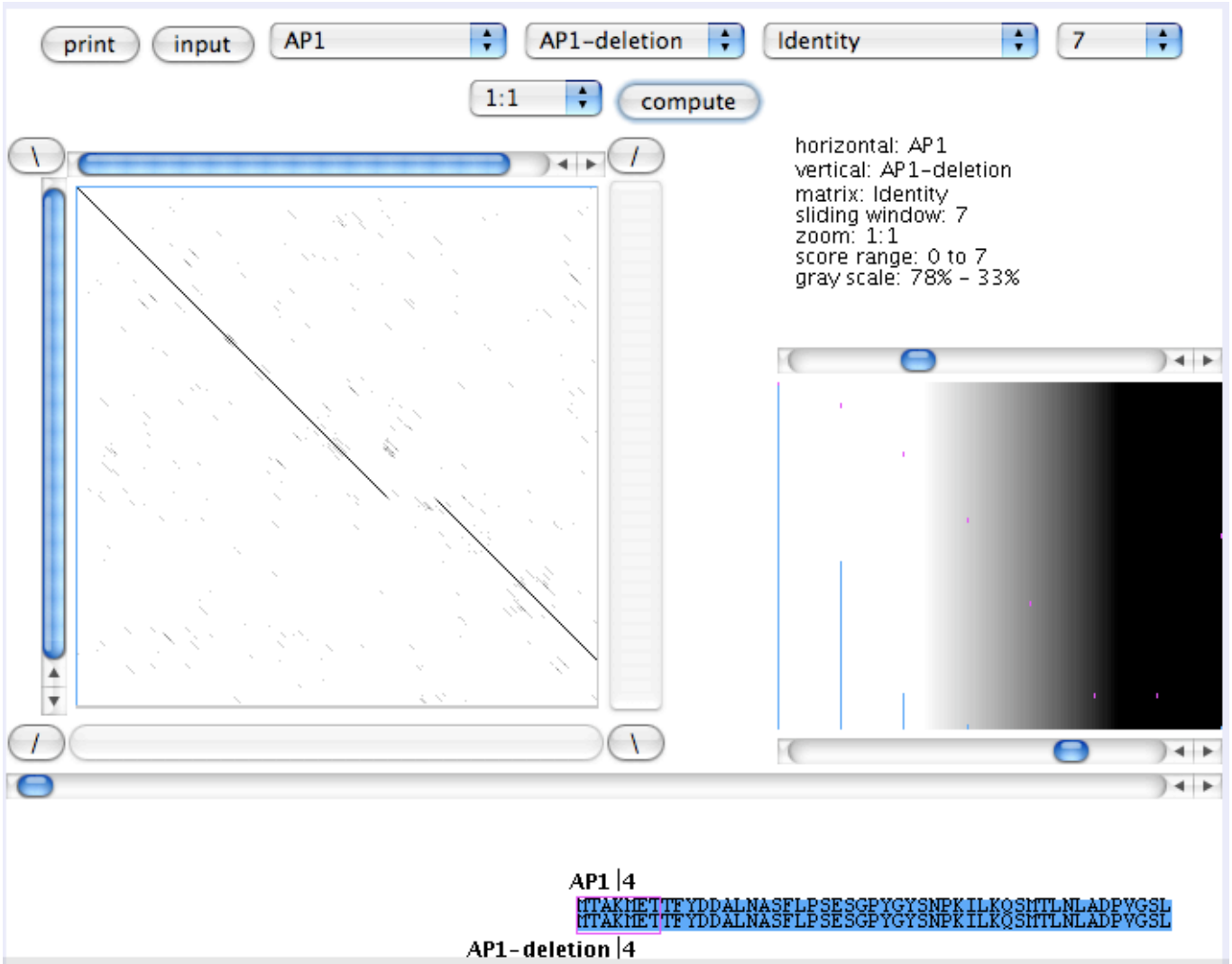
Στιγμοπίνακες insertion - deletions

Seq1	M	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	W	Y
	I	I	I	I	I	I	I					I	I	I	I	I	I	I	I	I
Seq2	M	A	C	D	E	F	G	–	–	–	–	N	P	Q	R	S	T	V	W	Y



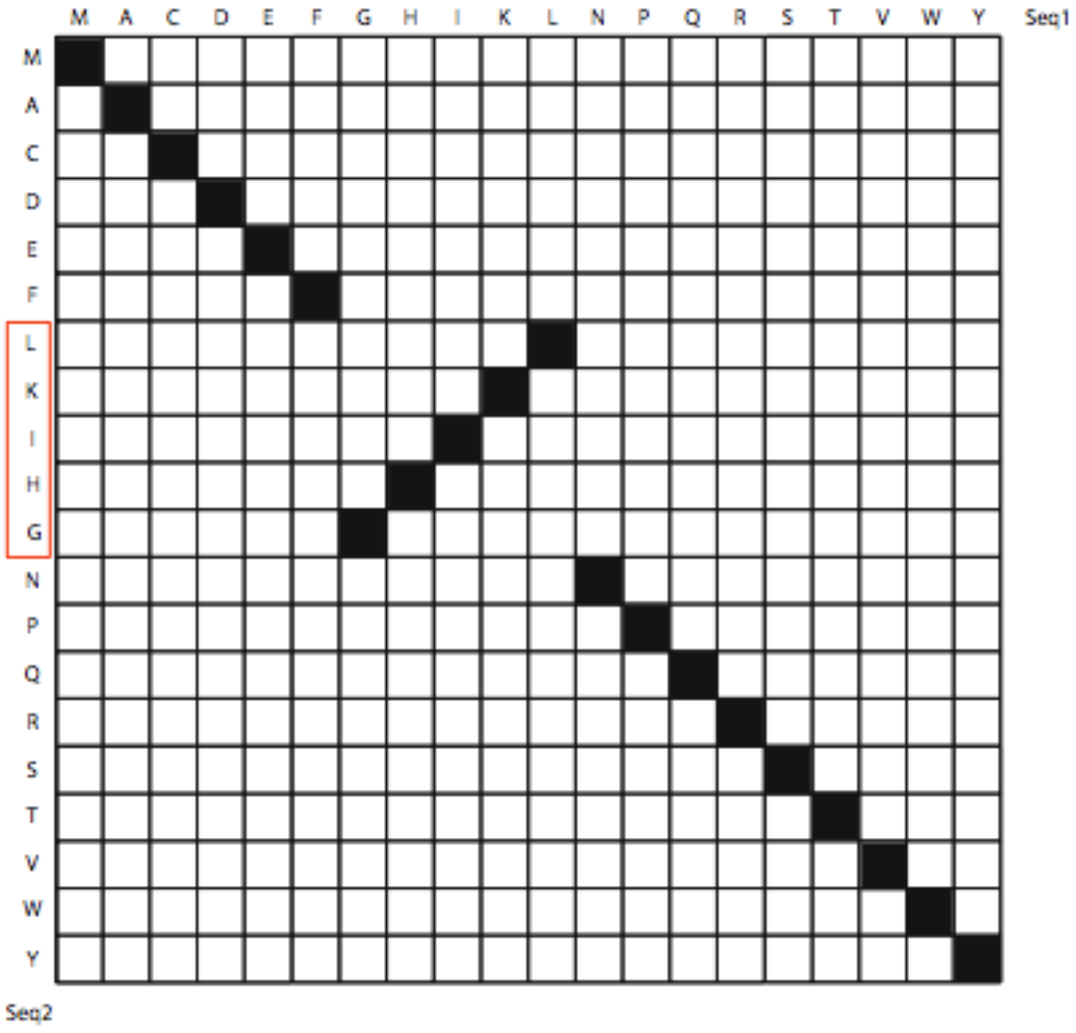
Στιγμοπίνακες

- Insertions/deletions (indels)



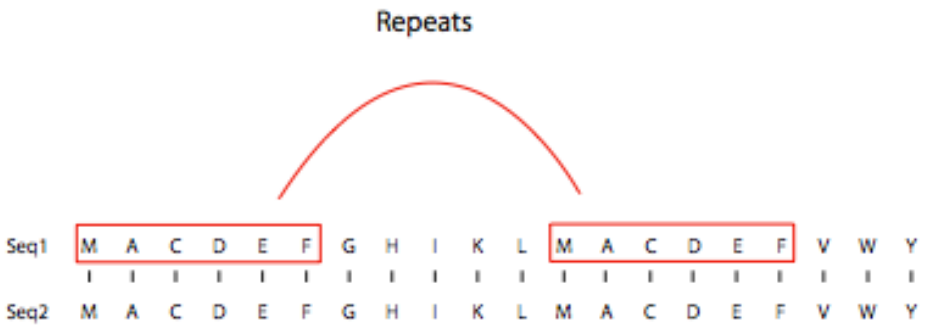
Στιγμοπίνακες - Αναστροφές

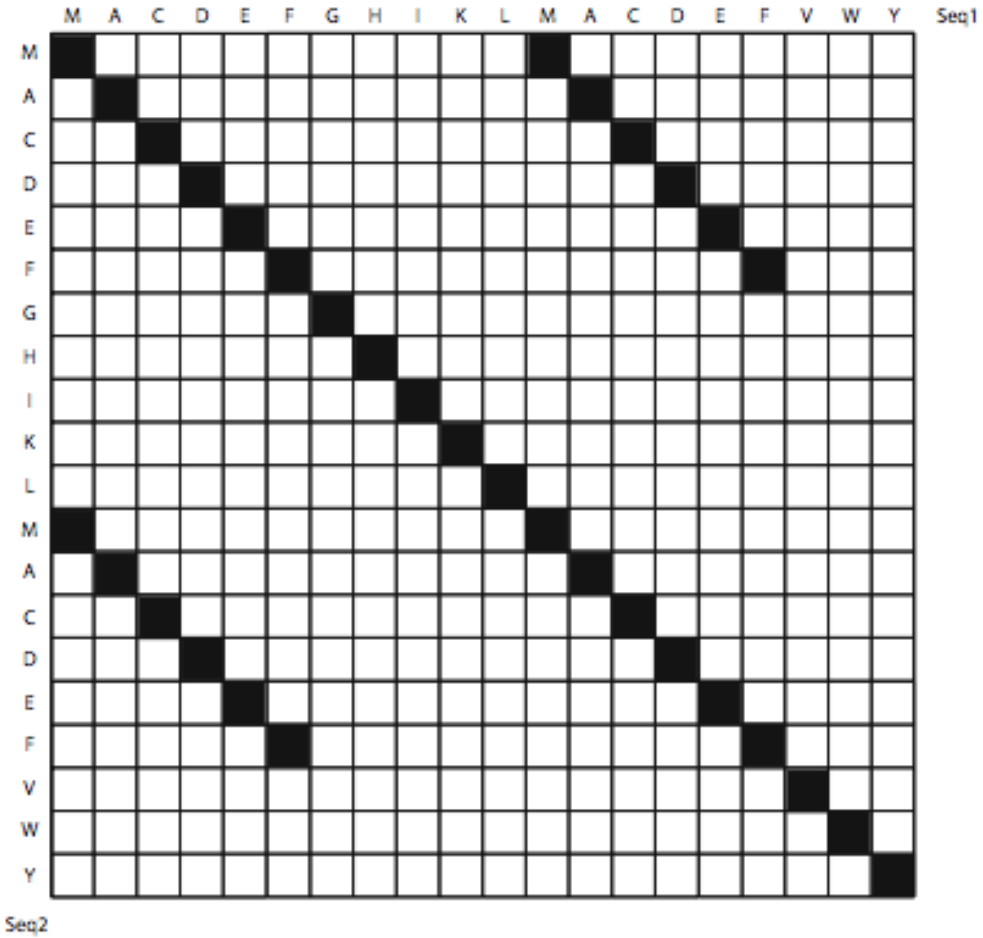
Seq1	M	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	W	Y
Seq2	M	A	C	D	E	F	L	K	I	H	G	N	P	Q	R	S	T	V	W	Y



Στιγμοπίνακες

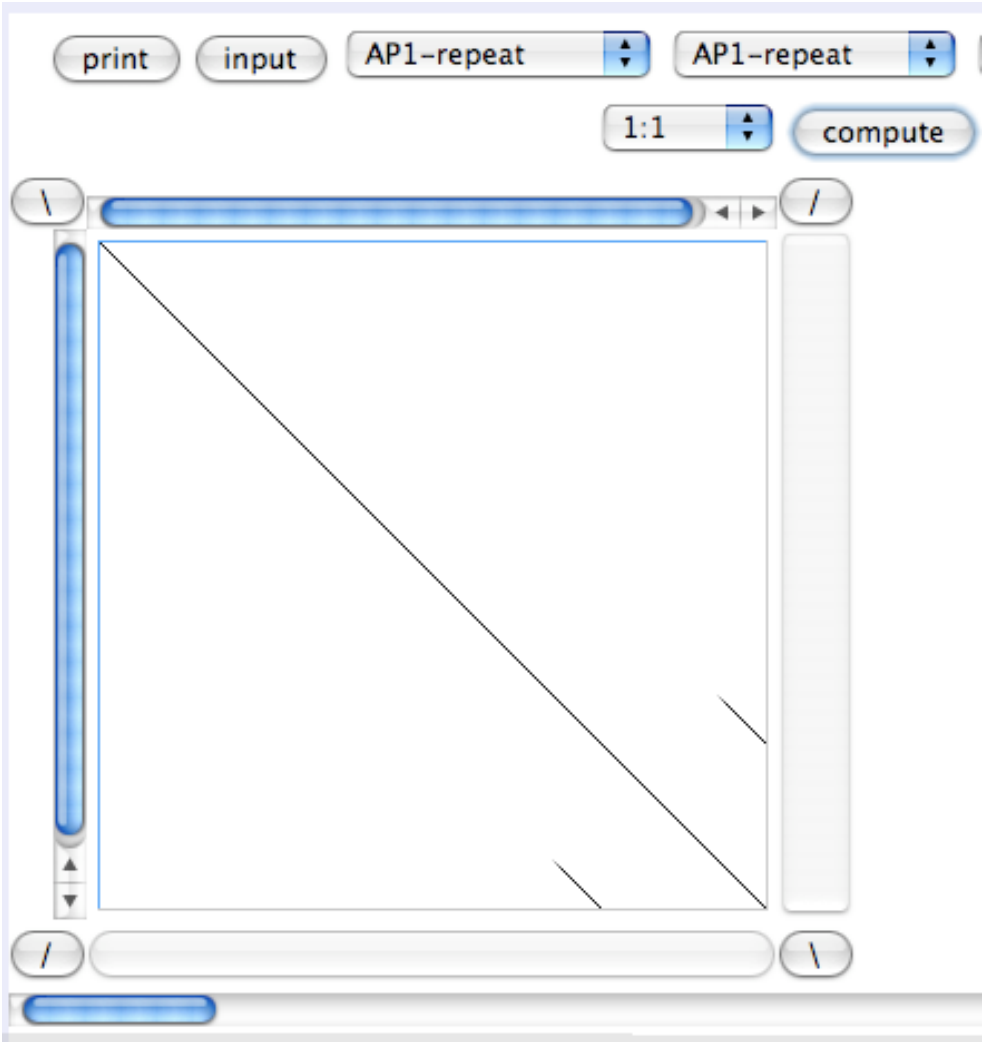
Επαναλήψεις





Στιγμοπίνακες

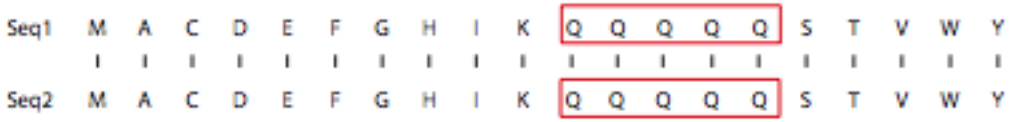
- Επαναλήψεις

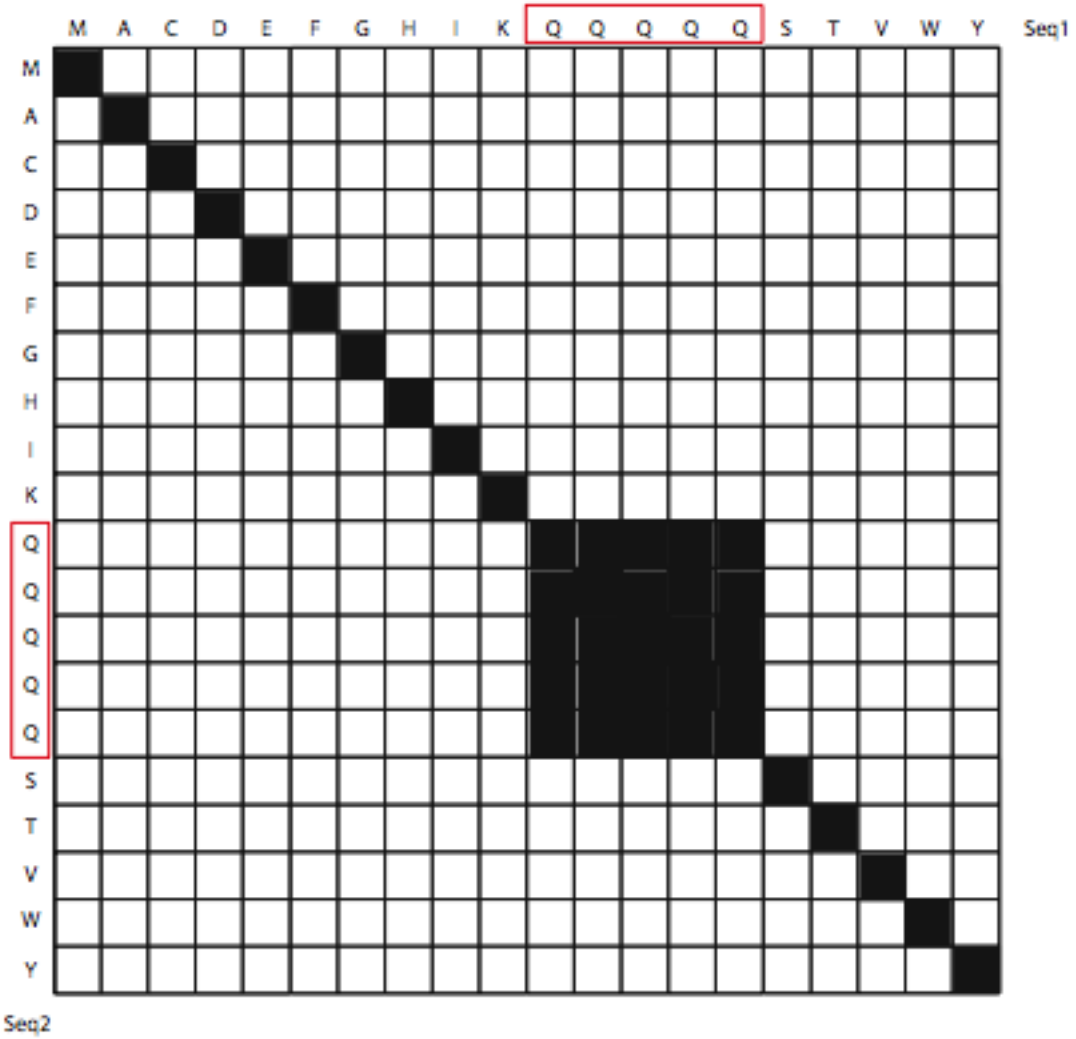


Στιγμοπίνακες

Επαναλήψεις

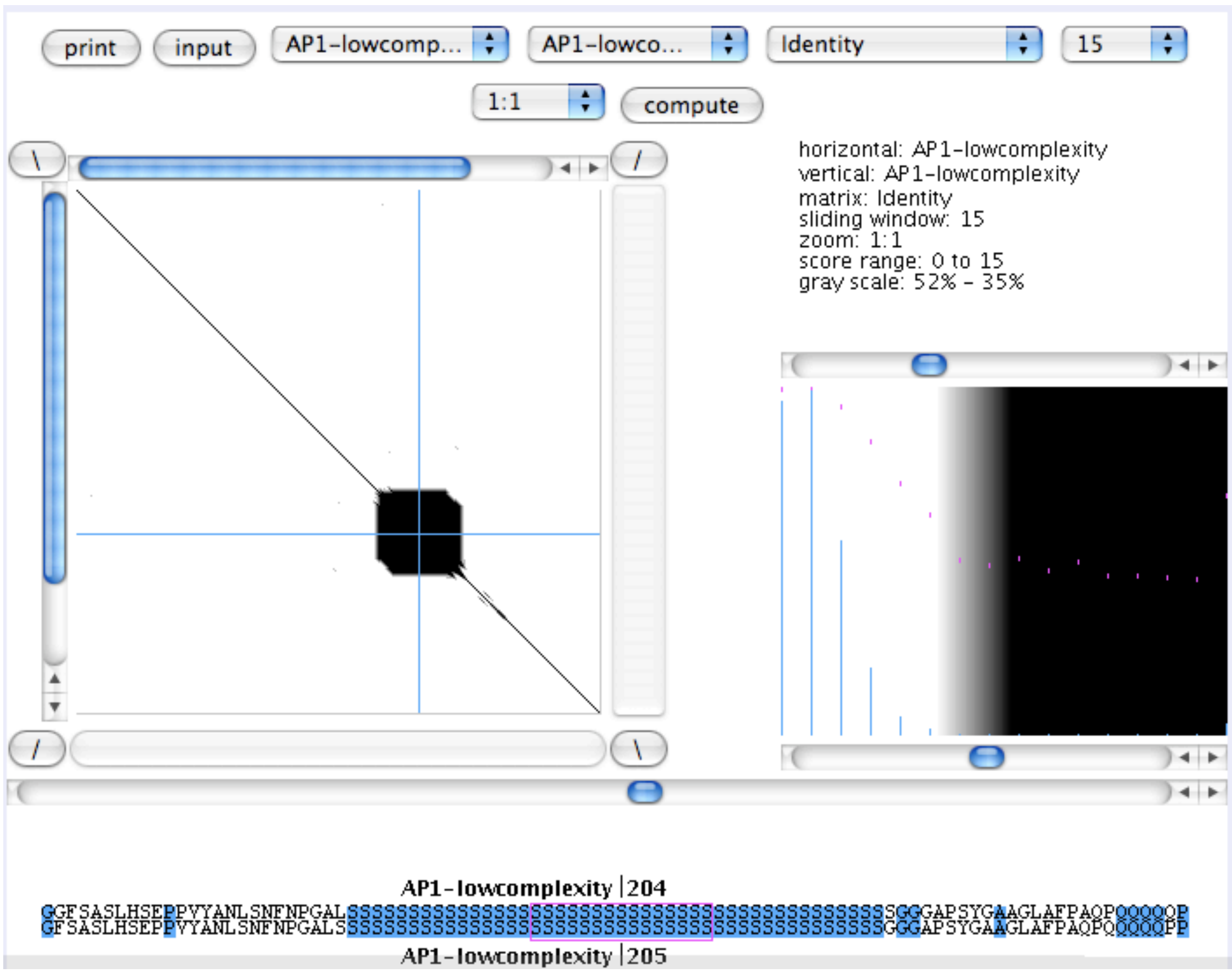
Περιοχές χαμηλής πολυπλοκότητας





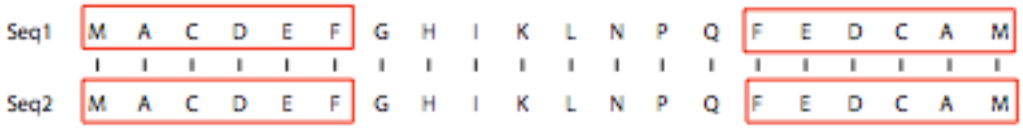
Στιγμοπίνακες

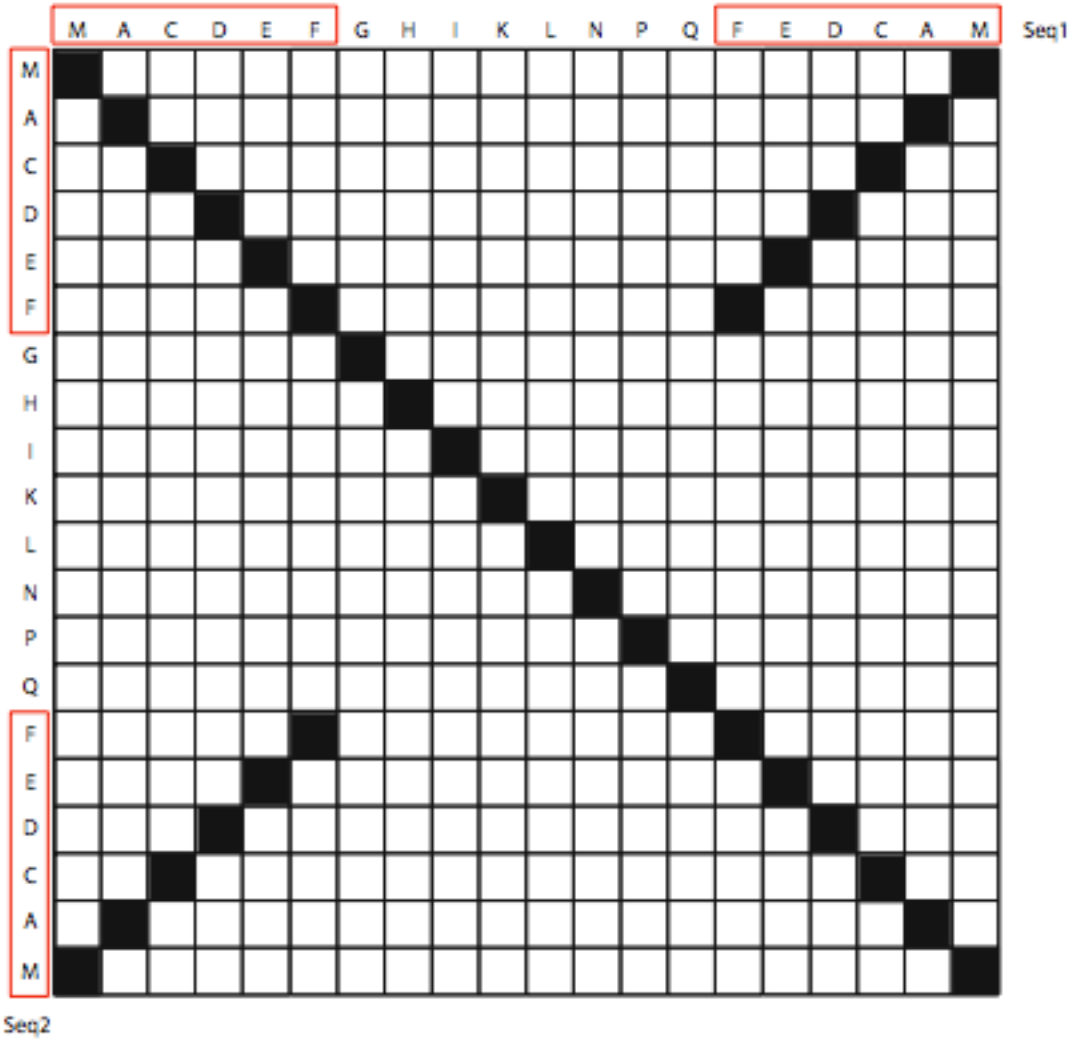
- Επαναλήψεις - Περιοχές χαμηλής πολυπλοκότητας



Στιγμοπίνακες

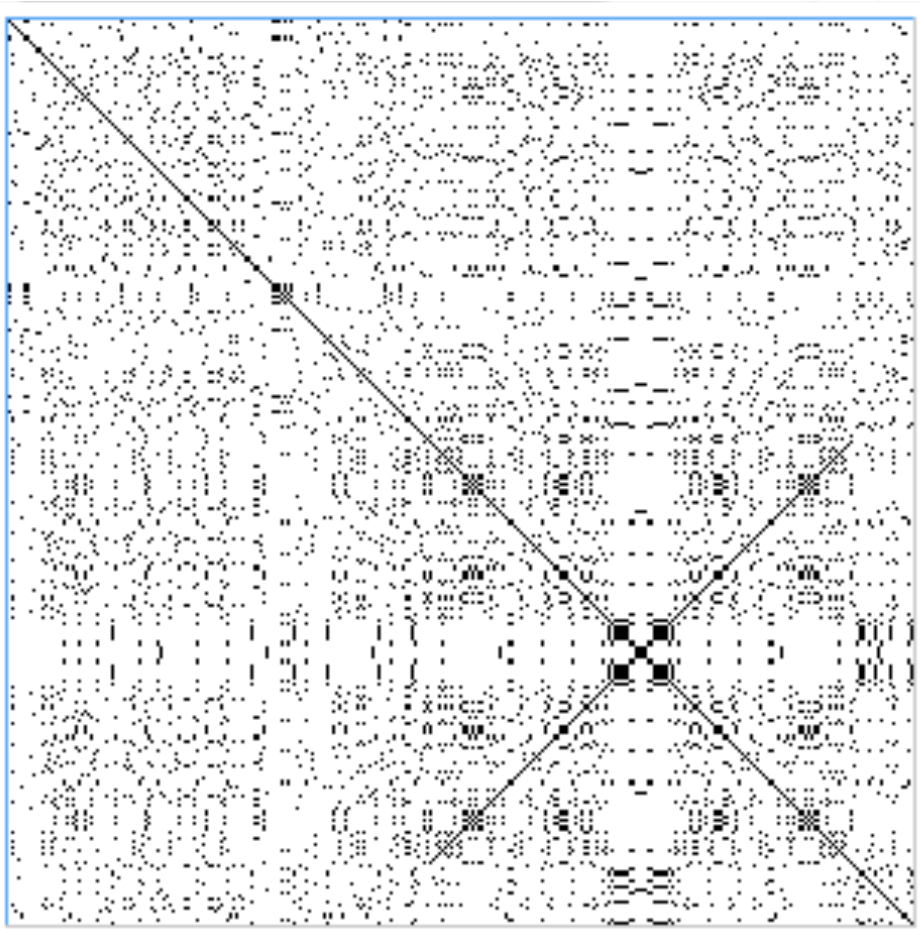
Ανεστραμμένες Επαναλήψεις





Στιγμοπίνακες

- Ανεστραμμένες επαναλήψεις



Dotplot: Πόσο σταθερή είναι η αρχιτεκτονική ενός γονιδιώματος;

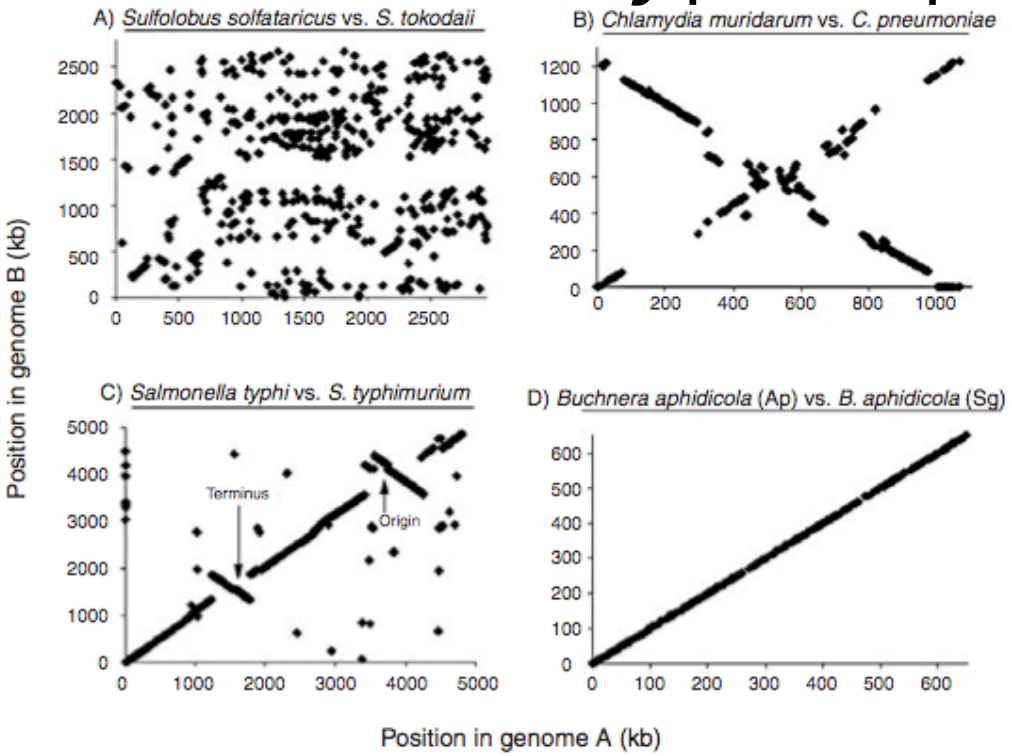


FIGURE 10.6 Gene position plots showing examples of both plasticity and stability in gene order between closely related species of prokaryotes. In these plots, the location of a given gene, measured as its distance from a given starting point in kilobases (kb), is plotted on one axis each for the two species being compared. Unless otherwise indicated, the origin of the axes represents the origin of replication in the chromosomes. (A) The archaeons *Sulfolobus solfataricus* and *S. tokodaii*, whose genomes share very little common gene order and are clearly extremely dynamic. (B) The bacteria *Chlamydia muridarum* and *C. pneumoniae*, which exhibit a clear "X-alignment," indicating a single, large, symmetrical inversion around the origin of replication (see also Eisen *et al.*, 2000; Hughes, 2000). (C) The bacteria *Salmonella typhi* and *S. typhimurium*, which show evidence of two smaller symmetrical inversions, one around the origin of replication and one around the replication terminus. (D) Two strains (or possibly species) of the endosymbiotic bacterium *Buchnera aphidicola* living in distantly related aphid hosts (Ap = *Acyrthosiphon pisum*; Sg = *Schizaphis graminum*). In this case, there has been remarkable stasis in gene order for 50–70 million years, despite considerable sequence divergence (see Tamas *et al.*, 2002). Based on a figure presented by Mira *et al.* (2002), reproduced by permission (© Elsevier Inc.).

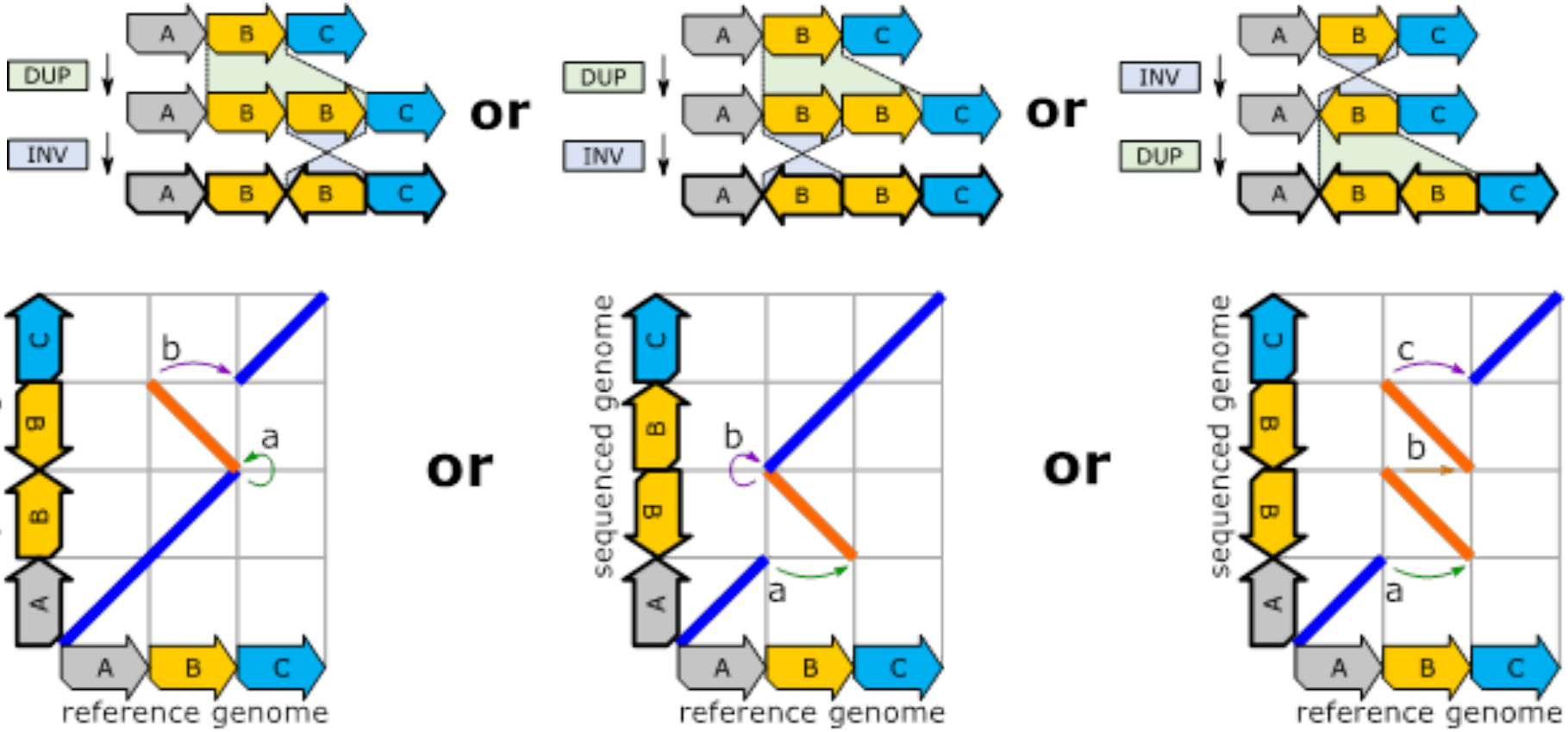
Εδώ, αντί να συγκρίνονται νουκλεοτίδια ή αμινοξέα, συγκρίνονται ολόκληρα γονίδια/ πρωτεΐνες.

Dotplot για ορθόλογα γονίδια μεταξύ δύο προκαρυωτών του ίδιου είδους.

Κάθε κουκίδα στο Dotplot είναι η θέση του ορθόλογου γονιδίου σε δύο διαφορετικά γονιδιώματα.

Κάποιοι οργανισμοί έχουν σταθερή γονιδιωματική αρχιτεκτονική και κάποιοι άλλοι όχι.

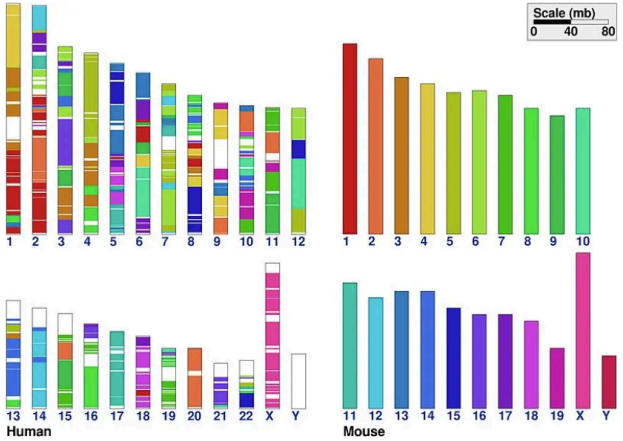
Dotplot: Πόσο σταθερή είναι η αρχιτεκτονική ενός γονιδιώματος;



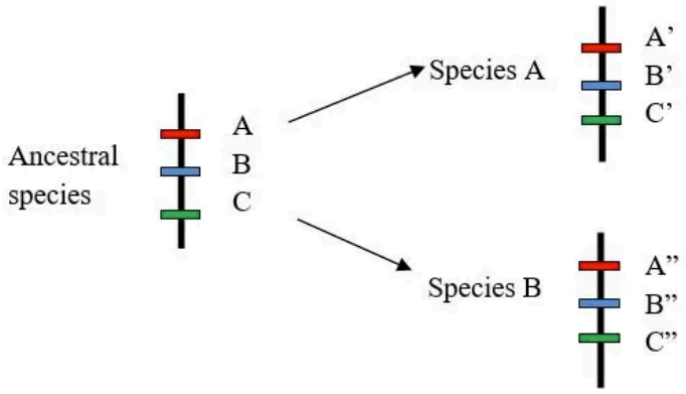
http://itbe.hanyang.ac.kr/ak/MSV/

Γονιδιωματικό Dotplot: συνταινικότητα και συγγραμικότητα;

Συνταινικότητα (synteny): Όταν μεταξύ δύο γονιδιωμάτων τα ορθόλογα παραμένουν στη ίδια γειτονιά (π.χ. ίδιο χρωμόσωμα). Δεν είναι απαραίτητο να παραμένουν με την ίδια σειρά.



Συγγραμικότητα (collinearity): Όταν μεταξύ δύο γονιδιωμάτων τα ορθόλογα παραμένουν στη ίδια γειτονιά με την ίδια σειρά.



https://www.differencebetween.com/difference-between-synteny-and-collinearity/

Στοίχιση ακολουθιών κατά ζεύγη (Pairwise alignment)

Στοίχιση κατά ζεύγη: Τι είναι

- Αντιστοίχιση των νουκλεοτιδίων/αμινοξέων δυο ακολουθιών, ώστε να εντοπιστούν οι ομοιότητες και οι διαφορές τους.
- Χρησιμοποιείται για:
  - Εντοπισμό μεταλλάξεων
  - αναζήτηση ομόλογων γονιδίων/πρωτεϊνών σε βάσεις δεδομένων.
  - Συναρμολόγιση γενωμάτων.
  - Έλεγχος εξειδίκευσης εκκινητών (primers) για PCR.

- Προσοχή!!!! Στους στιγμοπίνακες οι ακολουθίες συγκρίνονται ΧΩΡΙΣ να πραγματοποιηθεί στοίχιση.

Στοίχιση κατά ζεύγη: Τι είναι

- Τοποθετούνται οι ΑΝΤΙΣΤΟΙΧΟΙ ή αλλιώς ΟΜΟΛΟΓΟΙ χαρακτήρες ο ένας κάτω από τον άλλο και μπορεί να γίνει χρήση κενών (gaps)
- Δύο χαρακτήρες μπορεί να είναι:
  - Ίδιοι
  - Παρόμοιοι (κοινές φυσικοχημικές ιδιότητες, π.χ. Ισολευκίνη - βαλίνη)
  - Διαφορετικοί

```
Query  1    MKTPVSAAANLSIQNAGSSGATAIQIIPKTEPVGEEGPMSLDFQSPNLNTSTPNPNKRPG  60
            MKTPVSAAANLS  NAGSSGA AIQI+PKTEPVGEEGPMSLDFQSPNL+TSTPNPNKRPG
Sbjct  1    MKTPVSAAANLSNPNAGSSGAAAIQIVPKTEPVGEEGPMSLDFQSPNLSTSTPNPNKRPG  60

Query  61   SLDLNSKSAKNKRIFAPLVINSPDLSSKTVNTPDLEKILLSNNLMQTPQPGKVFPTKAGP  120
            SLDLNSK AKNKRIFAPLVINSPDL +KTVNTPDLEKILLSNNL+QTPQPGKVFPTKAGP
Sbjct  61   SLDLNSKGAKNKRIFAPLVINSPDLQAKTVNTPDLEKILLSNNLIQTPQPGKVFPTKAGP  120

Query  121  VTVEQLDFGRGFEEALHNLHTNSQAFPSANSAANSAANNTTAAAMTAVNNGISGGTFTYT  180
            VTVEQ DFGRGFEEAL NLHTNSQAFP    A NS ANNTT  AMTAVNNGISGGTFTY
Sbjct  121  VTVEQEDFGRGFEEALKNLHTNSQAFP----AVNSTANNTTGTAMTAVNNGISGGTFTY-  175
```

Στοίχιση κατά ζεύγη: Τι είναι

- Για δύο ακολουθίες με 95% ομοιότητα, η στοίχιση μπορεί να γίνει και με το μάτι.
- Τα διαθέσιμα προγράμματα αγγίζουν τα όρια των δυνατοτήτων τους όταν οι ακολουθίες έχουν 18-25% ομοιότητα (ζώνη του λυκόφωτος)

Είδη στοίχισης κατά ζεύγη (ι)

- Ολική στοίχιση (global alignment)
  - Προσπαθεί να στοιχίσει όσο το δυνατό περισσότερους χαρακτήρες σε ΟΛΟ το μήκος των δύο αλληλουχιών
  - Για ακολουθίες που δεν έχουν αποκλείνει σε μεγάλο βαθμό και επίσης έχουν παρόμοιο μέγεθος
  - Κλασσική μέθοδος: Needleman-Wunsch.
    - Βασίζεται στον δυναμικό προγραμματισμό

Είδη στοίχισης κατά ζεύγη (ii)

- Τοπική στοίχιση (local alignment)
  - Νησίδες στοίχισης.
  - Για ακολουθίες που έχουν αποκλείνει αρκετά και έχουν απομείνει συντηρημένες μόνο κάποιες περιοχές (domains)
  - Για αντιστοίχιση mRNA με γενωμικό DNA
  - Κλασσικές μέθοδοι:
    - Smith-Waterman (δυναμικός προγραμματισμός)
    - Blast (ευρετικές μέθοδοι-heuristics)

Είδη στοίχισης κατά ζεύγη

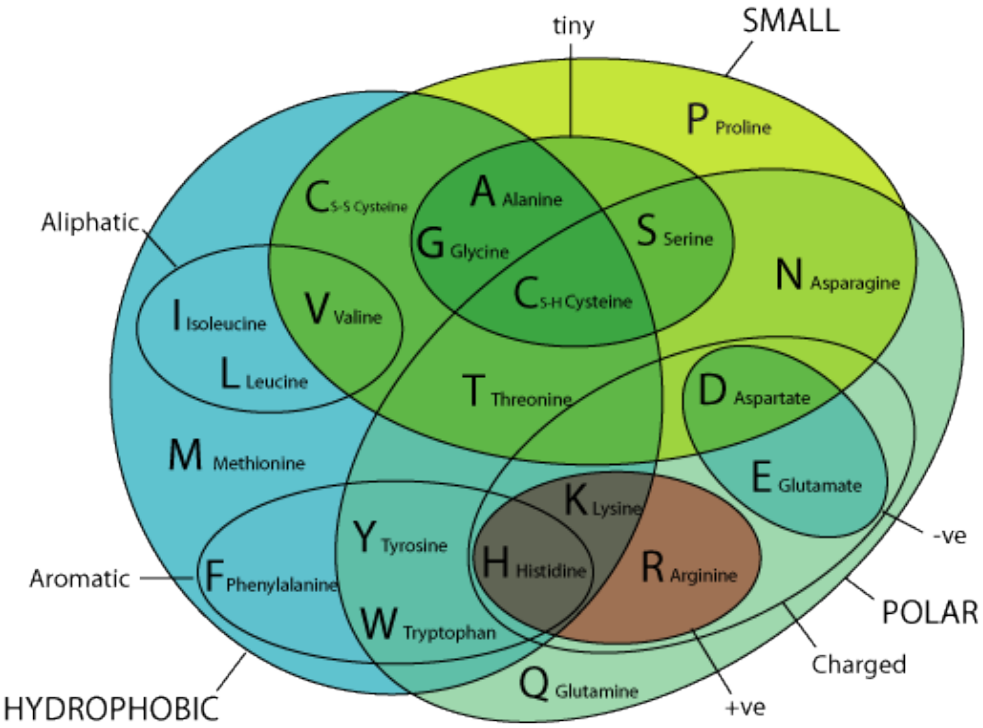
- Στοίχιση αλληλεπικάλυψης (overlap ή ends-free alignment) για συναρμολόγηση γονιδιώματος από μικρά αλληλεπικαλυπτόμενα κομμάτια DNA

Είδη στοίχισης κατά ζεύγη (iii)

```
Global FTFTALILLAVAV
       F--TAL-LLA-AV

Local  FTFTALILL-AVAV
       --FTAL-LLAAV--
```

Στιγμοπίνακες

- Αν συγκρίνουμε 2 πρωτεΐνες που έχουν αποκλίνει αρκετά, αντί να ελέγξουμε για ακριβές ταίριασμα των αμινοξέων, μπορούμε να ελέγξουμε για ταίριασμα αμινοξέων με παρόμοιες φυσικοχημικές ιδιότητες.
- Χρησιμοποιούμε πίνακες αντικατάστασης (π.χ. PAM, Blosum)
- Για το συρόμενο παράθυρο υπολογίζεται ένα σκορ με βάση τους χρησιμοποιούμενους πίνακες αντικατάστασης.



Δυναμικός προγραμματισμός

- Δίνει την βέλτιστη στοίχιση (Μαθηματικά αποδεδειγμένο).
- Και για ολικές και για τοπικές στοιχίσεις.
- Η στοίχιση εξαρτάται από το βαθμολογικό σύστημα που εφαρμόζεται.

Δυναμικός προγραμματισμός

- Το βαθμολογικό σύστημα πρέπει:
  - Να δίνει βαθμούς για κάθε θέση που οι χαρακτήρες ταιριάζουν απόλυτα
  - Να δίνει βαθμούς (λιγότερους) για κάθε θέση που οι χαρακτήρες έχουν παρόμοιες ιδιότητες
  - Να μην δίνει βαθμούς για μια θέση που οι χαρακτήρες είναι τελείως διαφορετικοί
  - Να βάζει ποινή για κάθε κενό που εισάγεται
  - Να βάζει ποινή (μικρότερη) για κάθε κενό που επεκτείνεται

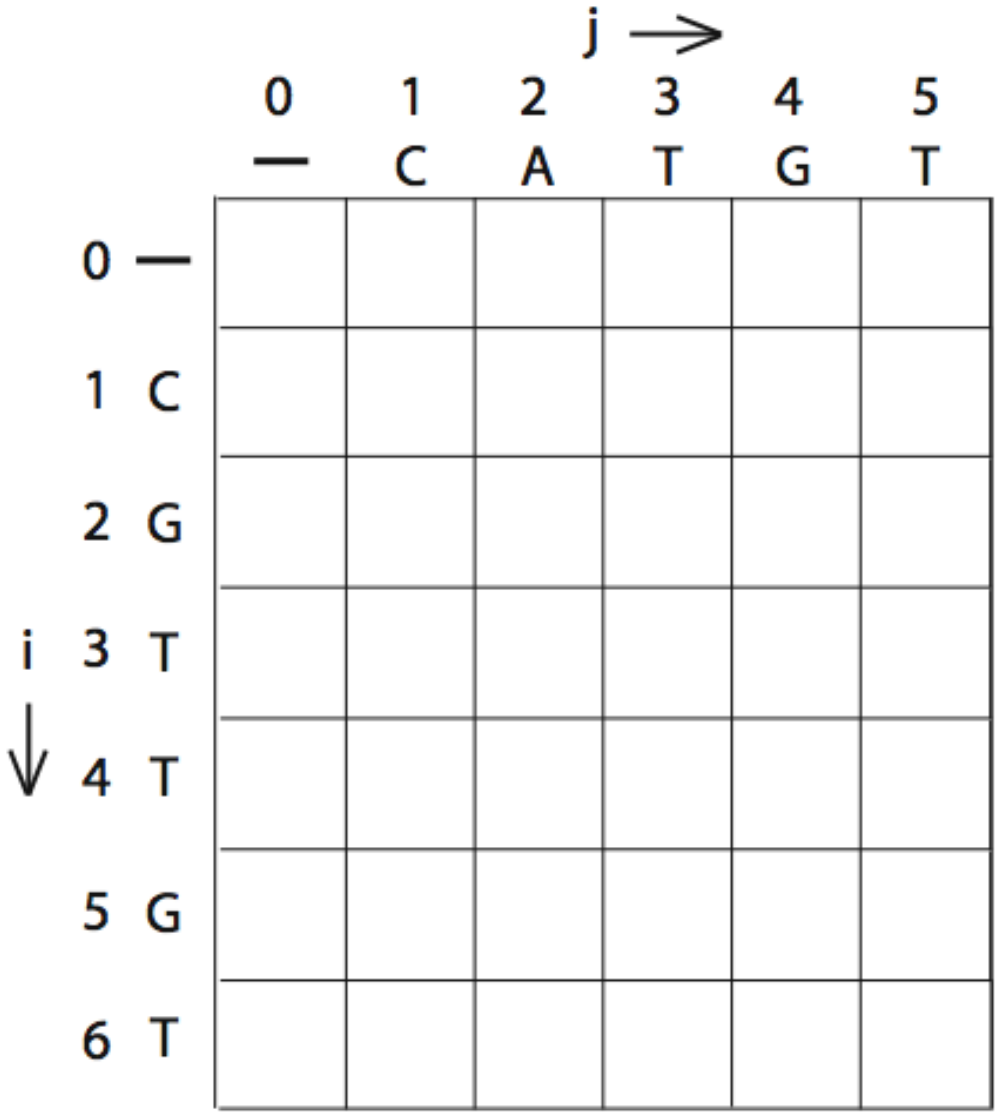
Δυναμικός προγραμματισμός

Το βαθμολογικό σύστημα

```
sequence 1   V   D   S   -   C   Y
sequence 2   V   E   S   L   C   Y
SCORE        4   2   4 -11   9   7   SCORE = SUM OF AMINO ACID PAIR SCORES
(26)                                 MINUS SINGLE GAP PENALTY (11) = 15
```

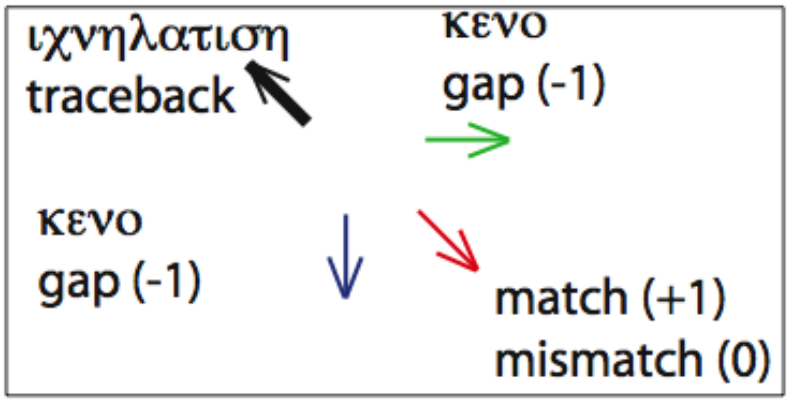
Figure 3.7. Example of scoring a sequence alignment with a gap penalty. The individual alignment scores are taken from an amino acid substitution matrix.

Δ.Π. Ολική στοίχιση παράδειγμα (i)



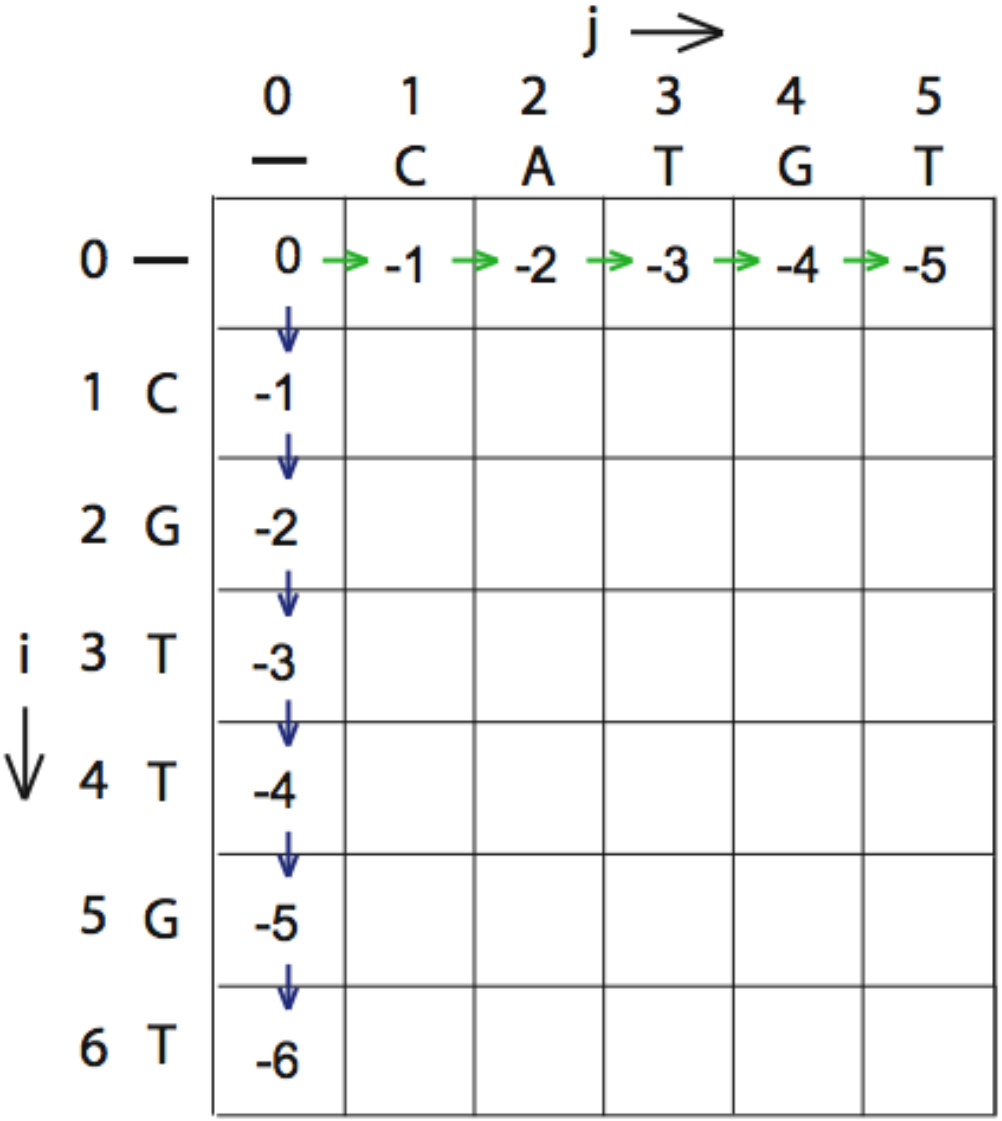
Scoring-System

match=1
mismatch=0
gap=-1



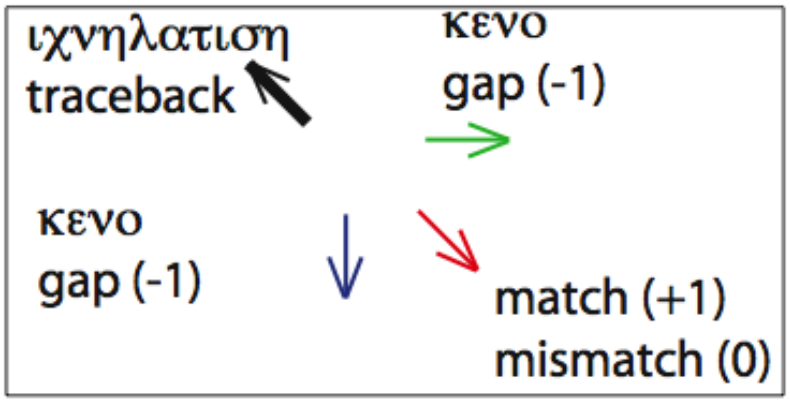
Δ.Π. Ολική στοίχιση παράδειγμα (ii)

Εκκίνηση του πίνακα



Scoring-System

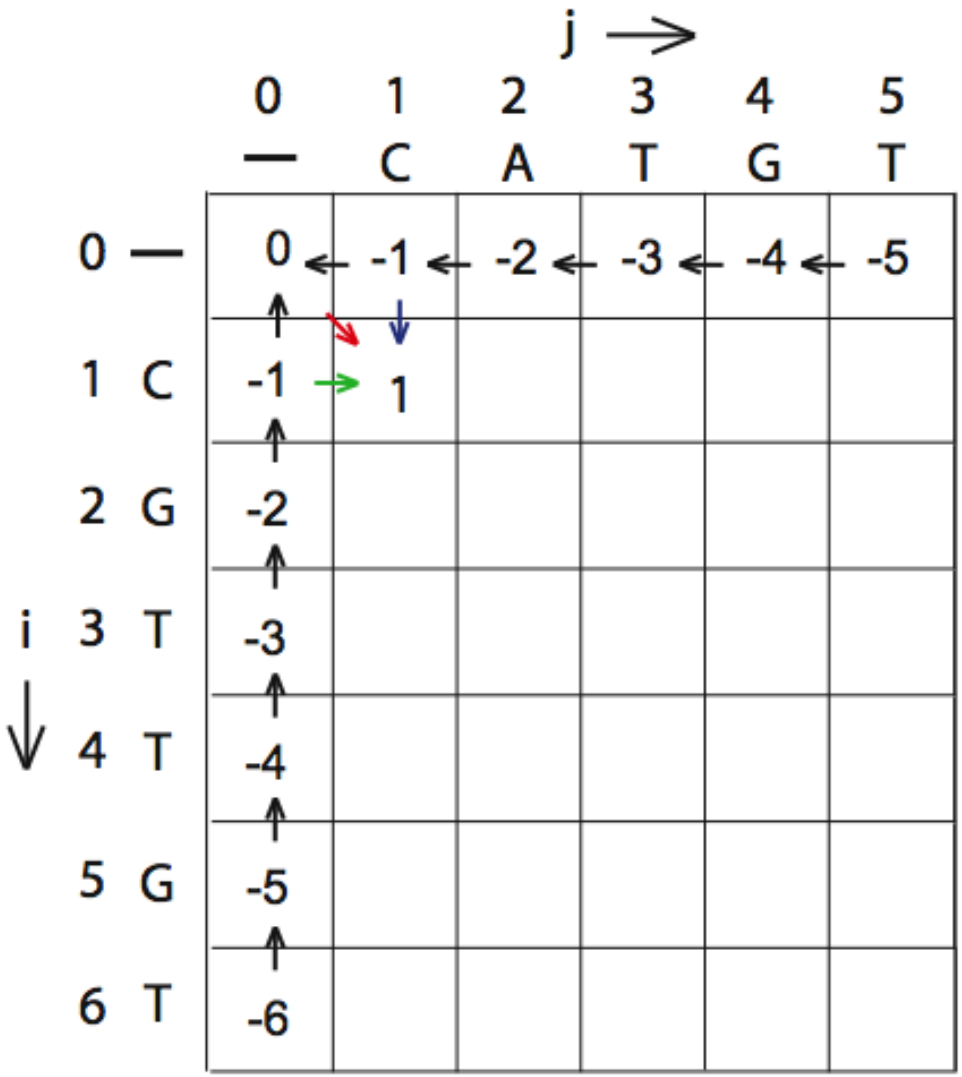
match=1
mismatch=0
gap=-1



Δ.Π. Ολική στοίχιση παράδειγμα (iii)

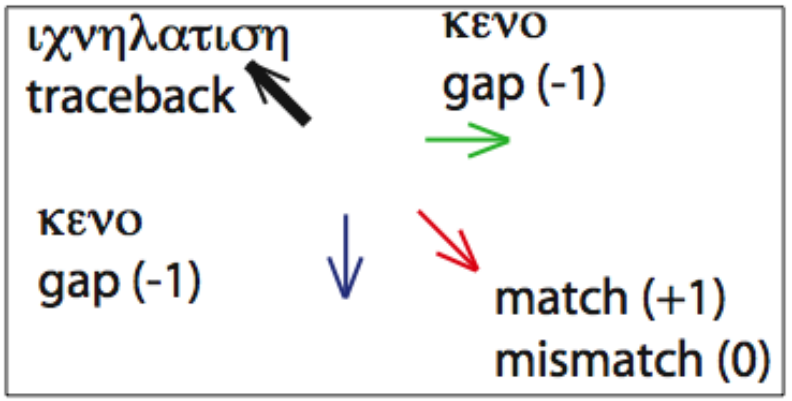
Συμπλήρωση πίνακα

$S_{1,1}$ = MAX [0+1, -1-1, -1-1] = 1



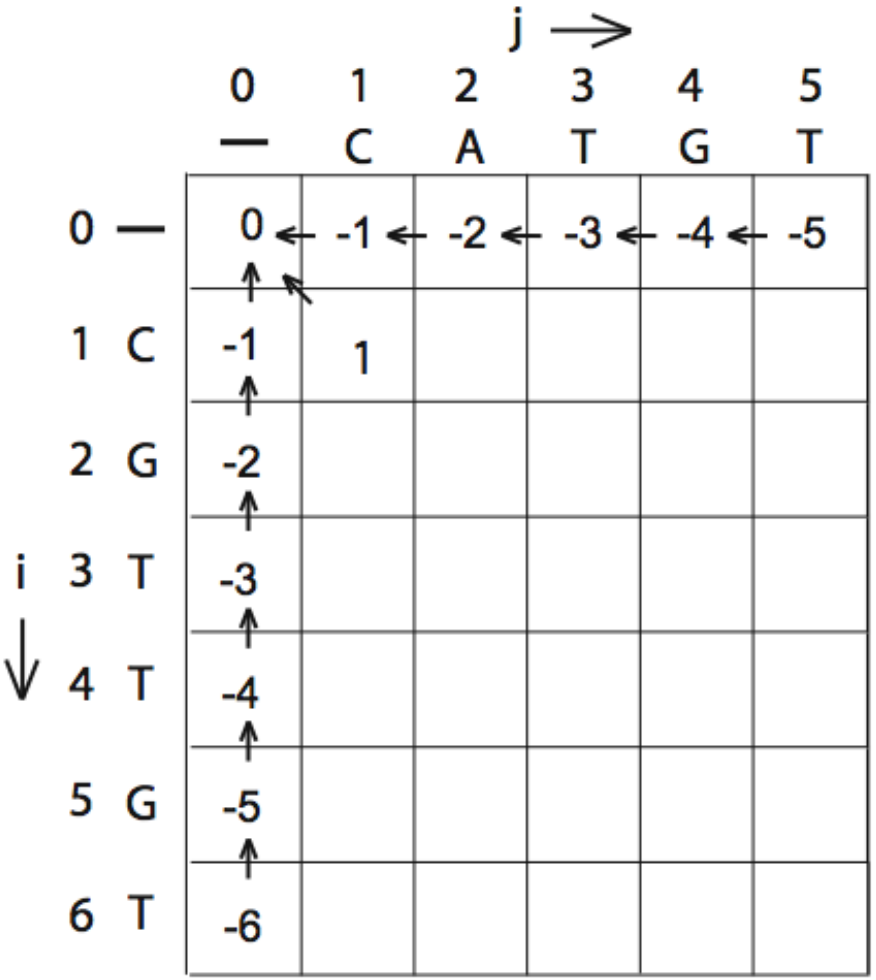
Scoring-System

match=1
mismatch=0
gap=-1



Δ.Π. Ολική στοίχιση παράδειγμα (iv)

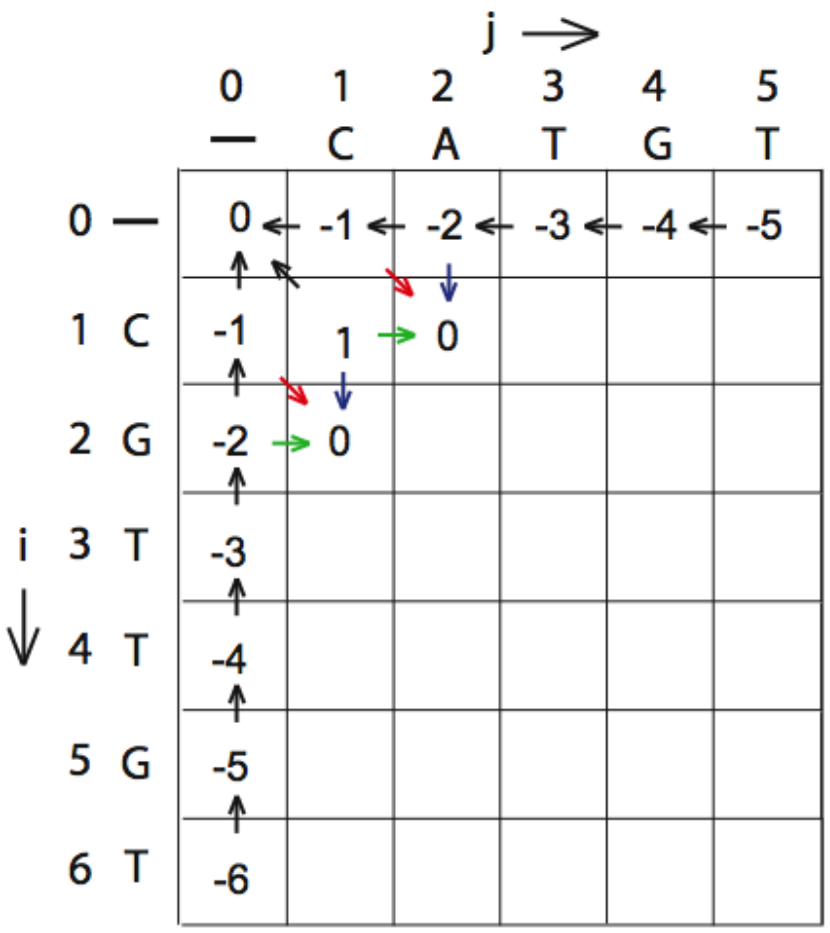
ιχνηλάτηση



Δ.Π. Ολική στοίχιση παράδειγμα (v)

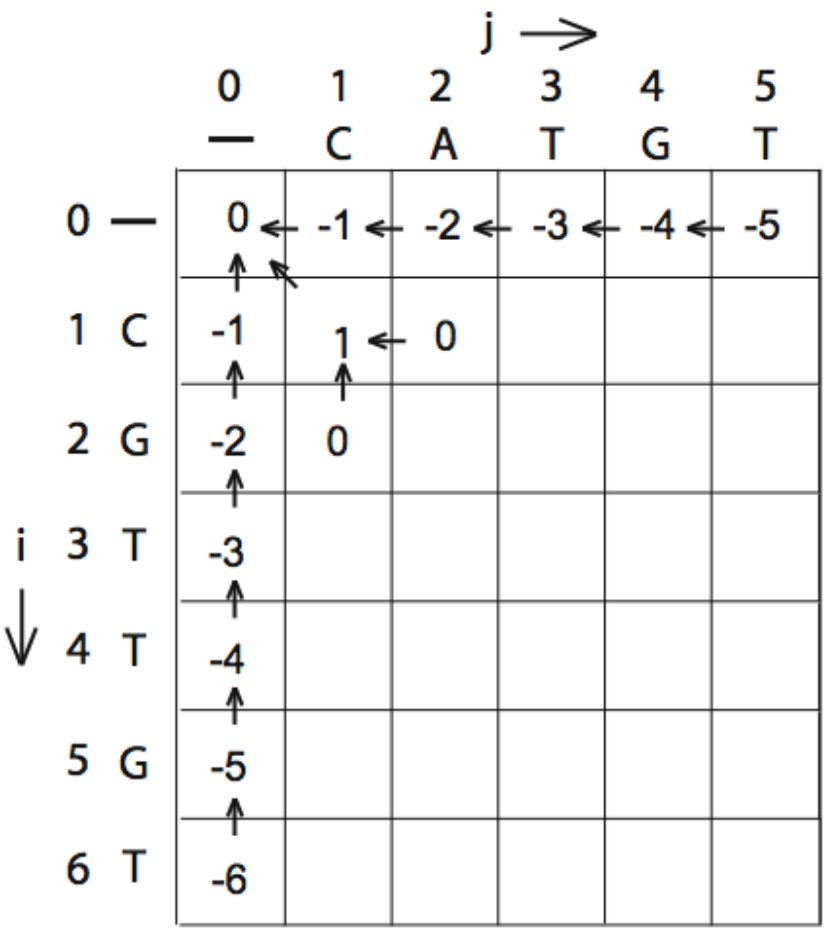
συμπλήρωση

$S_{1,2}$ = MAX [-1+0, 1-1, -2-1] = 0



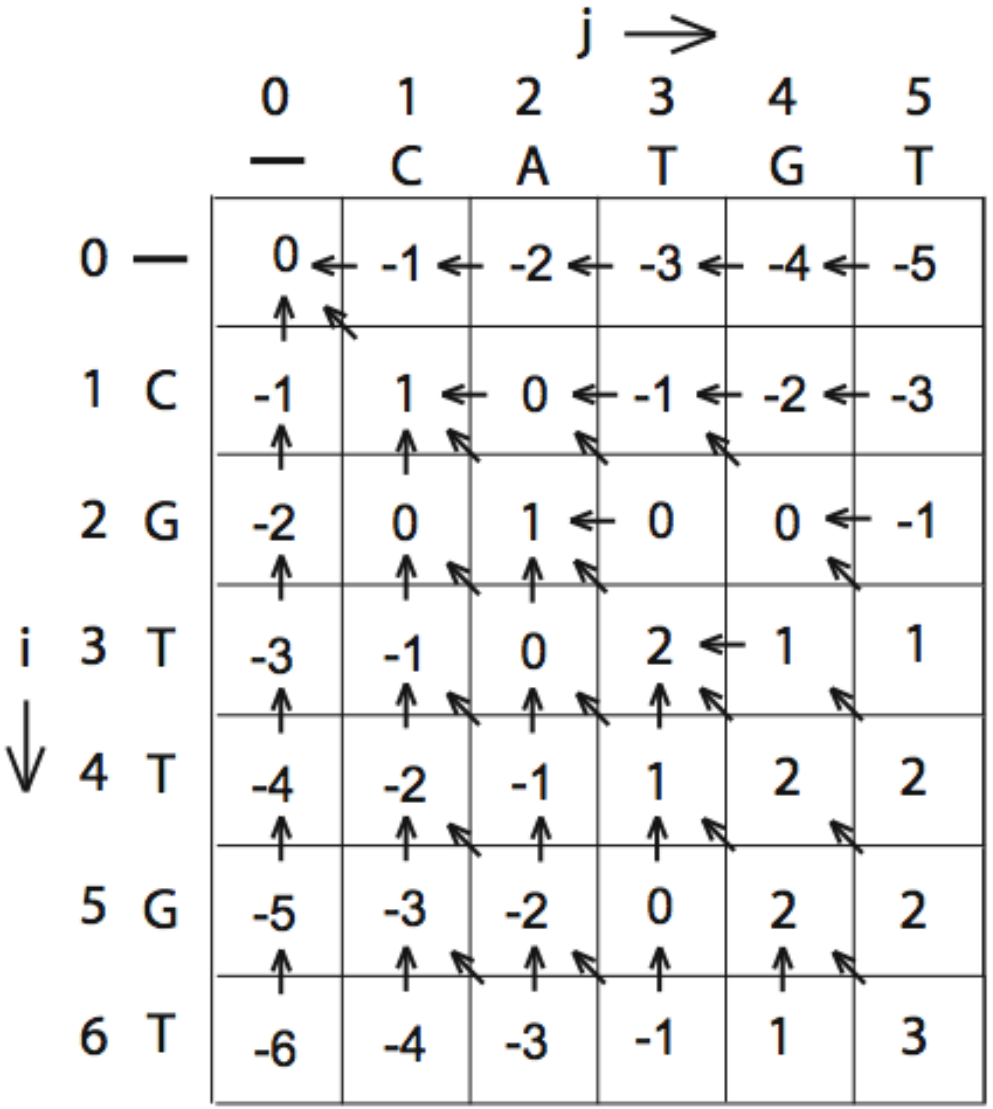
$S_{2,1}$ = MAX [-1+0, -2-1, 1-1] = 0

$S_{1,2}$ = MAX [-1+0, 1-1, -2-1] = 0

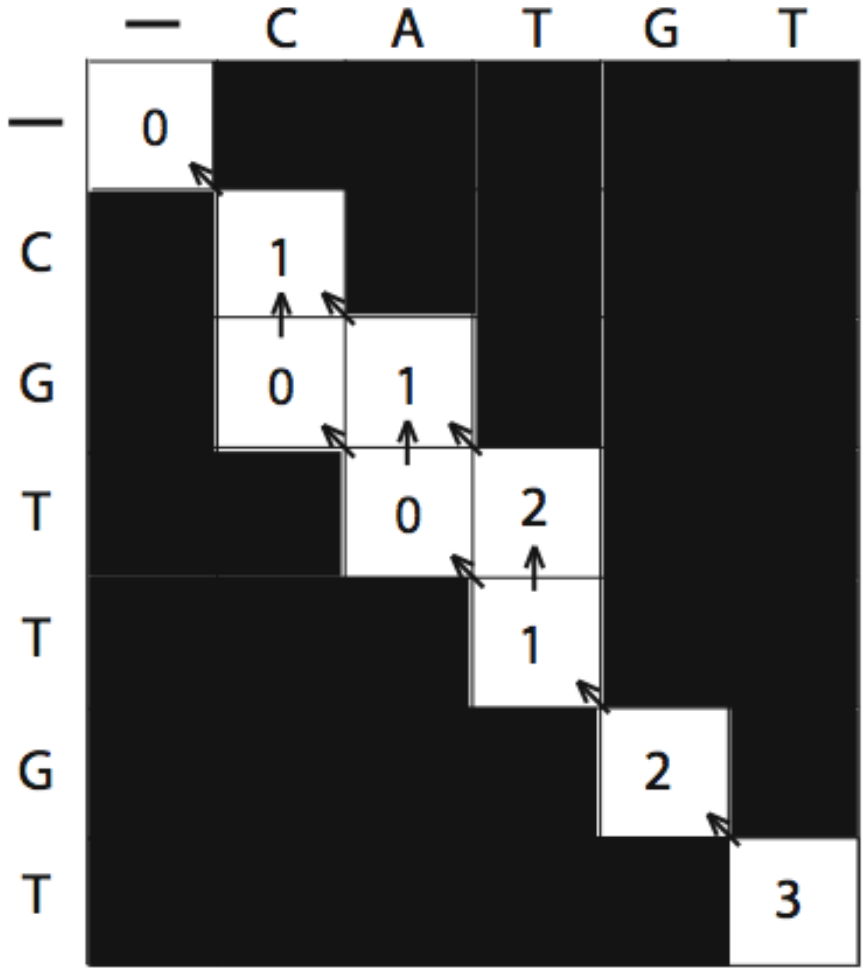


$S_{2,1}$ = MAX [-1+0, -2-1, 1-1] = 0

Δ.Π. Ολική στοίχιση παράδειγμα (vi)



Ολική στοίχιση: ιχνηλάτιση



0←1←1←2←1←2←3

```
C A T — G T   Seq1
I   I   I I
C G T T G T   Seq2
```

0←1←1←0←1←2←3

```
C A — T G T   Seq1
I     I I I
C G T T G T   Seq2
```

0←1←0←0←1←2←3

```
C — A T G T   Seq1
I   I I I I
C G T T G T   Seq2
```

Πρέπει να βρούμε όλες τις δυνατές πορείες από κάτω-δεξιά -> πάνω-αριστερά.

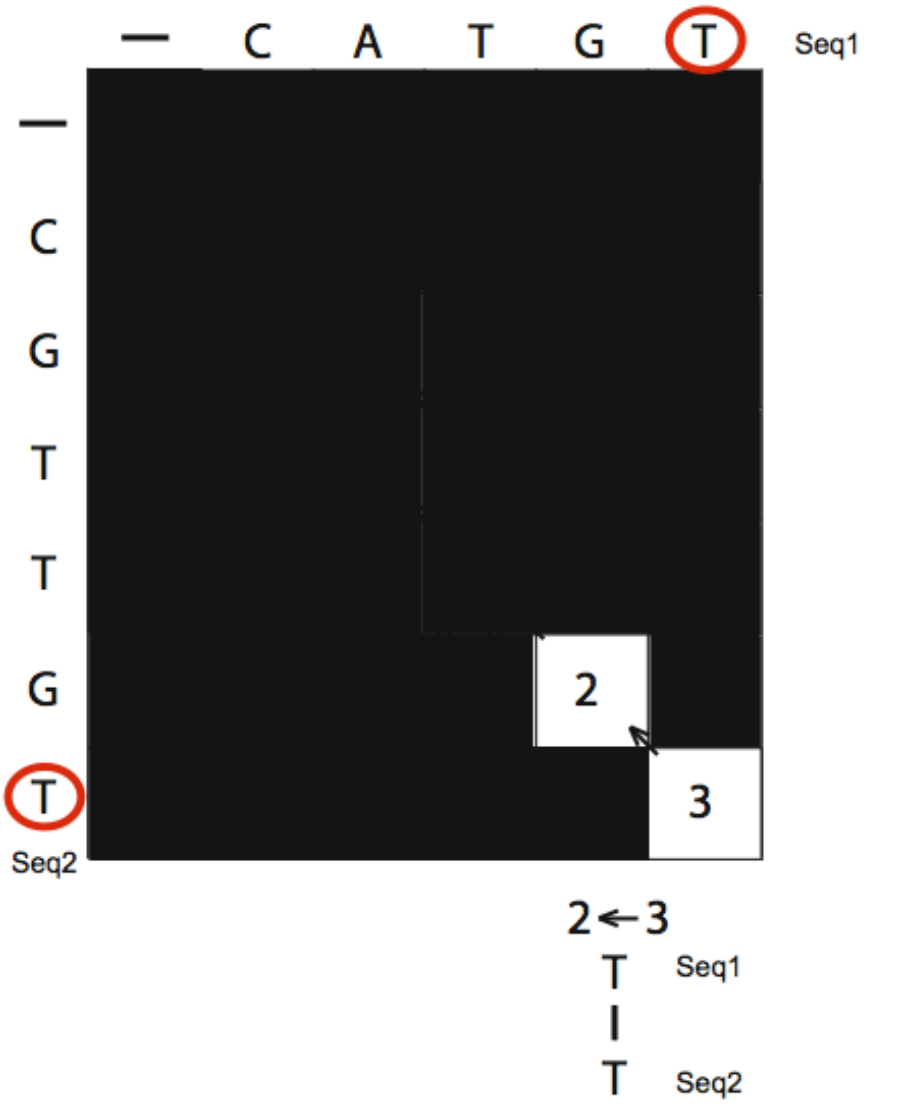
Εδώ: 3 πιθανές πορείες = 3 εξίσου καλές λύσεις

Πώς στοιχίζουμε

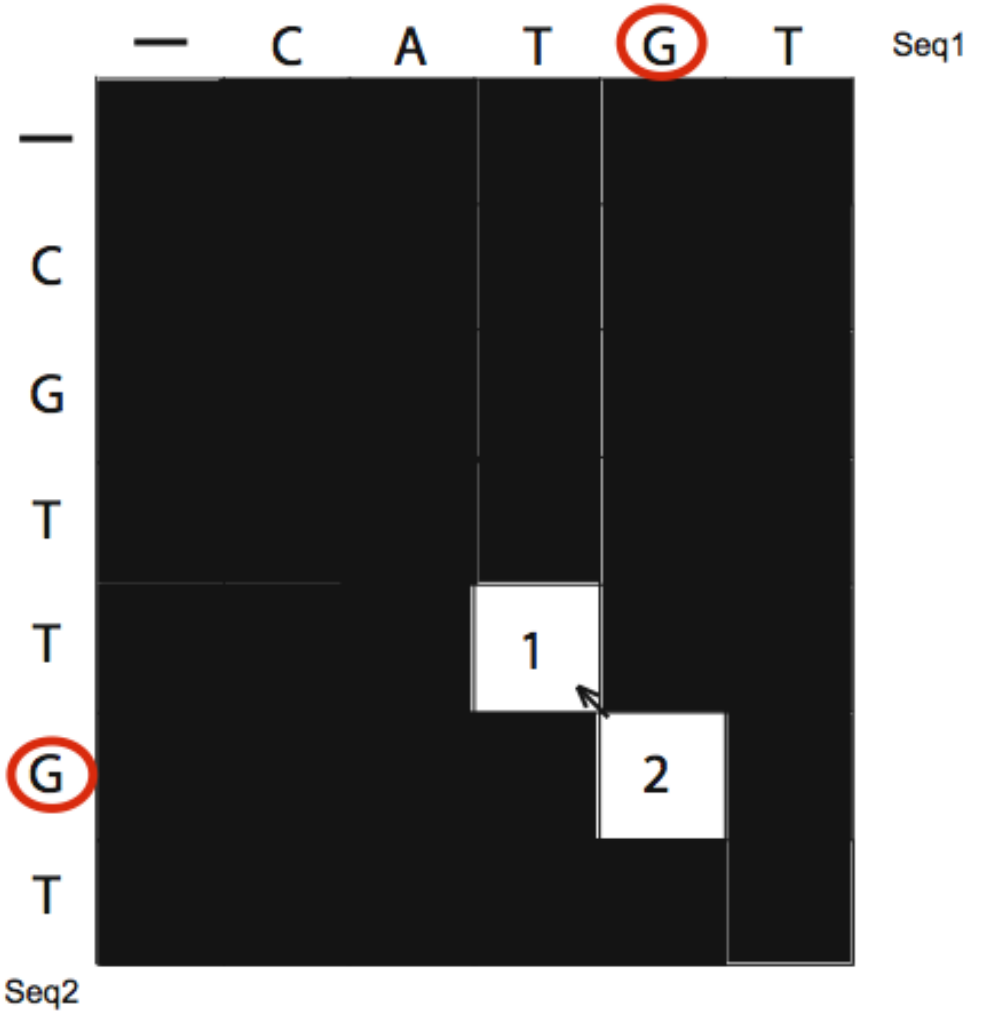
Για κάθε θέση:

- Αν κινηθούμε διαγώνια, τότε στοιχίζουμε τα 2 νουκλεοτίδια/ αμινοξέα που αντιστοιχούν για εκείνη την θέση (είτε ταιριάζουν είτε όχι).
- Αν κινηθούμε οριζόντια ή κάθετα βάζουμε κενό στην ακολουθία που δείχνει το βέλος

Πώς στοιχίζουμε



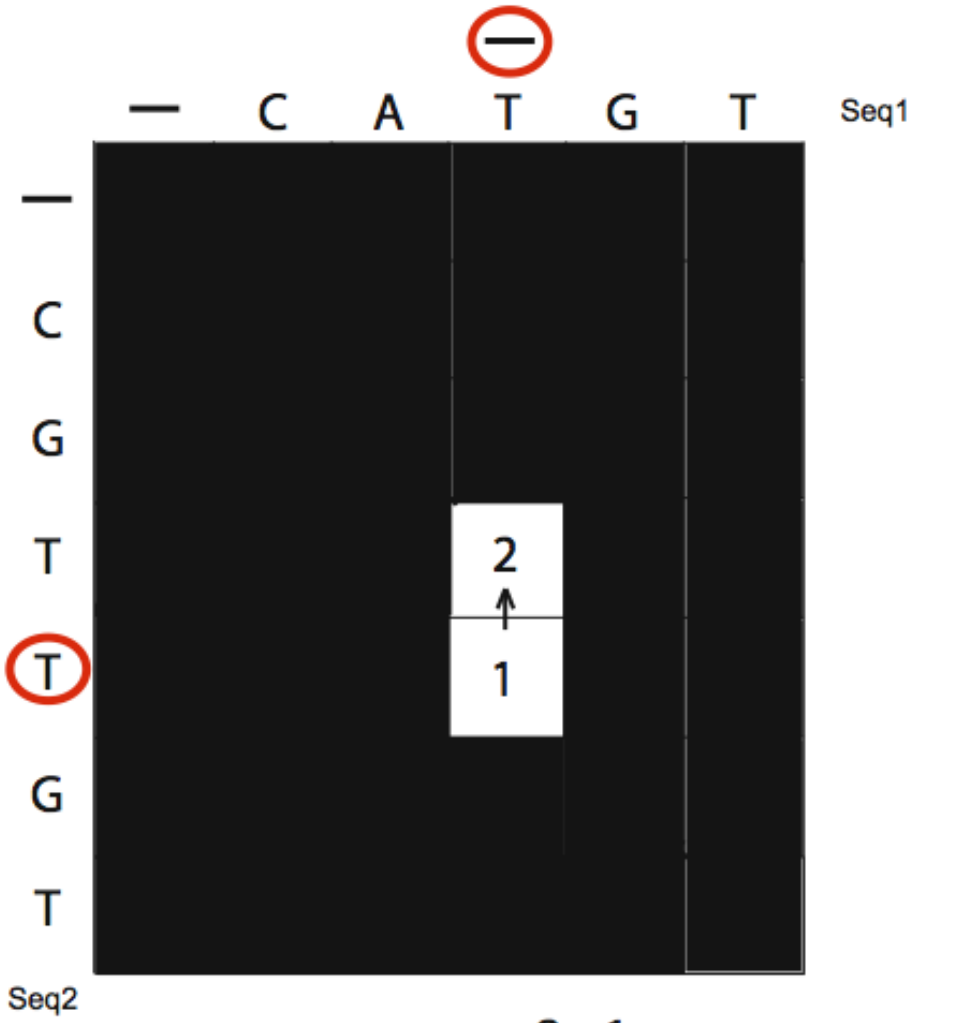
Πώς στοιχίζουμε



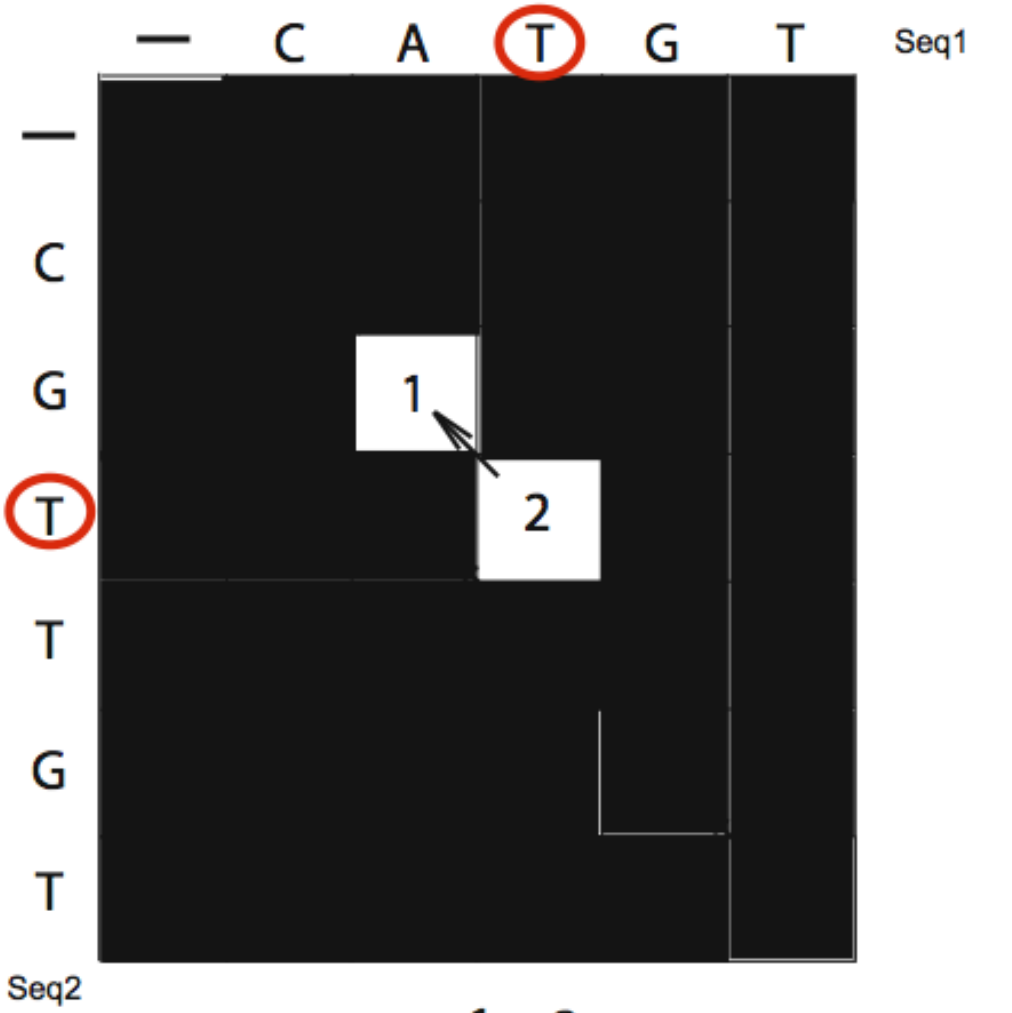
1←2

G	T	Seq1
I	I	
G	T	Seq2

Πώς στοιχίζουμε



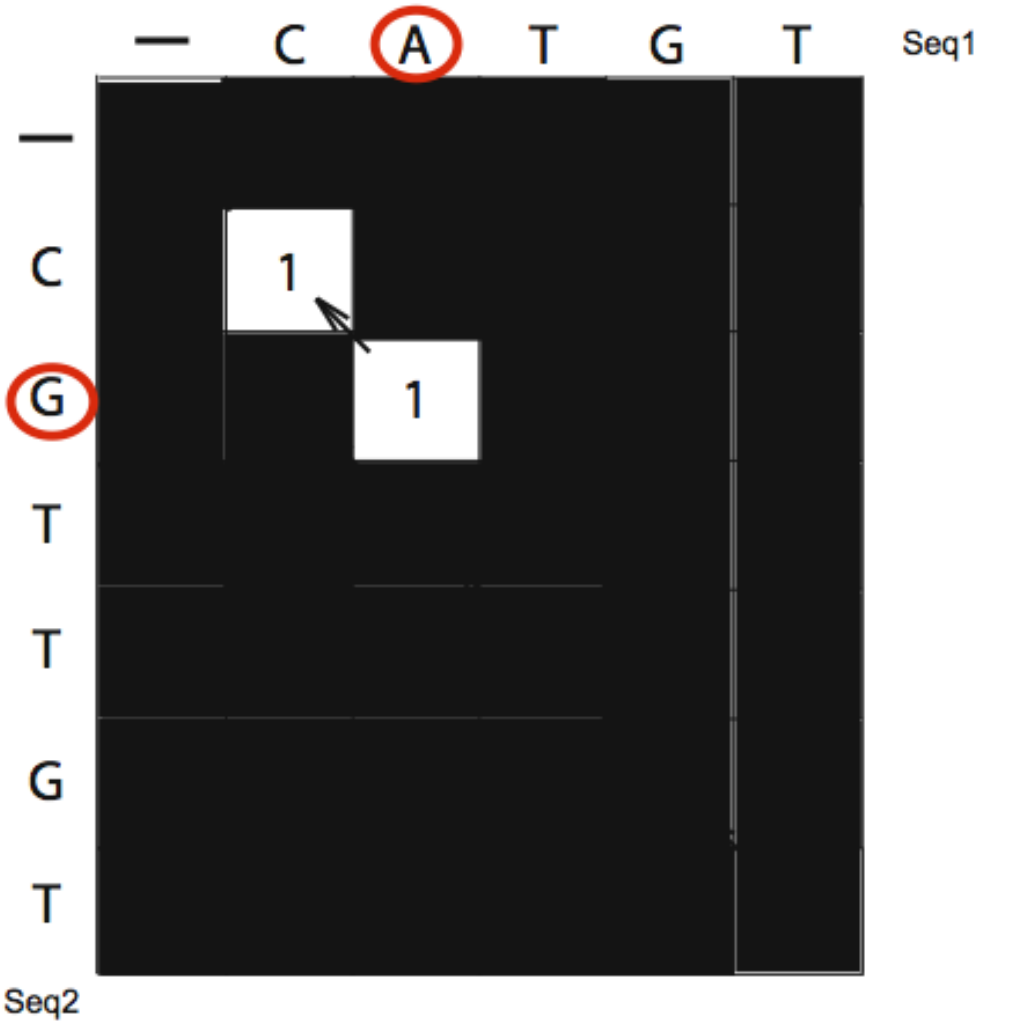
Πώς στοιχίζουμε



1←2

T	—	G	T	Seq1
I	I	I	I	
T	T	G	T	Seq2

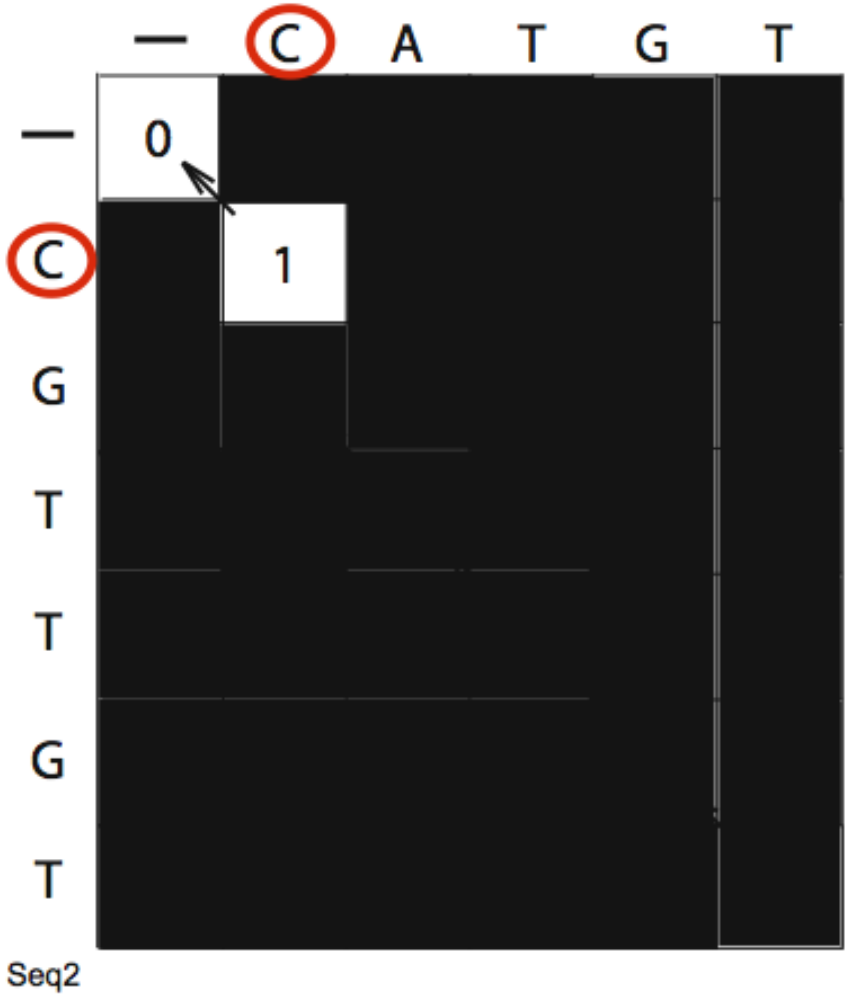
Πώς στοιχίζουμε



1←1

A	T	—	G	T	Seq1
	I		I	I	
G	T	T	G	T	Seq2

Πώς στοιχίζουμε



0←1

C A T — G T Seq1
I I I I
C G T T G T Seq2

Δυναμικός προγραμματισμός τοπική στοίχιση

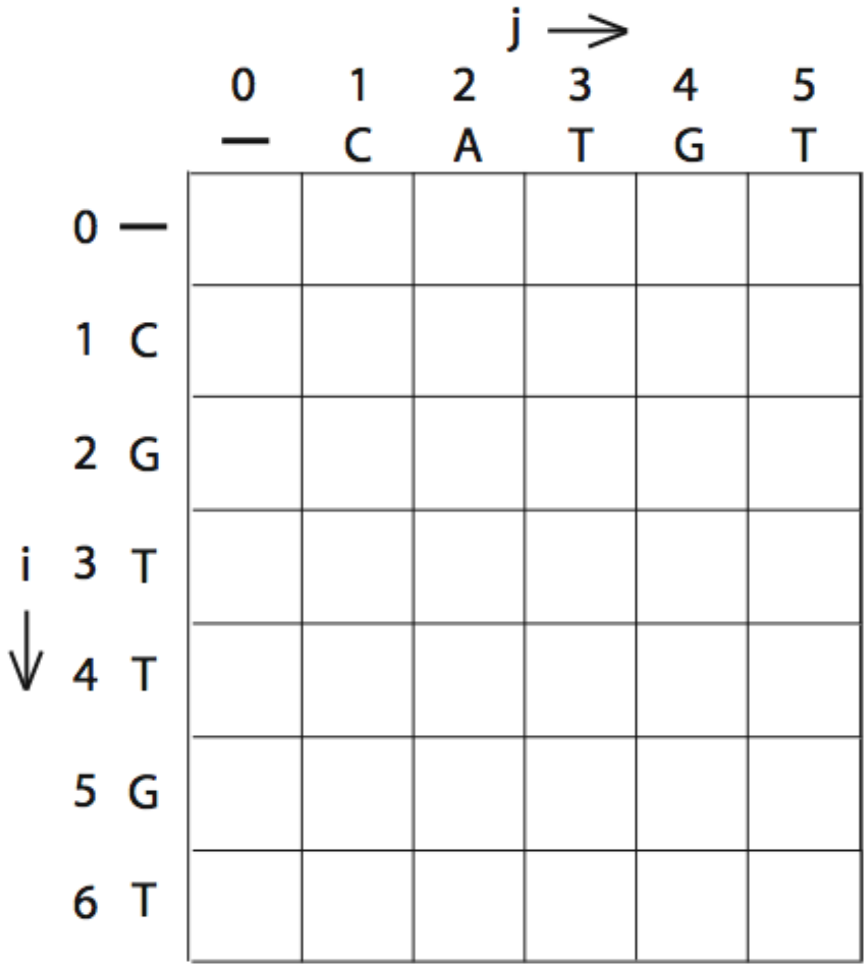
- Ενδείκνυται για
  - μακρομόρια διαφορετικού μεγέθους
  - Συντηρημένη μόνο μια μικρή περιοχή
  - Στοίχιση ώριμου mRNA με το γονίδιό του
  - 2 γονίδια με συντηρημένα εξόνια αλλά αποκλείνοντα ιντρόνια

- Αλγόριθμος Smith-Waterman (1981)

Δυναμικός προγραμματισμός τοπική στοίχιση

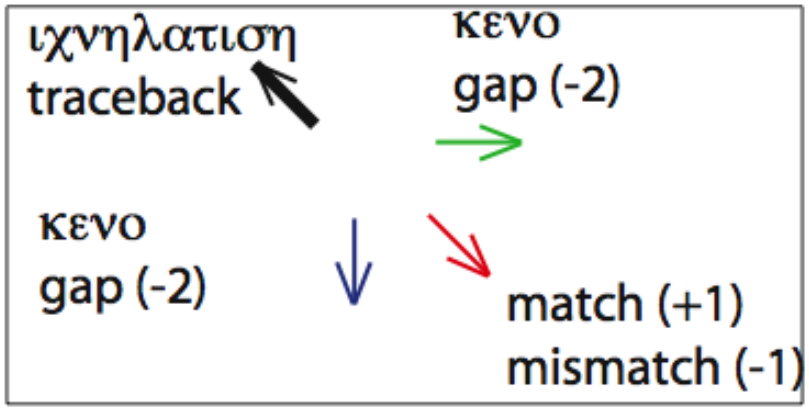
- Αλγόριθμος παρόμοιος με ολική στοίχιση
- Διαφορές:
  - Οι ασυμφωνίες δίνουν αρνητική βαθμολογία.
  - Όταν μια τιμή του πίνακα βγαίνει αρνητική, μηδενίζεται.
  - Βρίσκουμε την καλύτερη τοπική στοίχιση ξεκινώντας από το κουτάκι με την υψηλότερη τιμή και ακολουθούμε την ιχνιλάτηση μέχρι να καταλήξουμε σε ένα κουτάκι με τιμή 0.

Δ.Π τοπική στοίχιση παράδειγμα (i)

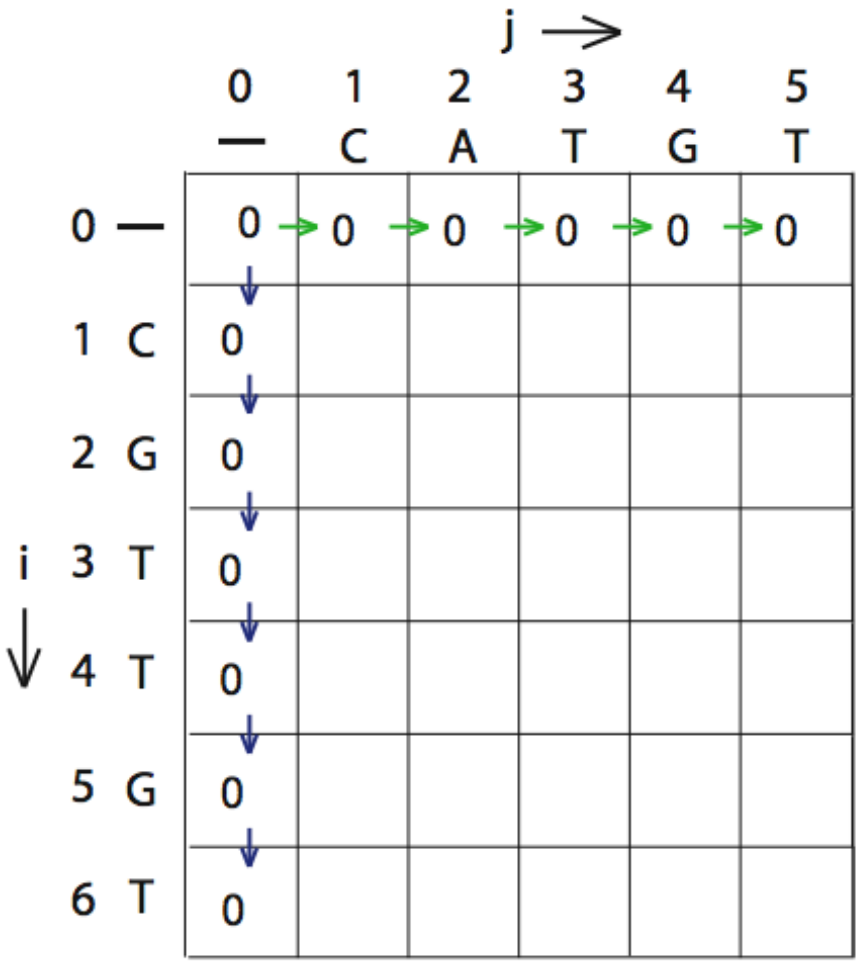


Scoring-System

match=1
mismatch=-1
gap=-2

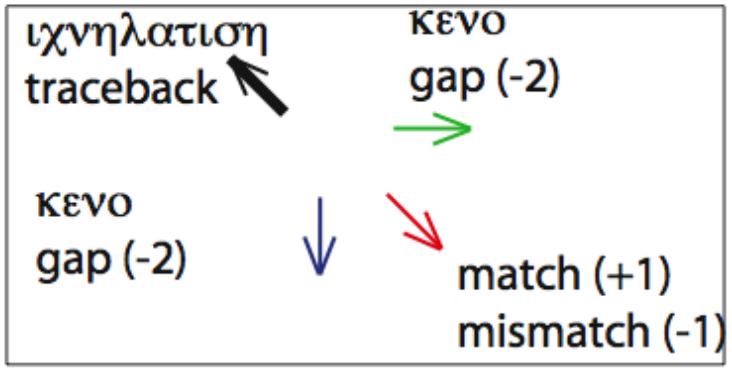


Δ.Π τοπική στοίχιση παράδειγμα (ii)



Scoring-System

match=1
mismatch=-1
gap=-2



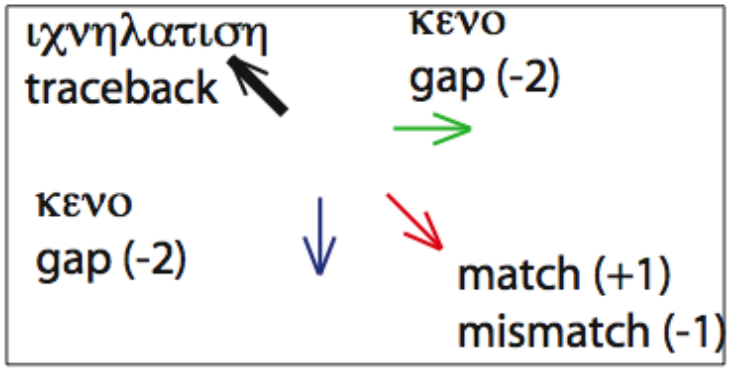
Δ.Π τοπική στοίχιση παράδειγμα (iii)

$S_{1,1}$ = MAX [0+1, 0-2, 0-2, 0] = 1

i \ j	0 —	1 C	2 A	3 T	4 G	5 T
0 —	0	0	0	0	0	0
1 C	0	1				
2 G	0					
3 T	0					
4 T	0					
5 G	0					
6 T	0					

Scoring-System

match=1
mismatch=-1
gap=-2



Δ.Π τοπική στοίχιση παράδειγμα (iv)

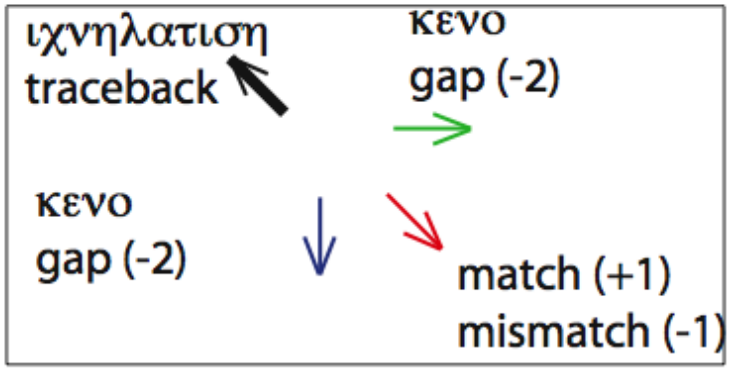
$S_{1,2}$ = MAX [0-1, 1-2, 0-2, 0] = 0

		j → 0	1	2	3	4	5
		—	C	A	T	G	T
0	—	0	0	0	0	0	0
1	C	0	1	0			
2	G	0	0				
i ↓ 3	T	0					
4	T	0					
5	G	0					
6	T	0					

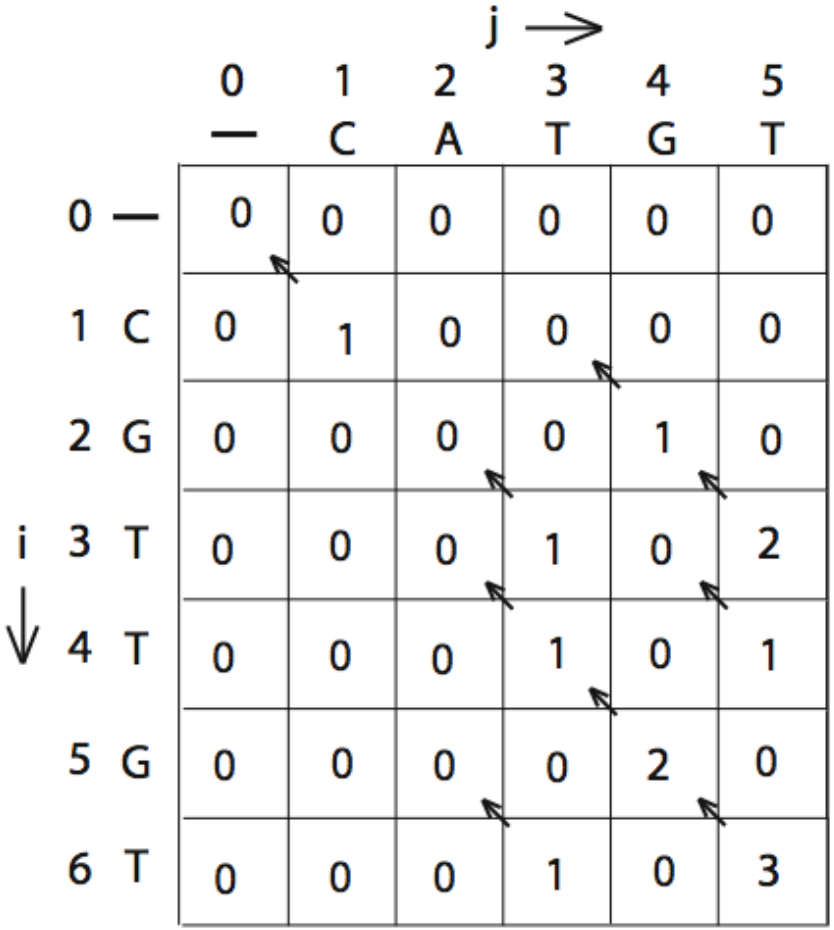
$S_{2,1}$ = MAX [0-1, 0-2, 1-2, 0] = 0

Scoring-System

match=1
mismatch=-1
gap=-2



Δ.Π τοπική στοίχιση παράδειγμα (v)



Βρίσκουμε την καλύτερη τοπική στοίχιση ξεκινώντας από το κουτάκι με την υψηλότερη τιμή και ακολουθούμε την ιχνιλάτηση μέχρι να καταλήξουμε σε ένα κουτάκι με τιμή 0.

0←1←2←3

T	G	T
I	I	I
T	G	T

Αν και στο συγκεκριμένο παράδειγμα ξεκινάμε από κάτω δεξιά, γενικά αυτό δεν είναι απαραίτητο, όπως είναι στην ΟΛΙΚΗ στοίχιση (Δυναμικό προγραμματισμό).

Πίνακες αντικατάστασης

- Στο παράδειγμα του Δυναμικού Προγραμματισμού, όλες οι συμφωνίες/ασυμφωνίες είχαν το ίδιο σκορ.
- Στην πράξη, πιο περίπλοκα συστήματα βαθμολόγισης. Μια ασυμφωνία μεταξύ δύο πουρινών δεν είναι το ίδιο με μια ασυμφωνία μεταξύ πουρίνης-πυριμιδίνης. Διαφορετικές συχνότητες μεταλλάξεων.
- Το ίδιο και για τις πρωτεΐνες.
- Χρειαζόμαστε πίνακες που βασίζονται σε συγκεκριμμένα εξελικτικά μοντέλα και λαμβάνουν υπόψην την συχνότητα του κάθε χαρακτήρα

Πίνακες αντικατάστασης

- Για πρωτεΐνες:
  - Πίνακες PAM
  - Πίνακες BLOSUM

Μεγαλύτερη πιθανότητα αντικατάστασης μεταξύ αμινοξέων με παρόμοιες φυσικοχημικές ιδιότητες, (συντηρητικές αντικαταστάσεις).

Λογαριθμικές πιθανότητες

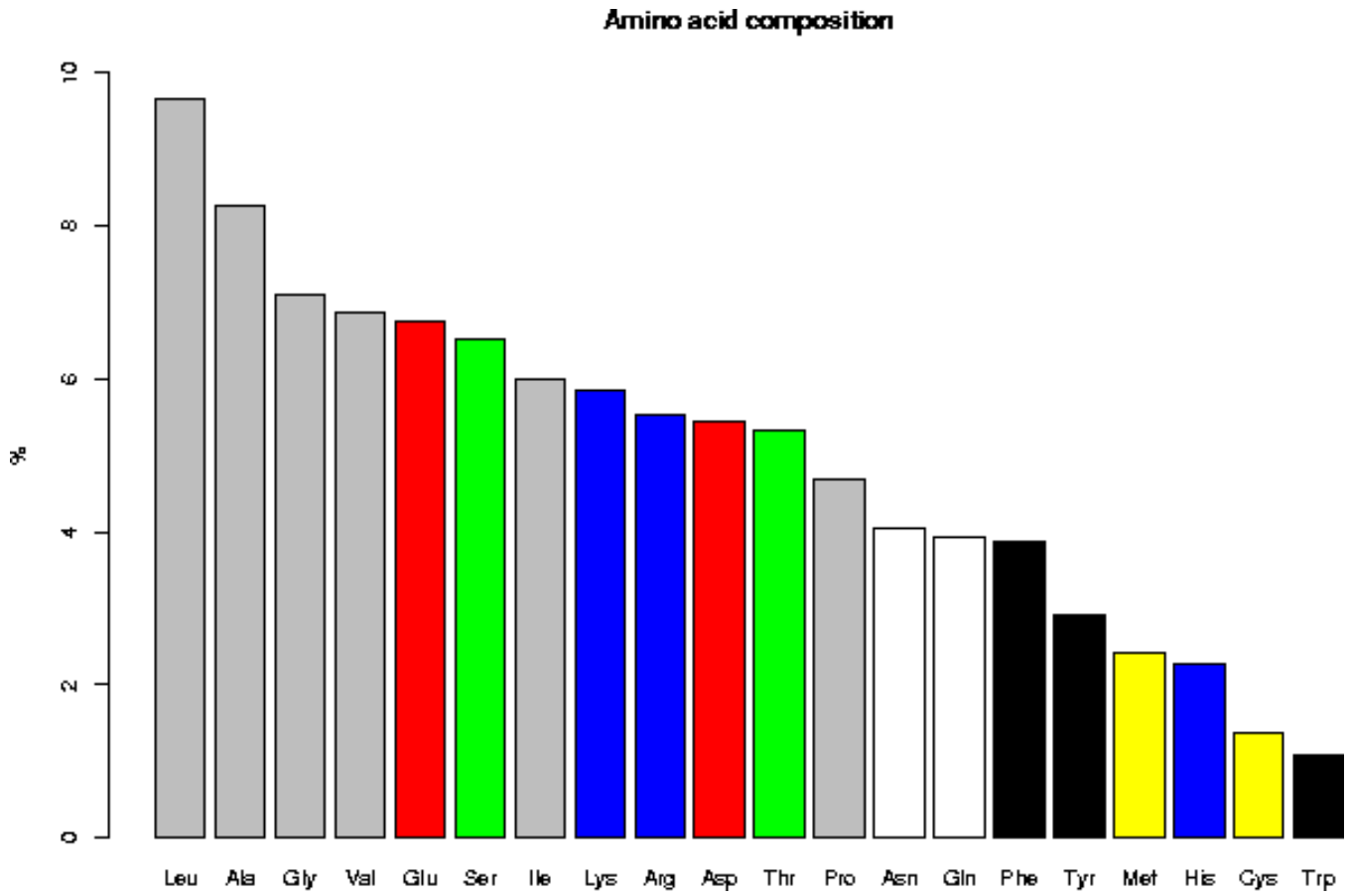
- Πρώτη χρήση από Dayhoff για πίνακες αντικατάστασης που χρησιμοποιούνται στη βαθμολόγιση στοιχίσεων.

- Βαθμολογία αντικατάστασης ~ log(συχνότητα στόχων / συχνότητα υποβάθρου)

- Συχνότητα στόχων: παρατηρηθείσες συχνότητες αντικατάστασης σε στοιχίσεις υπαρκτών και ομόλογων πρωτεΐνών. Χρησιμοποιούμε στοιχίσεις που έγιναν με το ‘μάτι’ και είμαστε σίγουροι ότι είναι σωστές.

- Συχνότητα υποβάθρου: προκύπτει από τις συνολικές συχνότητες των αμινοξέων στις πρωτεΐνες. Υποθέτουμε ότι δεν υπάρχει εξελικτική πίεση στις αντικαταστάσεις.

Συχνότητα αμινοξέων
από Swissprot



Πίνακες PAM

- Dayhoff *et al*., 1978
- PAM -> Percent Accepted Mutations
- Βασίστηκε σε 1572 αποδεκτές αντικαταστάσεις από 71 groups εξελικτικά ‘κοντινών’ ομόλογων ακολουθιών.
- 1 PAM -> μονάδα εξελικτικής απόκλισης, όπου 1% των αμινοξέων έχει αλλάξει.

C Cys	12																		
S Ser	0	2																	
T Thr	-2	1	3																
P Pro	-3	1	0	6															
A Ala	-2	1	1	1	2														
G Gly	-3	1	0	-1	1	5													
N Asn	-4	1	0	-1	0	0	2												
D Asp	-5	0	0	-1	0	1	2	4											
E Glu	-5	0	0	-1	0	0	1	3	4										
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4									
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6								
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Ανομοιογενής ρυθμός εξέλιξης για τις οικογένειες πρωτεϊνών. Άρα, 1 PAM σημαίνει διαφορετικό χρόνο εξέλιξης για την κάθε οικογένεια.

Για 250 μονάδες PAM, θα υπάρχει απόκλιση 100% μεταξύ δύο ομόλογων ακολουθιών;

Πίνακες PAM (ii)

- Όχι. Απόκλιση ~80%.
- Μερικές θέσεις μπορεί να έχουν υποστεί περισσότερες από μία αντικαταστάσεις, ή ακόμα και να έχουν επανέλθει στο αρχικό αμινοξύ!
- Το κάθε αμινοξύ θα έχει αποκλίνει σε διαφορετικό βαθμό. Π.χ. αμετάβλητες θα παραμείνουν 55% Trp, 6% Asn.

Πίνακες PAM (iii)

- Θετική τιμή στον πίνακα, μεταξύ δύο αμινοξέων -> πιο πιθανό να συναντήσουμε αυτό το ζεύγος σε μια στοίχιση μεταξύ ομόλογων ακολουθιών
- Αρνητική τιμή στον πίνακα, μεταξύ δύο αμινοξέων -> πιο απίθανο να συναντήσουμε αυτό το ζεύγος σε μια στοίχιση μεταξύ ομόλογων ακολουθιών
- Ζεύγη αμινοξέων με παρόμοιες ιδιότητες έχουν θετικές τιμές log-odds

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Πίνακες PAM (iv)

- Στις στοιχίσεις χρησιμοποιήθηκαν ακολουθίες που είχαν αποκλείνει πολύ λίγο μεταξύ τους (απόσταση 1 PAM).
- Αναγωγή σε απόσταση 250 PAM (Πίνακας PAM250). Πολλαπλασιάστηκε ο PAM1 X 250 φορές με τον εαυτό του
- Σειρά πινάκων. Εμπειρικά προτάθηκε για γενική χρήση ο PAM250
- Όσο μεγαλώνει το νούμερο, μεγαλώνει και η εξελικτική απόσταση.
- Για στοίχιση ακολουθιών με μικρή εξελικτική απόσταση, χρησιμοποιούμε πίνακες PAM με μικρά νούμερα.
- Οι πίνακες PAM δημιουργήθηκαν από ακολουθίες με μικρή εξελικτική απόσταση και επομένως είναι προτιμότερο να χρησιμοποιούνται για στοίχιση ‘κοντινών’ ακολουθιών

Πίνακες PAM (iv)

- Εγγενείς ατέλειες:
  - Δεν λαμβάνεται υπόψην ο διαφορετικός βαθμός συντήρησης των περιοχών μιας πρωτεΐνης.
  - Κάθε αντικατάσταση θεωρείται:
    - ανεξάρτητη από προηγούμενες αντικαταστάσεις στην ίδια θέση.
    - Ανεξάρτητη από τα γειτονικά αμινοξέα

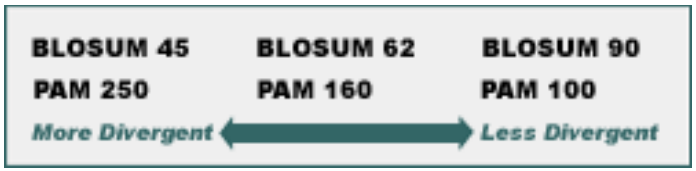
Πίνακες BLOSUM

- BLOcks SUbstitution Matrix
- Henikoff & Henikoff, 1992.
- Χρησιμοποίησαν τοπικές πολλαπλές στοιχίσεις από συντηρημένες περιοχές εξελικτικά απομακρυσμένων ακολουθιών (Β.Δ BLOCKS).
- Και εδώ σειρά πινάκων με διαφορετικά νούμερα.
- BLOSUM62 : Ακολουθίες με ομοιότητα 62% και παραπάνω ομαδοποιούνται.
- Δεν κάνουν αναγωγές στην εξελικτική απόσταση σε αντίθεση με τις PAM.

Βασικές διαφορές μεταξύ PAM-BLOSUM

- Ο κάθε πίνακας BLOSUM δημιουργείται από πραγματικά δεδομένα και όχι από αναγωγή ενός αρχικού πίνακα.

- Οι PAM δημιουργήθηκαν από ολική στοίχιση, ενώ οι BLOSUM από τοπική στοίχιση καλά συντηρημένων περιοχών.



Πίνακες αντικατάστασης νουκλεοτιδίων

- Μοντέλο Jukes-Cantor: Ενιαίοι ρυθμοί μετάλλαξης
- Μοντέλο Kimura: μεταπτώσεις (transitions) ποιό πιθανές από μεταστροφές (transversions)

Βαθμολόγιση Κενών

- Γραμμική ποινή για τα κενά (affine gap penalty)
  - Μια πολύ υψηλή τιμή για την εισαγωγή ενός κενού και χαμηλότερη τιμή για την επέκταση του κενού

- Επιλογή παραμέτρων εμπειρική!
- Θεωρείται σπάνιο γεγονός η εισαγωγή κενού, όταν όμως συμβαίνει, η επεκτασή του δεν είναι τόσο σπάνια
  - Π.χ. Για BLOSUM62: εισαγωγή κενού -> Ποινή 10-15. Επέκταση κενού -> ποινή 1-2

Βαθμολόγιση μιας στοίχισης με πίνακα αντικατάστασης και affine gap penalty

```
sequence 1   V   D   S   -   C   Y
sequence 2   V   E   S   L   C   Y
SCORE        4   2   4 -11   9   7   SCORE = SUM OF AMINO ACID PAIR SCORES
(26)                                 MINUS SINGLE GAP PENALTY (11) = 15
```

Figure 3.7. Example of scoring a sequence alignment with a gap penalty. The individual alignment scores are taken from an amino acid substitution matrix.

Οδηγίες χρήσης πινάκων

- Για οδηγίες χρήσης:
  - http://www.ebi.ac.uk/help/matrix.html

Guidelines for using matricies

Protein Query Length	Matrix	Open Gap	Extend Gap
>300	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
>300	PAM250	-10	-2
85-300	PAM120	-16	-4
35-85	MDM40	-12	-2
<=35	MDM20	-22	-4
<=10	MDM10	-23	-4

Στατιστική σημαντικότητα μιας στοίχισης κατά ζεύγη

- Περισσότερες πληροφορίες στο:

http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

- Στατιστική σημαντικότητα μιας στοίχισης πρακτικά σημαίνει ότι οι δύο ακολουθίες είναι ομόλογες (κοινή εξελικτική προέλευση)

```
sequence 1   V   D   S   -   C   Y
sequence 2   V   E   S   L   C   Y
SCORE        4   2   4 -11   9   7   SCORE = SUM OF AMINO ACID PAIR SCORES
(26)                                 MINUS SINGLE GAP PENALTY (11) = 15
```

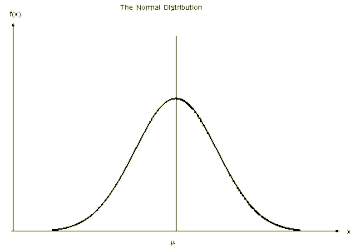
Στατιστική σημαντικότητα ολικής στοίχισης (i)

- Δεν μπορούμε να γνωρίζουμε την κατανομή τυχαίων τιμών μιας ολικής στοίχισης τυχαία επιλεγμένων (μη ομόλογων) ακολουθιών.

- Για κάθε στοίχιση, μπορούμε να πάρουμε την μια ακολουθία και να την ανακατέψουμε πολλές φορές (προσομοίωση). Έτσι διατηρείται η συχνότητα των αμινοξέων στην ακολουθία.

- Για το κάθε ανακάτεμα, υπολογίζουμε τη βαθμολογία της στοίχισης του τυχαίου ζεύγους.

- Θα ήταν λάθος να υποθέσουμε ότι η υπολογισμένη με προσομοιώσεις κατανομή τυχαίων τιμών είναι κανονική. Z-score δεν μπορεί να μετατραπεί σε P-value



Στατιστική σημαντικότητα ολικής στοίχισης (ii)

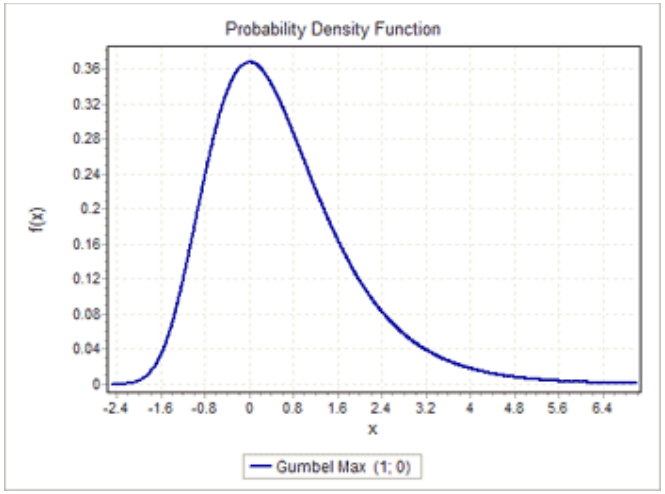
- Αν πραγματοποιηθεί το ανακάτεμα 100 φορές και η μέγιστη βαθμολογία στοίχισης δεν υπερβαίνει την βαθμολογία που παρατηρήσαμε για την στοίχιση των 2 πραγματικών ακολουθιών, τότε η στοίχιση είναι στατιστικά σημαντική σε επίπεδο P-value < 0.01

- Μεγάλο υπολογιστικό κόστος

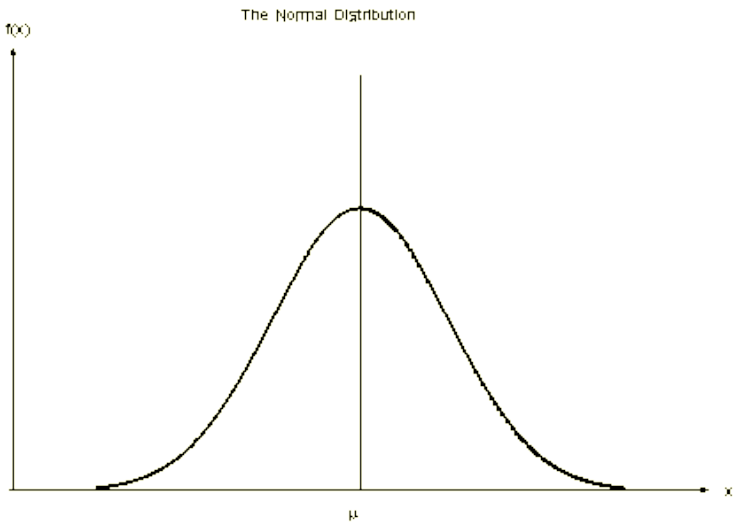
- Χρησιμοποιείται για ολικές στοιχίσεις,εντούτοις δεν ενδείκνυται η ολική στοίχιση για να αποφασίσουμε αν δύο ακολουθίες είναι ομόλογες

Στατιστική σημαντικότητα τοπικής στοίχισης (i)

- Για τοπικές στοιχίσεις χωρίς κενά:
  - αναλυτική μαθηματική θεωρία κατανομής τυχαίων βαθμολογιών.
  - Κατανομή ακραίων τιμών (Extreme value distribution - Gumbel).



- Γιατί όχι κανονική κατανομή;
  - Γιατί σε μια ομοπαράθεση δύο ακολουθιών χρησιμοποιούμε μόνο την βέλτιστη από όλες τις δυνατές στοιχίσεις



Στατιστική σημαντικότητα τοπικής στοίχισης (ii)

Κατανομή ακραίων τιμών Gumbel

- Οι παράμετροι της κατανομής πρέπει να προσαρμοστούν:
  - στο σύστημα βαθμολόγισης
  - Στα μήκη των δύο ακολουθιών
  - στις συχνότητες υποβάθρου των νουκλεοτιδίων/ αμινοξέων

Για τοπικές στοιχίσεις με κενά, δεν υπάρχει αναλυτική μαθηματική θεωρία, έχουν όμως αναπτυχθεί μέθοδοι υπολογισμού.

Στατιστική σημαντικότητα τοπικής στοίχισης (iii)

- Για μια δεδομένη τοπική στοίχιση (χωρίς κενά) δύο ακολουθιών με score S, πόσες τυχαίες στοιχίσεις θα μπορούσαν να δώσουν το ίδιο score ή καλύτερο;

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

- m,n μήκη των ακολουθιών
- S score στοίχισης
- K, λ εξαρτώνται από τη συχνότητα νουκλεοτιδίων/αμινοξέων και το σύστημα βαθμολόγισης.

- Τι σημαίνει για μια στοίχιση, E-value = 1;

- Συνήθως η σημαντικότητα ορίζεται: E-value < 10e-4

Στατιστική σημαντικότητα τοπικής στοίχισης (iv)

- Το raw score μιας τοπικής στοίχισης εξαρτάται από το βαθμολογικό σύστημα που χρησιμοποιήθηκε.
- Χρειάζεται να κανονικοποιηθεί (normalization). Είναι σαν να μιλάμε για απόσταση χωρίς να διευκρινίζουμε αν είναι σε μέτρα ή πόδια.
- Bit score S΄ είναι το κανονικοποιημένο raw score.

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- Το E-value για το κανονικοποιημένο score (bit score)

$$E = mn\,2^{-S'}$$

Αναζήτηση ομόλογων ακολουθιών σε βάσεις δεδομένων (i)

- Ομόλογες ακολουθίες πιθανόν να έχουν παρόμοιες λειτουργίες.

- Ακολουθία επερώτησης (query sequence)
- Υποκείμενες ακολουθίες στην βάση δεδομένων (subject sequences).

- 1 ακολουθία Χ Β.Δ
- Ν ακολουθίες Χ Β.Δ

- Αναζήτηση με δυναμικό προγραμματισμό: Smith-Waterman, SSearch

- Ευρετικοί αλγόριθμοι για ανίχνευση ομόλογων ακολουθιών.
  - FASTA
  - BLAST
- 50 φορές γρηγορότεροι από δυναμικό προγραμματισμό, αλλά ενδέχεται:
  - να μην εντοπίσουν κάποιες ‘απομακρυσμένες’ ομόλογες ακολουθίες.
  - να μη γίνει η βέλτιστη στοίχιση

Αναζήτηση ομόλογων ακολουθιών σε βάσεις δεδομένων (ii)

- Για κάθε στοίχιση μιας ακολουθίας Α με ακολουθίες από την Β.Δ., υπολογίζεται μια βαθμολογία S και κανονικοποιείται (bit score).

- Για μια αναζήτηση σε Β.Δ. γίνονται πολλές στοιχίσεις. Αυτό πρέπει να ληφθεί υπόψην στον υπολογισμό της στατιστικής σημαντικότητας (multiple testing correction).

- Διορθωμένο E-value = E-value X N
- (N=αριθμός ακολουθιών στην Β.Δ.)

- Υπάρχουν παραλλαγές του τρόπου υπολογισμού της στατιστικής σημαντικότητας, για το κάθε πρόγραμμα.

- Διαφορετικός υπολογισμός μεταξύ FASTA - BLAST.

Αλγόριθμος BLAST

http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=comgen&part=blast

- words: λέξεις μήκους W που
  - δεν απαιτείται να ταιριάζουν απόλυτα μεταξύ των πρωτεϊνικών ακολουθιών
  - πρέπει να ταιριάζουν απόλυτα μεταξύ των νουκλεοτιδικών ακολουθιών.

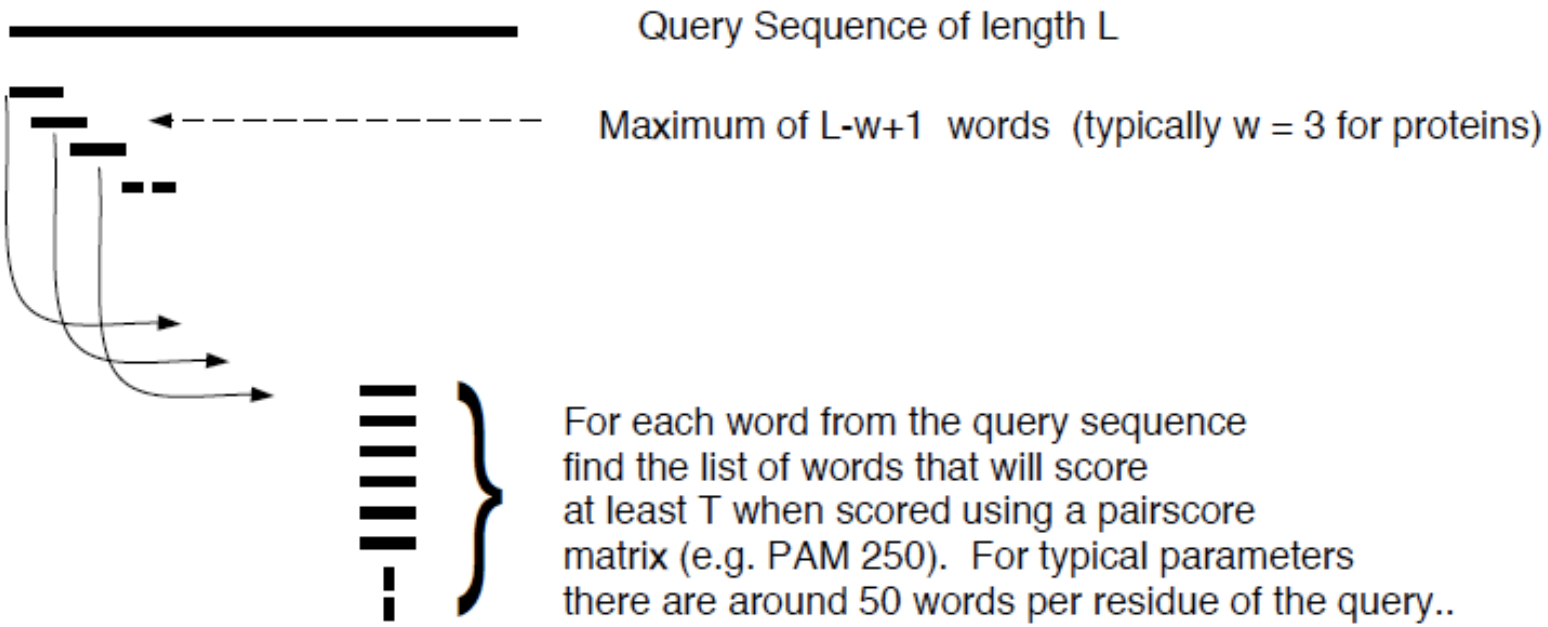
- Πρωτεΐνες: w=3

- Νουκλεϊκά οξέα: w=11

- E-value
  - Default: 10 (για να μη χαθούν ομόλογες ακολουθίες)
  - Συνήθως E-value < 1e-3 (για να απομείνουν ομόλογες ακολουθίες υψηλής εμπιστοσύνης)

Αλγόριθμος BLAST

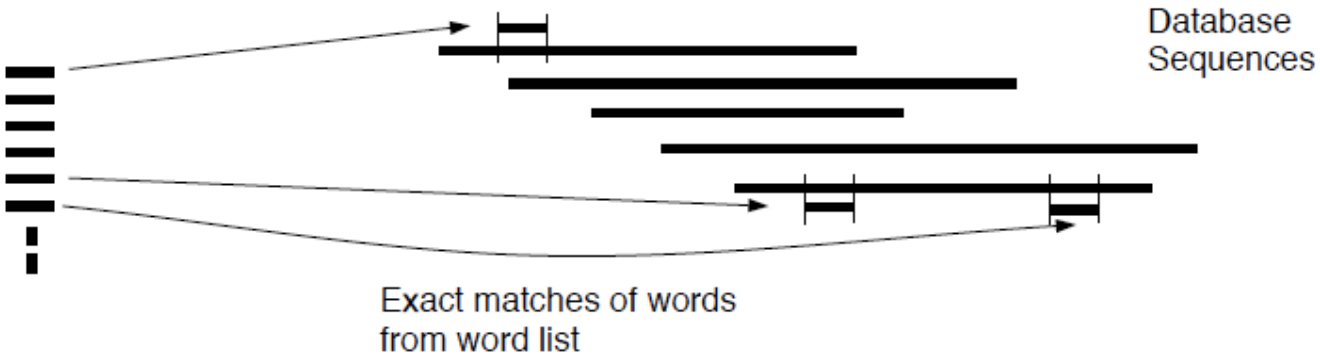
(1) For the query find the list of high scoring words of length w.



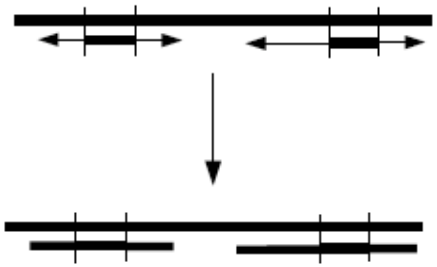
- PQG
- 20 X 20 X 20 = 8.000 words
- PQG X 8.000 words
- PQG X PEG = 7 + 2 + 6=15
- Όριο τιμής T

Αλγόριθμος BLAST

(2) Compare the word list to the database and identify exact matches.



(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.



Maximal Segment Pairs (MSPs)

Περιοχές χαμηλής πολυπλοκότητας (i)

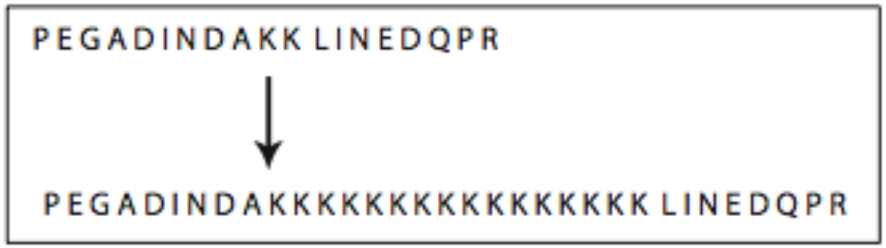
- Low complexity regions
- Επαναλήψεις:
  - poly-A tails
  - Poly-proline tracts
- Tandem repeats:
  KTPKTPKTPKTPKTP
- Interspersed repeats:
  KTPAKTPKTPKTP
- Προκύπτουν από λάθη:
  - Στην μιτωτική αντιγραφή (mitotic replication slippage)
  - Στον μειωτικό ανασυνδυασμό

Περιοχές χαμηλής πολυπλοκότητας (ii)

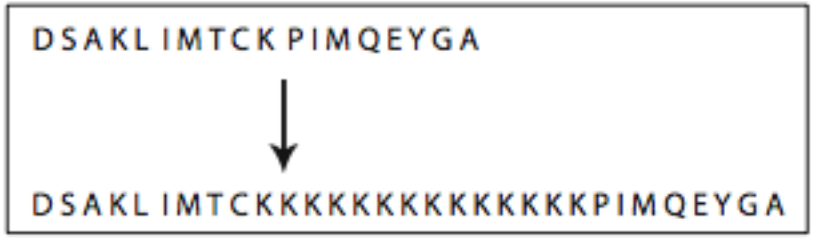
- 2 μη ομόλογες ακολουθίες.

PEGADINDAKK LINEDQPR
DSAKL IMTCK PIMQEYGA

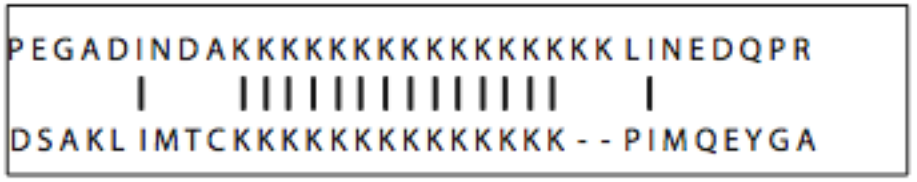
- Μεταλλάξεις στην ακολουθία 1.



- Μεταλλάξεις στην ακολουθία 2.

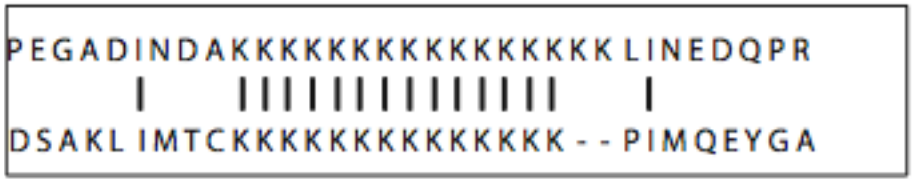


- Αν δεν φιλτραριστούν οι περιοχές χαμηλής πολυπλοκότητας:
  - Η στοίχιση θα δείξει ομολογία



Φιλτράρισμα περιοχών χαμηλής πολυπλοκότητας

- Φιλτράρισμα (masking)
- Και για BLAST και για FASTA.
- Φιλτράρεται η ακολουθία επερώτησης μόνο.
- Χ για πρωτεΐνες και Ν για νουκλεϊκά οξέα (ή μικρά γράμματα)
- Φίλτρα του Blast:
  - Dust: νουκλεοτίδια
  - Seg: πρωτεΐνες
- Άλλες ακολουθίες που μπορεί να φιλτράρονται:
  - Επαναλήψεις Alu
  - Φορείς κλωνοποίησης
  - Διαμεμβρανικές περιοχές
  - Coiled-coils

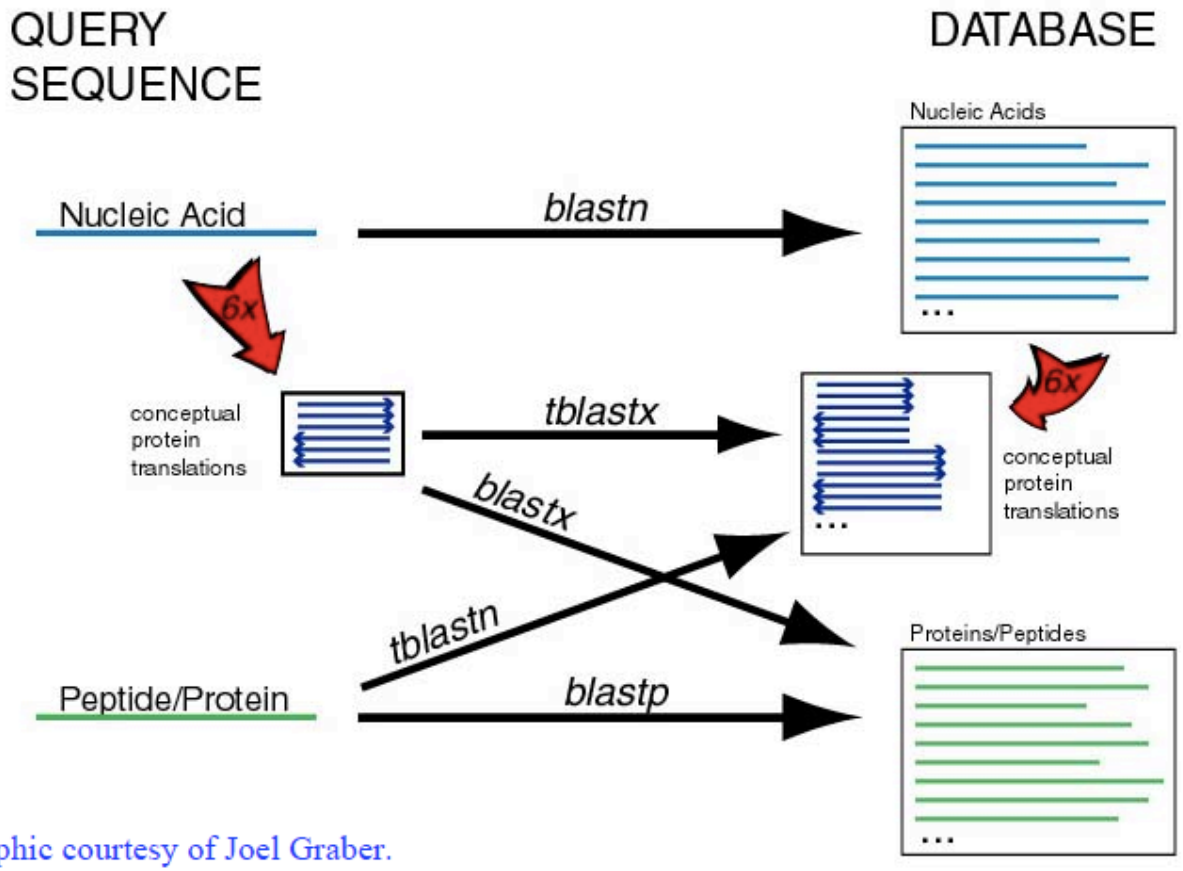


Blast

Blast

Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses
BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein
TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

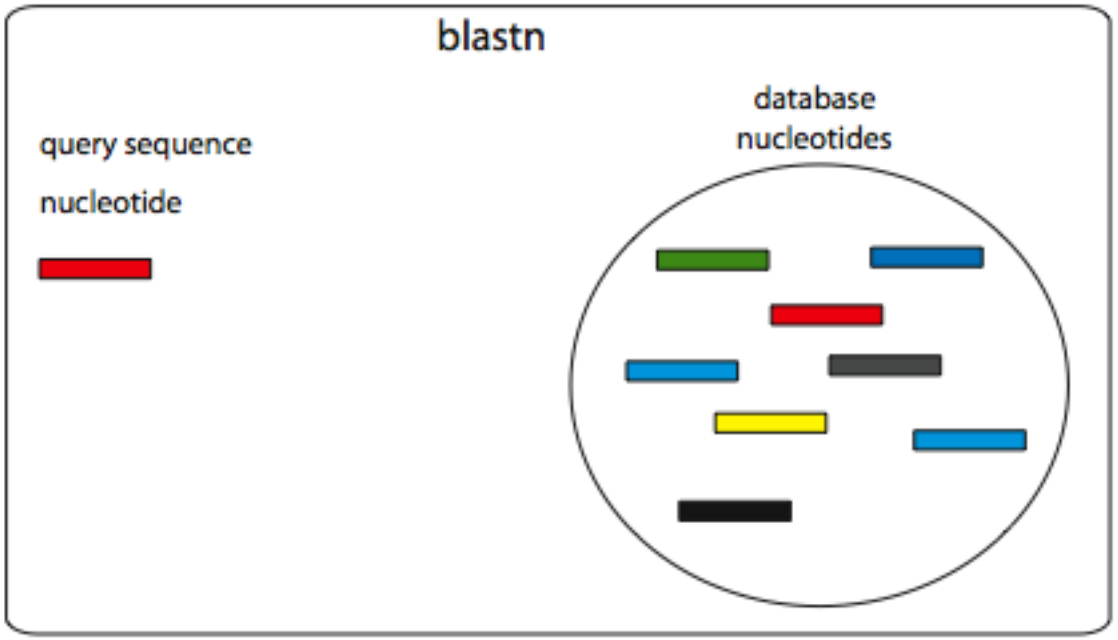
Blast



Graphic courtesy of Joel Graber.

Blastn / MegaBlast

- Blastn
  - Νουκλεοτίδια Χ νουκλεοτίδια
  - Για στοίχιση tRNA, rRNA, mRNA, γενωμικό DNA



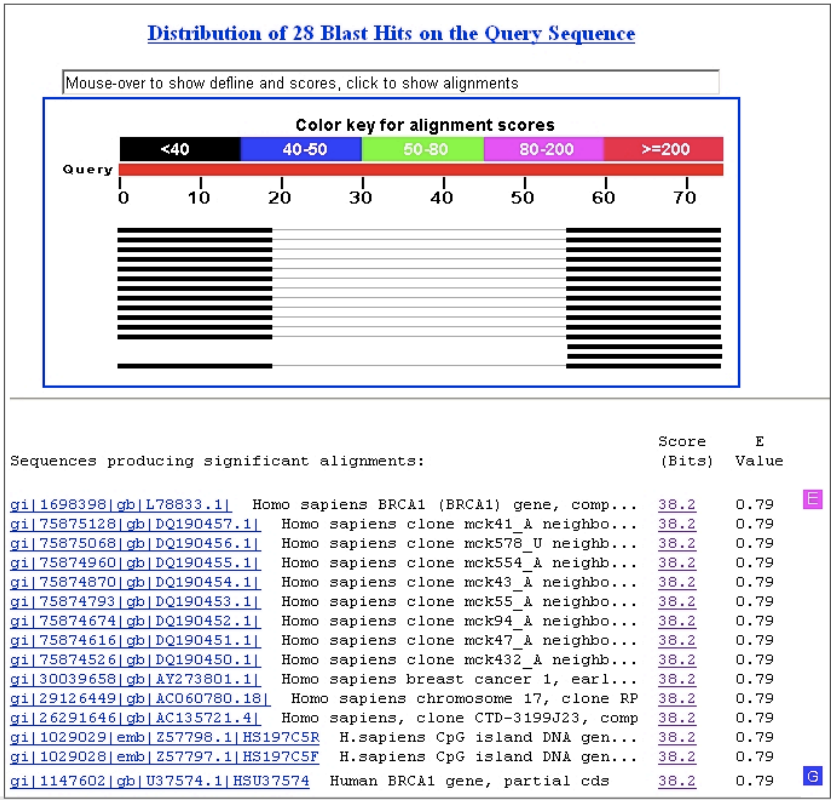
Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping

MegaBlast

- MegaBlast
  - 10X ταχύτερο από Blastn
  - Για στοίχιση ακολουθιών που διαφέρουν πολύ λίγο μεταξύ τους
  - Κυρίως για στοίχιση mRNA με ολόκληρο το γενωμικό DNA

Blastn

Παράδειγμα: Έλεγχος εξειδίκευσης ζεύγους εκκινητών (primers)



Sequences producing significant alignments:		Score (Bits)	E Value
gi|1698398|gb|L78833.1|	Homo sapiens BRCA1 (BRCA1) gene, comp...	38.2	0.79
gi|75875128|gb|DQ190457.1|	Homo sapiens clone mck41_A neighbo...	38.2	0.79
gi|75875068|gb|DQ190456.1|	Homo sapiens clone mck578_U neighb...	38.2	0.79
gi|75874960|gb|DQ190455.1|	Homo sapiens clone mck554_A neighb...	38.2	0.79
gi|75874870|gb|DQ190454.1|	Homo sapiens clone mck43_A neighbo...	38.2	0.79
gi|75874793|gb|DQ190453.1|	Homo sapiens clone mck55_A neighbo...	38.2	0.79
gi|75874674|gb|DQ190452.1|	Homo sapiens clone mck94_A neighbo...	38.2	0.79
gi|75874616|gb|DQ190451.1|	Homo sapiens clone mck47_A neighbo...	38.2	0.79
gi|75874526|gb|DQ190450.1|	Homo sapiens clone mck432_A neighb...	38.2	0.79
gi|30039658|gb|AY273801.1|	Homo sapiens breast cancer 1, earl...	38.2	0.79
gi|29126449|gb|AC060780.18|	Homo sapiens chromosome 17, clone RP	38.2	0.79
gi|26291646|gb|AC135721.4|	Homo sapiens, clone CTD-3199J23, comp	38.2	0.79
gi|1029029|emb|Z57798.1|HS197C5R	H.sapiens CpG island DNA gen...	38.2	0.79
gi|1029028|emb|Z57797.1|HS197C5F	H.sapiens CpG island DNA gen...	38.2	0.79
gi|1147602|gb|U37574.1|HSU37574	Human BRCA1 gene, partial cds	38.2	0.79

> gi|1698398|gb|L78833.1| Homo sapiens BRCA1 (BRCA1) gene, complete cds; ribosomal protein L21-like protein (rpL21) pseudogene, complete sequence; Rho7 (Rho7) and VatI (VatI) genes, complete cds; and unknown (ifp35) gene, exons 1 through 3 and partial cds
Length=117143

Score = 38.2 bits (19), Expect = 0.79
Identities = 19/19 (100%), Gaps = 0/19 (0%)
Strand=Plus/Plus

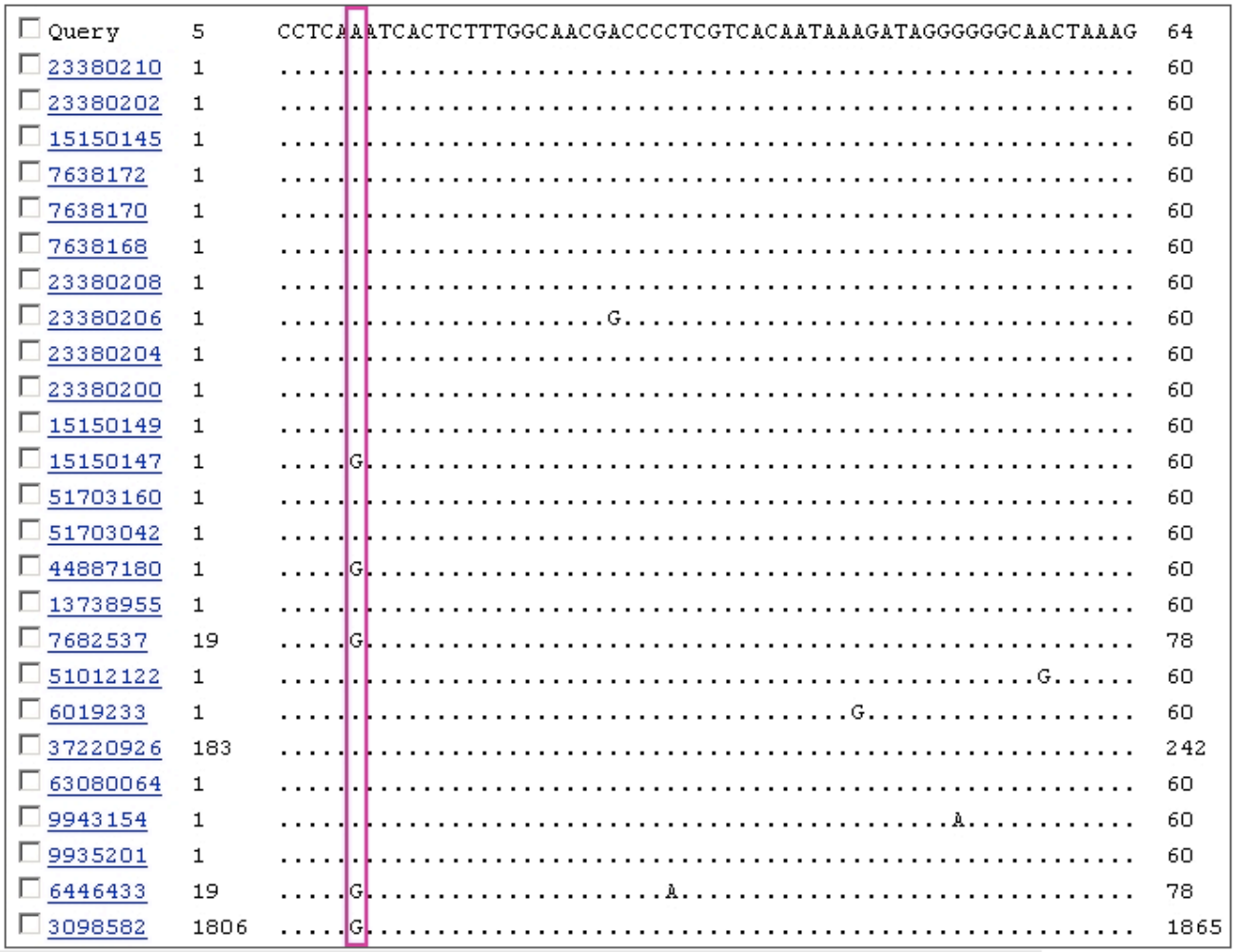
```
Query  1     GTACCTTGATTTCGTATTC  19
             ||||||||||||||||||
Sbjct  3252  GTACCTTGATTTCGTATTC  3270
```

Score = 38.2 bits (19), Expect = 0.79
Identities = 19/19 (100%), Gaps = 0/19 (0%)
Strand=Plus/Minus

```
Query  56    GACTCTACTACCTTTACCC  74
             ||||||||||||||||||
Sbjct  3475  GACTCTACTACCTTTACCC  3457
```

Blastn

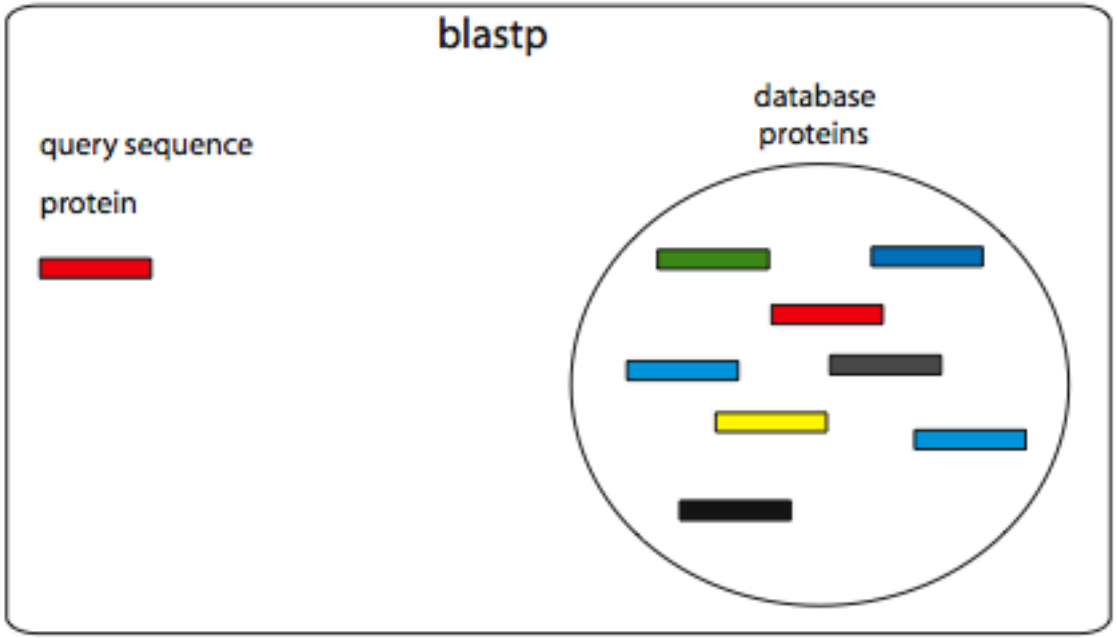
Παράδειγμα: Εντοπισμός SNPs σε ακολουθίες του ιού HIV-1 για ανθεκτικότητα σε φάρμακα



Blastp

- Πρωτεΐνη Χ πρωτεΐνες

- Παράδειγμα:
  - Πρόβλεψη λειτουργίας μιας άγνωστης πρωτεΐνης.
  - Εντοπισμός ορθόλογης πρωτεΐνης σε άλλα είδη.
  - Εντοπισμός όλων των μελών της πρωτεϊνικής οικογένειας στο ίδιο ή σε άλλα είδη



Translated Blast

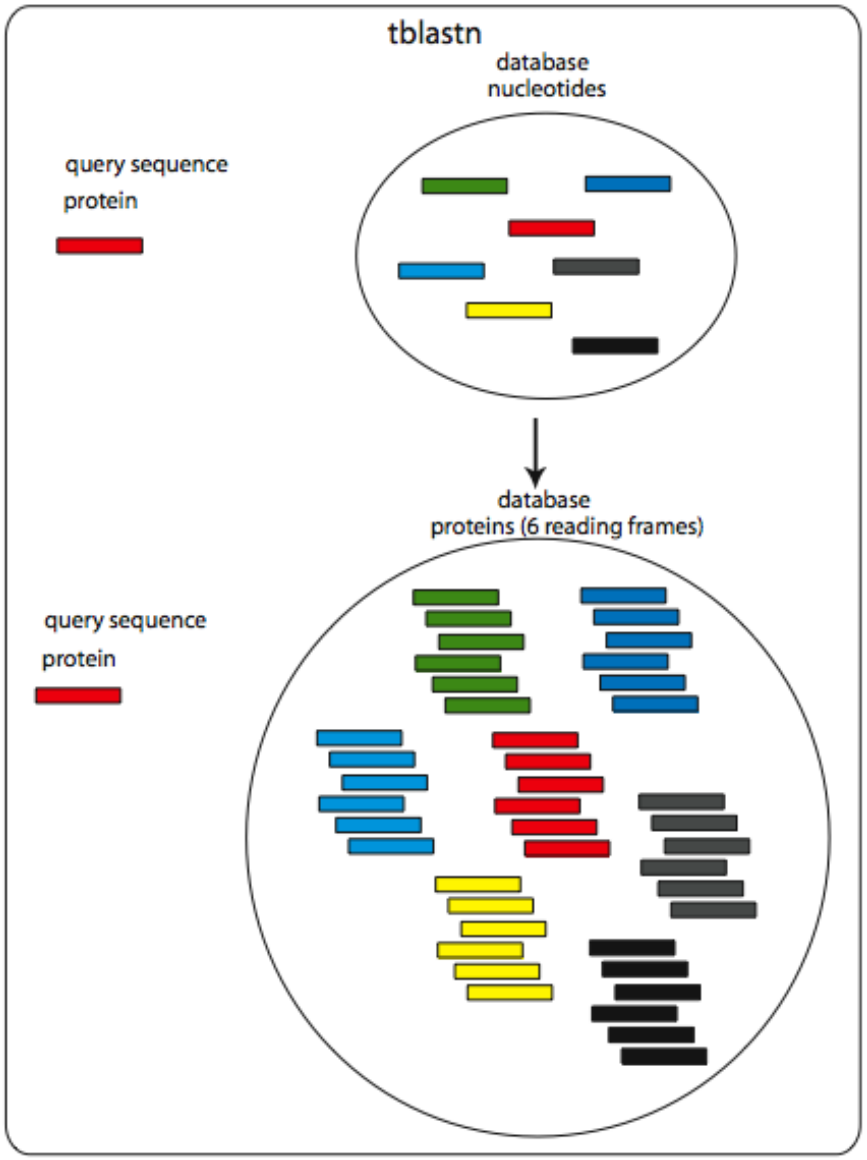
- Η νουκλεοτιδική ακολουθία ενός γονιδίου εμφανίζεται λιγότερο συντηρημένη από την αμινοξική ακολουθία της πρωτεΐνης του.

- Πιο ευαίσθητες μέθοδοι από Blastn για ανίχνευση ομόλογων περιοχών (για περιοχές που κωδικοποιούν πρωτεΐνες).

- Μετάφραση με συγκεκριμμένο γενετικό κώδικα
  - ακολουθίας επερώτησης (query sequence)
  - ακολουθιών στην Β.Δ.
  - και των δύο ταυτόχρονα

tblastn

Πρωτεΐνη (query) Χ Β.Δ. νουκλεοτιδικών ακολουθιών μεταφρασμένων και στα 6 αναγνωστικά πλαίσια.



TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA

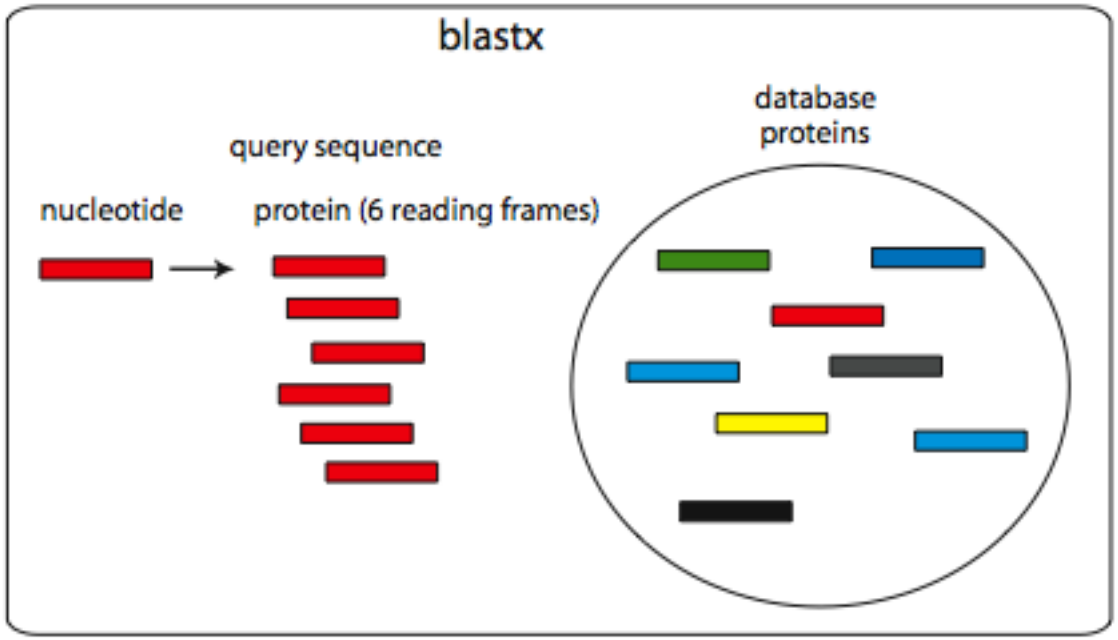
tblastn

- Χρήση
  - Η Β.Δ. περιέχει νουκλεοτιδικές ακολουθίες με άγνωστη λειτουργία (συλλογή ESTs ή αμορφοποίητα δεδομένα από την αλληλούχιση ενός γενώματος) ενός οργανισμού Α και θέλουμε να εντοπίσουμε μια πρωτεΐνη με συγκεκριμμένη λειτουργία στον οργανισμό Α. Ως ακολουθία επερώτησης χρησιμοποιούμε την πρωτεΐνη που είναι γνωστή στον οργανισμό Β.

- Αντιμετωπίζει το πρόβλημα λαθών στην αλληλούχιση, που θα μπορούσε να καταστρέψει το αναγνωστικό πλαίσιο.

Blastx

- Νουκλεοτιδική ακολουθία επερώτησης (query) που μεταφράζεται στα 6 αναγνωστικά πλαίσια και συγκρίνεται με Β.Δ. πρωτεϊνικών ακολουθιών.



BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein

Blastx

- Παράδειγμα: εντοπισμός μετάλλαξης που αλλάζει το αναγνωστικό πλαίσιο.
  - Στο παράδειγμα, υπάρχει αλλαγή αναγνωστικού πλαισίου (frame +2 -> frame +1) στη θέση 268 της πρωτεΐνης επερώτησης

Alignments

>gi|18538741|gb|AAL71647.1| envelope glycoprotein [Human immunodeficiency virus 1]
gi|18538703|gb|AAL71628.1| envelope glycoprotein [Human immunodeficiency virus 1]
Length=201

Score = 232 bits (591), Expect = 7e-60
Identities = 110/112 (98%), Positives = 110/112 (98%), Gaps = 1/112 (0%)
Frame = +1

```
Query  268  TIAFNQSSGGDPEIVMHSFNCGGEFFYCNTTQLFNSTWPTNK-KSTNKTGTITLPCRIKQ  444
            TIAFNQSSGGDPEIVMHSFNCGGEFFYCNTTQLFNSTWPTN  KSTNKTGTITLPCRIKQ
Sbjct  90   TIAFNQSSGGDPEIVMHSFNCGGEFFYCNTTQLFNSTWPTNNTKSTNKTGTITLPCRIKQ  149

Query  445  IINRWQEVGKAMYAPPIKGQIRCSSNITGIFLTRDGGNASDETETFRPGGGN  600
            IINRWQEVGKAMYAPPIKGQIRCSSNITGIFLTRDGGNASDETETFRPGGGN
Sbjct  150  IINRWQEVGKAMYAPPIKGQIRCSSNITGIFLTRDGGNASDETETFRPGGGN  201
```

Score = 181 bits (460), Expect = 1e-44
Identities = 89/89 (100%), Positives = 89/89 (100%), Gaps = 0/89 (0%)
Frame = +2

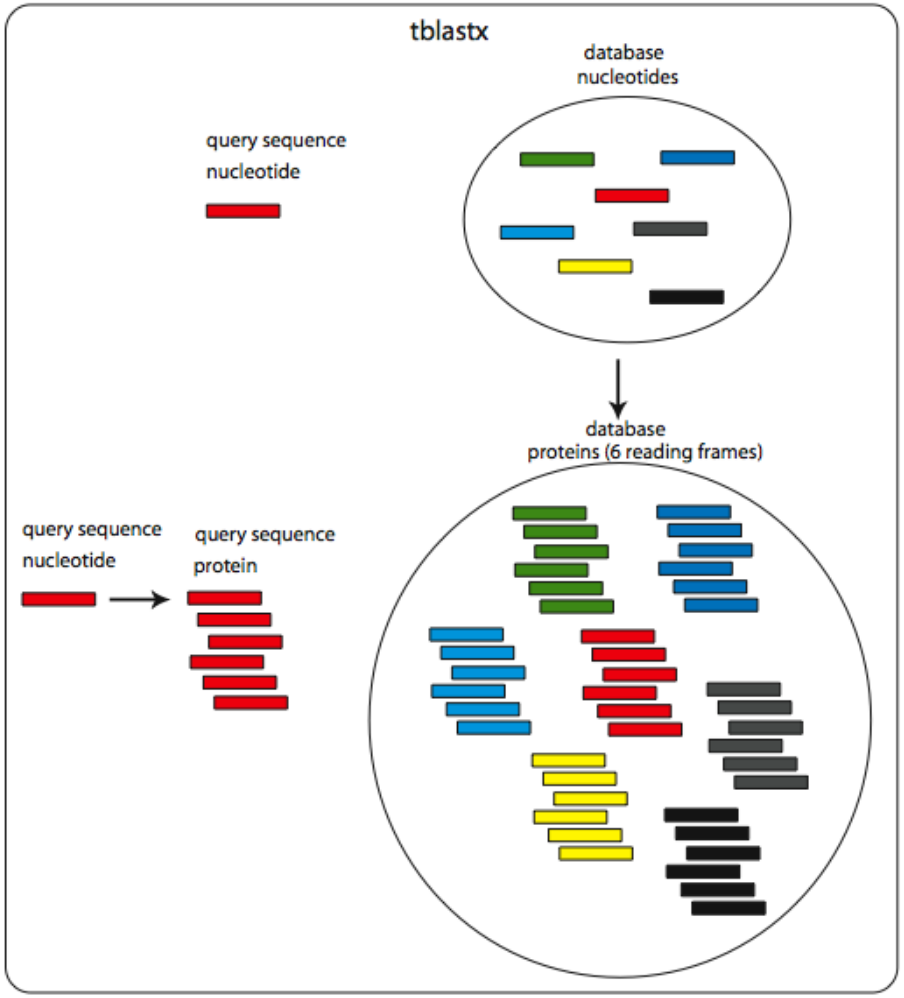
```
Query  2    EEDIVIRSENFTNNAKTIIVQLKESIKINCTRPNNNTRKSIPIATGGAIYATGDIIGDIR  181
            EEDIVIRSENFTNNAKTIIVQLKESIKINCTRPNNNTRKSIPIATGGAIYATGDIIGDIR
Sbjct  1    EEDIVIRSENFTNNAKTIIVQLKESIKINCTRPNNNTRKSIPIATGGAIYATGDIIGDIR  60

Query  182  QAHCNLSRDQWDNTLSQLVTKLREQFGNK  268
            QAHCNLSRDQWDNTLSQLVTKLREQFGNK
Sbjct  61   QAHCNLSRDQWDNTLSQLVTKLREQFGNK  89
```

>gi|40850479|gb|AAR95942.1| envelope glycoprotein [Human immunodeficiency virus 1]
gi|18538655|gb|AAL71604.1| envelope glycoprotein [Human immunodeficiency virus 1]
gi|18538613|gb|AAL71583.1| envelope glycoprotein [Human immunodeficiency virus 1]
Length=201

tblastx

- Νουκλεοτιδική ακολουθία επερώτησης (query) που μεταφράζεται στα 6 αναγνωστικά πλαίσια και συγκρίνεται με Β.Δ. νουκλεοτιδικών ακολουθιών μεταφρασμένων και στα 6 αναγνωστικά πλαίσια.

- 6X6 blastp



TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

tblastx

- Αναζήτηση (διαειδική) για άγνωστα μέχρι σήμερα γονίδια.

Blast και φυλογένεση

J Mol Evol (2001) 52:540–542
DOI: 10.1007/s002390010184

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

The Closest BLAST Hit Is Often Not the Nearest Neighbor

Liisa B. Koski, G. Brian Golding

Department of Biology, McMaster University, 1280 Main Street West, Hamilton, Ontario Canada, L8S 4K1

Received: 23 January 2001 / Accepted: 20 February 2001

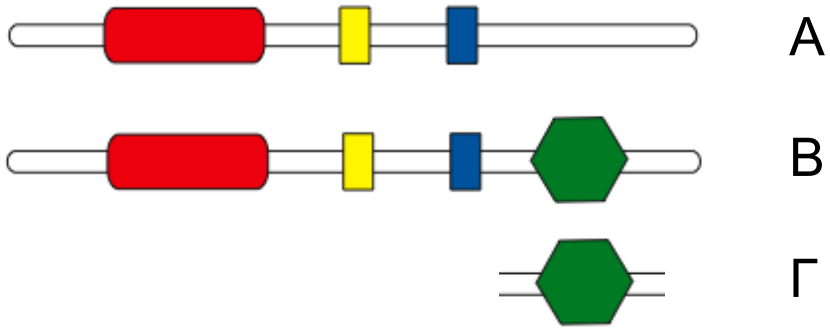
Επαλήθευση ομολογίας μέσω ενδιάμεσων ακολουθιών

- Έστω 2 ακολουθίες Α και Β είναι ομόλογες και στοιχίζονται σε όλο το μήκος τους.

- Αν μια ακολουθία Γ είναι ομόλογη με τη Β, τότε θα είναι ομόλογη και με την Α, έστω και εάν δεν παρατηρούμε στατιστικά σημαντική στοίχιση μεταξύ της Α και της Γ

Επαλήθευση ομολογίας μέσω ενδιάμεσων ακολουθιών

2 ακολουθίες Α και Β είναι ομόλογες αλλά ΔΕΝ στοιχίζονται σε όλο το μήκος τους.
Η Β είναι επίσης ομόλογη με την Γ.
Η Α είναι ομόλογη με την Γ;



PSI-Blast

PSI-Blast: τι είναι

- PSI-Blast: Position-specific iterated Blast
- Position specific scoring matrices (PSSMs) (Πίνακες αντικατάστασης θέσης)
- Altschul et al., 1997
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/pdf/253389.pdf
- Η αναζήτηση μακρινών ομολόγων σε Β.Δ. είναι πιο ευαίσθητη με τη χρήση αυτών των πινάκων.
- Για ομόλογες ακολουθίες το PSI-Blast βρίσκει μέχρι και 3 φορές περισσότερες μακρινές ομόλογες ακολουθίες (ομοιότητα < 30%) σε σχέση με το Blastp.

PSI-Blast: τι είναι

- Σε μια ακολουθία οι διάφορες θέσεις δεν είναι το ίδιο συντηρημένες/ευέλικτες λόγω δομικών/λειτουργικών περιορισμών.
- Χρησιμοποιώντας ομόλογες ακολουθίες από τον ίδιο ή άλλους οργανισμούς κατανοούμε την ευελιξία κάθε θέσης μιας ακολουθίας.
- Π.χ. Σε μια ακολουθία Α, στην θέση 123 (ενεργό κέντρο ενζύμου) βλέπουμε ένα μόνο αμινοξύ.
- Σε μια πολλαπλή στοίχιση της Α με ομόλογες ακολουθίες βλέπουμε για την ίδια θέση (123) ποιά άλλα αμινοξέα επιτρέπονται και σε τί συχνότητες.
- Το PSSM χρησιμοποιεί αυτή την πληροφορία για να αναζητήσει μακρινά ομόλογα σε μια Β.Δ.

PSSM

- Αρχικά γίνεται πολλαπλή στοίχιση των ακολουθιών

A. Sequence alignment[a]

urt-71	A	T	T	T	A	G	T	A	T	C	A	A	A	A	A	T	A	A	C	A	A	T	T	C
glnA-71	G	T	T	C	T	G	T	A	A	C	A	A	A	G	A	C	T	A	C	A	A	A	A	C
nirA-71	A	T	T	T	T	G	T	A	G	C	T	A	C	T	T	A	T	A	C	T	A	T	T	T
ntcB-71	A	A	G	C	T	G	T	A	A	C	A	A	A	A	T	C	T	A	C	C	A	A	A	T
devBCA-71	C	A	T	T	T	G	T	A	C	A	G	T	C	T	G	T	T	A	C	C	T	T	T	A

- Στη συνέχεια, για ακολουθία μήκους L δημιουργείται πίνακας:
  - L X 4 (nucleotides)
  - L X 20 (proteins)

PSSM

- Γίνεται καταμέτρηση των συχνοτήτων των χαρακτήρων για την κάθε θέση.

A. Sequence alignment[a]

urt-71	A	T	T	T	A	G	T	A	T	C	A	A	A	A	A	T	A	A	C	A	A	T	T	C
glnA-71	G	T	T	C	T	G	T	A	A	C	A	A	A	G	A	C	T	A	C	A	A	A	A	C
nirA-71	A	T	T	T	T	G	T	A	G	C	T	A	C	T	T	A	T	A	C	T	A	T	T	T
ntcB-71	A	A	G	C	T	G	T	A	A	C	A	A	A	A	T	C	T	A	C	C	A	A	A	T
devBCA-71	C	A	T	T	T	G	T	A	C	A	G	T	C	T	G	T	T	A	C	C	T	T	T	A

B. Table of occurrences[a]

A	3	2	0	0	1	0	0	5	2	1	3	4	3	2	2	1	1	5	0	2	4	2	2	1
C	1	0	0	2	0	0	0	0	1	4	0	0	2	0	0	2	0	0	5	2	0	0	0	2
G	1	0	1	0	0	5	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0
T	0	3	4	3	4	0	5	0	1	0	1	1	0	2	2	2	4	0	0	1	1	3	3	2

PSSM

- Ακολουθεί μια σειρά μετασχηματισμών
  - Συντελεστής βαρύτητας της κάθε ακολουθίας με βάση την ομοιότητά της με άλλες.
  - Pseudocounts
  - Λαμβάνεται υπόψην η συχνότητα υποβάθρου του κάθε χαρακτήρα
    - Υπολογισμός των odds (παρατηρούμενη συχνότητα / συχνότητα υποβάθρου).
  - Log-odds

- Ο πίνακας αυτός χρησιμοποιείται για τοπική στοίχιση με ακολουθίες σε μια Β.Δ. (αντικαθιστά την ακολουθία επερώτησης).

F. Position-specific scoring matrix: Log-odds form (B = 0.1)[c,d]

A	0.2	0.4	2.2	2.2	0.7	2.2	2.2	0.0	0.4	0.7	0.2	0.1	0.2	0.4	0.4	0.7	0.7	0.0	2.2	0.4	0.1	0.4	0.4	0.7
C	0.7	2.5	2.5	0.4	2.5	2.5	2.5	2.5	0.7	0.1	2.5	2.5	0.4	2.5	2.5	0.4	2.5	2.5	0.0	0.4	2.5	2.5	2.5	0.4
G	0.7	2.5	0.7	2.5	2.5	0.0	2.5	2.5	0.7	2.5	0.7	2.5	2.5	0.7	0.7	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
T	2.2	0.2	0.1	0.2	0.1	2.2	0.0	2.2	0.7	2.2	0.7	0.7	2.2	0.4	0.4	0.4	0.1	2.2	2.2	0.7	0.7	0.2	0.2	0.4

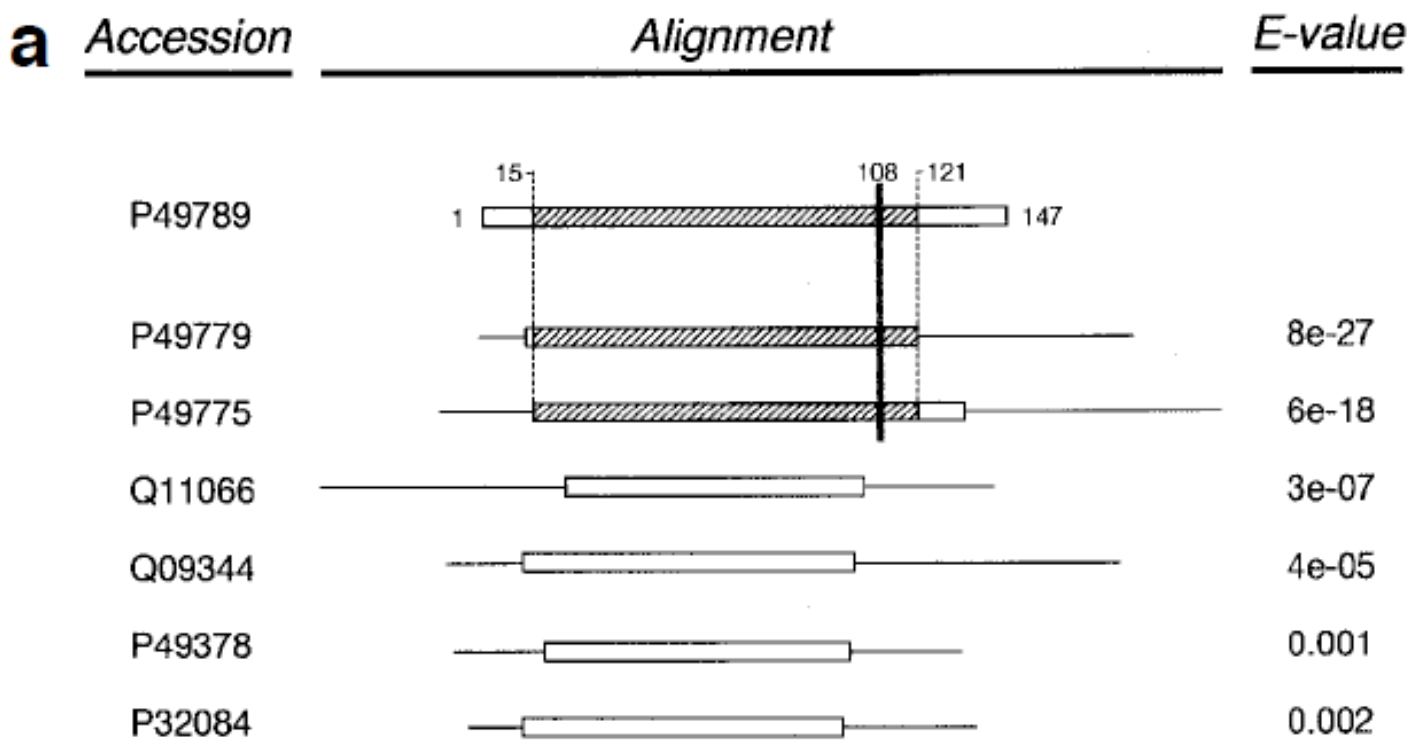
Στάδια του PSI-Blast

- Πρώτο στάδιο:
  - Blast με την ακολουθία επερώτησης σε μια Β.Δ. (E<0.001 default).
  - Οι τοπικές στοιχίσεις που βρέθηκαν (E-value < cutoff) χρησιμοποιούνται για τη δημιουργία μιας πολλαπλής στοίχισης M με σημείο αναφοράς την ακολουθία επερώτησης (L θέσεις).
    - Δεν επιτρέπονται κενά στην ακολουθία επερώτησης.
    - Αυτή η πολλαπλή στοίχιση (ακολουθία - σημείο αναφοράς) διαφέρει από τις τυπικές πολλαπλές στοιχίσεις
  - Απαλοιφή ακολουθιών με πολύ μεγάλη ομοιότητα.
  - Δημιουργία PSSM.

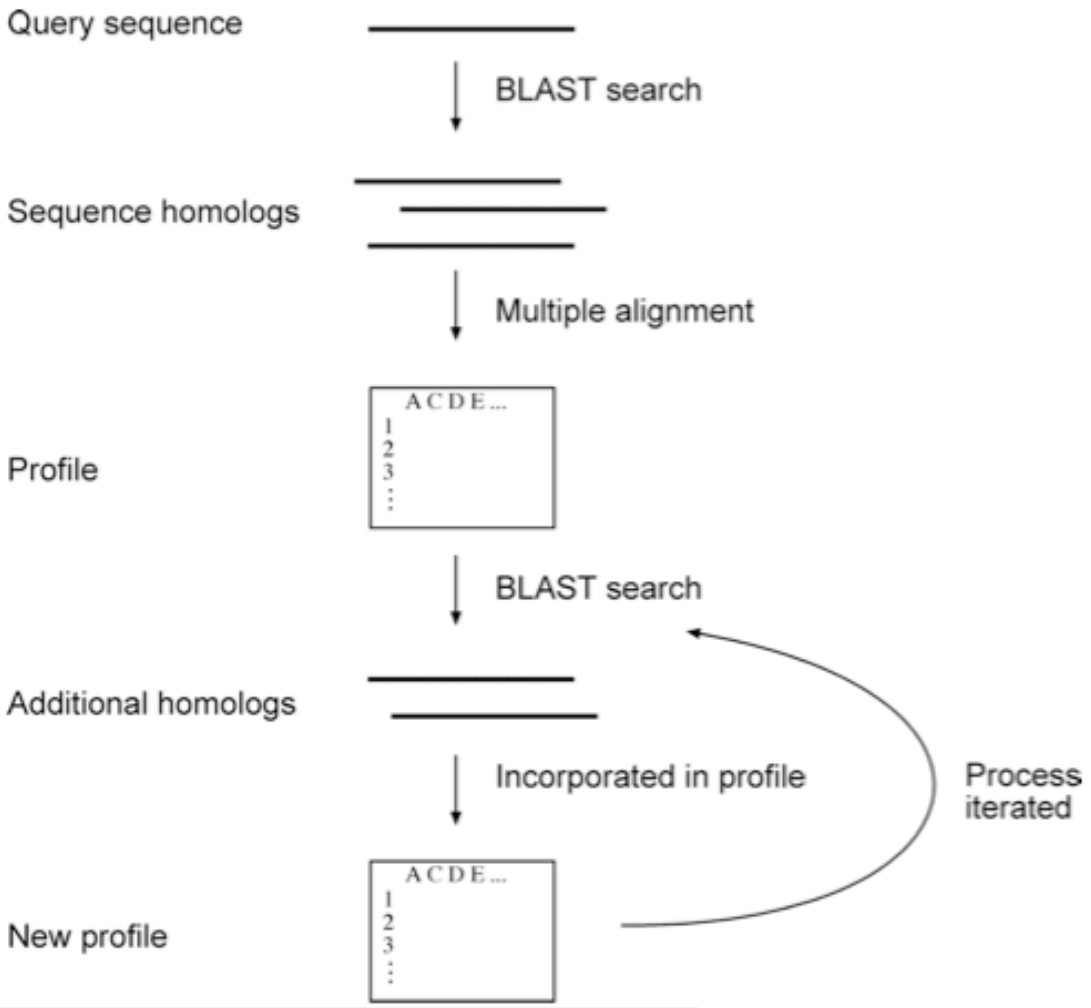
Στάδια του PSI-Blast

- Δεύτερο στάδιο:
  - Νέα αναζήτηση στη Β.Δ. με το PSSM αντί της αρχικής ακολουθίας επερώτησης.
  - Οι νέες ακολουθίες που βρέθηκαν και ξεπερνούν το κατώφλι E-value ανανεώνουν την πολλαπλή στοίχιση και δημιουργείται ένα νέο PSSM.
- Η διαδικασία επαναλαμβάνεται μέχρι να μη βρεθούν νέες ακολουθίες με Evalue < τιμή κατωφλίου (convergence).
- Συνήθως, 3-5 κύκλοι αρκούν για να βρεθούν τα περισσότερα μακρινά ομόλογα.

PSI-Blast


a
Accession
Alignment
E-value
15
108
121
1
147
P49789
P49779
8e-27
P49775
6e-18
Q11066
3e-07
Q09344
4e-05
P49378
0.001
P32084
0.002

PSI-Blast


Query sequence
BLAST search
Sequence homologs
Multiple alignment
A C D E ...
1
2
3
Profile
BLAST search
Additional homologs
Incorporated in profile
Process iterated
A C D E ...
1
2
3
New profile

PSI-Blast

- Πριν κάνουμε PSI-Blast πρέπει να ξέρουμε τι αναζητάμε!!!
  - αναζητούμε ομόλογες πρωτεΐνες με την ίδια αρχιτεκτονική επικρατειών (domain architecture);
  - Αναζητούμε πρωτεΐνες που να περιλαμβάνουν μια συγκεκριμμένη περιοχή; Χρησιμοποιούμε μόνο αυτή την περιοχή στην αρχική αναζήτηση.
  - Αν η περιοχή αυτή είναι γνωστή επικράτεια που υπάρχει σε Β.Δ. Πρωτεΐνικών επικρατειών (π.χ. PFAM), τότε καλύτερα να χρησιμοποιήσουμε αυτές τις Β.Δ.
  - Κάποιες περιοχές/επικράτειες συναντώνται σε πολλές πρωτεΐνες.
    - Προσοχή στην αναζήτηση όταν υπάρχουν τέτοιες περιοχές
  - Αν ξεκινήσουμε με άλλη ομόλογη ακολουθία επερώτησης δεν είναι σίγουρο ότι θα φτάσουμε στο ίδιο αποτέλεσμα!
  - Προσοχή ποιές ακολουθίες συμπεριλαμβάνουμε στο PSSM. Αν εισέλθουν λάθος ακολουθίες, το λάθος θα ανατροφοδοτείται σε κάθε κύκλο (profile drift)

Επικράτειες (Domains)

- Κάποιες επικράτειες συνδυάζονται πολύ συχνά με άλλες, στην ίδια πρωτεΐνη.
- http://genome.cshlp.org/content/18/3/449.full

Evolution of protein domain promiscuity in eukaryotes

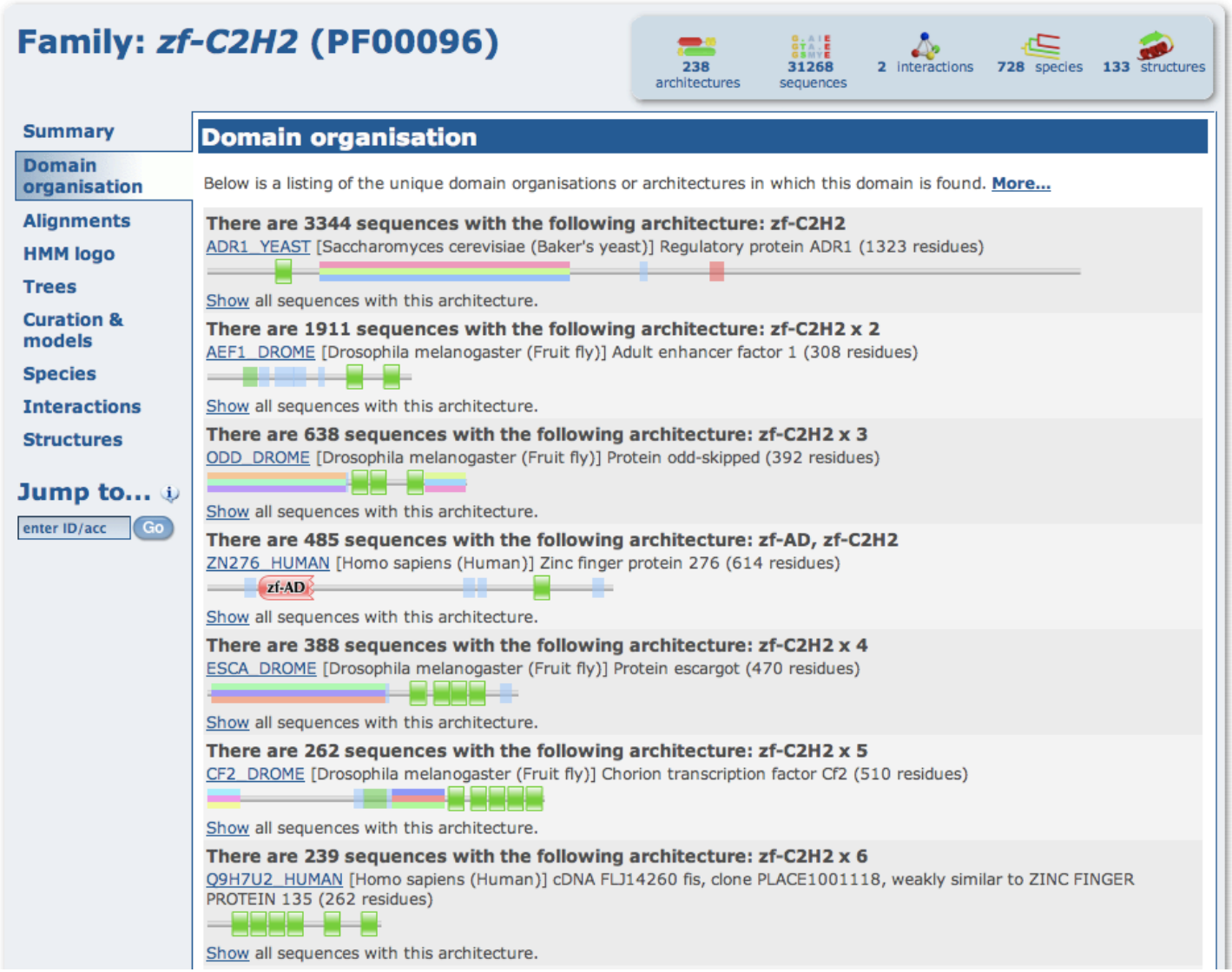
Click on image to view larger version.

Click on table to view larger version.

Table 2. The 10 most promiscuous domains in animals, fungi, and plants

Domain	Average promiscuity (π)	Most frequent bigram partner	No. of occurrenc
Animals			
PH (smart00233)	972.18	SH3 (smart00326)	96
PDZ (smart00228)	675.6	SH3 (smart00326)	166
SH3 (smart00326)	556.45	GuKc (smart00072)	197
C1 (smart00109)	479.35	C2 (smart00239)	85
PHD (smart00249)	464.83	BROMO (smart00297)	123
RING (smart00184)	441.26	BBOX (smart00336)	128
TyrKc (smart00219)	413.74	FN3 (smart00060)	223
EGF_CA (smart00179)	397.07	CUB (smart00042)	55
SAM (smart00454)	371.45	TyrKc (smart00219)	138
EGF (smart00181)	353.07	LamG (smart00282)	155

Επικράτειες και αναζήτηση σε Β.Δ.

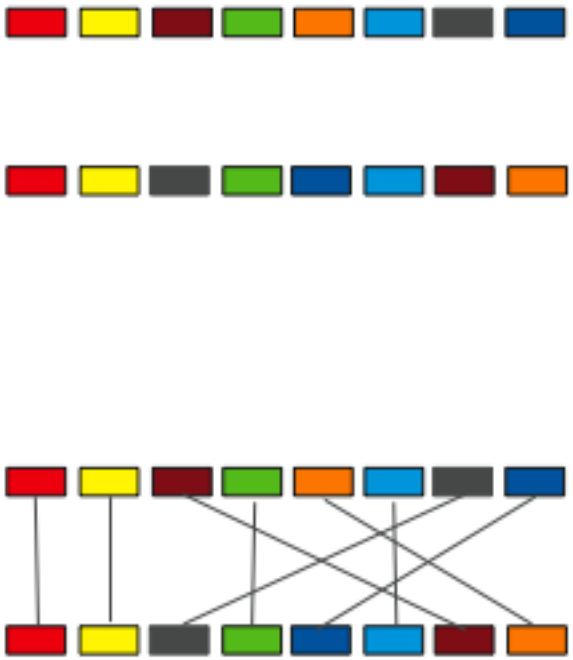


Ανταποδοτικό Blast
(Best reciprocal blast hit)

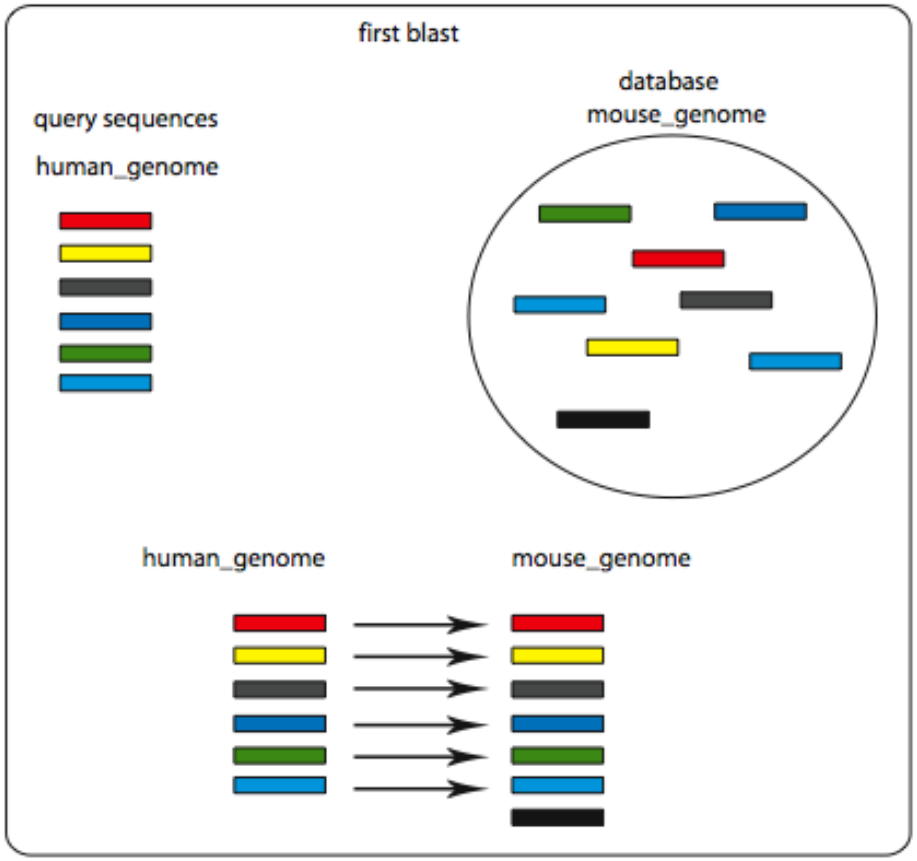
Ανταποδοτικό Blast (ι)

- Γρήγορη μέθοδος εντοπισμού ορθόλογων γονιδίων/πρωτεϊνών μεταξύ δύο γενωμάτων (π.χ. μόλις αλληλουχήθηκε ένα γένωμα).

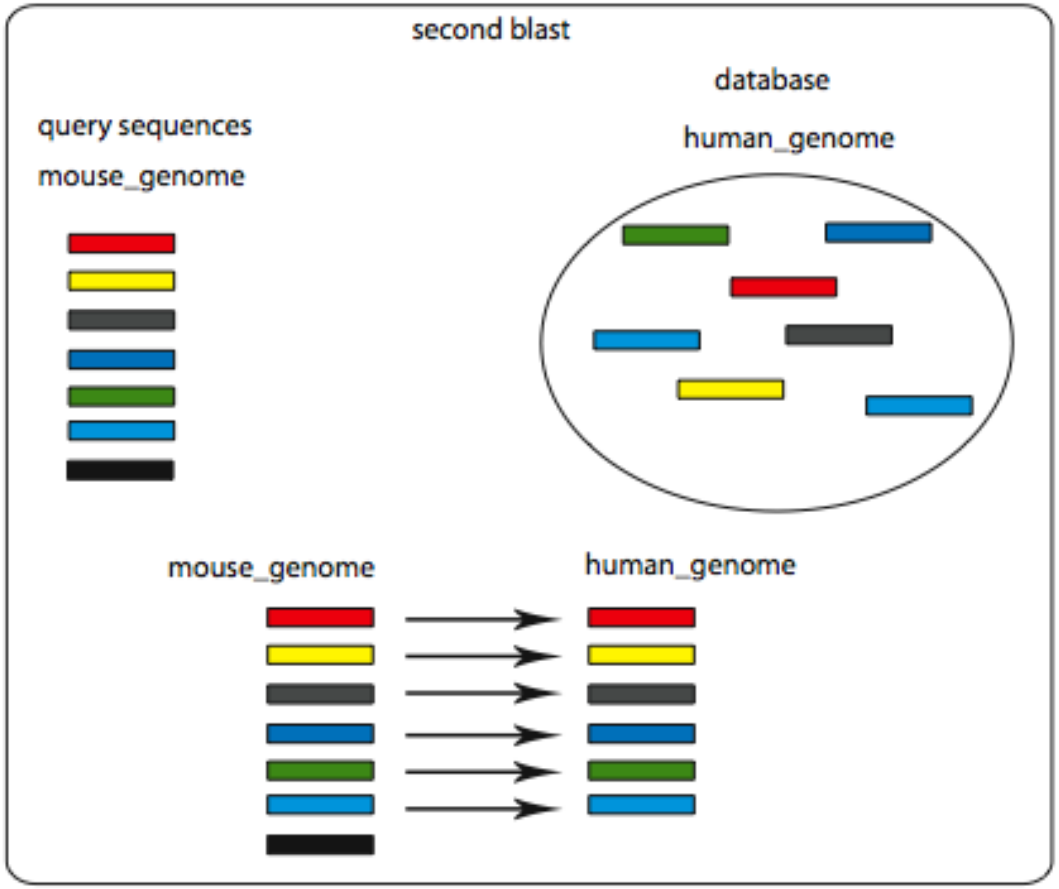
- Γιατί είναι σημαντικό να βρούμε το σωστό ορθόλογο;
  - Ορθόλογα συνήθως έχουν την ίδια λειτουργία
  - Παράλογα συνήθως αποκλείνουν στις λειτουργίες τους



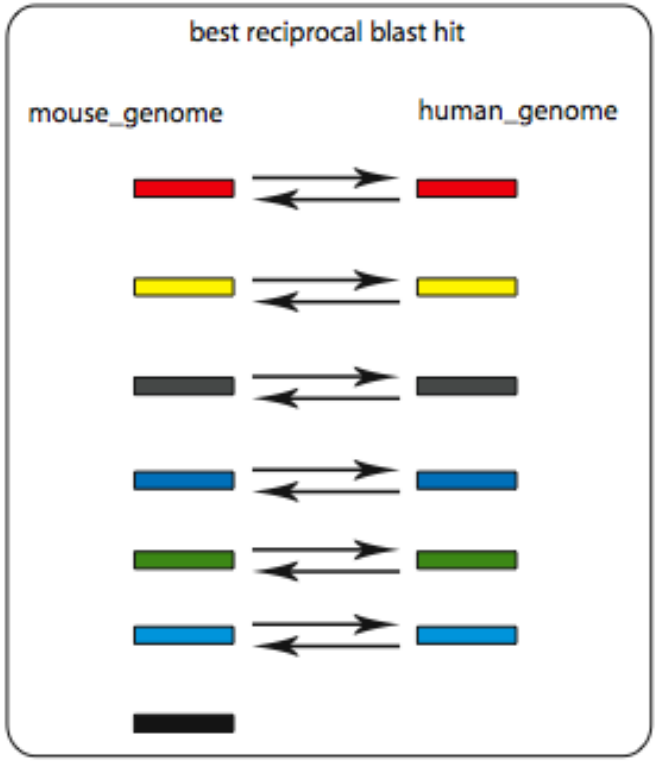
Ανταποδοτικό Blast (ii)



Ανταποδοτικό Blast (iii)

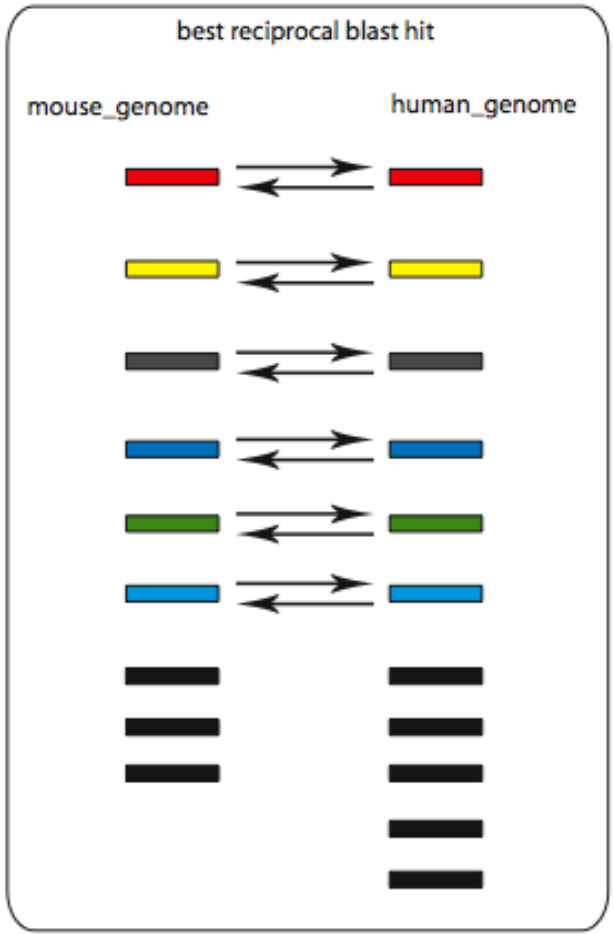


Ανταποδοτικό Blast (iv)

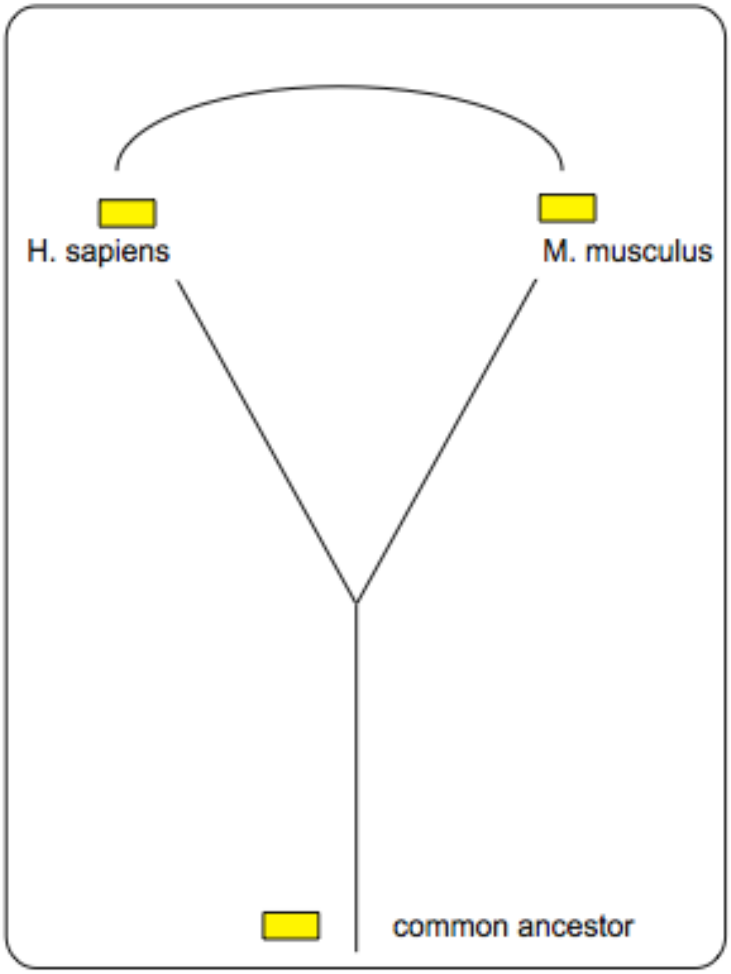


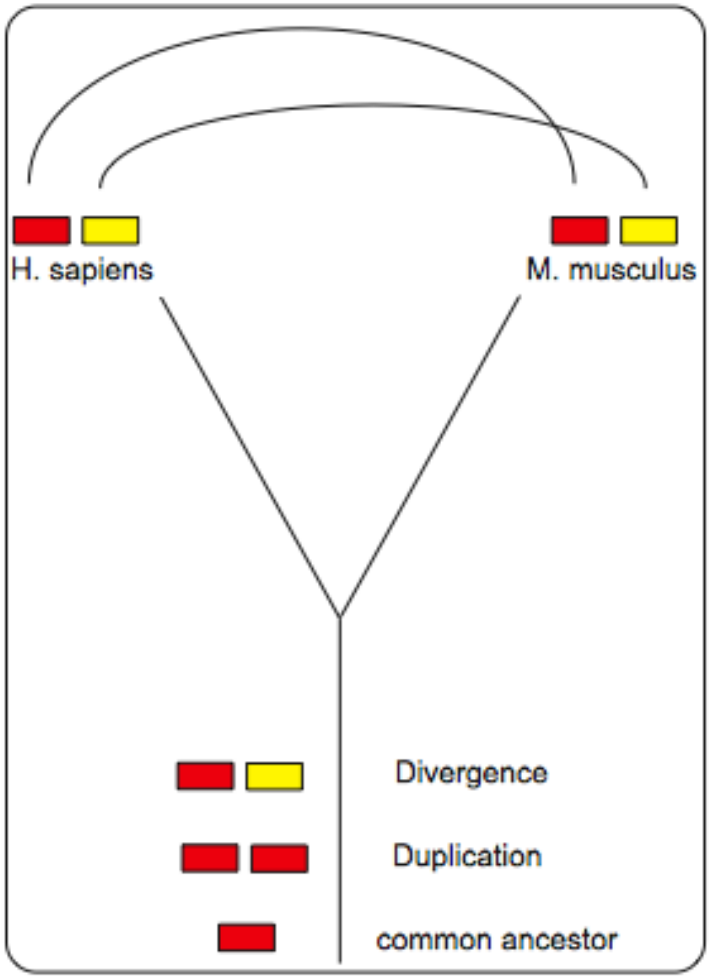
Ανταποδοτικό Blast (v)

Εκτεταμένος γονιδιακός διπλασιασμός

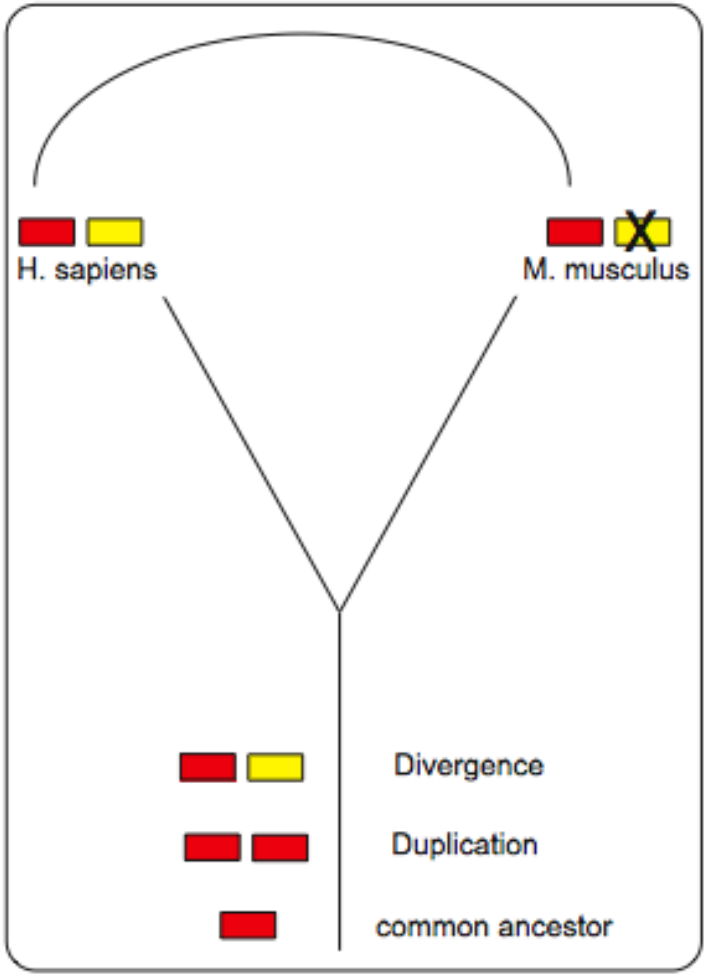


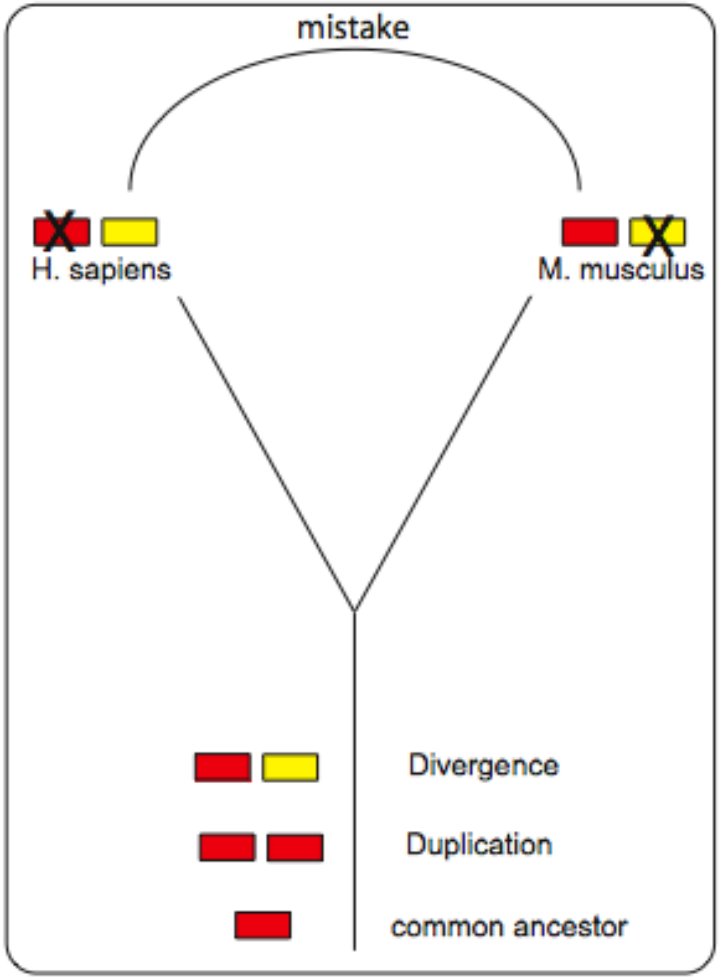
Ανταποδοτικό Blast (vi)





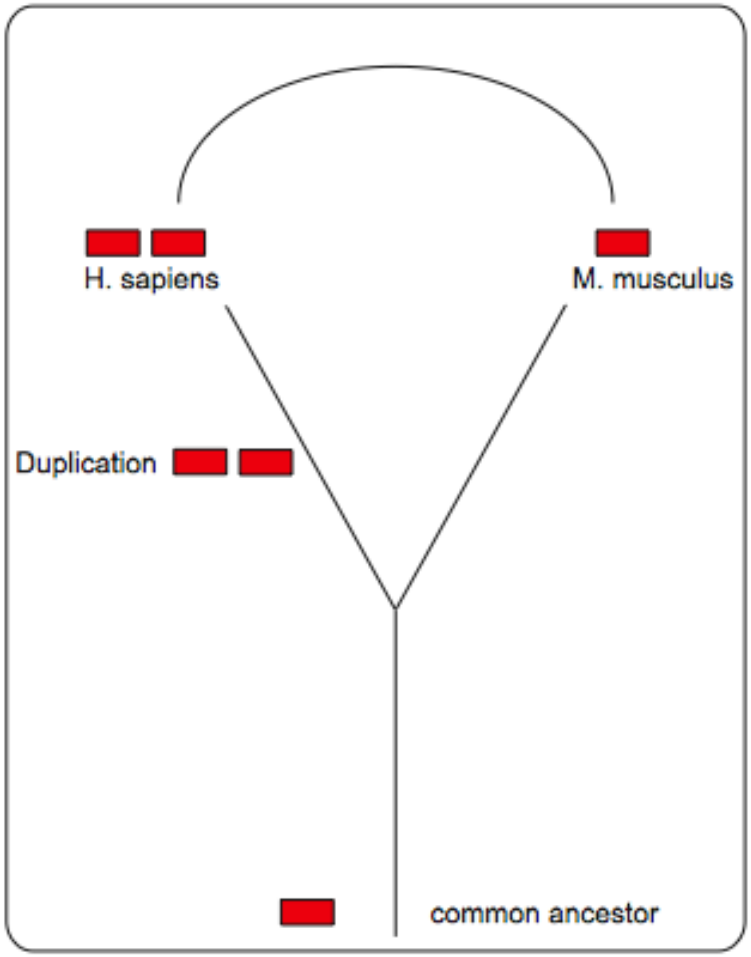
Ανταποδοτικό Blast (vii)

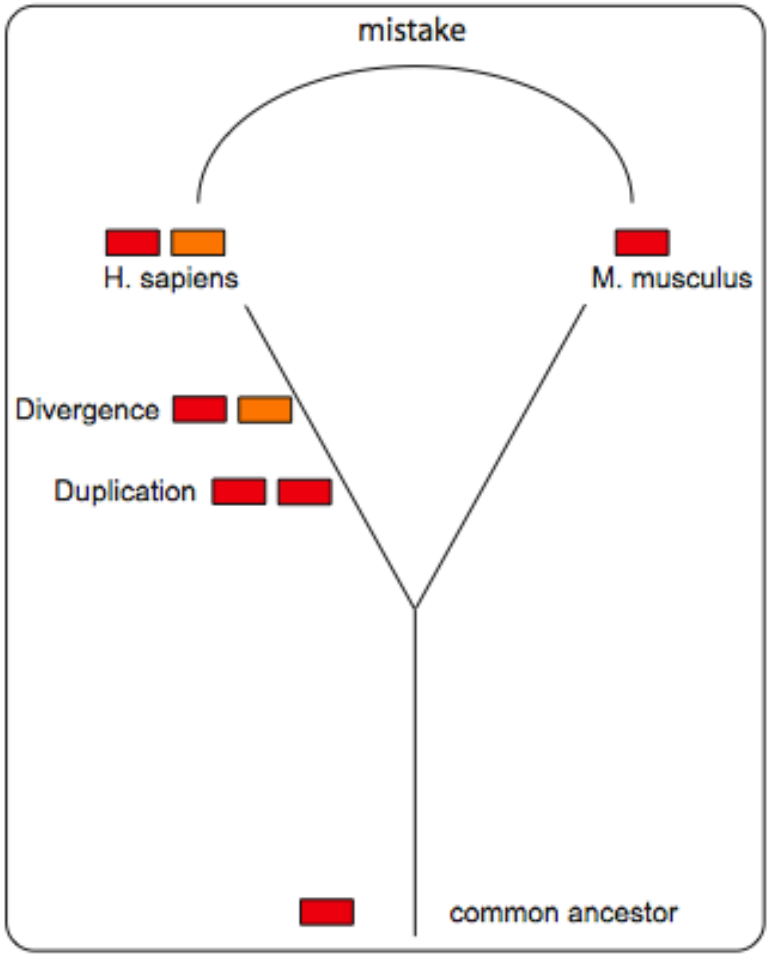




Πώς μπορεί να εντοπιστεί το λάθος;

Ανταποδοτικό Blast (viii)





Παράμετροι που επηρεάζουν την εύρεση ορθόλογων με ανταποδοτικό blast (i)

- Κυριότεροι παράμετροι που επηρεάζουν την εύρεση ορθόλογων
  - Είδος φιλτραρίσματος περιοχών χαμηλής πολυπλοκότητας
    - Soft filtering (φιλτράρισμα μόνο στην φάση αναζήτησης, όχι στην φάση τελικής στοίχισης) (default option)
    - Hard filtering (φιλτράρισμα και στις δύο φάσεις)
- Ο αλγόριθμος που κάνει την τελική στοίχιση
  - Blast (words με επέκταση) (default)
  - Smith-Waterman

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 3 2008, pages 319–324
doi:10.1093/bioinformatics/btm585

Sequence analysis

Choosing BLAST options for better detection of orthologs as reciprocal best hits

Gabriel Moreno-Hagelsieb* and Kristen Latimer
Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada, N2L 3C5

Received on August 29, 2007; revised on October 21, 2007; accepted on November 19, 2007
Advance Access publication November 26, 2007
Associate Editor: John Quackenbush

Παράμετροι που επηρεάζουν την εύρεση ορθόλογων με ανταποδοτικό blast (ii)

- Επιλεγμένο όριο τιμής E (E-value threshold) ή τιμής bit-score
- Κατώτατο όριο ποσοστού της ακολουθίας που συμμετέχει στην στοίχιση.
- Κατώτατο όριο ποσοστού ομοιότητας
- Διαφορετικές τιμές για την κάθε ανάλυση
- Π.χ. BioCyc: 10% identity, 40% similarity, E-value<1

- Το blast δεν δημιουργήθηκε για να μετράει την εξελικτική απόσταση δύο ακολουθιών, αλλά για να βρίσκει γρήγορα ομόλογες ακολουθίες

Πηγές λαθών για ανταποδοτικό blast

- Εκτεταμένος γονιδιακός διπλασιασμός που συνέβη πρόσφατα.
- Γονιδιακή σύντηξη
- Εκτεταμένες αναδιατάξεις της αρχιτεκτονικής των πρωτεϊνών (domain rearrangements)
  - Ανασυνδυασμός που οδηγεί στην εισδοχή μη ομόλογου domain

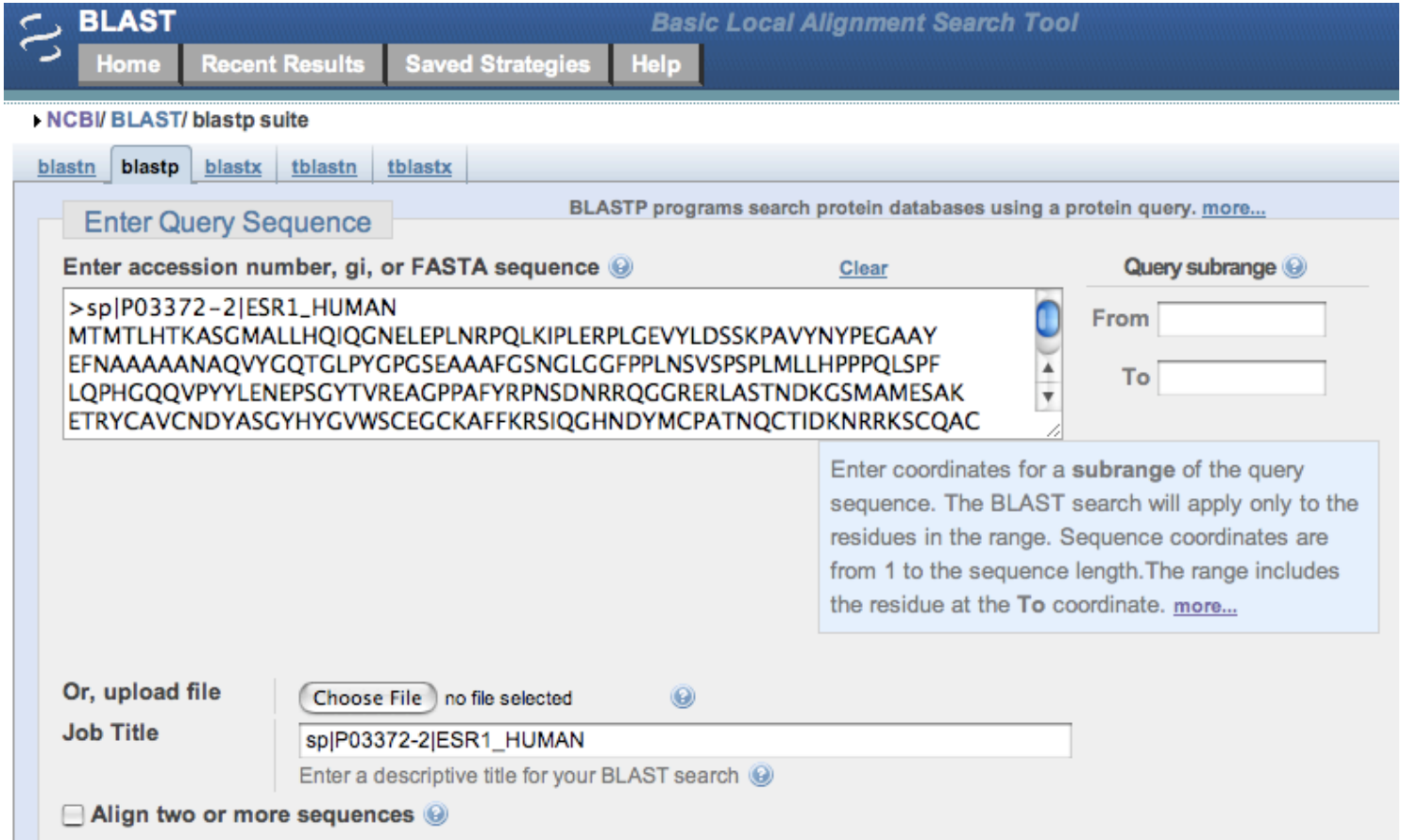
Ότι είναι θεωρητικώς δυνατόν να συμβεί, μάλλον έχει συμβεί κάπου!

Χρησιμοποιώντας το Blast

Χρησιμοποιώντας το Blast (i)

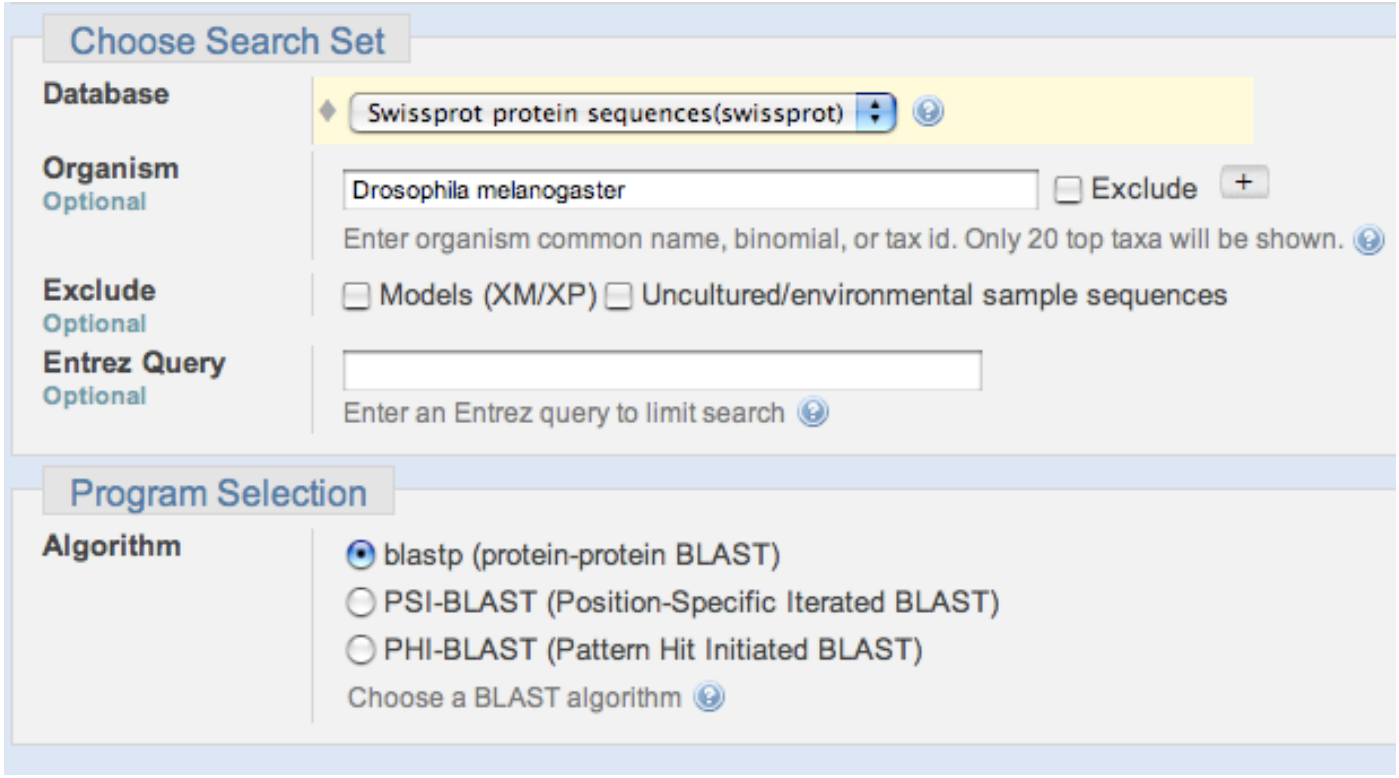
- Επεξηγήσεις στο σύνδεσμο:
  - http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

- Εισάγουμε την ακολουθία



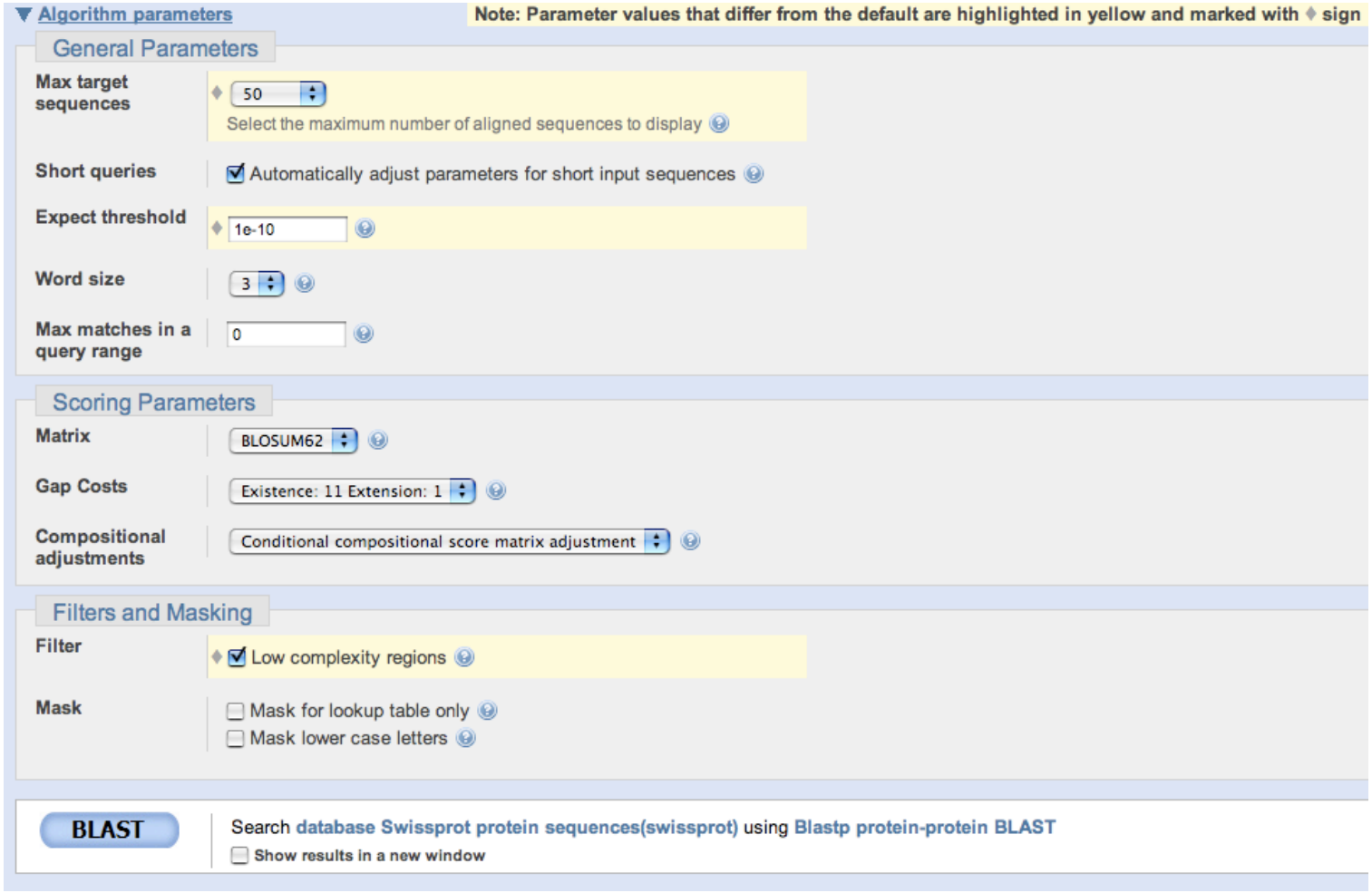
Χρησιμοποιώντας το Blast (ii)

- Επιλέγοντας:
  - τη βάση δεδομένων που θα γίνει η αναζήτηση
  - Τον οργανισμό που θα γίνει η αναζήτηση



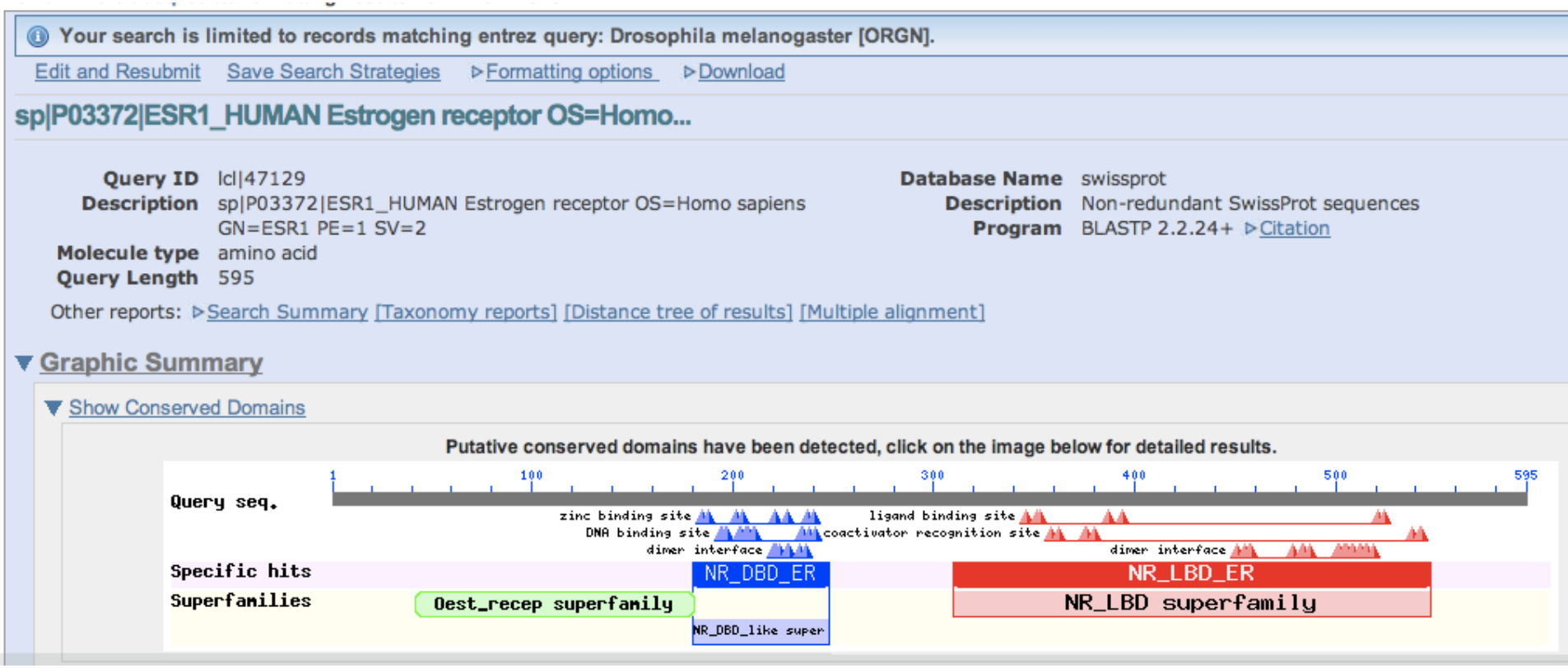
Χρησιμοποιώντας το Blast (iii)

- Παράμετροι του αλγόριθμου
- Expect threshold: ανάλογα με το τι αναζητούμε



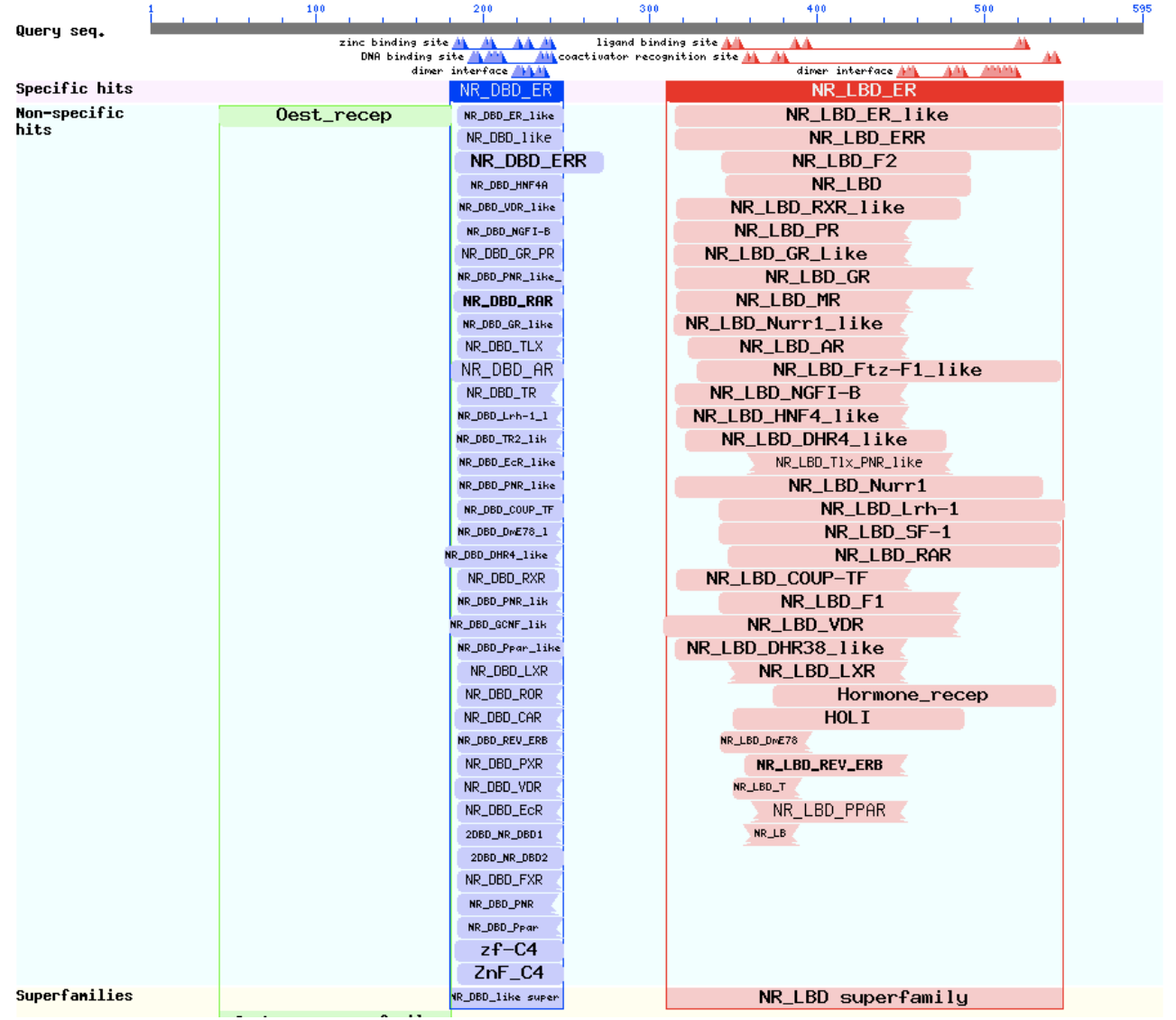
Χρησιμοποιώντας το Blast (iv)

- Αποτελέσματα για συντηρημένες επικράτειες (conserved domains)



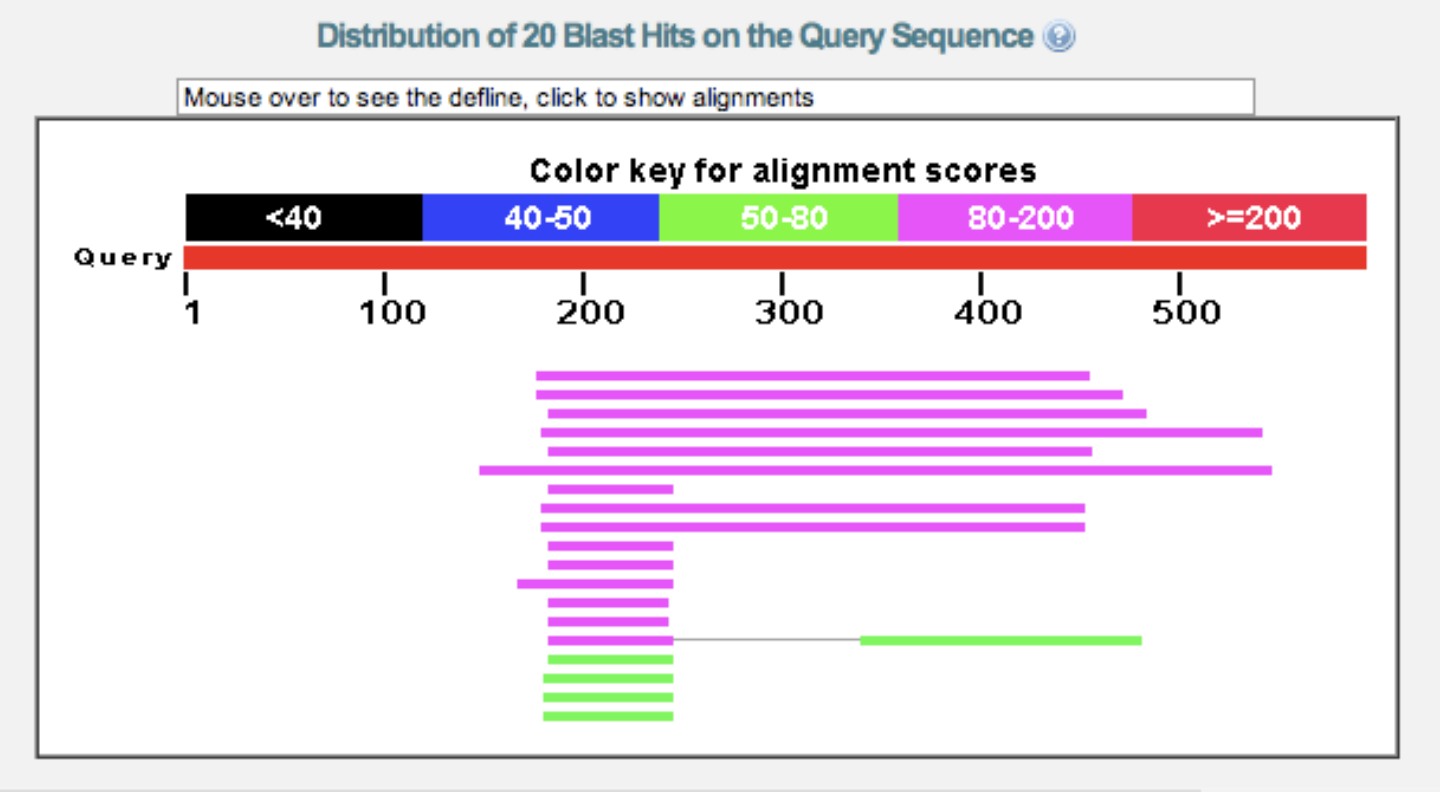
Χρησιμοποιώντας το Blast (v)

- Ομάδες συντηρημένων επικρατειών



Χρησιμοποιώντας το Blast (vi)

- Γράφημα των καλύτερων στοιχίσεων



Χρησιμοποιώντας το Blast (vii)

- Περιγραφές των αποτελεσμάτων (με φίλτρο)

Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
P16376.3	RecName: Full=Steroid receptor seven-up, isoform A; AltName	124	124	46%	2e-29	G
P16375.1	RecName: Full=Steroid receptor seven-up, isoforms B/C; AltN	124	124	49%	3e-29	G
P49866.2	RecName: Full=Transcription factor HNF-4 homolog; Short=dH	117	117	50%	3e-27	G
P49869.3	RecName: Full=Probable nuclear hormone receptor HR38; Sho	117	117	61%	3e-27	
P20153.1	RecName: Full=Protein ultraspiracle; AltName: Full=Chorion fa	113	113	46%	6e-26	G
P34021.1	RecName: Full=Ecdysone receptor; AltName: Full=20-hydroxy	105	105	67%	2e-23	G
P18102.1	RecName: Full=Protein tailless; AltName: Full=Nuclear recept	91.3	91.3	10%	2e-19	G
P13055.2	RecName: Full=Ecdysone-induced protein 75B, isoform B; AltN	90.1	90.1	45%	6e-19	G
P17671.2	RecName: Full=Ecdysone-induced protein 75B, isoforms C/D; A	89.7	89.7	45%	8e-19	G
P33244.2	RecName: Full=Nuclear hormone receptor FTZ-F1; AltName: F	88.2	88.2	10%	2e-18	G
Q9W539.4	RecName: Full=Hormone receptor 4; Short=dHR4; AltName: F	86.7	86.7	10%	7e-18	G
P31396.1	RecName: Full=Probable nuclear hormone receptor HR3; Short	85.5	85.5	13%	1e-17	G
P45447.3	RecName: Full=Ecdysone-induced protein 78C; Short=DR-78; A	84.0	84.0	10%	5e-17	
Q24142.2	RecName: Full=Nuclear hormone receptor HR78; Short=dHR78	82.0	82.0	10%	2e-16	G
Q05192.3	RecName: Full=Nuclear hormone receptor FTZ-F1 beta; AltNar	81.3	149	34%	2e-16	G
P10734.1	RecName: Full=Zygotic gap protein knirps; AltName: Full=Nuc	77.4	77.4	10%	4e-15	G
Q24143.1	RecName: Full=Nuclear hormone receptor HR96; Short=dHR96	76.3	76.3	11%	9e-15	G
P13054.1	RecName: Full=Knirps-related protein; AltName: Full=Nuclear	73.9	73.9	11%	5e-14	G
P15370.2	RecName: Full=Protein embryonic gonad; AltName: Full=Nucle	70.5	70.5	11%	4e-13	G

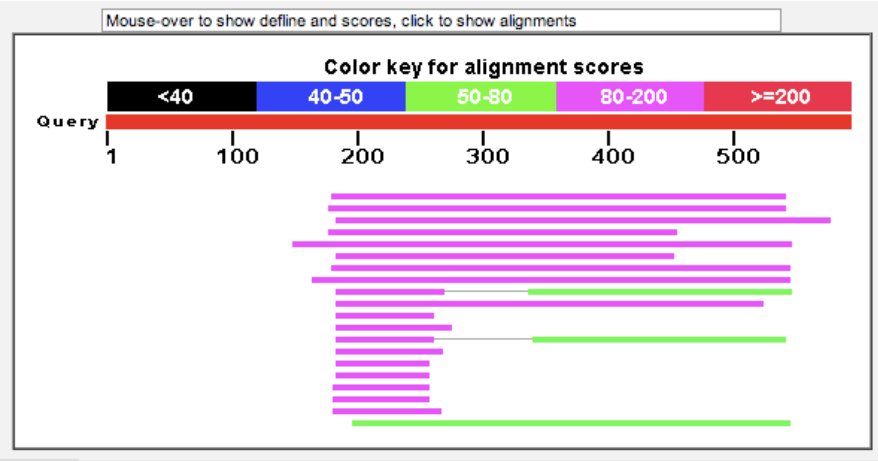
Χρησιμοποιώντας το Blast (viii)

- Στοιχίσεις (με φίλτρο - μικρά γράμματα)
- Identities (επί του αριθμού θέσεων στη στοίχιση)
- Positives (επί του αριθμού θέσεων στη στοίχιση)

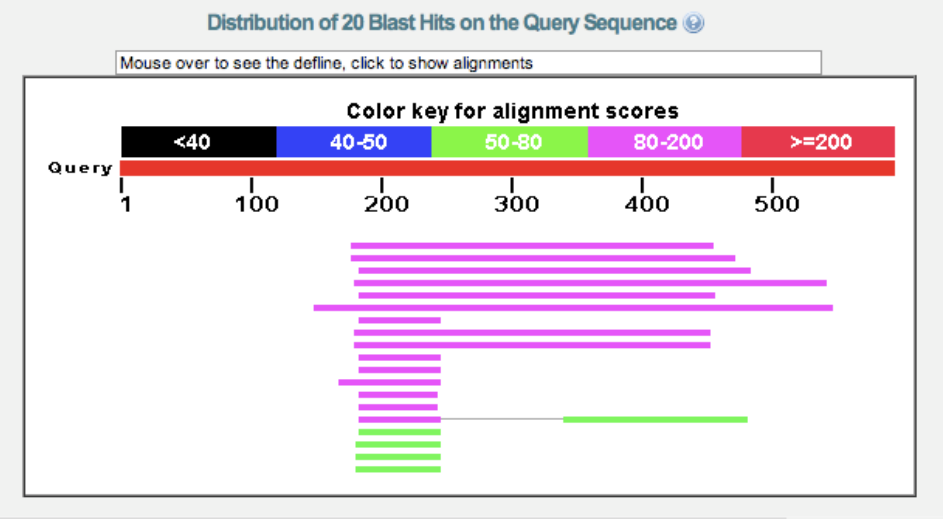
```
>sp|P16376.3|7UP2_DROME RecName: Full=Steroid receptor seven-up, isoform A; AltName:
Full=Nuclear receptor subfamily 2 group F member 3, isoform
A
Length=746

 GENE ID: 41491 svp | seven up [Drosophila melanogaster] (Over 100 PubMed links)

 Score =  124 bits (312),  Expect = 2e-29, Method: Compositional matrix adjust.
 Identities = 87/282 (31%), Positives = 133/282 (48%), Gaps = 25/282 (8%)

Query  179  AKETRYCAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQ  238
            +K+   C VC D +SG HYG ++CEGCK+FFKRS++ +  Y C  +  C ID++ R  CQ
Sbjct  194  SKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSRNCPIDQHHRNQCQ  253

Query  239  ACRLRKCYEVgmmkggirkdrrggrmLKHKRQRDDGEGRGEVGSAGDMRAANLWPSPLMI  298
             CRL+KC ++GM +  +             +R R      G  G  G  + AN    P+ I
Sbjct  254  YCRLKKCLKMGMRREAV------------QRGRVPPTQPGLAGMHGQYQIAN--GDPMGI  299

Query  299  KRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGL--LTNLADRELVH  356
                 +S   S      +S LL AEP   Y      +     ++MG+  +  LA R L
Sbjct  300  AGFNGHSYLSSY-----ISLLLRAEP---YPTSRYGQCMQPNNIMGIDNICELAARLLFS  351

Query  357  MINWAKRVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSME-HPGKLLFAPNLLLDRNQG  415
             + WAK +P F +L + DQV LL   W E+ ++     SM  H   LL A  L
Sbjct  352  AVEWAKNIPFFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAA  411

Query  416  KCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSG  457
              V   ++   +      + + +++   E+ CLK+I+L  +G
Sbjct  412  DRVVAFMDHIRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTG  453
```

Χρησιμοποιώντας το Blast (ix)

χωρίς φίλτρο



με φίλτρο



Η χρήση φίλτρου αλλάζει το score

Identities/Positives σταθερά

Χρησιμοποιώντας το Blast (x)

Χωρίς φίλτρο

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
P49869.3	RecName: Full=Probable nuclear hormone receptor HR38; Sho	144	144	61%	2e-35	
P16375.1	RecName: Full=Steroid receptor seven-up, isoforms B/C; AltN	144	144	61%	3e-35	G
P49866.2	RecName: Full=Transcription factor HNF-4 homolog; Short=dH	144	144	66%	3e-35	G
P16376.3	RecName: Full=Steroid receptor seven-up, isoform A; AltName	132	132	46%	1e-31	G
P34021.1	RecName: Full=Ecdysone receptor; AltName: Full=20-hydroxy	121	121	67%	2e-28	G
P18102.1	RecName: Full=Protein tailless; AltName: Full=Nuclear recept	121	121	45%	2e-28	G
P13055.2	RecName: Full=Ecdysone-induced protein 75B, isoform B; AltN	110	110	61%	6e-25	G
P17671.2	RecName: Full=Ecdysone-induced protein 75B, isoforms C/D;	110	110	64%	6e-25	G
P33244.2	RecName: Full=Nuclear hormone receptor FTZ-F1; AltName: F	103	171	49%	4e-23	G
P20153.1	RecName: Full=Protein ultraspiracle; AltName: Full=Chorion fa	101	101	57%	3e-22	G
Q9W539.4	RecName: Full=Hormone receptor 4; Short=dHR4; AltName: F	100	100	13%	5e-22	G
P31396.1	RecName: Full=Probable nuclear hormone receptor HR3; Short	99.8	99.8	15%	7e-22	G
Q05192.3	RecName: Full=Nuclear hormone receptor FTZ-F1 beta; AltNar	96.7	166	47%	6e-21	G
P45447.3	RecName: Full=Ecdysone-induced protein 78C; Short=DR-78;	93.2	93.2	14%	7e-20	
P10734.1	RecName: Full=Zygotic gap protein knirps; AltName: Full=Nuc	93.2	93.2	12%	7e-20	G
Q24142.2	RecName: Full=Nuclear hormone receptor HR78; Short=dHR78	89.0	89.0	12%	1e-18	G
P13054.1	RecName: Full=Knirps-related protein; AltName: Full=Nuclear	86.7	86.7	13%	7e-18	G
P15370.2	RecName: Full=Protein embryonic gonad; AltName: Full=Nucle	83.6	83.6	13%	6e-17	G
Q24143.1	RecName: Full=Nuclear hormone receptor HR96; Short=dHR96	82.0	82.0	14%	2e-16	G
P17672.2	RecName: Full=Ecdysone-induced protein 75B, isoform A; Sho	68.9	68.9	58%	2e-12	G

Με φίλτρο

▼ **Descriptions**

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
P16376.3	RecName: Full=Steroid receptor seven-up, isoform A; AltName	124	124	46%	2e-29	G
P16375.1	RecName: Full=Steroid receptor seven-up, isoforms B/C; AltN	124	124	49%	3e-29	G
P49866.2	RecName: Full=Transcription factor HNF-4 homolog; Short=dH	117	117	50%	3e-27	G
P49869.3	RecName: Full=Probable nuclear hormone receptor HR38; Sho	117	117	61%	3e-27	

Αλλάζει το score, E-value και η σειρά εμφάνισης

Χρησιμοποιώντας το Blast (xi)

Χωρίς φίλτρο

```
>sp|P16376.3|7UP2_DROME RecName: Full=Steroid receptor seven-up, isoform A; AltName:
Full=Nuclear receptor subfamily 2 group F member 3, isoform
A
Length=746

 GENE ID: 41491 svp | seven up [Drosophila melanogaster] (Over 100 PubMed links)

 Score =  132 bits (331),  Expect = 1e-31, Method: Compositional matrix adjust.
 Identities = 87/282 (31%), Positives = 133/282 (48%), Gaps = 25/282 (8%)

Query  179  AKETRYCAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQ  238
            +K+   C VC D +SG HYG ++CEGCK+FFKRS++ +  Y C  +  C ID++ R  CQ
Sbjct  194  SKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSRNCPIDQHHRNQCQ  253

Query  239  ACRLRKCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVGSAGDMRAANLWPSPLMI  298
             CRL+KC ++GM +  +            +R R      G  G  G  + AN    P+ I
Sbjct  254  YCRLKKCLKMGMRREAV----------QRGRVPPTQPGLAGMHGQYQIAN--GDPMGI  299

Query  299  KRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGL--LTNLADRELVH  356
                +S   S      +S LL AEP   Y      +     ++MG+  +  LA R L
Sbjct  300  AGFNGHSYLSSY-----ISLLLRAEP---YPTSRYGQCMQPNNIMGIDNICELAARLLFS  351

Query  357  MINWAKRVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSME-HPGKLLFAPNLLLDRNQG  415
             + WAK +P F +L + DQV LL   W E+ ++     SM  H   LL A  L
Sbjct  352  AVEWAKNIPFFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAA  411

Query  416  KCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSG  457
              V   ++   +      + + +++   E+ CLK+I+L  +G
Sbjct  412  DRVVAFMDHIRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTG  453
```

Με φίλτρο

```
GENE ID: 41491 svp | seven up [Drosophila melanogaster] (Over 100 PubMed links)

Score =  124 bits (312),  Expect = 2e-29, Method: Compositional matrix adjust.
Identities = 87/282 (31%), Positives = 133/282 (48%), Gaps = 25/282 (8%)
```

Identities & positives παραμένουν σταθερά

Χρησιμοποιώντας το Blast (xi)

- Αλλαγή στον Πίνακα αντικατάστασης και στις ποινές για κενά
  - Blosum 45 13:3, χωρίς φίλτρο

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
P49869.3	RecName: Full=Probable nuclear hormone receptor HR38; Sho	146	146	61%	6e-36	
P16375.1	RecName: Full=Steroid receptor seven-up, isoforms B/C; AltN	145	145	61%	1e-35	G
P49866.2	RecName: Full=Transcription factor HNF-4 homolog; Short=dH	139	139	66%	5e-34	G
P16376.3	RecName: Full=Steroid receptor seven-up, isoform A; AltName	138	138	46%	2e-33	G
P34021.1	RecName: Full=Ecdysone receptor; AltName: Full=20-hydroxy	116	116	67%	8e-27	G
P17671.2	RecName: Full=Ecdysone-induced protein 75B, isoforms C/D;	112	112	61%	7e-26	G
P13055.2	RecName: Full=Ecdysone-induced protein 75B, isoform B; AltN	112	112	47%	1e-25	G
P33244.2	RecName: Full=Nuclear hormone receptor FTZ-F1; AltName: F	110	177	49%	3e-25	G
P18102.1	RecName: Full=Protein tailless; AltName: Full=Nuclear recept	108	108	16%	1e-24	G
Q9W539.4	RecName: Full=Hormone receptor 4; Short=dHR4; AltName: F	108	108	13%	2e-24	G
P31396.1	RecName: Full=Probable nuclear hormone receptor HR3; Short	108	108	18%	2e-24	G
Q05192.3	RecName: Full=Nuclear hormone receptor FTZ-F1 beta; AltNan	102	172	37%	1e-22	G
P10734.1	RecName: Full=Zygotic gap protein knirps; AltName: Full=Nuc	99.8	99.8	13%	7e-22	G
P45447.3	RecName: Full=Ecdysone-induced protein 78C; Short=DR-78;	99.5	99.5	12%	9e-22	
Q24142.2	RecName: Full=Nuclear hormone receptor HR78; Short=dHR78	96.3	96.3	13%	7e-21	G
P13054.1	RecName: Full=Knirps-related protein; AltName: Full=Nuclear	94.8	94.8	13%	2e-20	G
P15370.2	RecName: Full=Protein embryonic gonad; AltName: Full=Nucle	90.7	90.7	14%	3e-19	G
Q24143.1	RecName: Full=Nuclear hormone receptor HR96; Short=dHR96	89.5	89.5	12%	9e-19	G
P17672.2	RecName: Full=Ecdysone-induced protein 75B, isoform A; Sho	67.3	67.3	55%	4e-12	G
P20153.1	RecName: Full=Protein ultraspiracle; AltName: Full=Chorion fa	65.2	65.2	31%	2e-11	G

Blosum 62 11:1, χωρίς φίλτρο

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
P49869.3	RecName: Full=Probable nuclear hormone receptor HR38; Sho	144	144	61%	2e-35	
P16375.1	RecName: Full=Steroid receptor seven-up, isoforms B/C; AltN	144	144	61%	3e-35	G
P49866.2	RecName: Full=Transcription factor HNF-4 homolog; Short=dH	144	144	66%	3e-35	G
P16376.3	RecName: Full=Steroid receptor seven-up, isoform A; AltName	132	132	46%	1e-31	G
P34021.1	RecName: Full=Ecdysone receptor; AltName: Full=20-hydroxy	121	121	67%	2e-28	G

Χρησιμοποιώντας το Blast (xii)

Blosum 45 13:3

>sp|P16376.3|7UP2_DROME RecName: Full=Steroid receptor seven-up, isoform A Full=Nuclear receptor subfamily 2 group F member 3, isoform A
Length=746

GENE ID: 41491 svp | seven up [Drosophila melanogaster] (Over 100 PubMed links

Score = 138 bits (456), Expect = 2e-33, Method: Compositional matrix adjust.
Identities = 87/282 (31%), Positives = 137/282 (49%), Gaps = 25/282 (8%)

```
Query  179  AKETRYCAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQ  238
            +K+   C VC D +SG HYG ++CEGCK+FFKRS++ +  Y C  +  C ID++ R  CQ
Sbjct  194  SKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSRNCPIDQHHRNQCQ  253

Query  239  ACRLRKCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVGSAGDMRAANLWPSPLMI  298
             CRL+KC ++GM +  +           +R R      G  G  G  + AN    P+ I
Sbjct  254  YCRLKKCLKMGMRREAV----------QRGRVPPTQPGLAGMHGQYQIAN--GDPMGI  299

Query  299  KRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGL--LTNLADRELVH  356
                 +S   S      +S LL AEP   Y      +     ++MG+  +  LA R L
Sbjct  300  AGFNGHSYLSS-----YISLLLRAEP---YPTSRYGQCMQPNNIMGIDNICELAARLLFS  351

Query  357  MINWAKRVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSME-HPGKLLFAPNLLLDRNQG  415
             + WAK +P F +L + DQV LL   W E++++     SM  H   LL A  L
Sbjct  352  AVEWAKNIPFFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAA  411

Query  416  KCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSG  457
              V   ++   ++     +++ +++   E+ CLK+I+L+ +G
Sbjct  412  DRVVAFMDHIRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTG  453
```

Blosum 62 11:1

>sp|P16376.3|7UP2_DROME RecName: Full=Steroid receptor seven-up, isoform A; Full=Nuclear receptor subfamily 2 group F member 3, isoform A
Length=746

GENE ID: 41491 svp | seven up [Drosophila melanogaster] (Over 100 PubMed links)

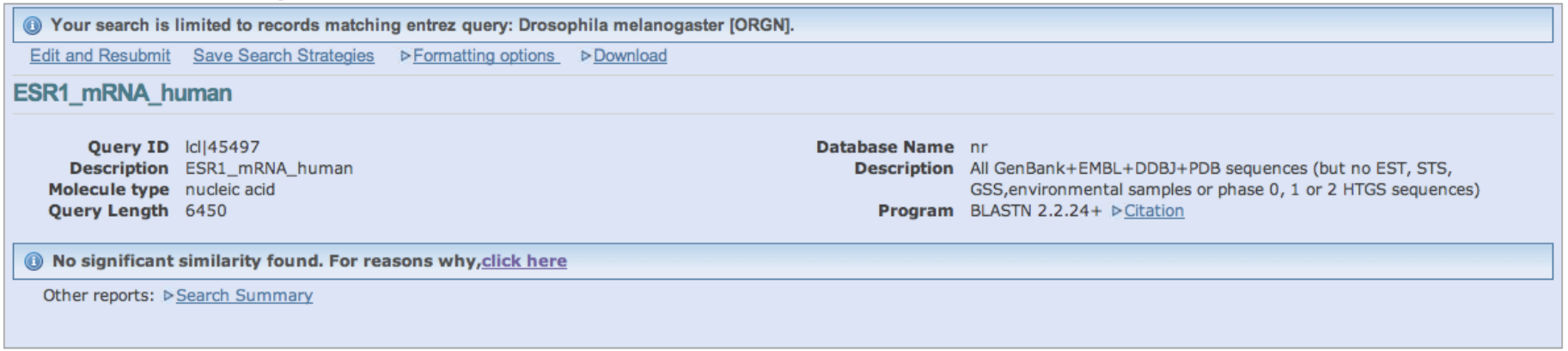
Score = 132 bits (331), Expect = 1e-31, Method: Compositional matrix adjust.
Identities = 87/282 (31%), Positives = 133/282 (48%), Gaps = 25/282 (8%)

```
Query  179  AKETRYCAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQ  238
            +K+   C VC D +SG HYG ++CEGCK+FFKRS++ +  Y C  +  C ID++ R  CQ
Sbjct  194  SKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSRNCPIDQHHRNQCQ  253

Query  239  ACRLRKCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVGSAGDMRAANLWPSPLMI  298
             CRL+KC ++GM +  +           +R R      G  G  G  + AN    P+ I
Sbjct  254  YCRLKKCLKMGMRREAV----------QRGRVPPTQPGLAGMHGQYQIAN--GDPMGI  299

Query  299  KRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGL--LTNLADRELVH  356
                 +S   S      +S LL AEP   Y      +     ++MG+  +  LA R L
Sbjct  300  AGFNGHSYLSSY-----ISLLLRAEP---YPTSRYGQCMQPNNIMGIDNICELAARLLFS  351

Query  357  MINWAKRVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSME-HPGKLLFAPNLLLDRNQG  415
             + WAK +P F +L + DQV LL   W E+ ++     SM  H   LL A  L
Sbjct  352  AVEWAKNIPFFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAA  411

Query  416  KCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSG  457
              V   ++   +     + + +++   E+ CLK+I+L+ +G
Sbjct  412  DRVVAFMDHIRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTG  453
```

Μικρές διαφορές στη στοίχιση, στο score & E-value

Χρησιμοποιώντας το Blast (xiii)

- Αν για το ίδιο γονίδιο (ESR1_Human) χρησιμοποιούσαμε το mRNA του (X03635.1 Homo sapiens mRNA for estrogen receptor α) και όχι την πρωτεΐνη για την αναζήτηση στην Drosophila:
  - Blastn (nr database): κανένας στόχος. Γιατί;
  - Ποιό πρόγραμμα του Blast θα έπρεπε να χρησιμοποιήσουμε;

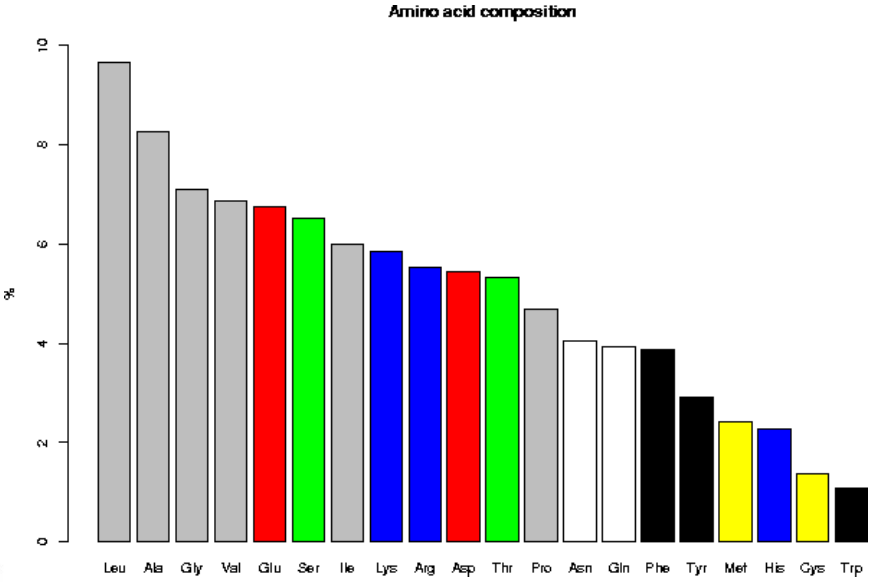


Search Parameters	
Program	blastn
Word size	11
Expect value	1e-05
Hitlist size	100
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

- Εκφυλισμός του γενετικού κώδικα
- Αμινοξέα με κοινές φυσικοχημικές ιδιότητες

Second Letter

First Letter	T	C	A	G	Third Letter
T	TTT Phe TTC Phe TTA Leu TTG Leu	TCT Ser TCC Ser TCA Ser TCG Ser	TAT Tyr TAC Tyr TAA Stop TAG Stop	TGT Cys TGC Cys TGA Stop TGG Trp	T C A G
C	CTT Leu CTC Leu CTA Leu CTG Leu	CCT Pro CCC Pro CCA Pro CCG Pro	CAT His CAC His CAA Gln CAG Gln	CGT Arg CGC Arg CGA Arg CGG Arg	T C A G
A	ATT Ile ATC Ile ATA Ile ATG Met	ACT Thr ACC Thr ACA Thr ACG Thr	AAT Asn AAC Asn AAA Lys AAG Lys	AGT Ser AGC Ser AGA Arg AGG Arg	T C A G
G	GTT Val GTC Val GTA Val GTG Val	GCT Ala GCC Ala GCA Ala GCG Ala	GAT Asp GAC Asp GAA Glu GAG Glu	GGT Gly GGC Gly GGA Gly GGG Gly	T C A G

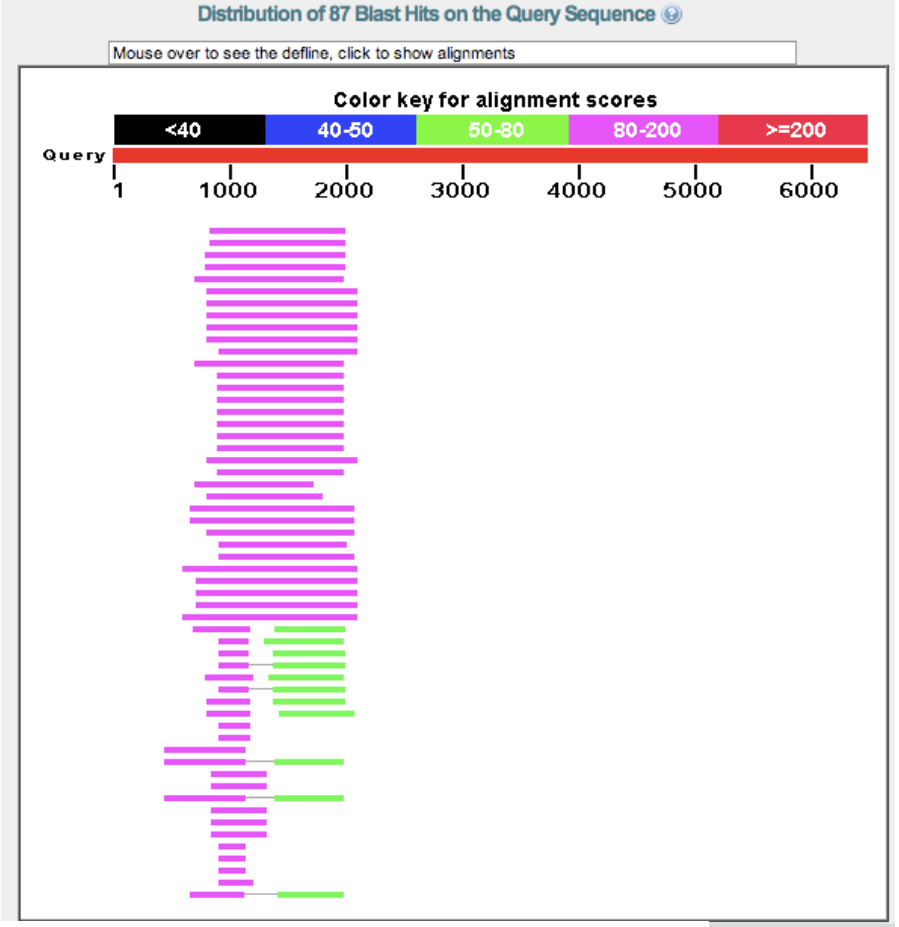


	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Χρησιμοποιώντας το Blast (xiv)

Για το ίδιο mRNA

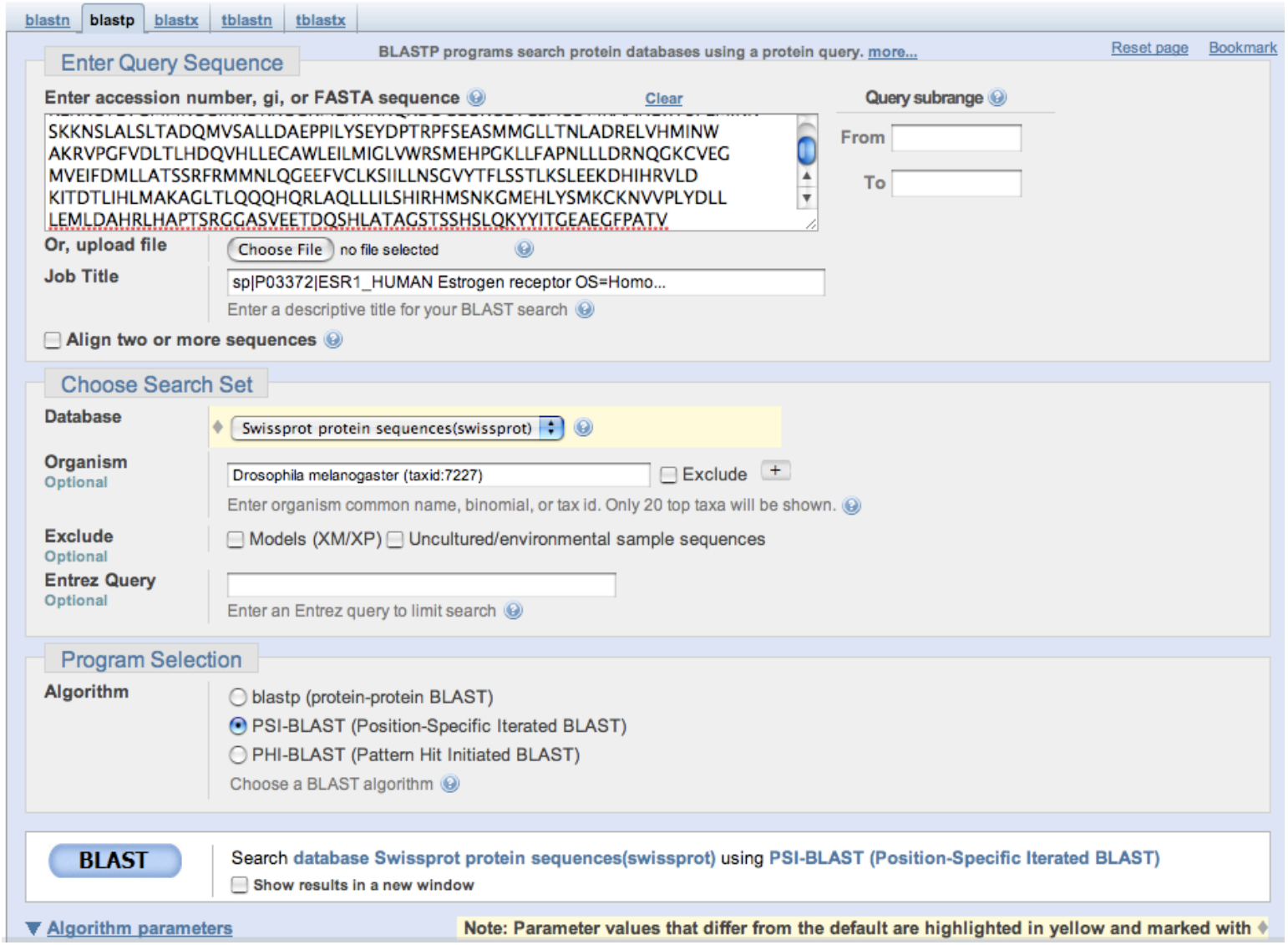
- Blastx (nr database)



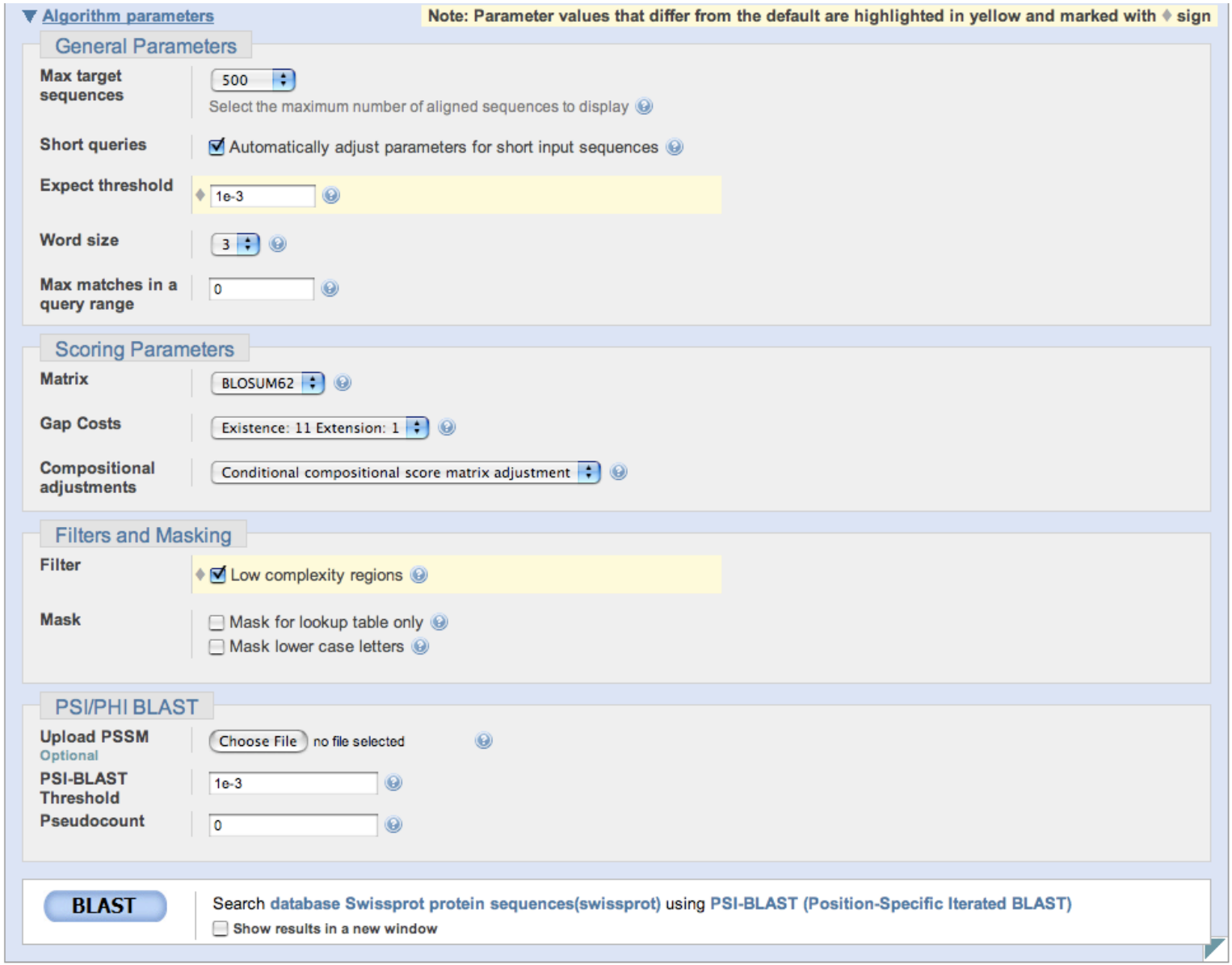
Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NP_729340.1	estrogen-related receptor, isoform A [Drosophila melanogas	198	198	17%	5e-50	31%	U G
NP_648183.3	estrogen-related receptor, isoform B [Drosophila melanogas	198	198	17%	5e-50	31%	U G
CAA36827.1	unnamed protein product [Drosophila melanogaster]	154	154	18%	1e-36	26%	
NP_476781.1	ultraspiracle [Drosophila melanogaster] >sp|P20153.1|USP_	151	151	18%	7e-36	25%	U G
NP_524325.1	seven up, isoform B [Drosophila melanogaster] >sp|P16375	143	143	19%	3e-33	27%	U G
ACS68165.1	FI04795p [Drosophila melanogaster]	140	140	19%	2e-32	25%	
NP_001097126.1	hepatocyte nuclear factor 4, isoform D [Drosophila melanog	140	140	19%	2e-32	25%	G
NP_723413.1	hepatocyte nuclear factor 4, isoform B [Drosophila melanog	140	140	19%	2e-32	25%	U G
NP_476887.2	hepatocyte nuclear factor 4, isoform A [Drosophila melanog	140	140	19%	2e-32	25%	G
NP_723414.1	hepatocyte nuclear factor 4, isoform C [Drosophila melanog	140	140	19%	2e-32	25%	U G
AAB09592.1	hepatocyte nuclear factor 4 homolog [Drosophila melanogas	139	139	18%	4e-32	26%	
AAM76194.1	RE08410p [Drosophila melanogaster]	139	139	19%	5e-32	27%	

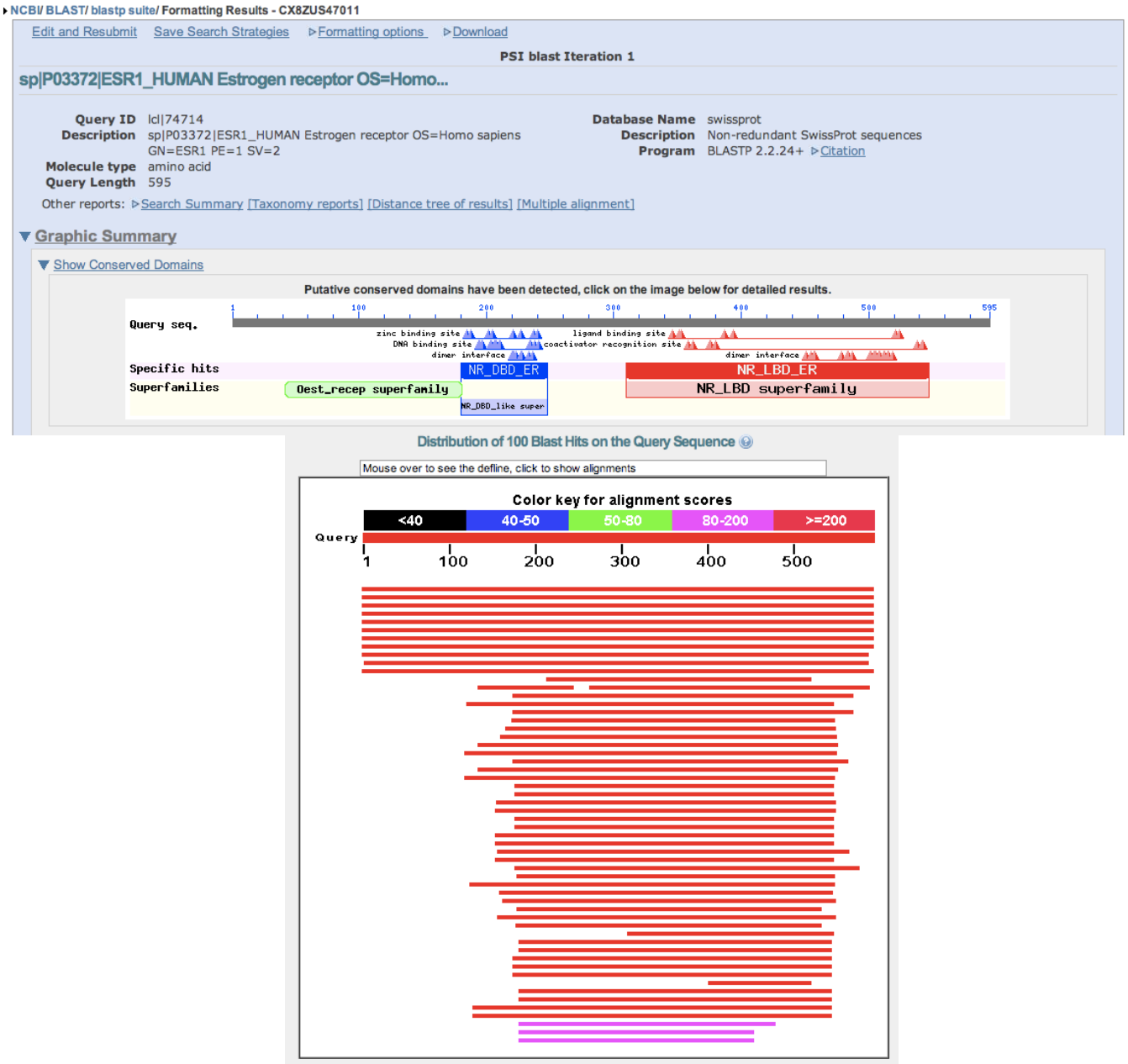
Χρησιμοποιώντας το PSI-Blast (i)



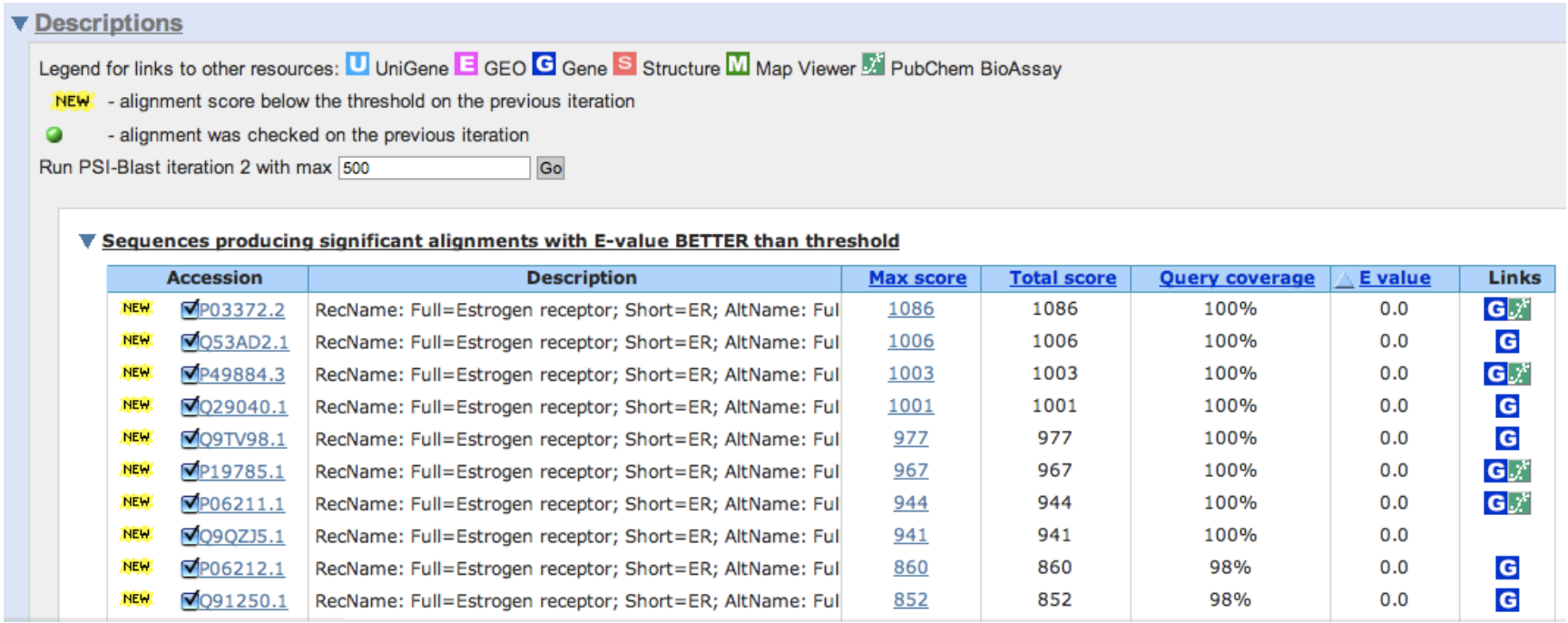
Χρησιμοποιώντας το PSI-Blast (ii)



Χρησιμοποιώντας το PSI-Blast (iii)



Χρησιμοποιώντας το PSI-Blast (iv)



Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

NEW - alignment score below the threshold on the previous iteration

- alignment was checked on the previous iteration

Run PSI-Blast iteration 2 with max 500 Go

Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Links
NEW P03372.2	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	1086	1086	100%	0.0	G
NEW Q53AD2.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	1006	1006	100%	0.0	G
NEW P49884.3	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	1003	1003	100%	0.0	G
NEW Q29040.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	1001	1001	100%	0.0	G
NEW Q9TV98.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	977	977	100%	0.0	G
NEW P19785.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	967	967	100%	0.0	G
NEW P06211.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	944	944	100%	0.0	G
NEW Q9QZJ5.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	941	941	100%	0.0	
NEW P06212.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	860	860	98%	0.0	G
NEW Q91250.1	RecName: Full=Estrogen receptor; Short=ER; AltName: Ful	852	852	98%	0.0	G

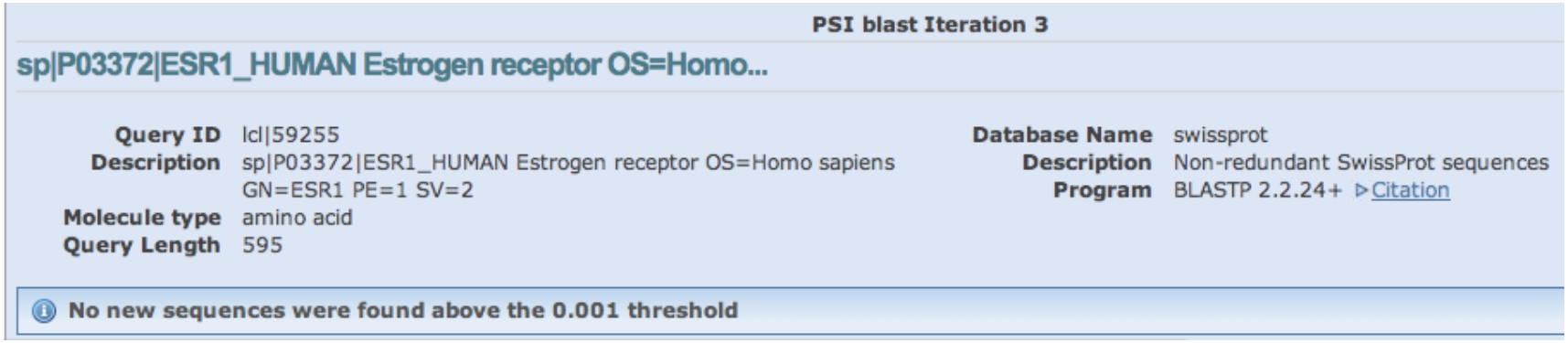
Χρησιμοποιώντας το PSI-Blast (v)

- Πράσινο σφαιρίδιο για ακολουθίες που είχαν βρεθεί σε προηγούμενο γύρο αναζήτησης.
- Μπορούμε να επιλέξουμε τον αποκλεισμό κάποιων ακολουθιών

	☑ O16662.3	RecName: Full=Nuclear hormone receptor family member	95.2	95.2	33%	1e-18	G
	☑ P41933.1	RecName: Full=Nuclear hormone receptor family member	93.7	93.7	21%	4e-18	G
NEW	☑ O57568.1	RecName: Full=Thyroid hormone receptor alpha; AltName:	85.2	85.2	21%	1e-15	
NEW	☑ O17573.1	RecName: Full=Nuclear hormone receptor family member	80.9	80.9	49%	2e-14	G
NEW	☑ P79404.1	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	80.6	80.6	19%	3e-14	G
NEW	☑ Q15466.2	RecName: Full=Nuclear receptor subfamily 0 group B mem	78.3	78.3	43%	2e-13	G
	☑ P35547.1	RecName: Full=Glucocorticoid receptor; Short=GR; AltNam	76.7	76.7	11%	5e-13	G
NEW	☑ Q9TXJ1.2	RecName: Full=Nuclear hormone receptor family member	74.8	74.8	11%	2e-12	G
NEW	☑ P70503.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	73.2	73.2	34%	4e-12	G
NEW	☑ Q9BG94.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	70.9	70.9	33%	2e-11	
NEW	☑ Q61066.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	70.9	70.9	43%	2e-11	G
NEW	☑ P79386.2	RecName: Full=Nuclear receptor subfamily 0 group B mem	70.5	70.5	38%	3e-11	G
NEW	☑ Q8QHI2.1	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	70.2	70.2	17%	4e-11	G
NEW	☑ Q9BG97.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.8	69.8	33%	5e-11	G
NEW	☑ P51843.2	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.8	69.8	33%	6e-11	G
NEW	☑ Q9BG93.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.8	69.8	33%	6e-11	G
NEW	☑ O17025.2	RecName: Full=Nuclear hormone receptor family member	69.4	69.4	11%	7e-11	G
NEW	☑ P97947.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.4	69.4	37%	7e-11	G
NEW	☑ O16360.3	RecName: Full=Nuclear hormone receptor family member	69.4	69.4	33%	7e-11	G
NEW	☑ Q9BG96.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	67.5	67.5	33%	3e-10	
NEW	☑ O02305.2	RecName: Full=Nuclear hormone receptor family member	66.3	66.3	45%	6e-10	G
NEW	☑ O17934.1	RecName: Full=Nuclear hormone receptor family member	65.9	65.9	52%	8e-10	G
NEW	☑ Q62227.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	65.2	65.2	33%	1e-09	G
NEW	☑ O45907.1	RecName: Full=Nuclear hormone receptor family member	55.9	55.9	34%	7e-07	G
NEW	☐ P20659.4	RecName: Full=Histone-lysine N-methyltransferase trithora	55.1	55.1	18%	1e-06	G
NEW	☐ Q24742.1	RecName: Full=Histone-lysine N-methyltransferase trithora	55.1	55.1	19%	1e-06	
NEW	☑ O16354.1	RecName: Full=Nuclear hormone receptor family member	53.6	53.6	33%	4e-06	G
NEW	☑ Q9PUA8.1	RecName: Full=Thyroid hormone receptor alpha; AltName:	45.9	45.9	3%	9e-04	

Run PSI-Blast iteration 3 with max 5000 Go

Χρησιμοποιώντας το PSI-Blast (vi)



Χρησιμοποιώντας το PSI-Blast (vii)

- Αν περιλαμβάνονταν οι 2 μεθυλ-τρανσφεράσες…

	Q24742.1	RecName: Full=Histone-lysine N-methyltransferase trithora	85.2	85.2	19%	1e-15	
	P79404.1	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	84.4	84.4	19%	2e-15	G
	O57568.1	RecName: Full=Thyroid hormone receptor alpha; AltName:	84.4	84.4	21%	2e-15	
	P35547.1	RecName: Full=Glucocorticoid receptor; Short=GR; AltNam	79.4	79.4	11%	7e-14	G
	Q9TXJ1.2	RecName: Full=Nuclear hormone receptor family member	79.4	79.4	54%	7e-14	G
	Q8QHI2.1	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	76.0	76.0	17%	8e-13	G
	P20659.4	RecName: Full=Histone-lysine N-methyltransferase trithora	74.4	74.4	18%	2e-12	G
	Q9PUA8.1	RecName: Full=Thyroid hormone receptor alpha; AltName:	47.1	47.1	11%	4e-04	
NEW	P16356.2	RecName: Full=DNA-directed RNA polymerase II subunit RI	45.5	45.5	19%	0.001	G

Run PSI-Blast iteration 4 with max 5000 Go

	P20659.4	RecName: Full=Histone-lysine N-methyltransferase trithora	71.0	71.0	18%	2e-11	G
	P16356.2	RecName: Full=DNA-directed RNA polymerase II subunit RI	66.7	66.7	19%	4e-10	G
NEW	P35074.2	RecName: Full=DNA-directed RNA polymerase II subunit RI	61.7	61.7	19%	1e-08	G
NEW	P24928.2	RecName: Full=DNA-directed RNA polymerase II subunit RI	50.6	50.6	18%	3e-05	G
NEW	P08775.3	RecName: Full=DNA-directed RNA polymerase II subunit RI	50.6	50.6	18%	3e-05	G
	Q9PUA8.1	RecName: Full=Thyroid hormone receptor alpha; AltName:	46.3	46.3	11%	6e-04	

Run PSI-Blast iteration 5 with max 5000 Go

Χρησιμοποιώντας το PSI-Blast (viii)

- Αποθήκευση αποτελεσμάτων