Στοίχιση ακολουθιών κατά ζεύγη (Pairwise alignment)

&

Blast

Στοίχιση κατά ζεύγη

- 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

Στοίχιση κατά ζεύγη

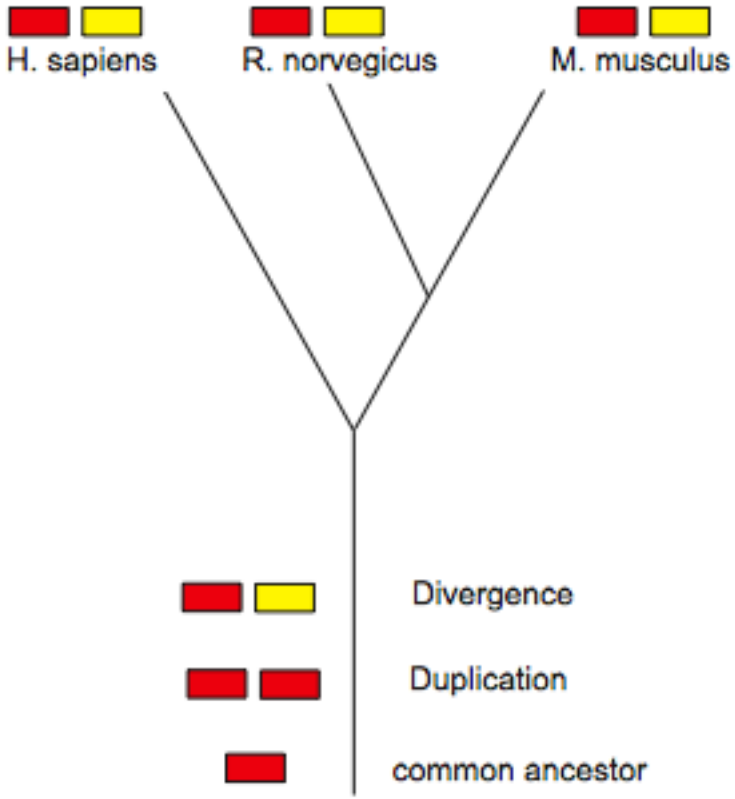
- Τοποθετούνται οι αντίστοιχοι χαρακτήρες ο ένας κάτω από τον άλλο και μπορεί να γίνει χρήση κενών (gaps)
- Δύο χαρακτήρες μπορεί να είναι:
  - Ίδιοι
  - Παρόμοιοι (κοινές φυσικοχημικές ιδιότητες, π.χ. Ισολευκίνη - βαλίνη)
  - Διαφορετικοί

```
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
```

Λίγη εξέλιξη: ομολογία

- Ομόλογα γονίδια: κοινός εξελικτικός πρόγονος.
- 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)

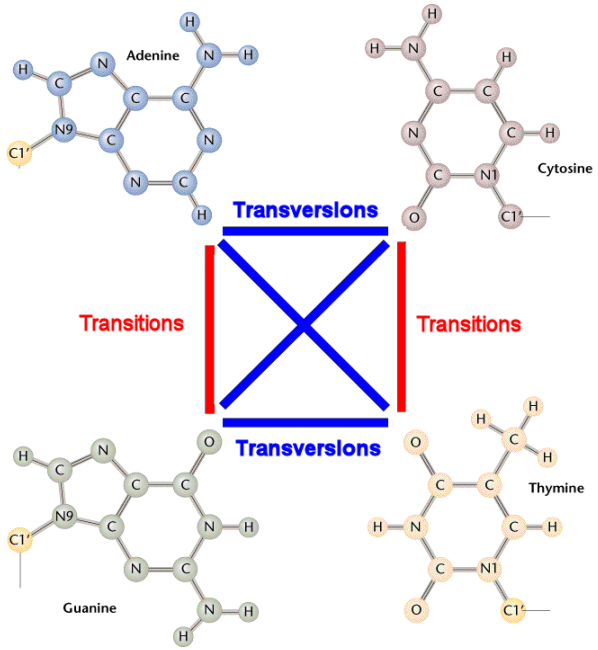


Βασικότερα είδη μεταλλάξεων

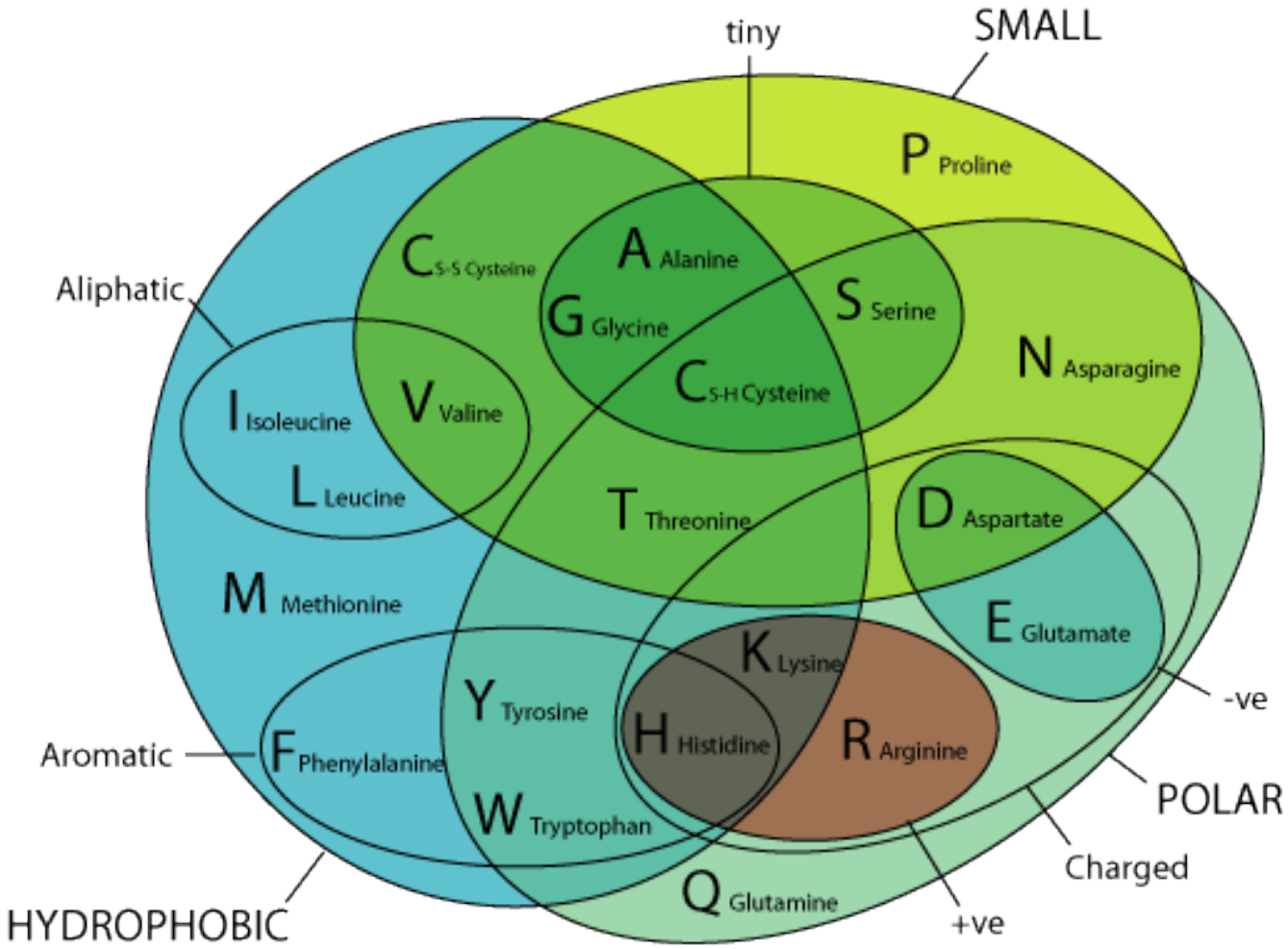
- Μεταλλάξεις σημείου (point mutations)
  - Συνώνυμες (synonymous)
  - Μη-συνώνυμες (non-synonymous)
    - Αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες
    - Αμινοξέα με διαφορετικές φυσικοχημικές ιδιότητες
  - Κωδικόνια τερματισμού

Μεταπτώσεις-μεταστροφές

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



Κατηγοριοποίηση αμινοξέων

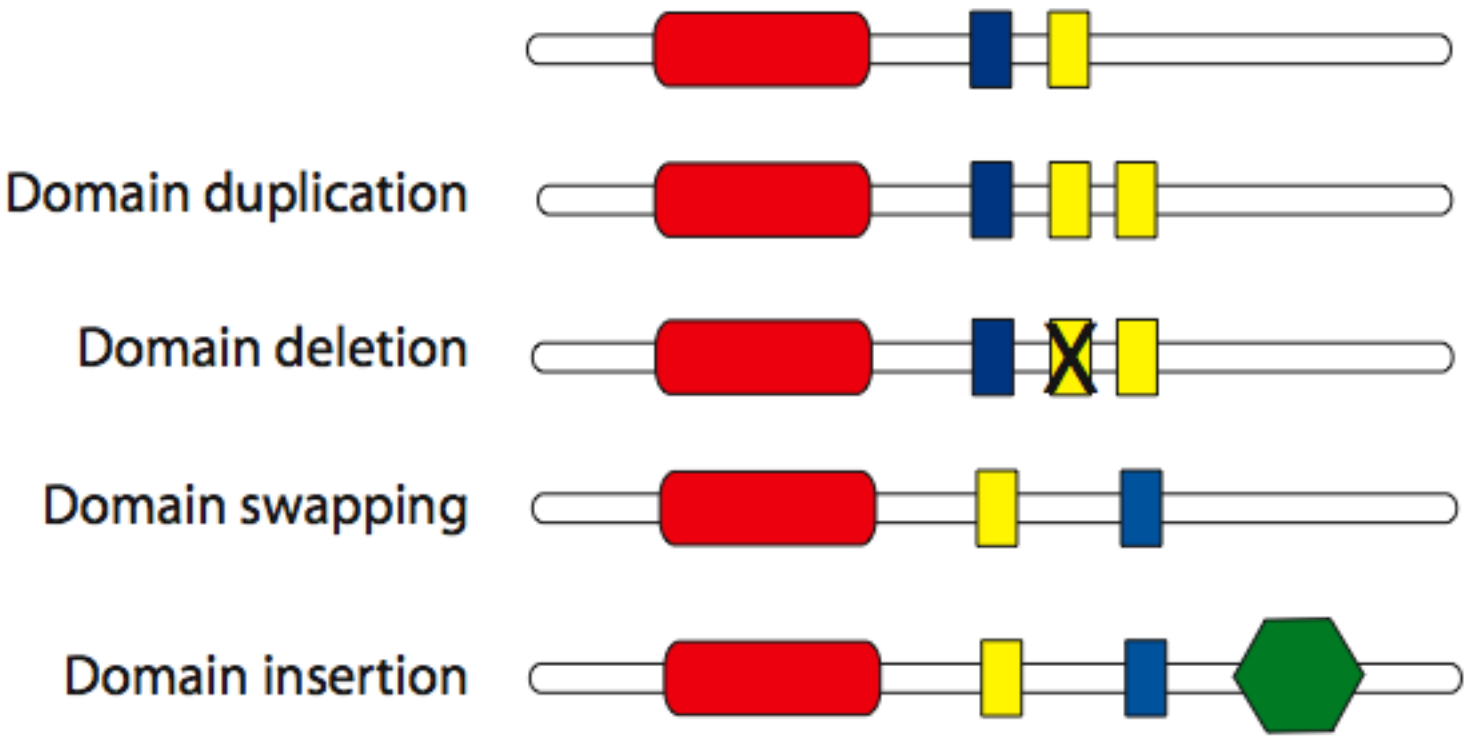


Βασικότερα είδη μεταλλάξεων

- 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): 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)

Είδη στοίχισης (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 (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 (iii)

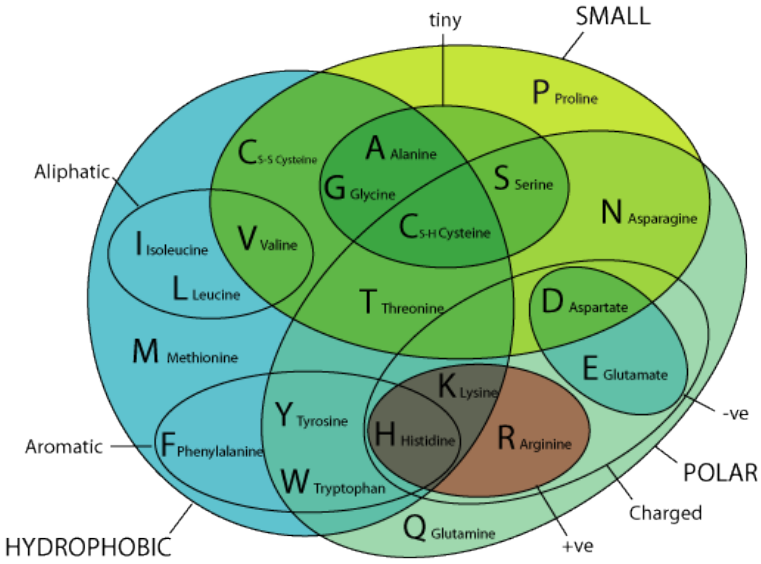
```
Global FTFTALILLAVAV
       F--TAL-LLA-AV

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

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




```
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.

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 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.

Πίνακες αντικατάστασης

- Στο παράδειγμα, 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.

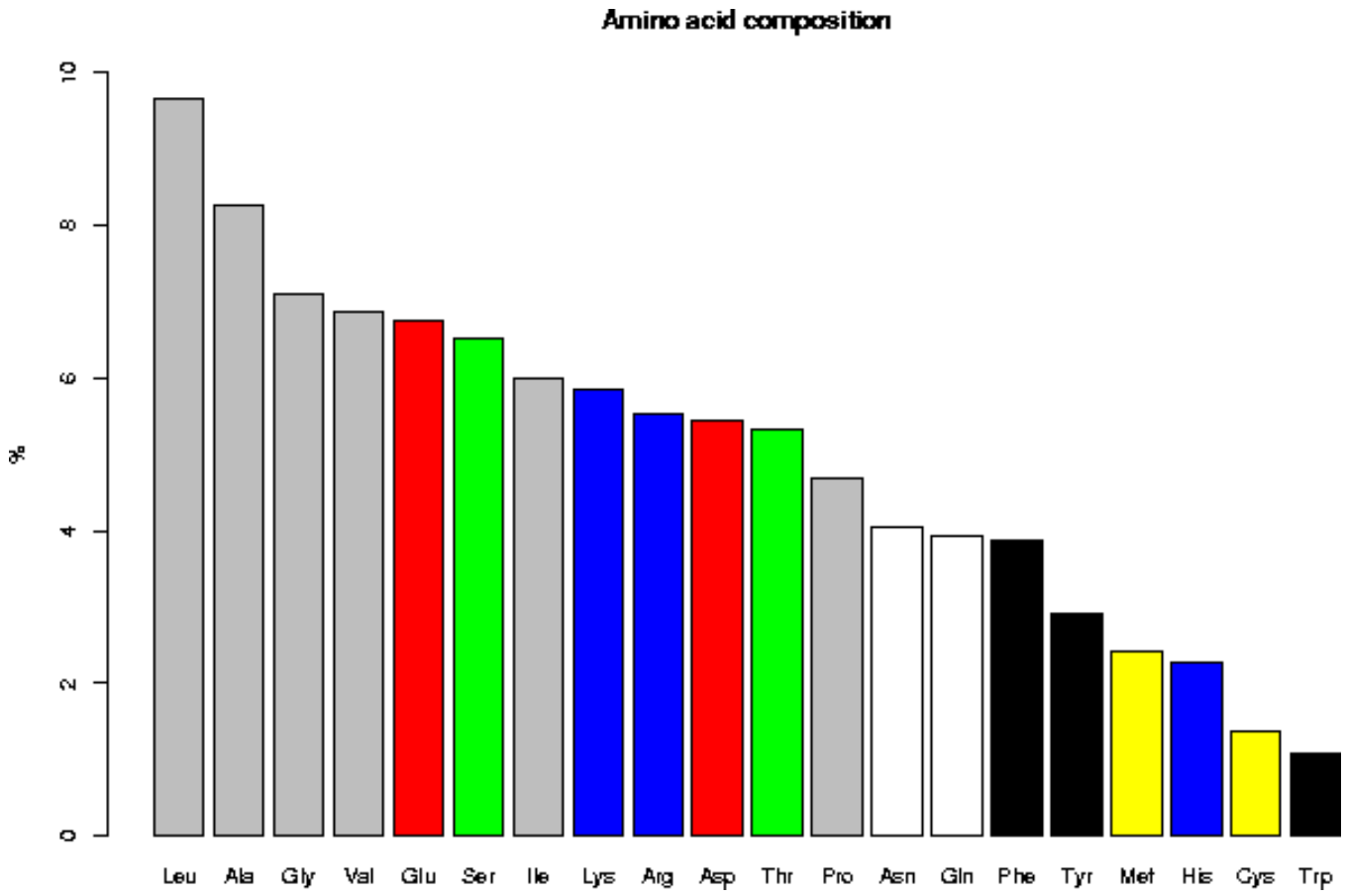
- Το ίδιο και για τις πρωτεΐνες.

- Χρειαζόμαστε πίνακες που βασίζονται σε συγκεκριμμένα εξελικτικά μοντέλα και λαμβάνουν υπόψην την συχνότητα του κάθε χαρακτήρα

Πίνακες αντικατάστασης

- Για πρωτεΐνες:
  - Για αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες, μεγαλύτερη πιθανότητα αντικατάστασης (συντηρητικές αντικαταστάσεις).
  - Πίνακες PAM
  - Πίνακες BLOSUM

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
- Αμινοξέα 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

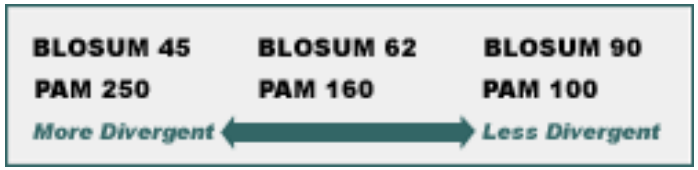
Πίνακες BLOSUM

- 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.



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.

Στατιστική σημαντικότητα μιας στοίχισης κατά ζεύγη

- Περισσότερες πληροφορίες στο:

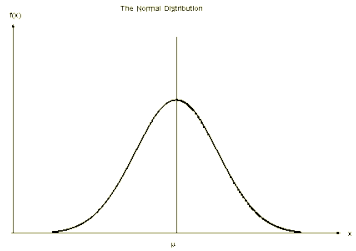
http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

- Στατιστική σημαντικότητα μιας στοίχισης 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)

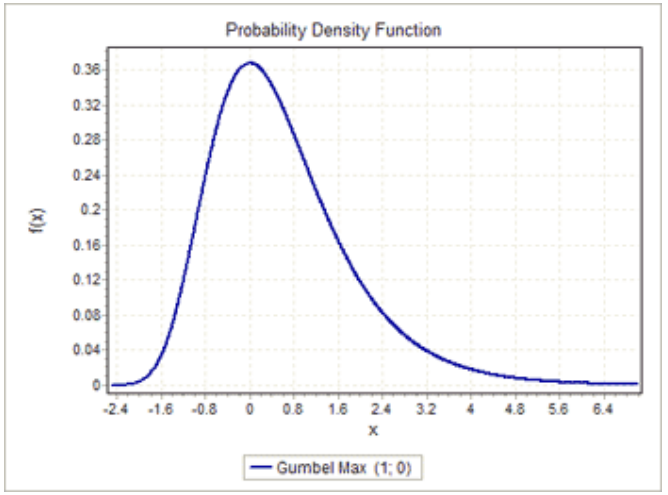
- Αν πραγματοποιηθεί το ανακάτεμα 100 φορές και η μέγιστη βαθμολογία στοίχισης δεν υπερβαίνει την βαθμολογία που παρατηρήσαμε για την στοίχιση των 2 πραγματικών ακολουθιών, τότε η στοίχιση είναι στατιστικά σημαντική σε επίπεδο P-value < 0.01

- Μεγάλο υπολογιστικό κόστος

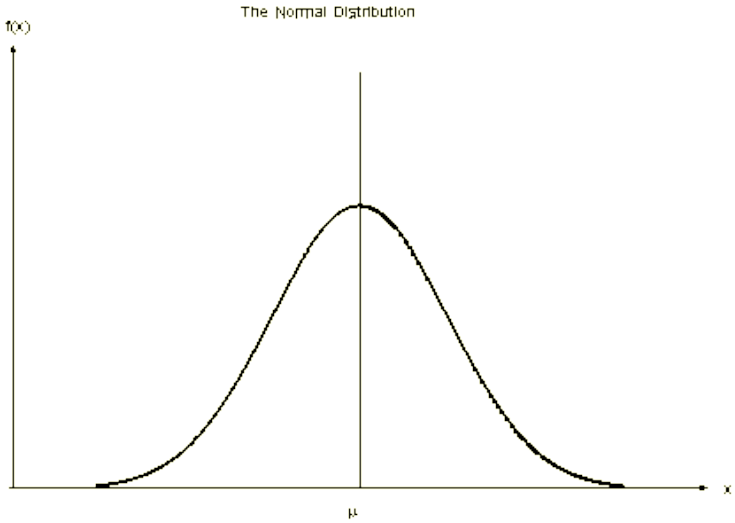
- Χρησιμοποιείται για ολικές στοιχίσεις,εντούτοις δεν ενδείκνυται η ολική στοίχιση για να αποφασίσουμε αν δύο ακολουθίες είναι ομόλογες

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 (iii)

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

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

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.

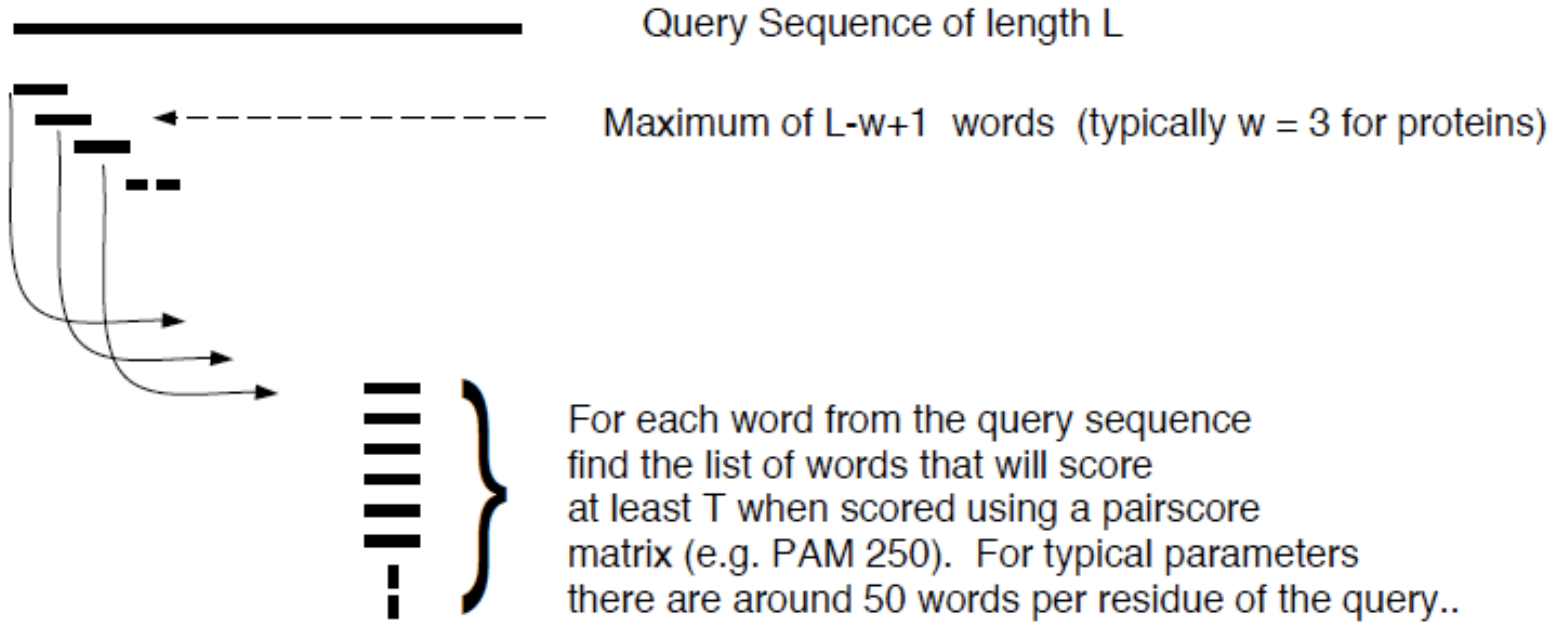
- Πρωτεΐνες: w=3

- 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

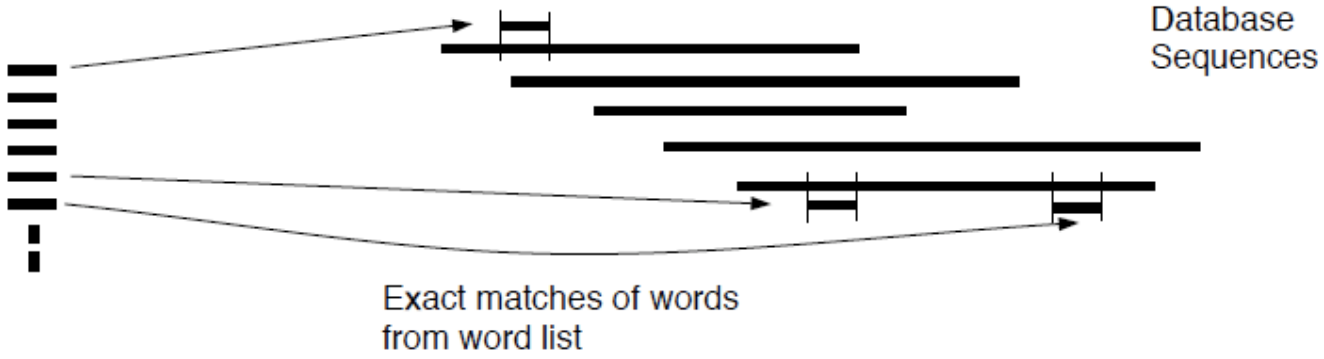
(1) For the query find the list of high scoring words of length w.



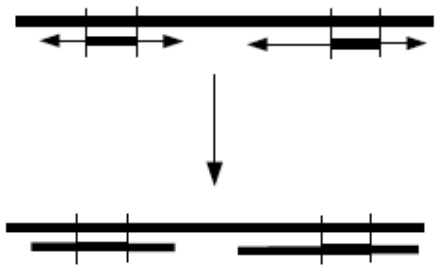
- PQG
- 20 X 20 X 20 = 8.000 words
- PQG X 8.000 words
- PQG X PEG = 7 + 2 + 6=15
- Όριο τιμής T

Αλγόριθμος BLAST

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



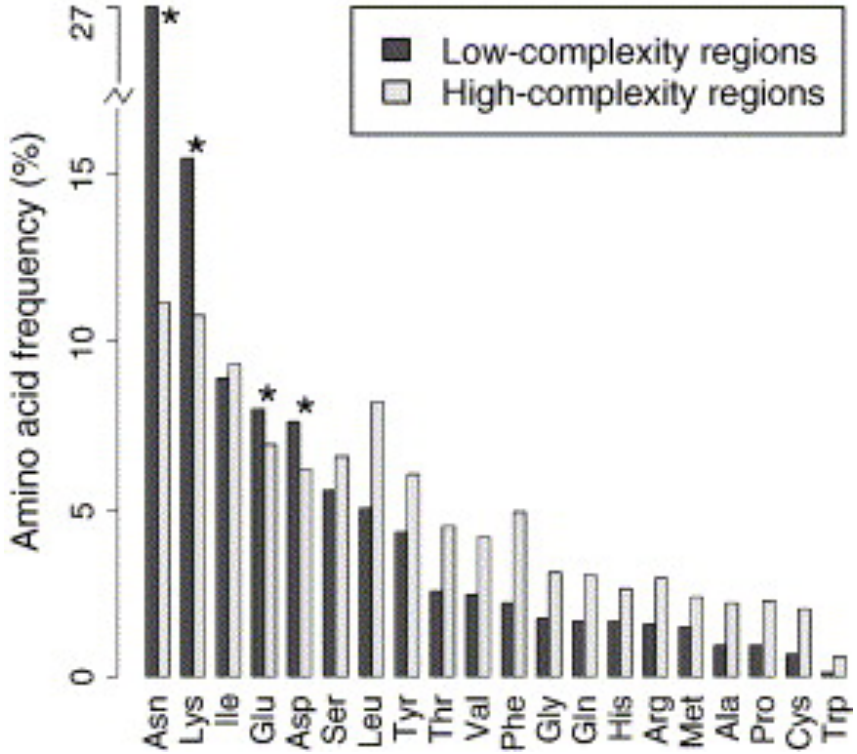
(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.



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

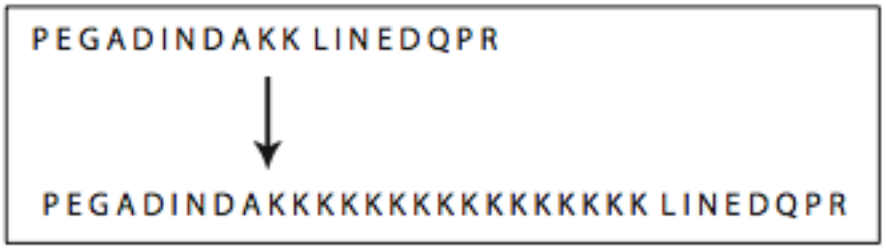


Περιοχές χαμηλής πολυπλοκότητας (ii)

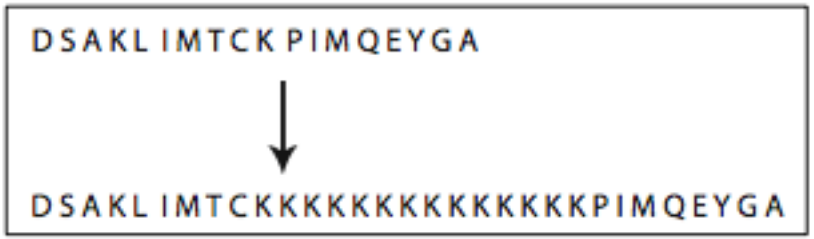
- 2 μη ομόλογες ακολουθίες.

PEGADINDAKK LINEDQPR
DSAKL IMTCK PIMQEYGA

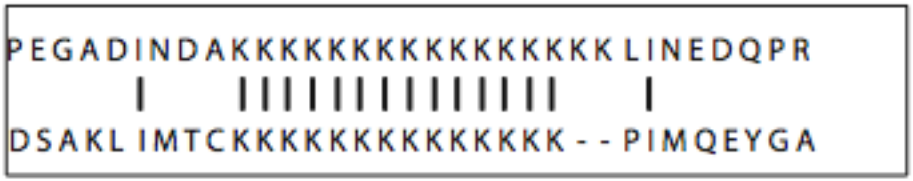
- Μεταλλάξεις στην ακολουθία 1.



- Μεταλλάξεις στην ακολουθία 2.

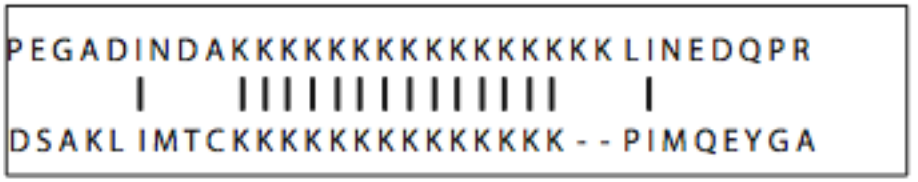


- Αν δεν φιλτραριστούν οι περιοχές χαμηλής πολυπλοκότητας:
  - Η στοίχιση θα δείξει ομολογία



Φιλτράρισμα περιοχών χαμηλής πολυπλοκότητας

- Φιλτράρισμα (masking)
- Και για BLAST και για FASTA.
- Φιλτράρεται η ακολουθία επερώτησης μόνο.
- X για πρωτεΐνες και N για νουκλεϊκά οξέα



- Άλλες ακολουθίες που φιλτράρονται:
  - Επαναλήψεις Alu
  - Φορείς κλωνοποίησης
- Φίλτρα του Blast:
  - Dust: νουκλεοτίδια
  - Seg: πρωτεΐνες

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

Blastn / MegaBlast

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

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

Blastn

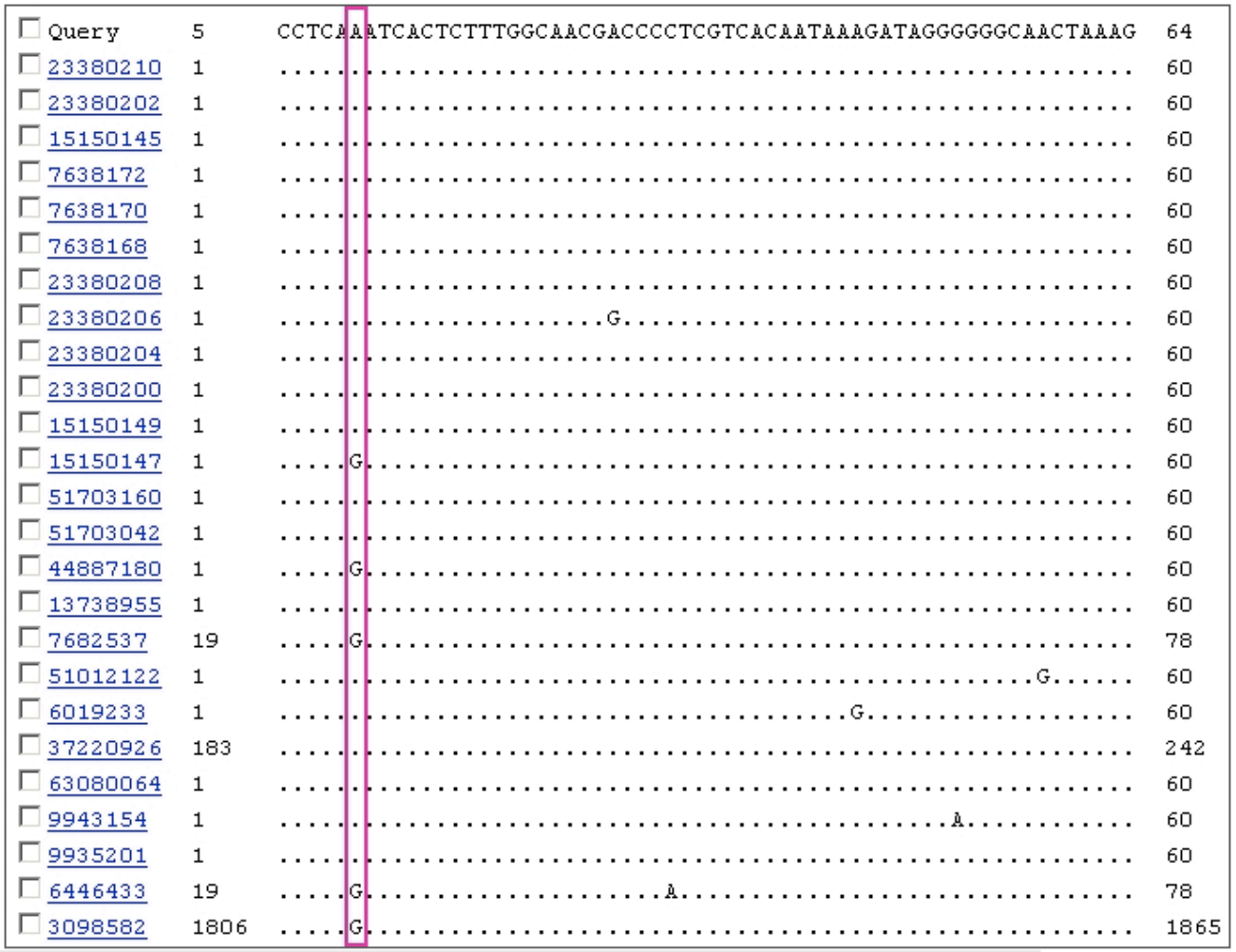
Παράδειγμα: Έλεγχος εξειδίκευσης ζεύγους εκκινητών (primers)

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

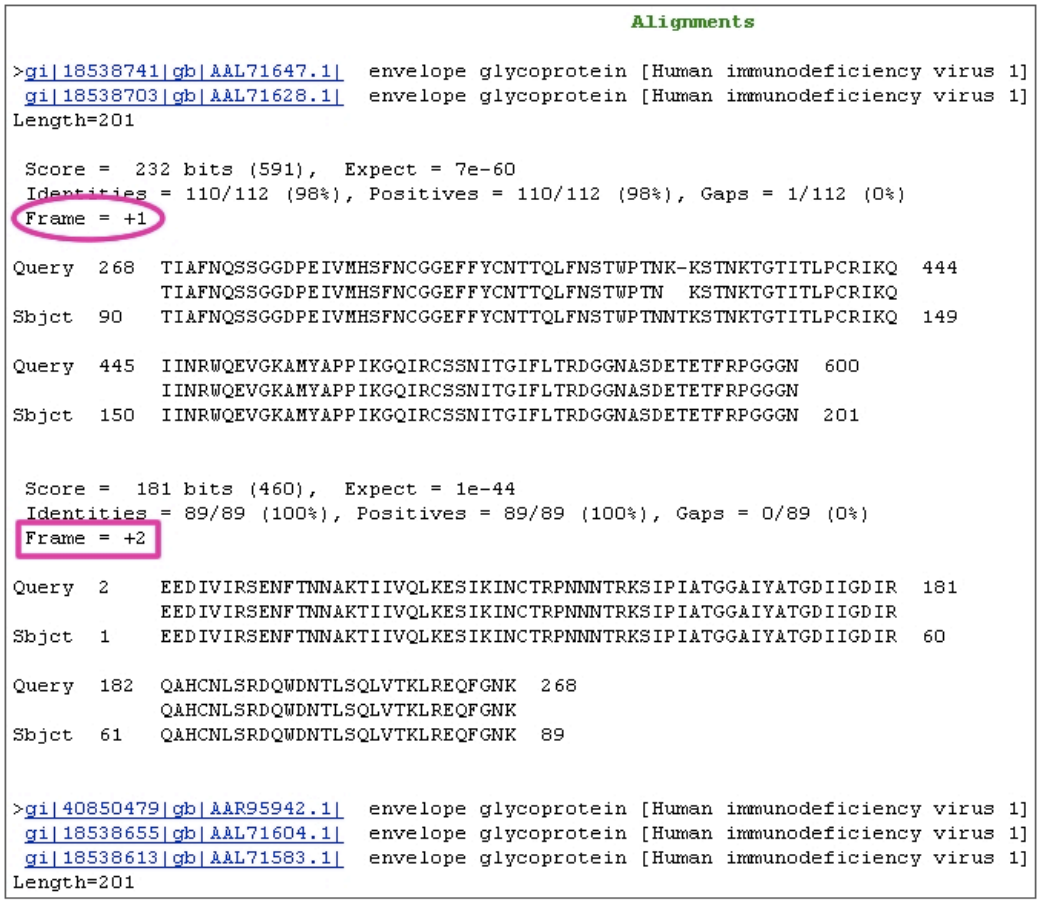
tblastn

- Πρωτεΐνη (query) Χ Β.Δ. νουκλεοτιδικών ακολουθιών μεταφρασμένων και στα 6 αναγνωστικά πλαίσια.
- Η Β.Δ. περιέχει νουκλεοτιδικές ακολουθίες με άγνωστη λειτουργία.
- Π.χ. Η Β.Δ. μπορεί να είναι μια συλλογή ESTs ή αμορφοποίητα δεδομένα από την αλληλούχιση ενός γενώματος (draft genome records)
- 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.

- Παράδειγμα: εντοπισμός μετάλλαξης που αλλάζει το αναγνωστικό πλαίσιο.



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.

- 6Χ6 blastp

- Αναζήτηση (διαειδική) για γονίδια που δεν έχουν εντοπιστεί με τις κλασσικές μεθόδους ή για γονίδια που οι πρωτεΐνες τους δεν υπάρχουν στην Β.Δ.

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

PSI-Blast

PSI-Blast

- 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

- 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

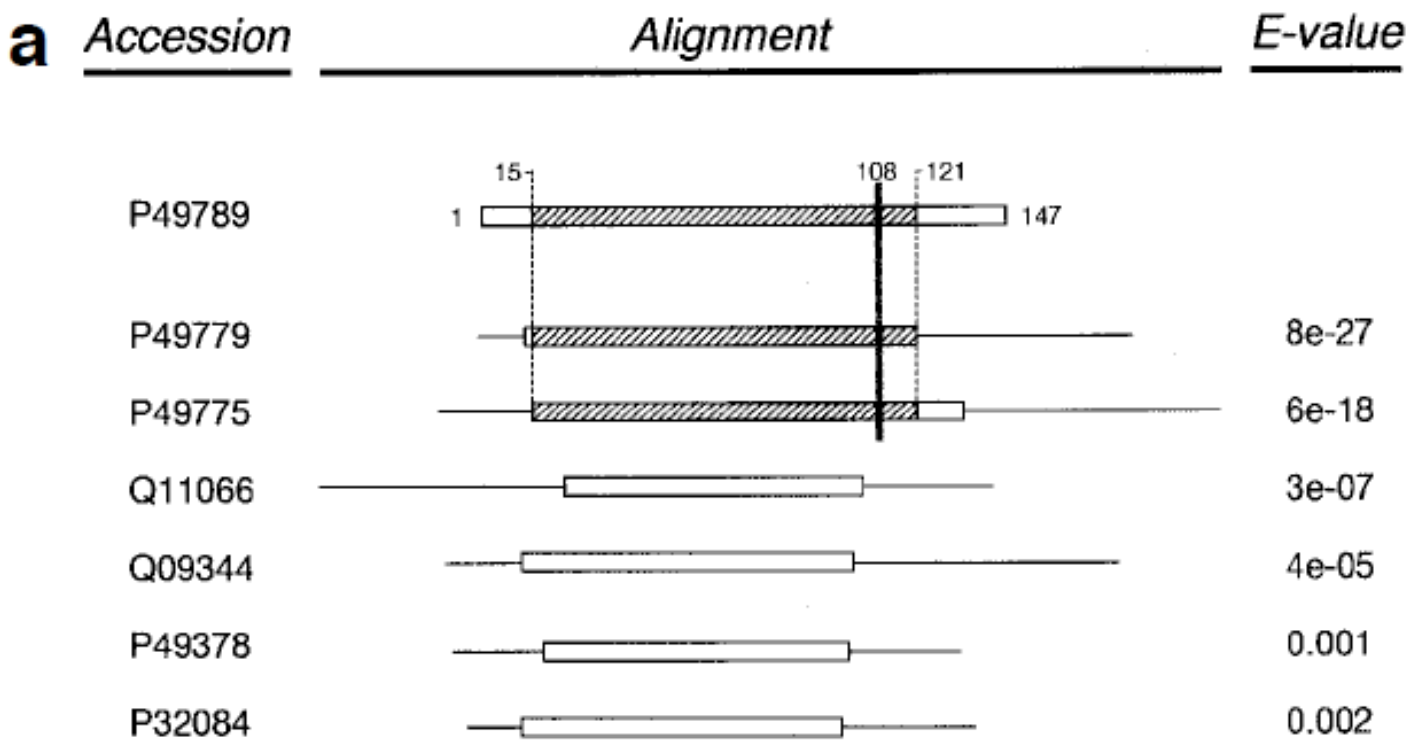
PSI-Blast

- 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.

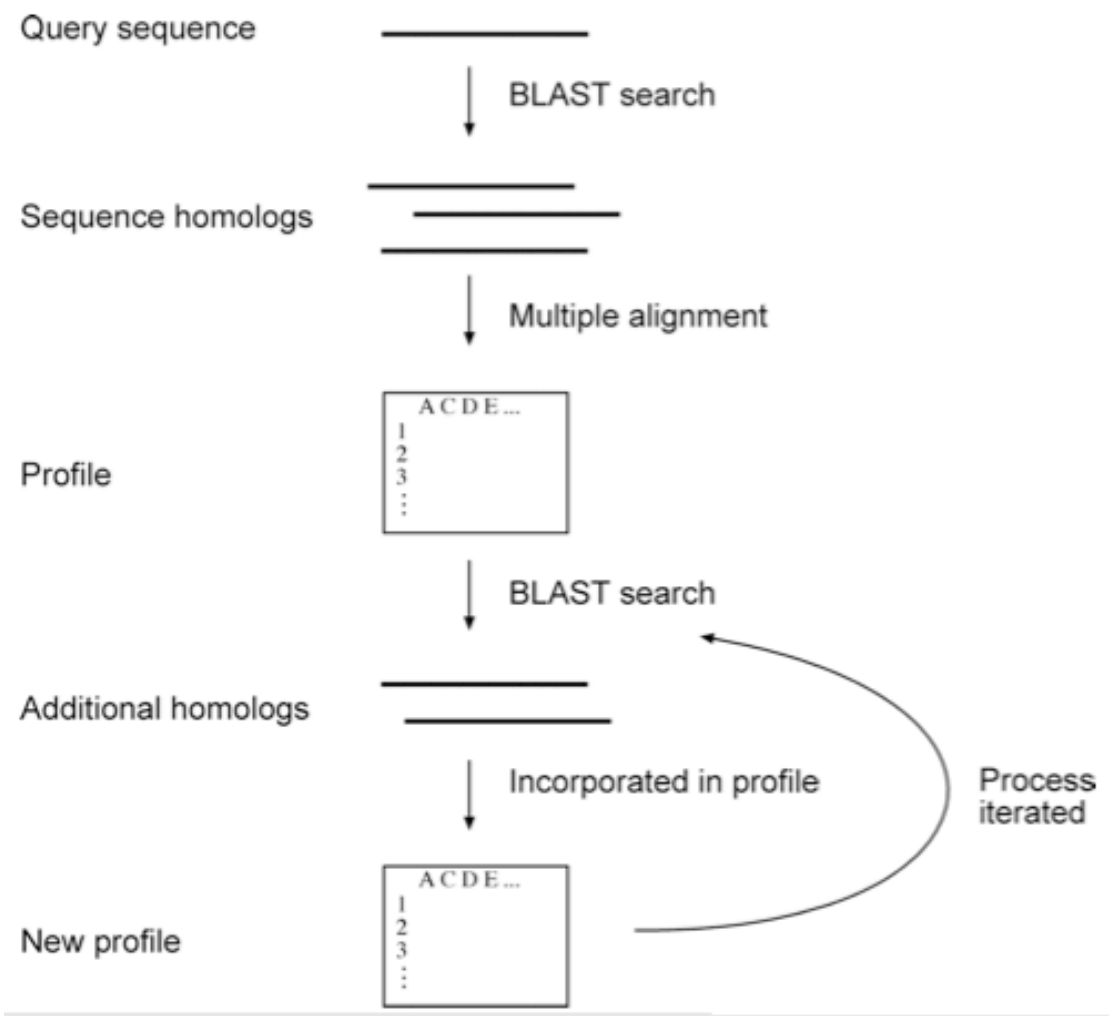
PSI-Blast

- 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
Profile
A C D E ...
1
2
3
BLAST search
Additional homologs
Incorporated in profile
Process iterated
New profile
A C D E ...
1
2
3

PSI-Blast

- Πριν κάνουμε PSI-Blast πρέπει να ξέρουμε τι αναζητάμε!!!
  - 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

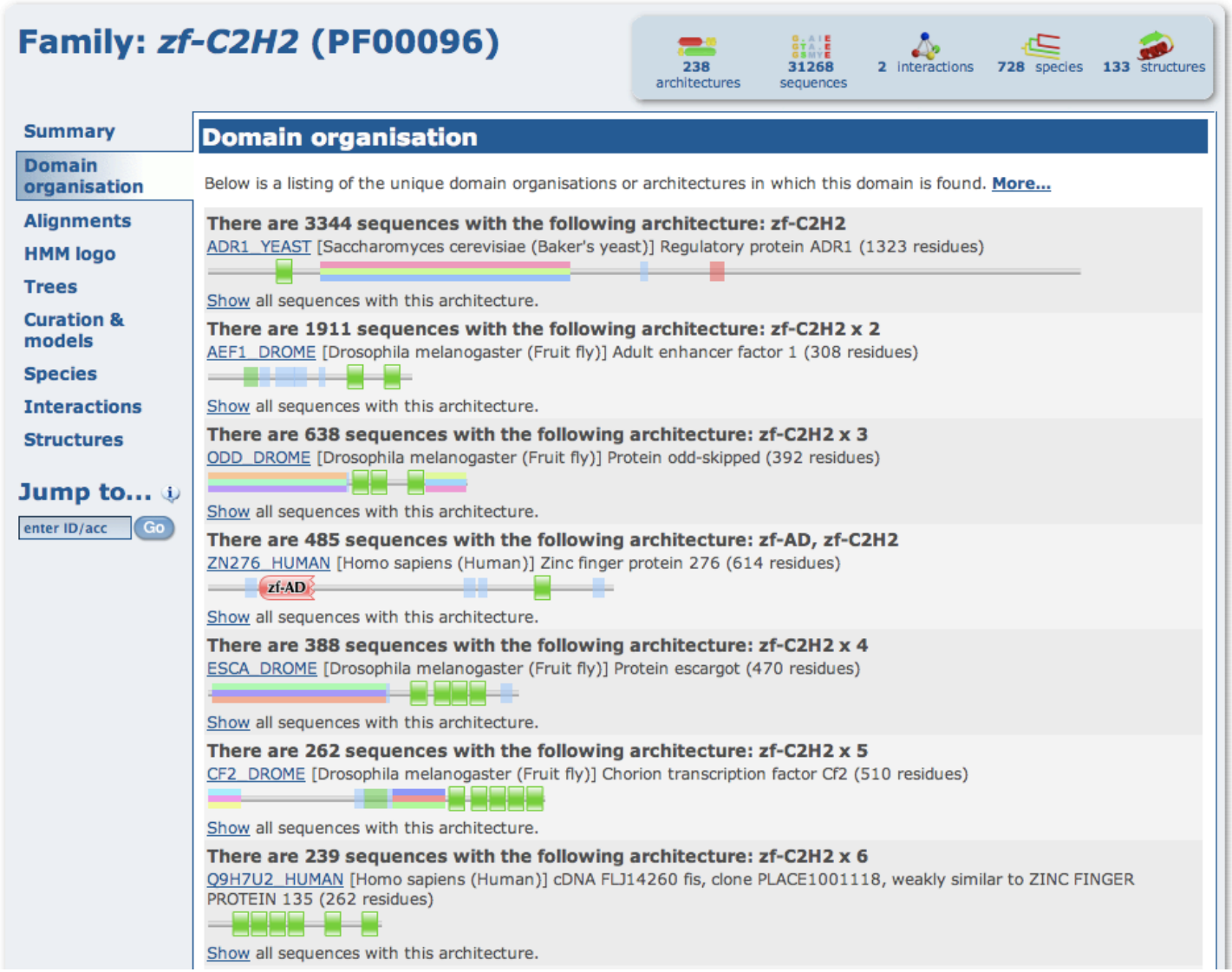
Click on image to view larger version.

Click on table to view larger version.

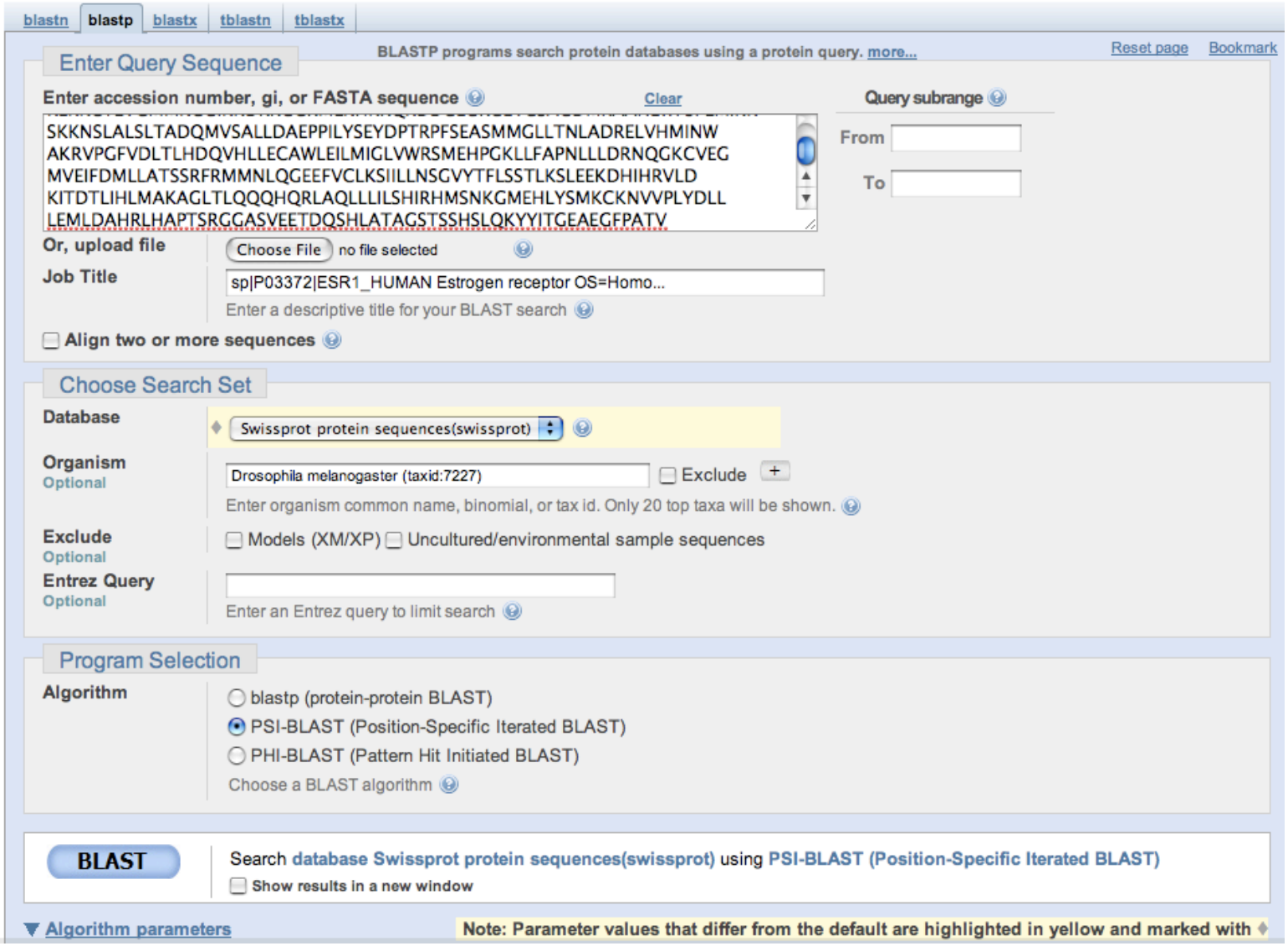
Table 2. The 10 most promiscuous domains in animals, fungi, and plants

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

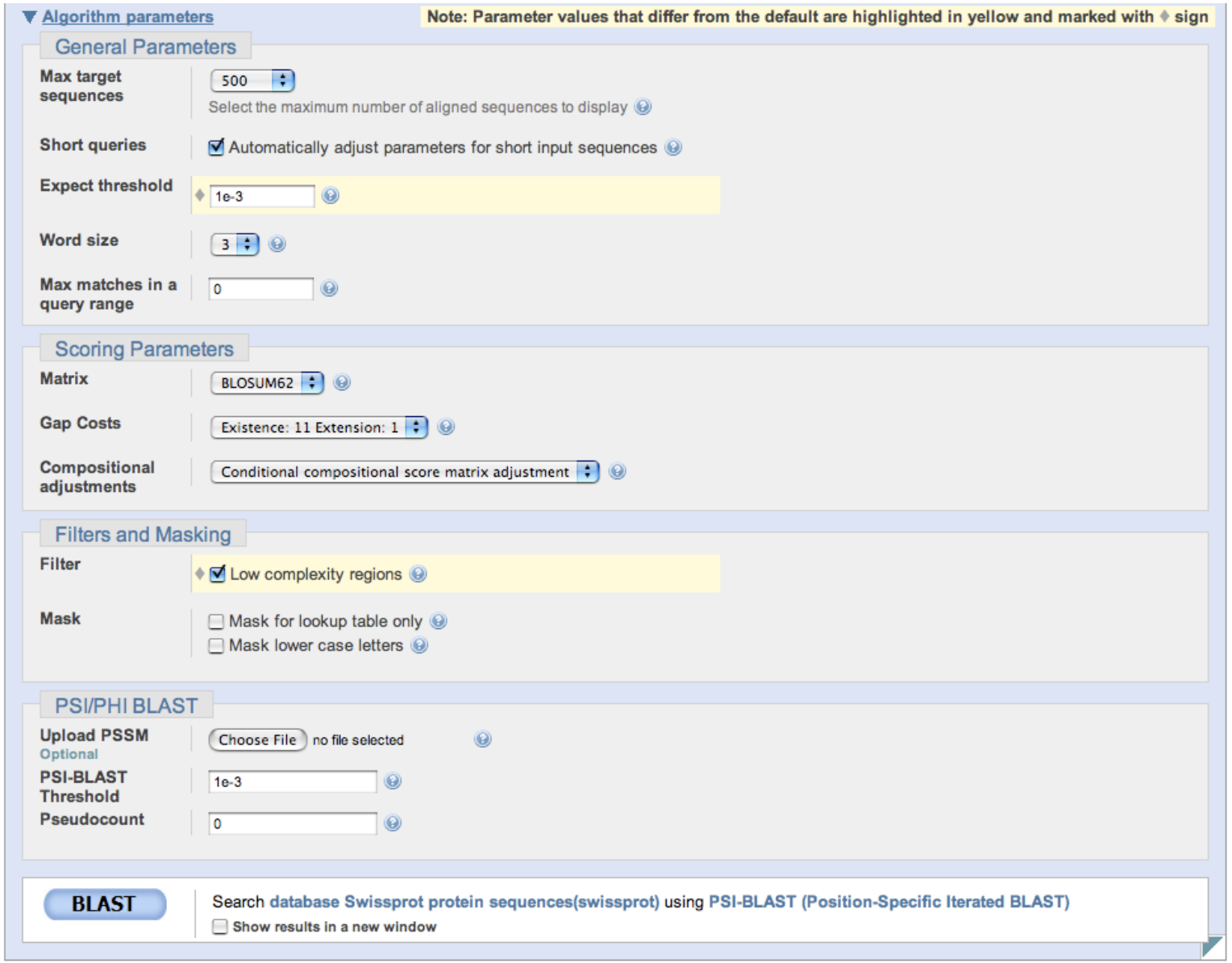
Επικράτειες και αναζήτηση σε Β.Δ.



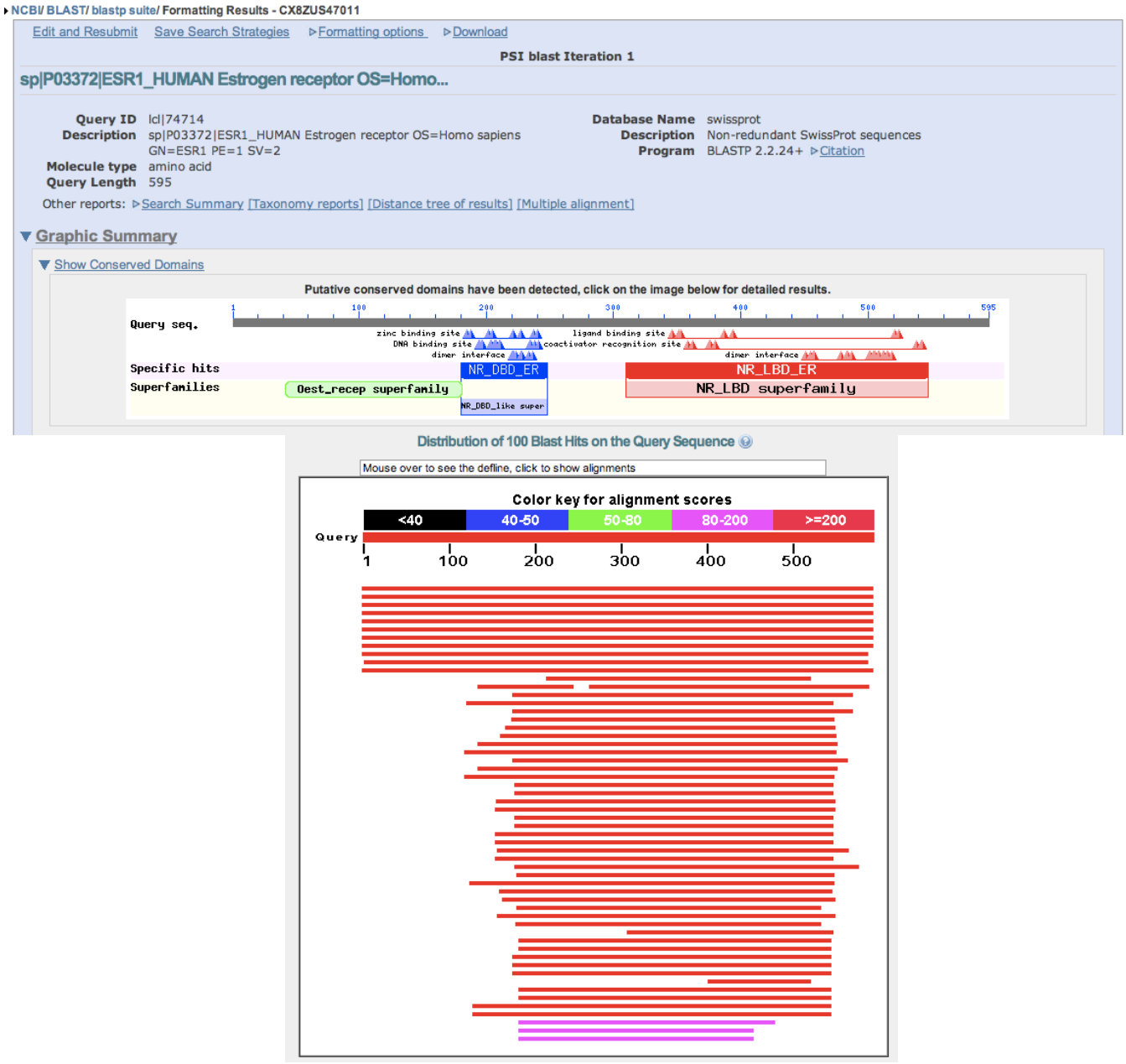
Χρησιμοποιώντας το PSI-Blast



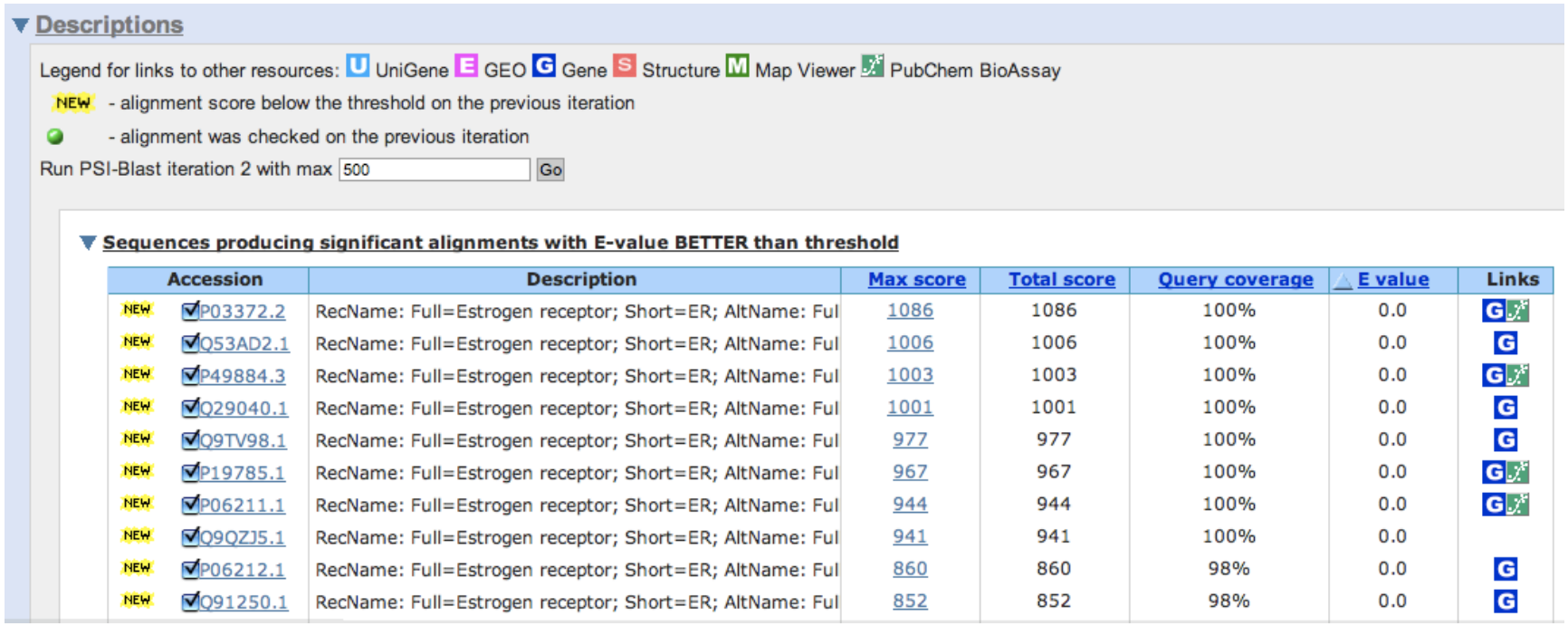
Χρησιμοποιώντας το PSI-Blast



Χρησιμοποιώντας το PSI-Blast



Χρησιμοποιώντας το PSI-Blast



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

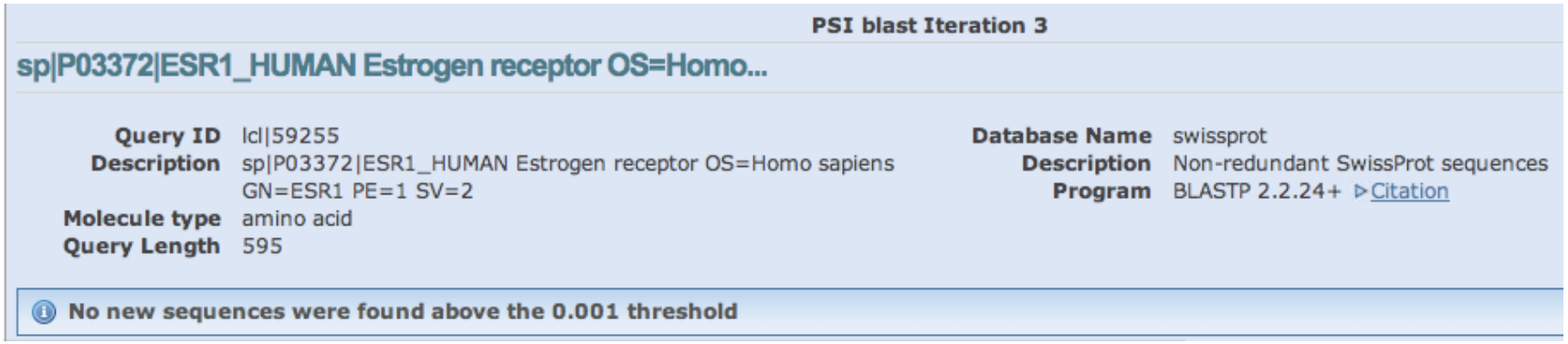
Χρησιμοποιώντας το PSI-Blast

- 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
	P41933.1	RecName: Full=Nuclear hormone receptor family member	93.7	93.7	21%	4e-18
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
NEW	P79404.1	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	80.6	80.6	19%	3e-14
NEW	Q15466.2	RecName: Full=Nuclear receptor subfamily 0 group B mem	78.3	78.3	43%	2e-13
	P35547.1	RecName: Full=Glucocorticoid receptor; Short=GR; AltNam	76.7	76.7	11%	5e-13
NEW	Q9TXJ1.2	RecName: Full=Nuclear hormone receptor family member	74.8	74.8	11%	2e-12
NEW	P70503.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	73.2	73.2	34%	4e-12
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
NEW	P79386.2	RecName: Full=Nuclear receptor subfamily 0 group B mem	70.5	70.5	38%	3e-11
NEW	Q8QHI2.1	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	70.2	70.2	17%	4e-11
NEW	Q9BG97.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.8	69.8	33%	5e-11
NEW	P51843.2	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.8	69.8	33%	6e-11
NEW	Q9BG93.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.8	69.8	33%	6e-11
NEW	O17025.2	RecName: Full=Nuclear hormone receptor family member	69.4	69.4	11%	7e-11
NEW	P97947.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	69.4	69.4	37%	7e-11
NEW	O16360.3	RecName: Full=Nuclear hormone receptor family member	69.4	69.4	33%	7e-11
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
NEW	O17934.1	RecName: Full=Nuclear hormone receptor family member	65.9	65.9	52%	8e-10
NEW	Q62227.1	RecName: Full=Nuclear receptor subfamily 0 group B mem	65.2	65.2	33%	1e-09
NEW	O45907.1	RecName: Full=Nuclear hormone receptor family member	55.9	55.9	34%	7e-07
NEW	P20659.4	RecName: Full=Histone-lysine N-methyltransferase trithora	55.1	55.1	18%	1e-06
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
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



Χρησιμοποιώντας το PSI-Blast

- 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

- Αποθήκευση αποτελεσμάτων

Άσκηση
Μέρος 1ο

Blast (1a)

- Βρείτε την ακολουθία του Estrogen receptor alpha (σε μορφή FASTA) ως:
  - mRNA από την EMBL bank (accesion number: X03635).
  - ως πρωτεΐνη από την Uniprot (accesion number: P03372).

Blast (1a)

Τα προγράμματα του Blast θα τα βρείτε στο:

http://blast.ncbi.nlm.nih.gov/Blast.cgi

Θέλετε να βρείτε τις ομόλογες πρωτεΐνες του ανθρώπινου estrogen receptor alpha (πρωτεΐνη) στη μύγα Drosophila melanogaster, χρησιμοποιώντας τη ΒΔ Swissprot.

Ποιό πρόγραμμα του Blast πρέπει να χρησιμοποιήσετε;

Οι παράμετροι της αναζήτησης:

- ΒΔ Swissprot
- οργανισμός: Drosophila melanogaster
- Expect threshold: 1e-5
- Low-complexity filtering

Blast (1a)

- Δείτε τα συντηρημένα domains. Ποιά είναι;
- Ποιό είναι το καλύτερο blast hit; με ποιό score & Evalue; Τι πρωτεΐνη είναι;
- Για το καλύτερο blast hit, δείτε στην τοπική στοίχιση:
  - Identities
  - Positives
  - Low complexity regions

Blast (1a)

- Βρείτε την πρωτεϊνική ακολουθία (σε μορφή FASTA) του καλύτερου blast hit και με αυτή κάνετε την αντίστροφη διαδικασία.

- Δηλαδή, blast έναντι της ΒΔ Swissprot, για τον οργανισμό Homo sapiens, χρησιμοποιώντας ως ακολουθία επερώτησης (query sequence) το προηγούμενο καλύτερο Blast hit. Όλες οι προηγούμενες παράμετροι του blast παραμένουν ίδιες.

- Βρίσκετε ως νέο καλύτερο blast hit το estrogen receptor alpha; Είναι ανταποδοτικό το blast; Τι σημαίνει αυτό για τις εξελικτικές σχέσεις μεταξύ των δύο ακολουθιών;

Blast (1b)

- Χρησιμοποιώντας ως ακολουθία επερώτησης το mRNA του estrogen receptor alpha από τον άνθρωπο (EMBL-bank accession: X03635), βρείτε αν υπάρχουν ομόλογες νουκλεοτιδικές ακολουθίες στη Drosophila melanogaster, χρησιμοποιώντας τη νουκλεοτιδική ΒΔ nucleotide collection (nr/nt).
- Ποιό πρόγραμμα του Blast πρέπει να χρησιμοποιήσετε;

- Παράμετροι του blast που θα κάνετε:
  - νουκλεοτιδική ΒΔ nucleotide collection (nr/nt)
  - Οργανισμό Drosophila melanogaster
  - Optimize for somewhat similar sequences
  - Expect threshold 1e-5
  - Filter low-complexity regions

- Βρέθηκαν ομόλογες νουκλεοτιδικές ακολουθίες στη Drosophila;
- Γιατί;

Blast (1c)

- Ποιό άλλο πρόγραμμα του Blast πρέπει να χρησιμοποιήσετε, για να δείτε αν υπάρχουν ομόλογες πρωτεΐνες για το mRNA σας, στη Drosophila melanogaster;
- Παράμετροι του Blast.
  - Genetic code standard
  - Database: non-redundant protein sequences (nr)
  - Οργανισμός: Drosophila melanogaster
  - Expectation threshold 1e-5
  - Low complexity regions filtering

- Τι βρίσκετε;

Άσκηση
Μέρος 2ο

Άσκηση (2)

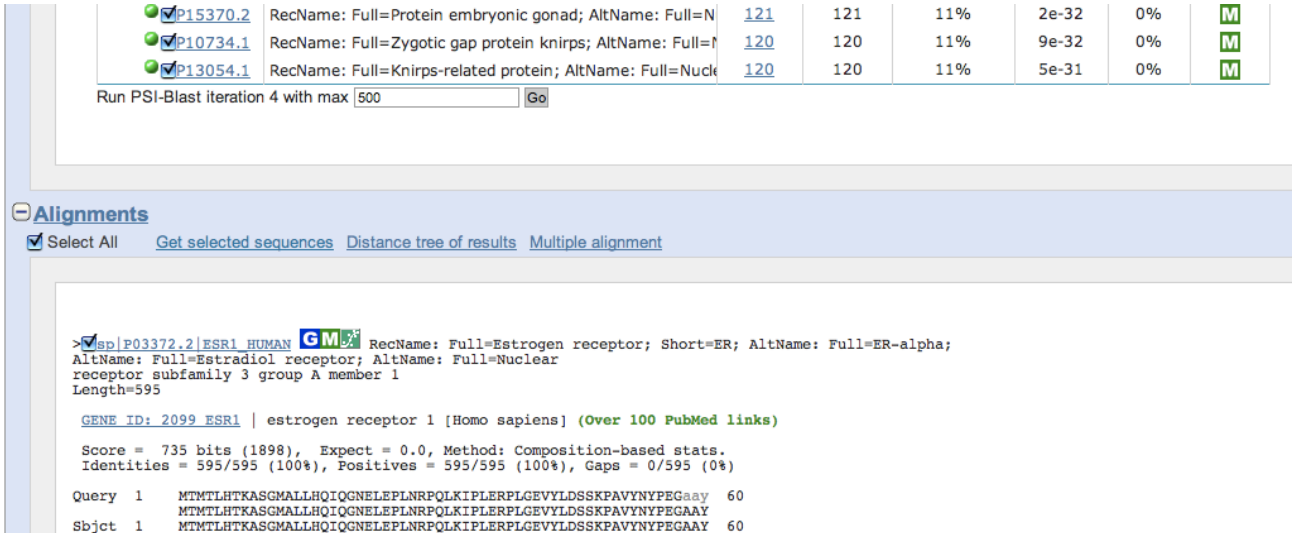
- Βρείτε την πρωτεϊνική ακολουθία του human estrogen receptor alpha (Uniprot id: P03372) σε μορφή FASTA.

- Με την ακολουθία αυτή (P03372), βρείτε τις ομόλογες πρωτεϊνικές ακολουθίες της, στη Drosophila melanogaster και στον άνθρωπο ταυτόχρονα, με τη βοήθεια του PSI-BLAST. Κάνετε το PSI-Blast στην ιστοσελίδα του NCBI, χρησιμοποιώντας την Swissprot, expectation value 1e-10 και low-complexity filtering. Επαναλάβετε τους κύκλους του PSI-blast μέχρι να συγκλίνει ο αλγόριθμος.

- Αποθηκεύεστε σε ένα αρχείο (με όνομα sequences.fasta) με μορφή FASTA τις ακολουθίες από την παραπάνω αναζήτηση.

Αποθήκευση ακολουθιών από το Blast

- Select all
- Get selected sequences



Αποθήκευση ακολουθιών από το Blast

- Send to ->
- File ->
- Format: FASTA ->
- Creat file