

Στοίχιση ακολουθιών κατά ζεύγη  
(Pairwise alignment)

&

Blast

# Στοιχίση κατά ζεύγη

- Αντιστοιχίση 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

# Στοιχίση 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
Sbjct 1 MKTPVSAAANLS NAGSSGA AIQI+PKTEPVGEEGPMSLDFQSPNL+TSTPNPNKRPG 60

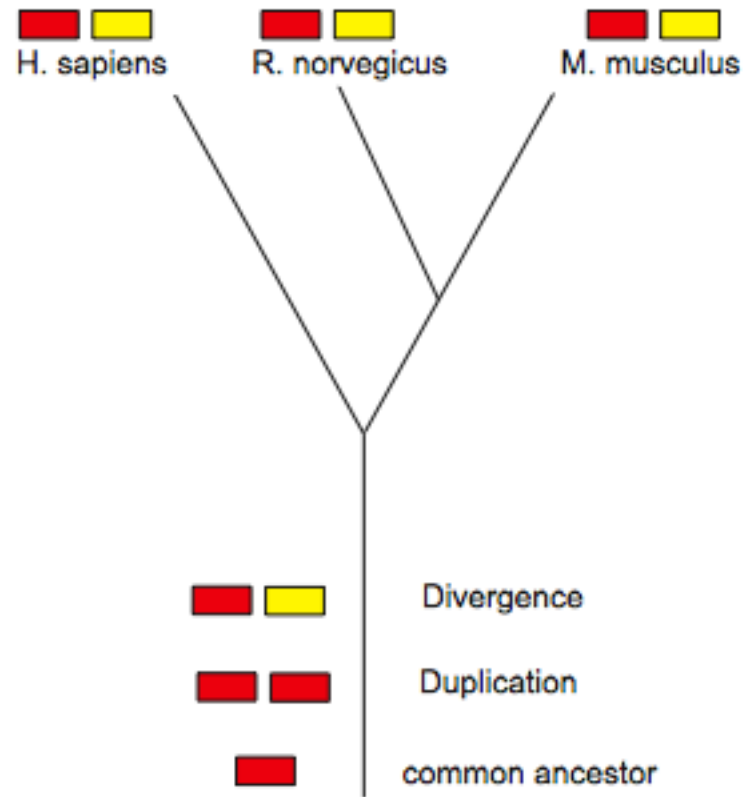
Query 61 SLDLNSKSAKNKRIFAPLVINSPDLSSKTVNTPDLEKILLSNNLMQTPQPGKVFPTKAGP 120
Sbjct 61 SLDLNSK AKNKRIFAPLVINSPDL +KTVNTPDLEKILLSNNL+QTPQPGKVFPTKAGP 120

Query 121 VTVEQLDFGRGFEEALHNLHTNSQAFPSANSAANNTTAAAMTAVNNGISGGTFTYT 180
Sbjct 121 VTVEQ DFGRGFEEAL NLHTNSQAFP A NS ANNTT AMTAVNNGISGGTFTY 175
```

# Λίγη εξέλιξη: ομολογία

- Ομόλογα γονίδια: κοινός εξελικτικός πρόγονος.
- Ορθόλογα γονίδια: προέρχονται από ειδογένεση. Ουσιαστικά, ένα γονίδιο  $\alpha$  (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)

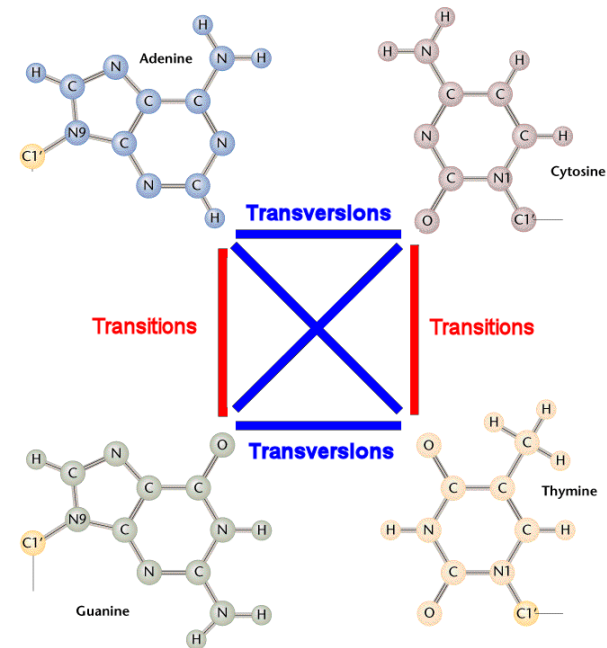


# Βασικότερα είδη μεταλλάξεων

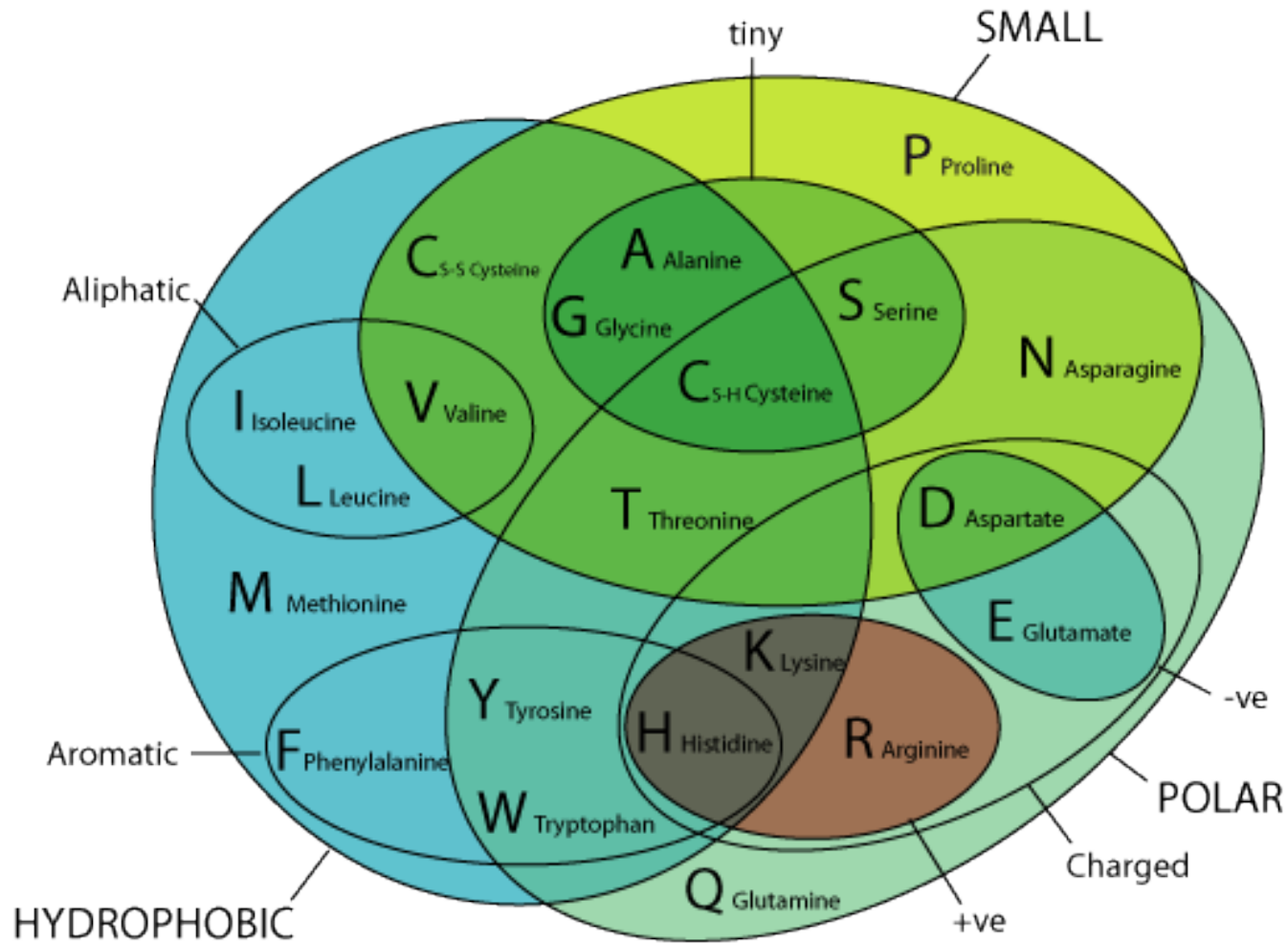
- Μεταλλάξεις σημείου (point mutations)
  - Συνώνυμες (synonymous)
  - Μη-συνώνυμες (non-synonymous)
    - Αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες
    - Αμινοξέα με διαφορετικές φυσικοχημικές ιδιότητες
  - Κωδικόνια τερματισμού

# Μεταπτώσεις-μεταστροφές

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



# Κατηγοριοποίηση αμινοξέων



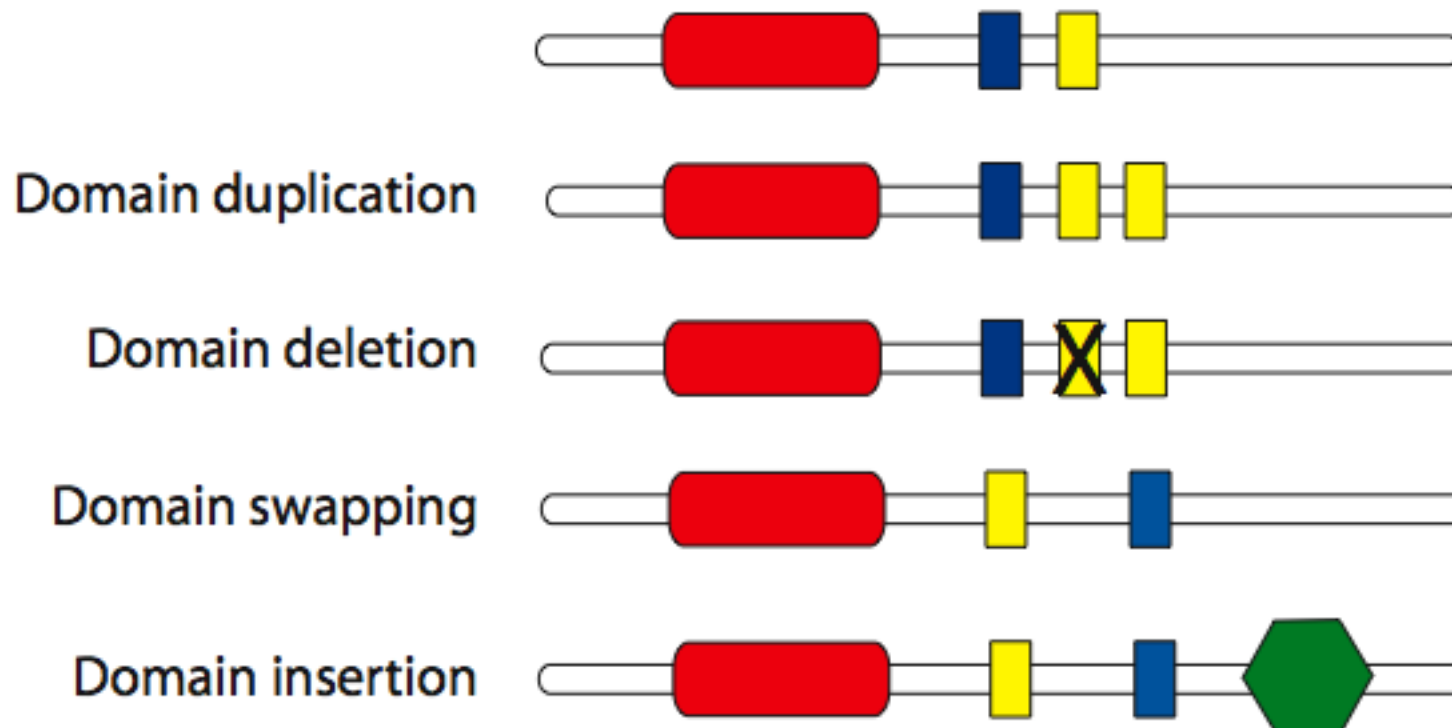


# Βασικότερα είδη μεταλλάξεων

- 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 μεταλλάσσονται 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.
  - Μεταλλάσσονται 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 (I)

- 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--

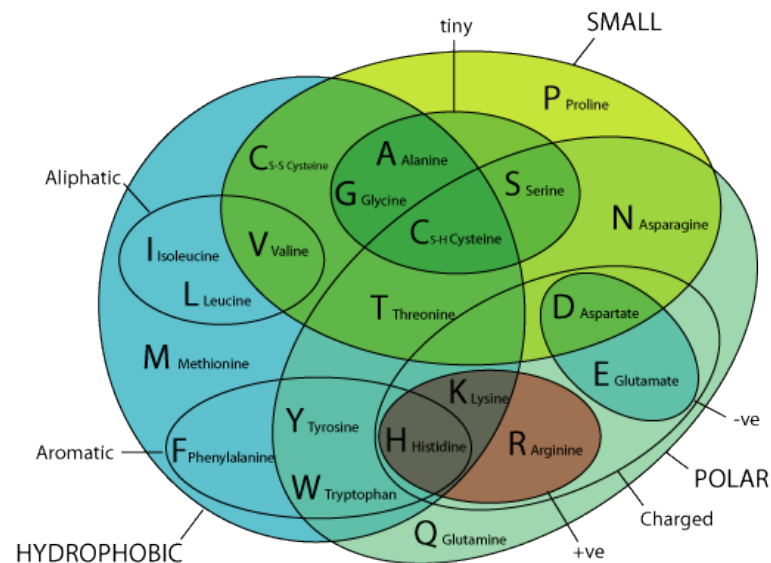
# Δυναμικός προγραμματισμός

- 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



<b>sequence 1</b>	<b>V</b>	<b>D</b>	<b>S</b>	<b>-</b>	<b>C</b>	<b>Y</b>
<b>sequence 2</b>	<b>V</b>	<b>E</b>	<b>S</b>	<b>L</b>	<b>C</b>	<b>Y</b>
<b>SCORE</b>	<b>4</b>	<b>2</b>	<b>4</b>	<b>-11</b>	<b>9</b>	<b>7</b>
<b>(26)</b>						

**SCORE = SUM OF AMINO ACID PAIR SCORES  
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 ΤΟΠΙΚΗ ΣΤΟΙΧΙΣΗ

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

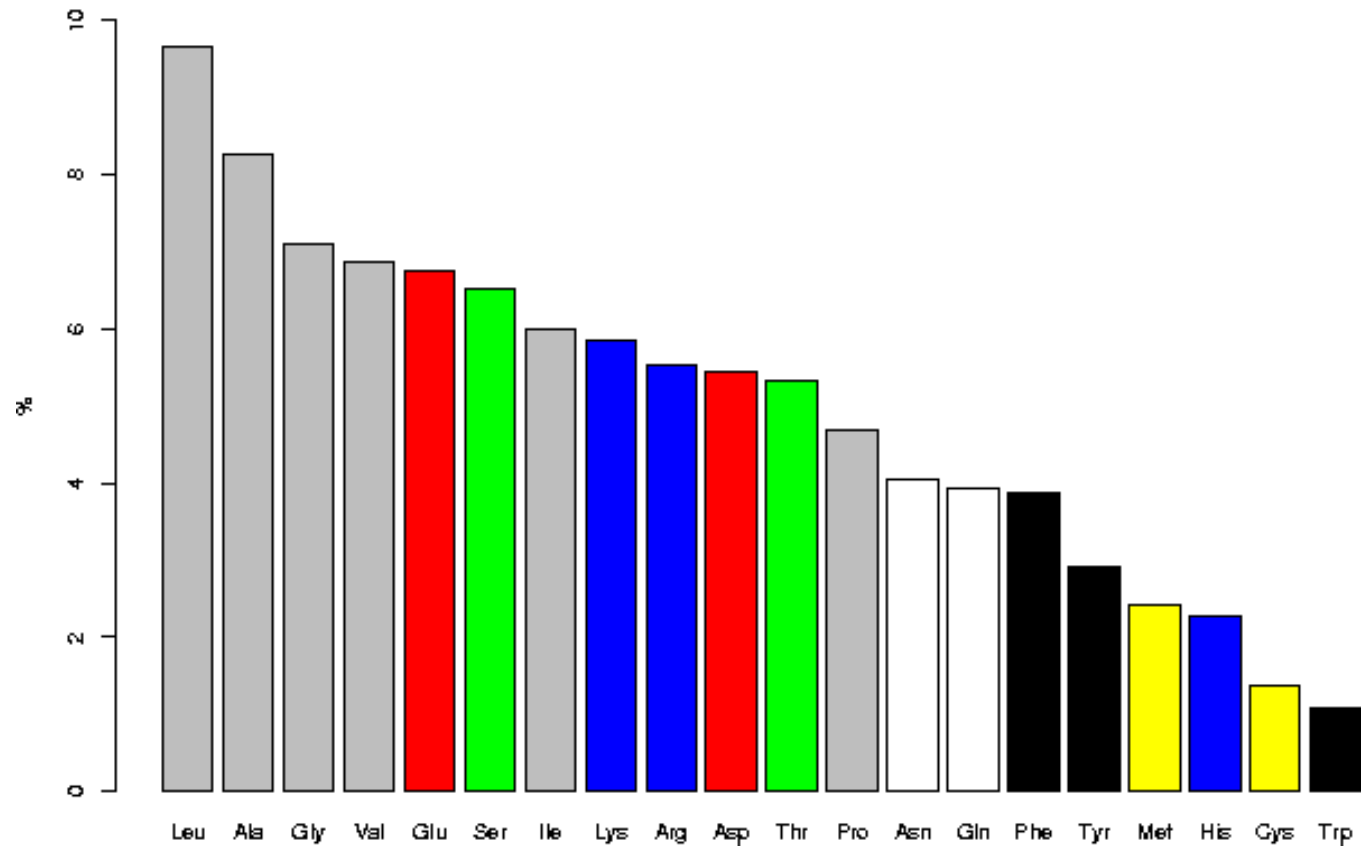
- Στο παράδειγμα, όλες οι συμφωνίες/ασυμφωνίες είχαν το ίδιο σκορ.
- Στην πράξη, πιο περίπλοκα συστήματα βαθμολόγησης. Μια ασυμφωνία μεταξύ δύο πουρινών δεν είναι το ίδιο με μια ασυμφωνία μεταξύ πουρίνης-πυριμιδίνης. Διαφορετικές συχνότητες μεταλλάξεων.
- Το ίδιο και για τις πρωτεΐνες.
- Χρειαζόμαστε πίνακες που βασίζονται σε συγκεκριμένα εξελικτικά μοντέλα και λαμβάνουν υπόψη 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

Amino acid composition







# 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 (iv)

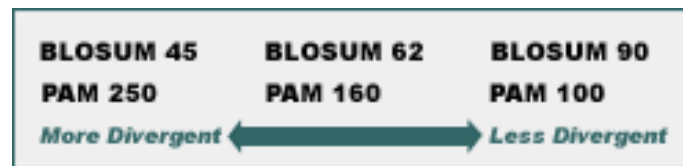
- Στις στοιχίσεις χρησιμοποιήθηκαν ακολουθίες που είχαν αποκλείει πολύ λίγο μεταξύ τους (απόσταση 1 PAM).
- Αναγωγή σε απόσταση 250 PAM (Πίνακας PAM250). Πολλαπλασιάστηκε ο PAM1 X 250 φορές με τον εαυτό του
- Σειρά πινάκων. Εμπειρικά προτάθηκε για γενική χρήση ο PAM250
- Όσο μεγαλώνει το νούμερο, μεγαλώνει 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 BLOSUM

- BLOcks SUbstitution Matrix
- Henikoff & Henikoff, 1992.
- Χρησιμοποίησαν τοπικές πολλαπλές στοιχίσεις από συντηρημένες περιοχές εξελικτικά απομακρυσμένων ακολουθιών (B.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.



# Βαθμολόγηση 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 επέκτασή του δεν είναι τόσο σπάνια
  - Π.χ. Για BLOSUM62: εισαγωγή κενού -> Ποινή 10-15. Επέκταση κενού -> ποινή 1-2

# Βαθμολόγηση 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 MINUS SINGLE GAP PENALTY (11) = 15
(26)							

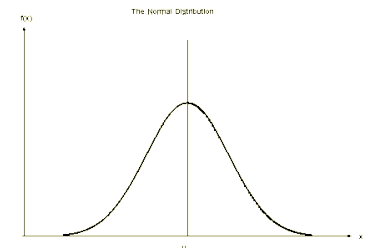
**Figure 3.7.** Example of scoring a sequence alignment with a gap penalty. The individual alignment scores are taken from an amino acid substitution matrix.





# Στατιστική σημαντικότητα 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

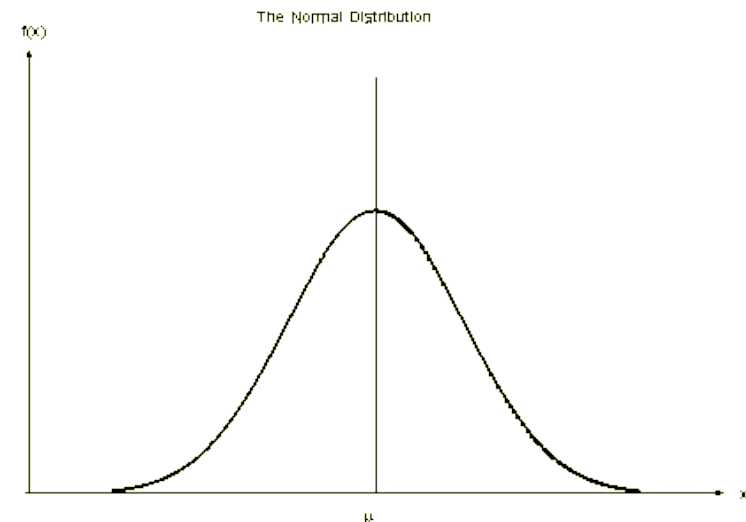
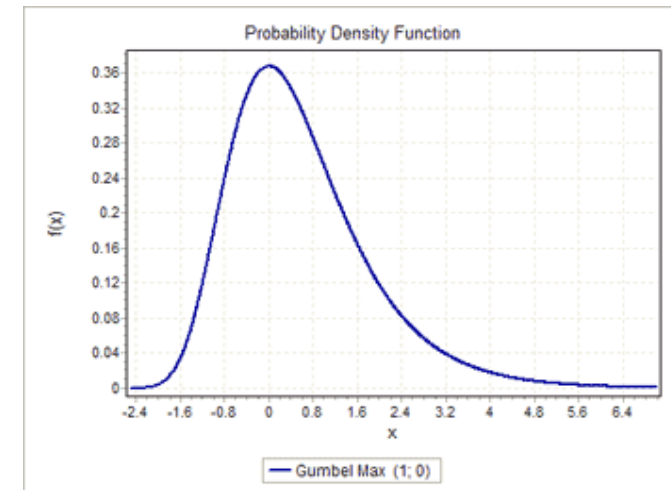


# Στατιστική σημαντικότητα ολικής στοίχισης (ii)

- 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  $P\text{-value} < 0.01$
- 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 στοίχισεις 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 στοίχισεις



# Στατιστική σημαντικότητα τοπικής στοίχισης (iii)

- Για μια δεδομένη τοπική στοίχιση (χωρίς κενά) δύο ακολουθιών με score  $S$ , πόσες τυχαίες στοίχισεις θα μπορούσαν να δώσουν το ίδιο score;
- $E = Kmne^{-\lambda S}$  (E-value)
- $m, n$  μήκη των ακολουθιών
- $S$  score στοίχισης
- $K, \lambda$  εξαρτώνται από τη συχνότητα νουκλεοτιδίων/αμινοξέων και το σύστημα βαθμολόγησης.
- 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 X B.Δ
- N ακολουθίες X B.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 A 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 υπόψη 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

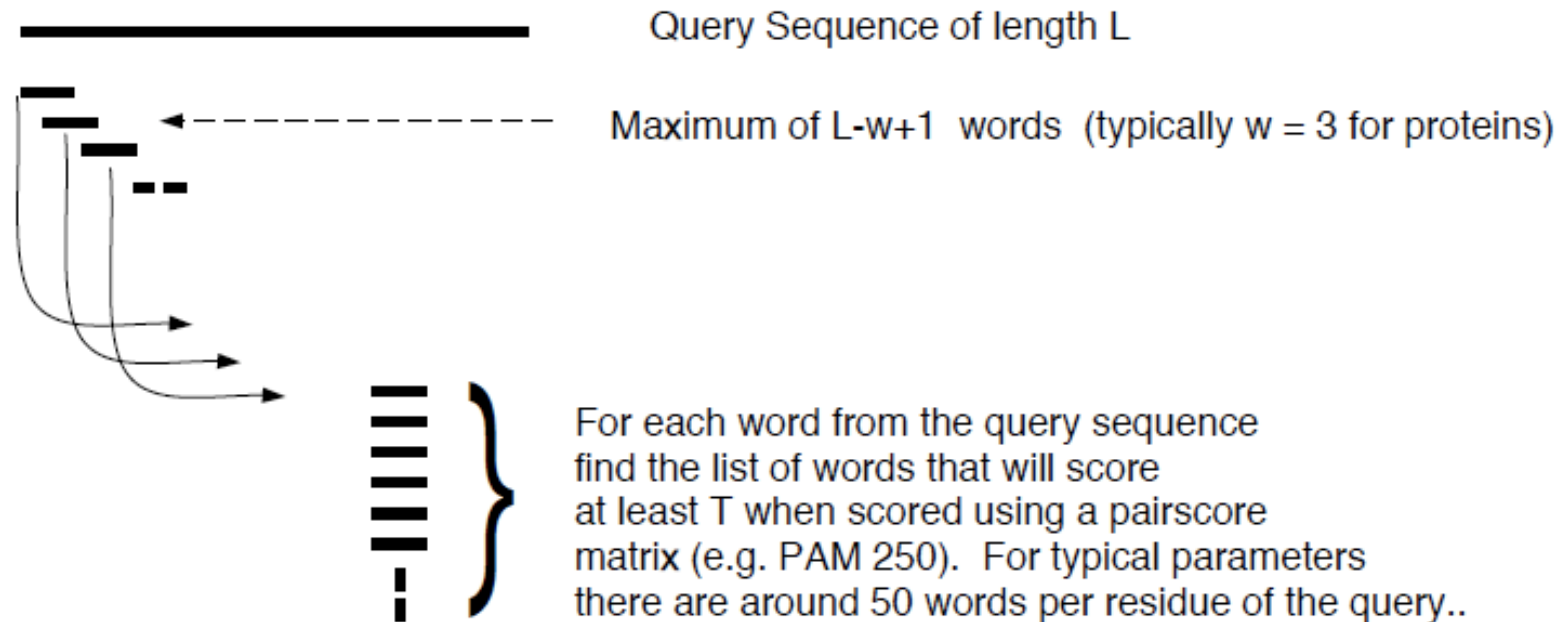
<http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=comgen&part=blast>

- 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

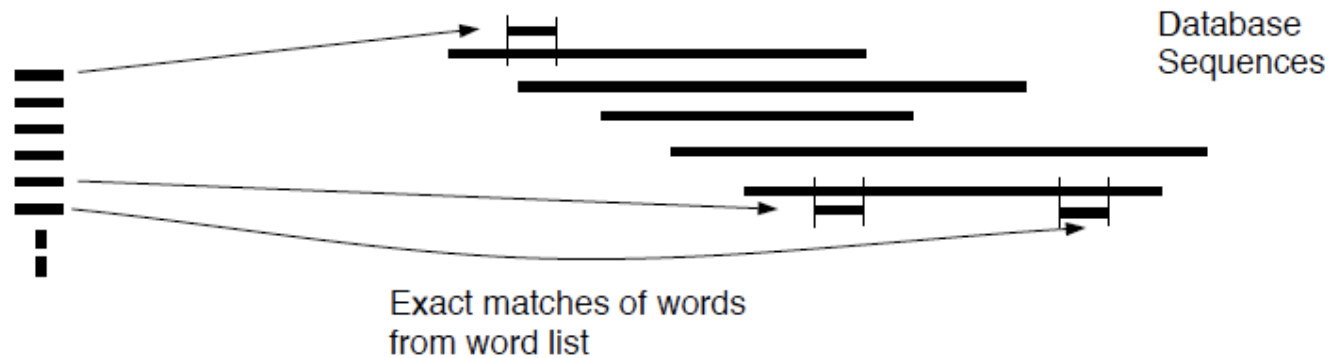
(1) For the query find the list of high scoring words of length  $w$ .



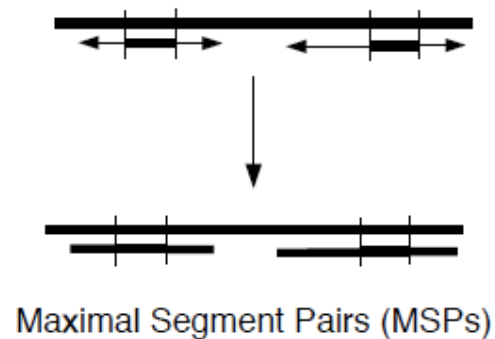
- PQG
- $20 \times 20 \times 20 = 8.000$  words
- PQG  $\times$  8.000 words
- PQG  $\times$  PEG =  $7 + 2 + 6 = 15$
- Όριο τιμής  $T$

# Αλγόριθμος BLAST

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

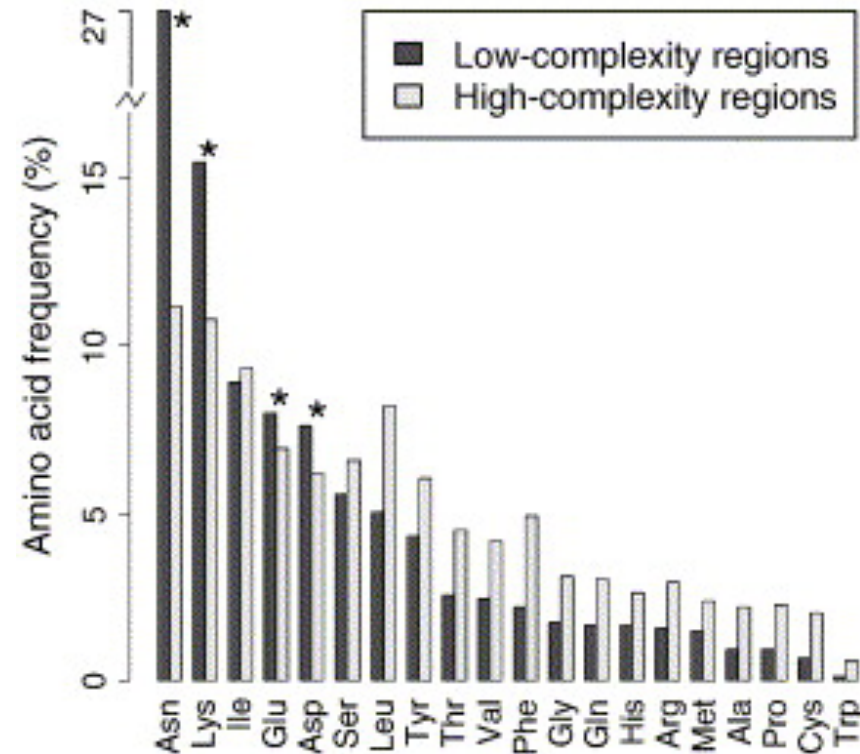


- (3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold  $S$ .



# Περιοχές χαμηλής πολυπλοκότητας (i)

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



# Περιοχές χαμηλής πολυπλοκότητας (ii)

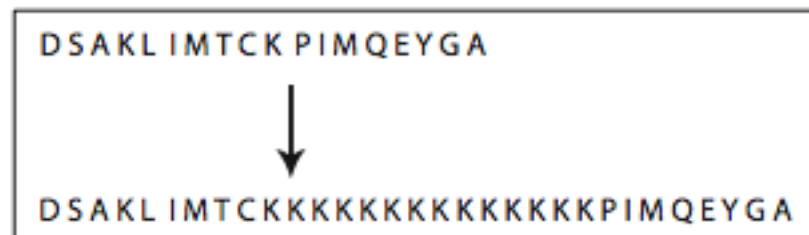
- 2 μη ομόλογες ακολουθίες.

PEGADINDAKK LINEDQPR  
DSAKL IMTCK PIMQEYGA

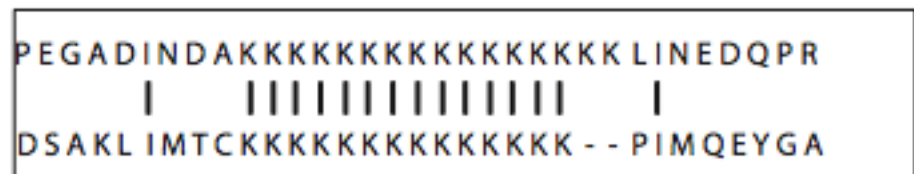
- Μεταλλάξεις στην ακολουθία 1.



- 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

```
PEGADINDAKKKKKKKKKKKKKKKKKKKKKK LINEDQPR
      |  | | | | | | | | | | | | | | | | | | | |
DSAKL IMTCCKKKKKKKKKKKKKKKKKKK - - PIMQEYGA
```

```
PEGADINDAXXXXXXXXXXXXXXXXXXXXXX LINEDQPR
DSAKL IMTCXXXXXXXXXXXXXXXXXXXXX - - PIMQEYGA
```

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

# Blast

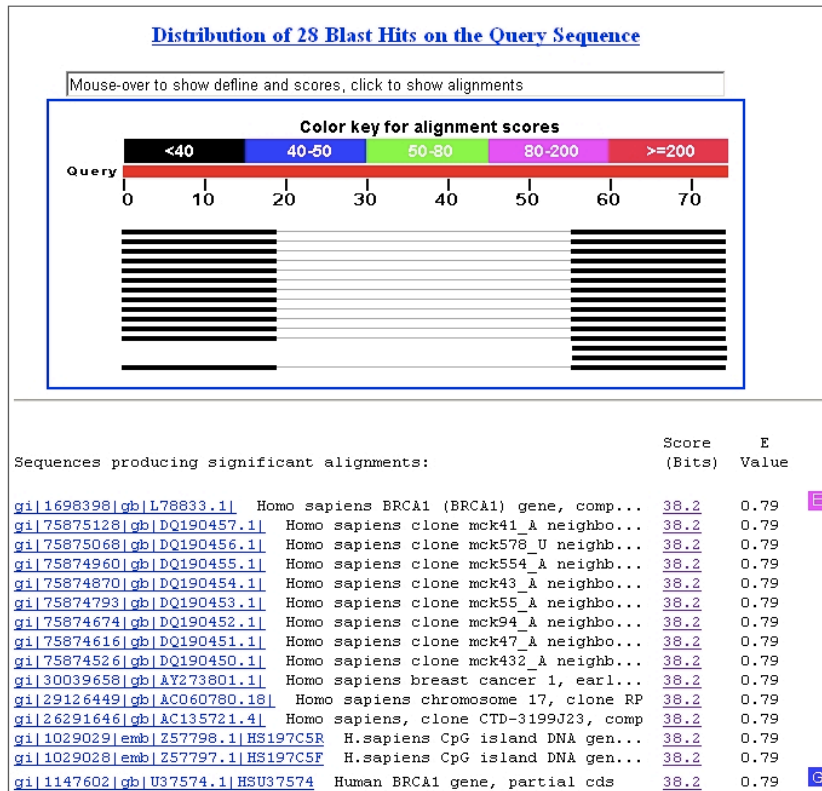
Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping
BLASTP	Protein	Protein	Identifying common 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 X νουκλεοτίδια
  - Για στοίχιση tRNA, rRNA, mRNA, γενωμικό DNA
- 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)



```
> gi|1698398|gb|L78833.1 E D 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      GTACCTTGATTTTCGTATTC 19
            |||
Sbjct 3252  GTACCTTGATTTTCGTATTC 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

<input type="checkbox"/> Query	5	CCTCMAATCACTCTTTGGCAACGACCCCTCGTCACAATAAAAGATAGGGGGGCAACTAAAG	64
<input type="checkbox"/> <a href="#">23380210</a>	1	.....	60
<input type="checkbox"/> <a href="#">23380202</a>	1	.....	60
<input type="checkbox"/> <a href="#">15150145</a>	1	.....	60
<input type="checkbox"/> <a href="#">7638172</a>	1	.....	60
<input type="checkbox"/> <a href="#">7638170</a>	1	.....	60
<input type="checkbox"/> <a href="#">7638168</a>	1	.....	60
<input type="checkbox"/> <a href="#">23380208</a>	1	.....	60
<input type="checkbox"/> <a href="#">23380206</a>	1	.....G.....	60
<input type="checkbox"/> <a href="#">23380204</a>	1	.....	60
<input type="checkbox"/> <a href="#">23380200</a>	1	.....	60
<input type="checkbox"/> <a href="#">15150149</a>	1	.....	60
<input type="checkbox"/> <a href="#">15150147</a>	1	.....G.....	60
<input type="checkbox"/> <a href="#">51703160</a>	1	.....	60
<input type="checkbox"/> <a href="#">51703042</a>	1	.....	60
<input type="checkbox"/> <a href="#">44887180</a>	1	.....G.....	60
<input type="checkbox"/> <a href="#">13738955</a>	1	.....	60
<input type="checkbox"/> <a href="#">7682537</a>	19	.....G.....	78
<input type="checkbox"/> <a href="#">51012122</a>	1	.....G.....	60
<input type="checkbox"/> <a href="#">6019233</a>	1	.....G.....	60
<input type="checkbox"/> <a href="#">37220926</a>	183	.....	242
<input type="checkbox"/> <a href="#">63080064</a>	1	.....	60
<input type="checkbox"/> <a href="#">9943154</a>	1	.....A.....	60
<input type="checkbox"/> <a href="#">9935201</a>	1	.....	60
<input type="checkbox"/> <a href="#">6446433</a>	19	.....G.....A.....	78
<input type="checkbox"/> <a href="#">3098582</a>	1806	.....G.....	1865

# Blastp

- Πρωτεΐνη X 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 συγκεκριμένο γενετικό κώδικα
  - ακολουθίας επερώτησης (query sequence)
  - ακολουθιών στην Β.Δ.
  - και των δύο ταυτόχρονα

# tblastn

- Πρωτεΐνη (query) X 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-KSTNKTGTITLPCRIQ 444
                TIAFNQSSGGDPEIVMHSFNCGGEFFYCNTTQLFNSTWPTN KSTNKTGTITLPCRIQ
Sbjct 90    TIAFNQSSGGDPEIVMHSFNCGGEFFYCNTTQLFNSTWPTNNTKSTNKTGTITLPCRIQ 149

Query 445  IINRWQEVGKAMYAPPIKQIRCSSNITGIFLTRDGGNASDETETFRPGGGN 600
                IINRWQEVGKAMYAPPIKQIRCSSNITGIFLTRDGGNASDETETFRPGGGN
Sbjct 150  IINRWQEVGKAMYAPPIKQIRCSSNITGIFLTRDGGNASDETETFRPGGGN 201

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

Query 2    EEDIVIRSENFMTNAKTIIIVQLKESIKINCTRPMMNTRKSIPIATGGAIYATGDIIGDIR 181
                EEDIVIRSENFMTNAKTIIIVQLKESIKINCTRPMMNTRKSIPIATGGAIYATGDIIGDIR
Sbjct 1    EEDIVIRSENFMTNAKTIIIVQLKESIKINCTRPMMNTRKSIPIATGGAIYATGDIIGDIR 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
- 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**

© Springer-Verlag New York Inc. 2001

---

*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 A, στην θέση 123 (ενεργό κέντρο ενζύμου) βλέπουμε ένα μόνο αμινοξύ.
- Σε μια πολλαπλή στοίχιση της A 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<sup>a</sup>

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<sup>a</sup>

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<sup>a</sup>

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 υπόψη 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 ( $B = 0.1$ )<sup>c,d</sup>

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 στοίχισεις που βρέθηκαν ( $E\text{-value} < \text{cutoff}$ ) 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 στοίχισεις
  - Απαλοιφή ακολουθιών με πολύ μεγάλη ομοιότητα.
  - Δημιουργία PSSM.

# PSI-Blast

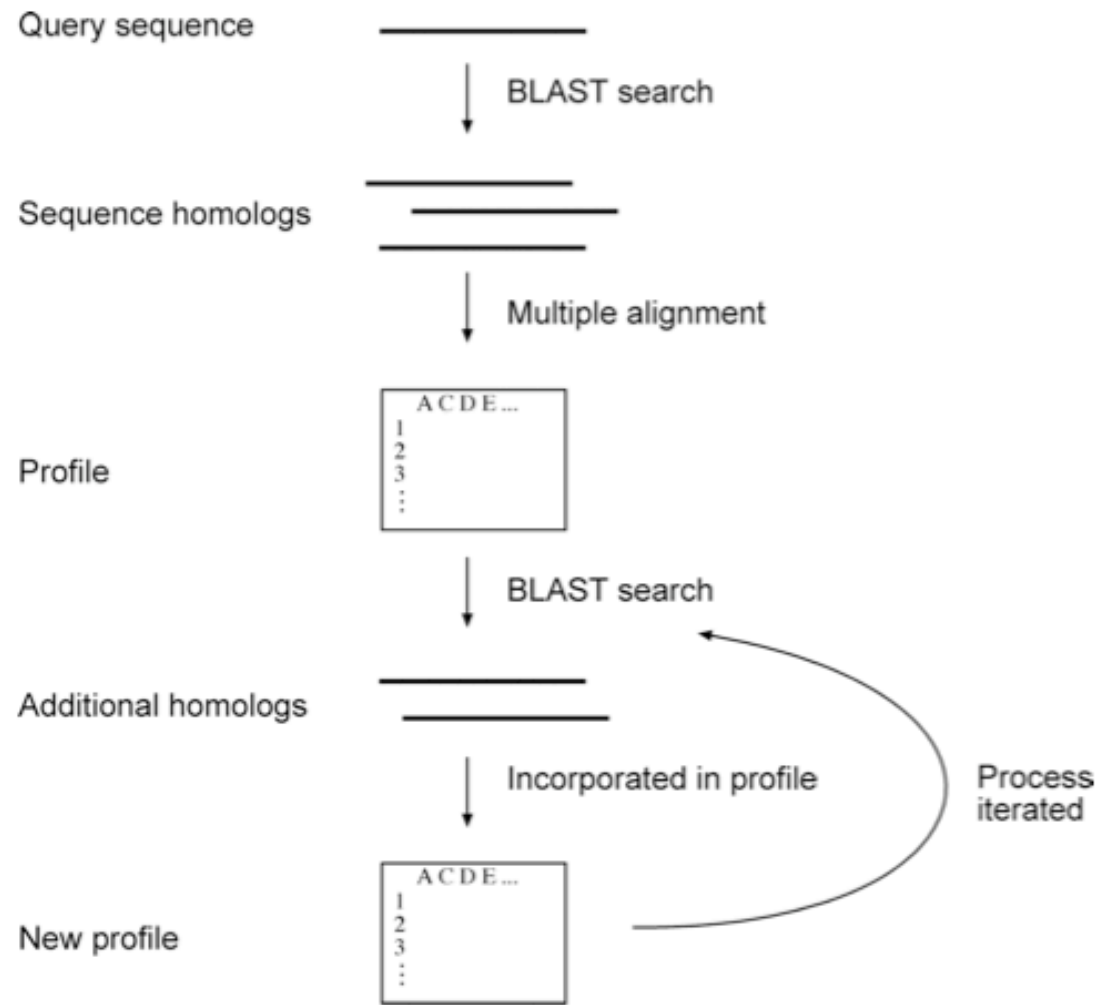
- Δεύτερο στάδιο:
  - Νέα αναζήτηση στη Β.Δ. με το PSSM αντί της αρχικής ακολουθίας επερώτησης.
  - Οι νέες ακολουθίες που βρέθηκαν και ξεπερνούν το κατώφλι E-value ανανεώνουν την πολλαπλή στοίχιση και δημιουργείται ένα νέο PSSM.
- Η διαδικασία επαναλαμβάνεται μέχρι να μη βρεθούν νέες ακολουθίες με Evalue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

# PSI-Blast

<b>a</b> <u>Accession</u>	<u>Alignment</u>	<u>E-value</u>
P49789		
P49779		8e-27
P49775		6e-18
Q11066		3e-07
Q09344		4e-05
P49378		0.001
P32084		0.002



# PSI-Blast



# 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 ( $\pi$ )	Most frequent bigram partner	No. of occurrences
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.

**Family: *zf-C2H2* (PF00096)**

238 architectures   31268 sequences   2 interactions   728 species   133 structures

**Summary**

**Domain organisation**

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

**There are 3344 sequences with the following architecture: *zf-C2H2***  
[ADR1\\_YEAST](#) [Saccharomyces cerevisiae (Baker's yeast)] Regulatory protein ADR1 (1323 residues)  
  
[Show](#) all sequences with this architecture.

**There are 1911 sequences with the following architecture: *zf-C2H2 x 2***  
[AEF1\\_DROME](#) [Drosophila melanogaster (Fruit fly)] Adult enhancer factor 1 (308 residues)  
  
[Show](#) all sequences with this architecture.

**There are 638 sequences with the following architecture: *zf-C2H2 x 3***  
[ODD\\_DROME](#) [Drosophila melanogaster (Fruit fly)] Protein odd-skipped (392 residues)  
  
[Show](#) all sequences with this architecture.

**There are 485 sequences with the following architecture: *zf-AD, zf-C2H2***  
[ZN276\\_HUMAN](#) [Homo sapiens (Human)] Zinc finger protein 276 (614 residues)  
  
[Show](#) all sequences with this architecture.

**There are 388 sequences with the following architecture: *zf-C2H2 x 4***  
[ESCA\\_DROME](#) [Drosophila melanogaster (Fruit fly)] Protein escargot (470 residues)  
  
[Show](#) all sequences with this architecture.

**There are 262 sequences with the following architecture: *zf-C2H2 x 5***  
[CF2\\_DROME](#) [Drosophila melanogaster (Fruit fly)] Chorion transcription factor Cf2 (510 residues)  
  
[Show](#) all sequences with this architecture.

**There are 239 sequences with the following architecture: *zf-C2H2 x 6***  
[Q9H7U2\\_HUMAN](#) [Homo sapiens (Human)] cDNA FLJ14260 fis, clone PLACE1001118, weakly similar to ZINC FINGER PROTEIN 135 (262 residues)  
  
[Show](#) all sequences with this architecture.

**Jump to...**

# Χρησιμοποιώντας το PSI-Blast

blastn blastp **blastx** tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

### Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

```
SKKNSLALSILTADQMVSALLDAEPPILYSEYDPTTRPFSEASMMGLLTNLADRELVHMINW
AKRVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSMHPGKLLFAPNLLDRNQGKCEVG
MVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLD
KITDTLIHLMAKAGLTQQQHQRQAQLLLSHIRHMSNKGMEHLYSMKCKNVVPLYDLL
LEMLDAHRLHAPTSRGGASVEETDQSHLATAGSTSSHSLQKYYITGEAEGFPATV
```

Query subrange [From](#)   
[To](#)

Or, upload file  no file selected

Job Title   
Enter a descriptive title for your BLAST search

Align two or more sequences

### Choose Search Set

Database

Organism Optional   Exclude   
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional  Models (XM/XP)  Uncultured/environmental sample sequences

Entrez Query Optional   
Enter an Entrez query to limit search

### Program Selection

Algorithm  blastp (protein-protein BLAST)  PSI-BLAST (Position-Specific Iterated BLAST)  PHI-BLAST (Pattern Hit Initiated BLAST)  
Choose a BLAST algorithm

Search database **Swissprot protein sequences(swissprot)** using **PSI-BLAST (Position-Specific Iterated BLAST)**  
 Show results in a new window

[Algorithm parameters](#) **Note: Parameter values that differ from the default are highlighted in yellow and marked with**

# Χρησιμοποιώντας το PSI-Blast

Algorithm parameters Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

### General Parameters

Max target sequences: 500  
Select the maximum number of aligned sequences to display

Short queries:  Automatically adjust parameters for short input sequences

Expect threshold: ♦ 1e-3

Word size: 3

Max matches in a query range: 0

### Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

### Filters and Masking

Filter: ♦  Low complexity regions

Mask:  Mask for lookup table only  
 Mask lower case letters

### PSI/PHI BLAST

Upload PSSM: Choose File no file selected

PSI-BLAST Threshold: 1e-3

Pseudocount: 0

**BLAST** Search database **Swissprot protein sequences(swissprot)** using **PSI-BLAST (Position-Specific Iterated BLAST)**  
 Show results in a new window

# Χρησιμοποιώντας το PSI-Blast

NCBI/BLAST/blastp suite/ Formatting Results - CX8ZUS47011

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

**PSI blast Iteration 1**

**sp|P03372|ESR1\_HUMAN Estrogen receptor OS=Homo...**

<b>Query ID</b>	lc 74714	<b>Database Name</b>	swissprot
<b>Description</b>	sp P03372 ESR1_HUMAN Estrogen receptor OS=Homo sapiens GN=ESR1 PE=1 SV=2	<b>Description</b>	Non-redundant SwissProt sequences
<b>Molecule type</b>	amino acid	<b>Program</b>	BLASTP 2.2.24+ <a href="#">Citation</a>
<b>Query Length</b>	595		

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Multiple alignment](#)

**Graphic Summary**

[Show Conserved Domains](#)

Putative conserved domains have been detected, click on the image below for detailed results.

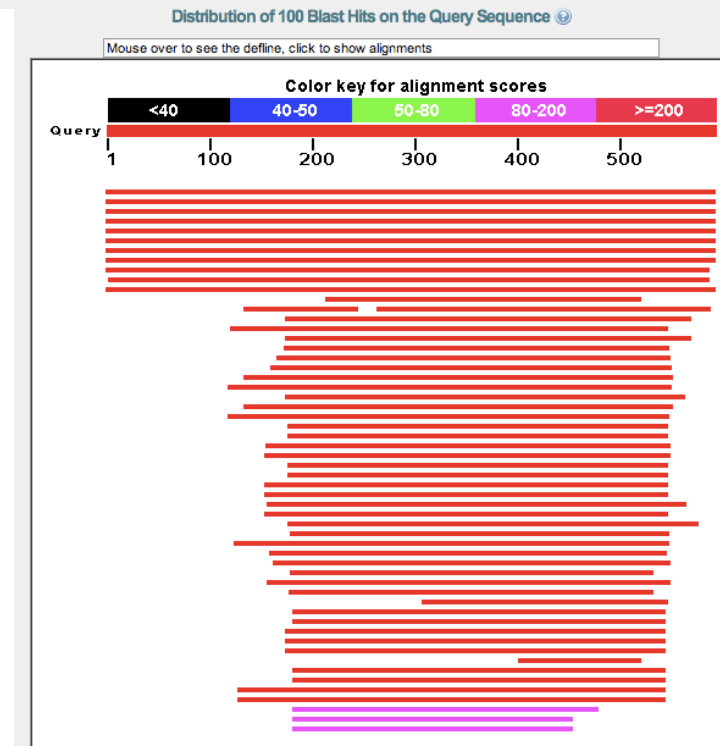
Query seq. 1 100 200 300 400 500 595

zinc binding site  
DNA binding site  
dimer interface  
NR\_DBD\_ER  
NR\_DBD\_like super

ligand binding site  
coactivator recognition site  
dimer interface  
NR\_LBD\_ER

Specific hits  
Superfamilies

Oest\_recep superfamily  
NR\_DBD\_like super  
NR\_LBD superfamily



# Χρησιμοποιώντας το PSI-Blast

▼ **Descriptions**

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [P](#) 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

▼ **Sequences producing significant alignments with E-value BETTER than threshold**

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



# Χρησιμοποιώντας το PSI-Blast

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

<input checked="" type="checkbox"/>	<a href="#">O16662.3</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">95.2</a>	95.2	33%	1e-18		
<input checked="" type="checkbox"/>	<a href="#">P41933.1</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">93.7</a>	93.7	21%	4e-18		
NEW	<input checked="" type="checkbox"/>	<a href="#">O57568.1</a>	RecName: Full=Thyroid hormone receptor alpha; AltName:	<a href="#">85.2</a>	85.2	21%	1e-15	
NEW	<input checked="" type="checkbox"/>	<a href="#">O17573.1</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">80.9</a>	80.9	49%	2e-14	
NEW	<input checked="" type="checkbox"/>	<a href="#">P79404.1</a>	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	<a href="#">80.6</a>	80.6	19%	3e-14	
NEW	<input checked="" type="checkbox"/>	<a href="#">O15466.2</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">78.3</a>	78.3	43%	2e-13	
<input checked="" type="checkbox"/>	<a href="#">P35547.1</a>	RecName: Full=Glucocorticoid receptor; Short=GR; AltNam	<a href="#">76.7</a>	76.7	11%	5e-13		
NEW	<input checked="" type="checkbox"/>	<a href="#">O9TXJ1.2</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">74.8</a>	74.8	11%	2e-12	
NEW	<input checked="" type="checkbox"/>	<a href="#">P70503.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">73.2</a>	73.2	34%	4e-12	
NEW	<input checked="" type="checkbox"/>	<a href="#">O9BG94.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">70.9</a>	70.9	33%	2e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">O61066.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">70.9</a>	70.9	43%	2e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">P79386.2</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">70.5</a>	70.5	38%	3e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">O8QHI2.1</a>	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	<a href="#">70.2</a>	70.2	17%	4e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">O9BG97.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">69.8</a>	69.8	33%	5e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">P51843.2</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">69.8</a>	69.8	33%	6e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">O9BG93.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">69.8</a>	69.8	33%	6e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">O17025.2</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">69.4</a>	69.4	11%	7e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">P97947.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">69.4</a>	69.4	37%	7e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">O16360.3</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">69.4</a>	69.4	33%	7e-11	
NEW	<input checked="" type="checkbox"/>	<a href="#">O9BG96.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">67.5</a>	67.5	33%	3e-10	
NEW	<input checked="" type="checkbox"/>	<a href="#">O02305.2</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">66.3</a>	66.3	45%	6e-10	
NEW	<input checked="" type="checkbox"/>	<a href="#">O17934.1</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">65.9</a>	65.9	52%	8e-10	
NEW	<input checked="" type="checkbox"/>	<a href="#">O62227.1</a>	RecName: Full=Nuclear receptor subfamily 0 group B mem	<a href="#">65.2</a>	65.2	33%	1e-09	
NEW	<input checked="" type="checkbox"/>	<a href="#">O45907.1</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">55.9</a>	55.9	34%	7e-07	
NEW	<input type="checkbox"/>	<a href="#">P20659.4</a>	RecName: Full=Histone-lysine N-methyltransferase trithora	<a href="#">55.1</a>	55.1	18%	1e-06	
NEW	<input type="checkbox"/>	<a href="#">O24742.1</a>	RecName: Full=Histone-lysine N-methyltransferase trithora	<a href="#">55.1</a>	55.1	19%	1e-06	
NEW	<input checked="" type="checkbox"/>	<a href="#">O16354.1</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">53.6</a>	53.6	33%	4e-06	
NEW	<input checked="" type="checkbox"/>	<a href="#">O9PUA8.1</a>	RecName: Full=Thyroid hormone receptor alpha; AltName:	<a href="#">45.9</a>	45.9	3%	9e-04	


Run PSI-Blast iteration 3 with max

# Χρησιμοποιώντας το PSI-Blast

**PSI blast Iteration 3**

**sp|P03372|ESR1\_HUMAN Estrogen receptor OS=Homo...**

<b>Query ID</b>	lc 59255	<b>Database Name</b>	swissprot
<b>Description</b>	sp P03372 ESR1_HUMAN Estrogen receptor OS=Homo sapiens GN=ESR1 PE=1 SV=2	<b>Description</b>	Non-redundant SwissProt sequences
<b>Molecule type</b>	amino acid	<b>Program</b>	BLASTP 2.2.24+ <a href="#">▶ Citation</a>
<b>Query Length</b>	595		

 **No new sequences were found above the 0.001 threshold**

# Χρησιμοποιώντας το PSI-Blast

- Αν περιλαμβάνονταν οι 2 μεθυλ-τρανσφεράσες...

<input checked="" type="checkbox"/>	<a href="#">Q24742.1</a>	RecName: Full=Histone-lysine N-methyltransferase trithora	<a href="#">85.2</a>	85.2	19%	1e-15		
<input checked="" type="checkbox"/>	<a href="#">P79404.1</a>	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	<a href="#">84.4</a>	84.4	19%	2e-15		
<input checked="" type="checkbox"/>	<a href="#">O57568.1</a>	RecName: Full=Thyroid hormone receptor alpha; AltName:	<a href="#">84.4</a>	84.4	21%	2e-15		
<input checked="" type="checkbox"/>	<a href="#">P35547.1</a>	RecName: Full=Glucocorticoid receptor; Short=GR; AltNam	<a href="#">79.4</a>	79.4	11%	7e-14		
<input checked="" type="checkbox"/>	<a href="#">Q9TXJ1.2</a>	RecName: Full=Nuclear hormone receptor family member	<a href="#">79.4</a>	79.4	54%	7e-14		
<input checked="" type="checkbox"/>	<a href="#">O8QHI2.1</a>	RecName: Full=Mineralocorticoid receptor; Short=MR; AltN	<a href="#">76.0</a>	76.0	17%	8e-13		
<input checked="" type="checkbox"/>	<a href="#">P20659.4</a>	RecName: Full=Histone-lysine N-methyltransferase trithora	<a href="#">74.4</a>	74.4	18%	2e-12		
<input checked="" type="checkbox"/>	<a href="#">Q9PUA8.1</a>	RecName: Full=Thyroid hormone receptor alpha; AltName:	<a href="#">47.1</a>	47.1	11%	4e-04		
	<input checked="" type="checkbox"/>	<a href="#">P16356.2</a>	RecName: Full=DNA-directed RNA polymerase II subunit R	<a href="#">45.5</a>	45.5	19%	0.001	

Run PSI-Blast iteration 4 with max


<input checked="" type="checkbox"/>	<a href="#">P20659.4</a>	RecName: Full=Histone-lysine N-methyltransferase trithora	<a href="#">71.0</a>	71.0	18%	2e-11		
<input checked="" type="checkbox"/>	<a href="#">P16356.2</a>	RecName: Full=DNA-directed RNA polymerase II subunit R	<a href="#">66.7</a>	66.7	19%	4e-10		
	<input checked="" type="checkbox"/>	<a href="#">P35074.2</a>	RecName: Full=DNA-directed RNA polymerase II subunit R	<a href="#">61.7</a>	61.7	19%	1e-08	
	<input checked="" type="checkbox"/>	<a href="#">P24928.2</a>	RecName: Full=DNA-directed RNA polymerase II subunit R	<a href="#">50.6</a>	50.6	18%	3e-05	
	<input checked="" type="checkbox"/>	<a href="#">P08775.3</a>	RecName: Full=DNA-directed RNA polymerase II subunit R	<a href="#">50.6</a>	50.6	18%	3e-05	
<input checked="" type="checkbox"/>	<a href="#">Q9PUA8.1</a>	RecName: Full=Thyroid hormone receptor alpha; AltName:	<a href="#">46.3</a>	46.3	11%	6e-04		

Run PSI-Blast iteration 5 with max

# Χρησιμοποιώντας το PSI-Blast

- Αποθήκευση αποτελεσμάτων

[Edit and Resubmit](#) [Save Search Strategies](#) [▶ Formatting options](#) [▽ Download](#)

Download						
<b>Alignment</b>				<b>Search Strategies</b>	<b>Bloseq</b>	<b>PssmWithParameters</b>
<a href="#">Text</a>	<a href="#">XML</a>	<a href="#">ASN.1</a>	<a href="#">Hit Table(text)</a>	<a href="#">Hit Table(csv)</a>	<a href="#">ASN.1</a>	<a href="#">ASN.1</a> 

# Άσκηση

## Μέρος 1ο

# Blast (1a)

- Βρείτε την ακολουθία του Estrogen receptor alpha (σε μορφή FASTA) ως:
  - mRNA από την EMBL bank (accession number: X03635).
  - ως πρωτεΐνη από την Uniprot (accession 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 estrogen receptor alpha (πρωτεΐνη) στη μύγα *Drosophila melanogaster*, χρησιμοποιώντας ως δόλωμα την ανθρώπινη πρωτεΐνη estrogen receptor alpha (βρείτε την ακολουθία της ανθρώπινης πρωτεΐνης από 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 (από το website του NCBI), 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) στο NCBI.
- 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.
- Αποθηκεύστε 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

<input checked="" type="checkbox"/>	<a href="#">P15370.2</a>	RecName: Full=Protein embryonic gonad; AltName: Full=N	<a href="#">121</a>	121	11%	2e-32	0%	
<input checked="" type="checkbox"/>	<a href="#">P10734.1</a>	RecName: Full=Zygotic gap protein knirps; AltName: Full=n	<a href="#">120</a>	120	11%	9e-32	0%	
<input checked="" type="checkbox"/>	<a href="#">P13054.1</a>	RecName: Full=Knirps-related protein; AltName: Full=Nuck	<a href="#">120</a>	120	11%	5e-31	0%	

Run PSI-Blast iteration 4 with max

### Alignments

Select All [Get selected sequences](#) [Distance tree of results](#) [Multiple alignment](#)

>  [sp|P03372.2|ESR1\\_HUMAN](#) RecName: Full=Estrogen receptor; Short=ER; AltName: Full=ER-alpha; AltName: Full=Estradiol receptor; AltName: Full=Nuclear receptor subfamily 3 group A member 1 Length=595

[GENE ID: 2099 ESR1](#) | estrogen receptor 1 [Homo sapiens] ([Over 100 PubMed links](#))

Score = 735 bits (1898), Expect = 0.0, Method: Composition-based stats.  
Identities = 595/595 (100%), Positives = 595/595 (100%), Gaps = 0/595 (0%)

```
Query 1  MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGEVYLDSSKPAVYNYPEGaay 60
          MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGEVYLDSSKPAVYNYPEGAAY
Sbjct 1  MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGEVYLDSSKPAVYNYPEGAAY 60
```

# Αποθήκευση ακολουθιών από το Blast

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

The screenshot shows the NCBI Blast search results page. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus, and a 'My NCBI Sign In' link. Below this is a search bar with 'Protein' selected in the dropdown and a 'Search' button. The page displays search results for a query, showing two results. A 'Choose Destination' dialog box is open over the first result, allowing the user to select a destination (File or Clipboard), format (FASTA), and a 'Create File' button. The first result is for 'Full=Nuclear receptor subfamily 3 group A member 1' (595 aa protein) with accession P03372.2. The second result is for 'Retinoic acid receptor beta' (455 aa protein) with accession P10826.2. The page also includes a 'Display Settings' section and a 'Filter your results' section.

NCBI Resources How To My NCBI Sign In

Protein Protein Search Limits Advanced Help

Display Settings: Summary, 20 per page, Sorted by Default order Send to: Filter your results:

Results: 1 to 20 of 67

1. [RecName: Full=Estrogen receptor; Short=ER; AltName: Full=ER-alpha; AltName: Full=Estradiol r](#)  
[Full=Nuclear receptor subfamily 3 group A member 1](#)  
595 aa protein  
Accession: P03372.2 GI: 544257  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

2. [RecName: Full=Retinoic acid receptor beta; Short=RAR-beta; AltName: Full=HBV-activated prote](#)  
[receptor subfamily 1 group B member 2; AltName: Full=RAR-epsilon](#)  
455 aa protein  
Accession: P10826.2 GI: 17380507  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

Choose Destination  
 File  Clipboard  
 Collections  
Download 67 items.  
Format  
FASTA  
Create File

Organisms [Tree]  
Drosophila melanogaster (20)