Πίνακες αντικατάστασης

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

Πίνακες αντικατάστασης

- Για πρωτεΐνες:
  - Πίνακες PAM
  - Πίνακες BLOSUM

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

Λογαριθμικές πιθανότητες

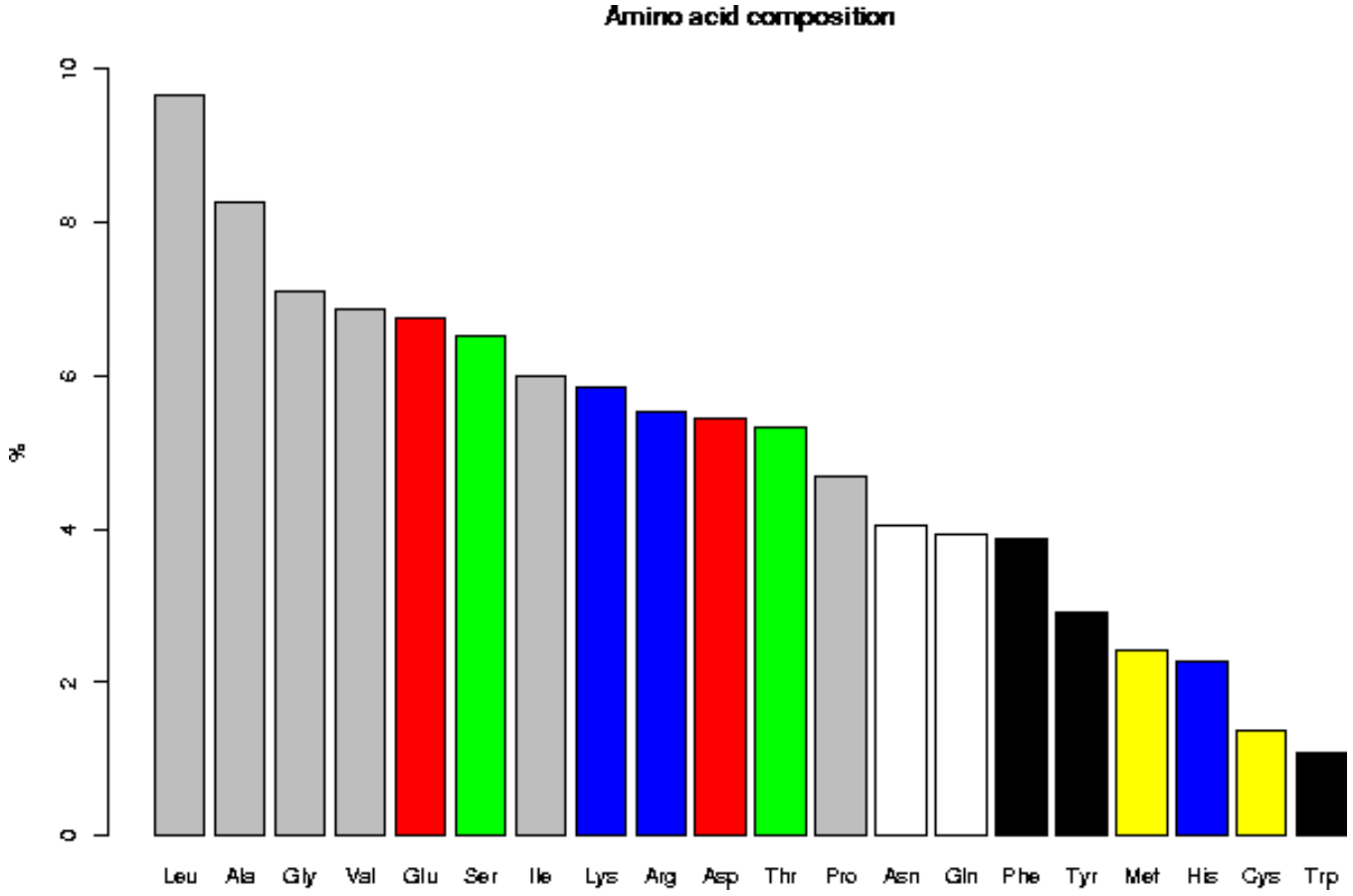
- Πρώτη χρήση από Dayhoff για πίνακες αντικατάστασης που χρησιμοποιούνται στη βαθμολόγιση στοιχίσεων.

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

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

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

Συχνότητα αμινοξέων
από Swissprot



Πίνακες PAM

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

	C	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

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

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

Πίνακες PAM (ii)

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

Πίνακες PAM (iii)

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

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

Πίνακες PAM (iv)

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

Πίνακες PAM (iv)

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

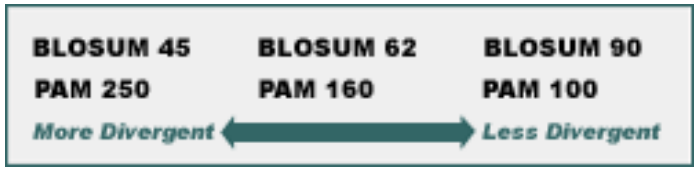
Πίνακες BLOSUM

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

Βασικές διαφορές μεταξύ PAM-BLOSUM

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

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



Πίνακες αντικατάστασης νουκλεοτιδίων

- Μοντέλο Jukes-Cantor: Ενιαίοι ρυθμοί μετάλλαξης

- Μοντέλο Kimura: μεταπτώσεις (transitions) ποιό πιθανές από μεταστροφές (transversions)

Βαθμολόγιση Κενών

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

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

Βαθμολόγιση μιας στοίχισης με πίνακα αντικατάστασης και affine gap penalty

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

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

Οδηγίες χρήσης πινάκων

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

Guidelines for using matricies

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

Στατιστική σημαντικότητα μιας στοίχισης κατά ζεύγη

- Περισσότερες πληροφορίες στο:

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

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

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

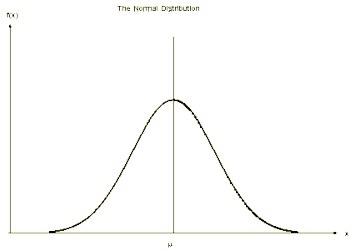
Στατιστική σημαντικότητα ολικής στοίχισης (i)

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

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

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

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



Στατιστική σημαντικότητα ολικής στοίχισης (ii)

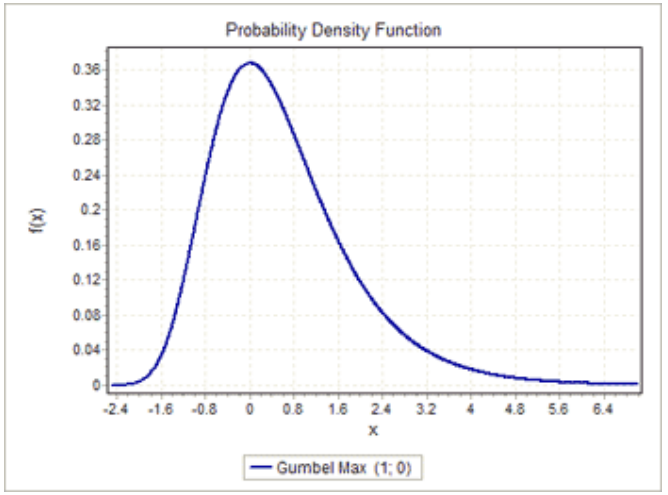
- Αν πραγματοποιηθεί το ανακάτεμα 100 φορές και η μέγιστη βαθμολογία στοίχισης δεν υπερβαίνει την βαθμολογία που παρατηρήσαμε για την στοίχιση των 2 πραγματικών ακολουθιών, τότε η στοίχιση είναι στατιστικά σημαντική σε επίπεδο P-value < 0.01

- Μεγάλο υπολογιστικό κόστος

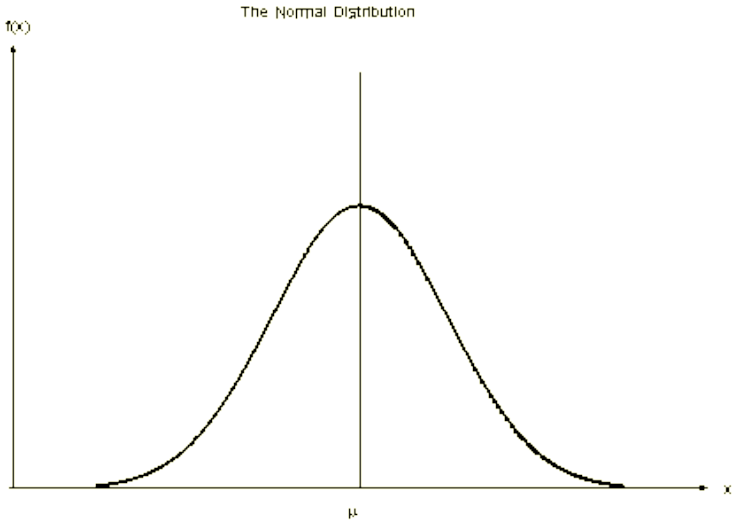
- Χρησιμοποιείται για ολικές στοιχίσεις,εντούτοις δεν ενδείκνυται η ολική στοίχιση για να αποφασίσουμε αν δύο ακολουθίες είναι ομόλογες

Στατιστική σημαντικότητα τοπικής στοίχισης (i)

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



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



Στατιστική σημαντικότητα τοπικής στοίχισης (ii)

Κατανομή ακραίων τιμών Gumbel

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

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

Στατιστική σημαντικότητα τοπικής στοίχισης (iii)

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

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

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

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

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

Στατιστική σημαντικότητα τοπικής στοίχισης (iv)

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

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

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

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

Αναζήτηση ομόλογων ακολουθιών σε βάσεις δεδομένων (i)

- Ομόλογες ακολουθίες πιθανόν να έχουν παρόμοιες λειτουργίες.

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

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

- Αναζήτηση με δυναμικό προγραμματισμό: Smith-Waterman, SSearch

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

Αναζήτηση ομόλογων ακολουθιών σε βάσεις δεδομένων (ii)

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

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

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

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

- Διαφορετικός υπολογισμός μεταξύ FASTA - BLAST.

Αλγόριθμος FASTA

- Ktuples: λέξεις μήκους κ που ταιριάζουν απόλυτα μεταξύ των ακολουθιών.

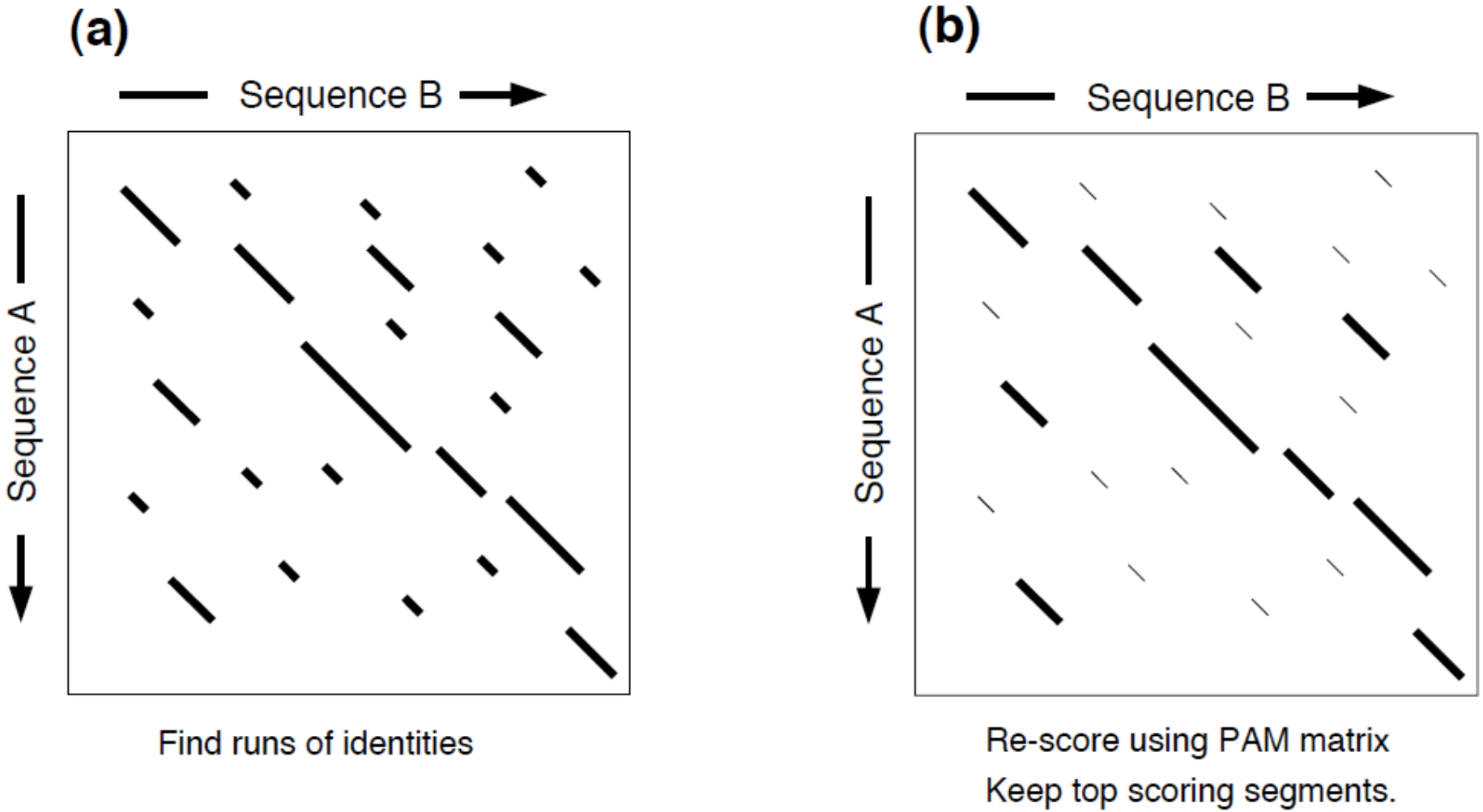
- Για πρωτεΐνες:
  - Ktup 1-2. (20 αμινοξέα)

- Για DNA:
  - Ktup 4-6. (μόνο 4 νουκλεοτίδια)

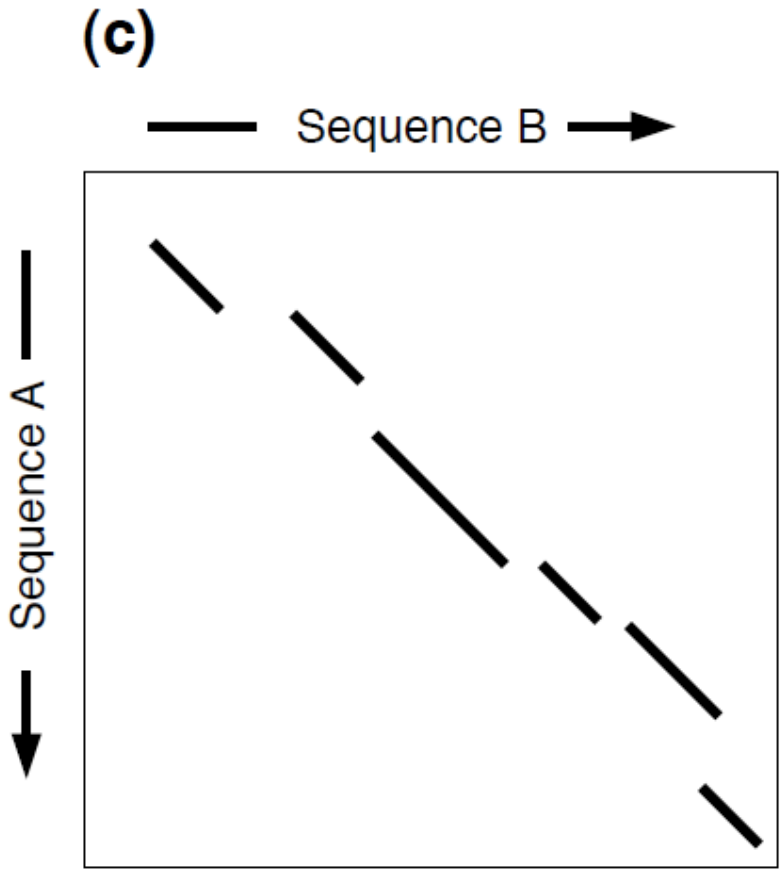
Αλγόριθμος FASTA: εν συντομία

- Ο αλγόριθμος ψάχνει γρήγορα για μικρές περιοχές με μεγάλη ομοιότητα.
- Αν εντοπίσει τέτοιες περιοχές, προσπαθεί να βελτιώσει την στοίχιση τοπικά.
- Αν η γρήγορη τοπική στοίχιση ξεπεράσει κάποια οριακή τιμή, τότε γίνεται κανονική τοπική στοίχιση Smith-Waterman

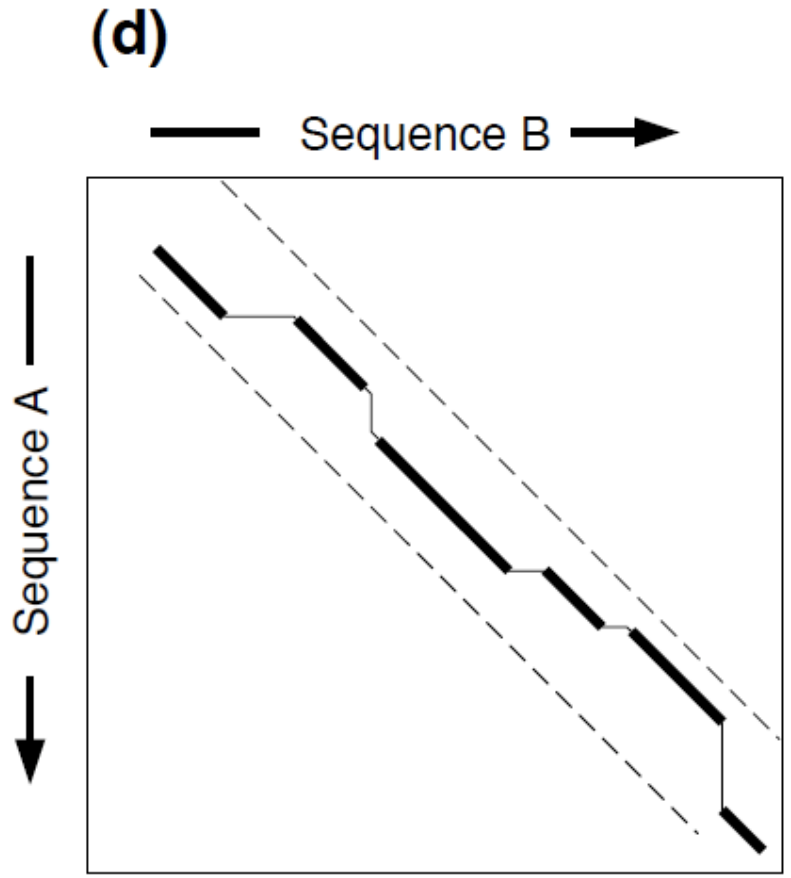
Αλγόριθμος FASTA



Αλγόριθμος FASTA



Apply "joining threshold"
to eliminate segments that
are unlikely to be part of the alignment
that includes highest scoring segment.



Use dynamic programming
to optimise the alignment in a
narrow band that encompasses
the top scoring segments.

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.

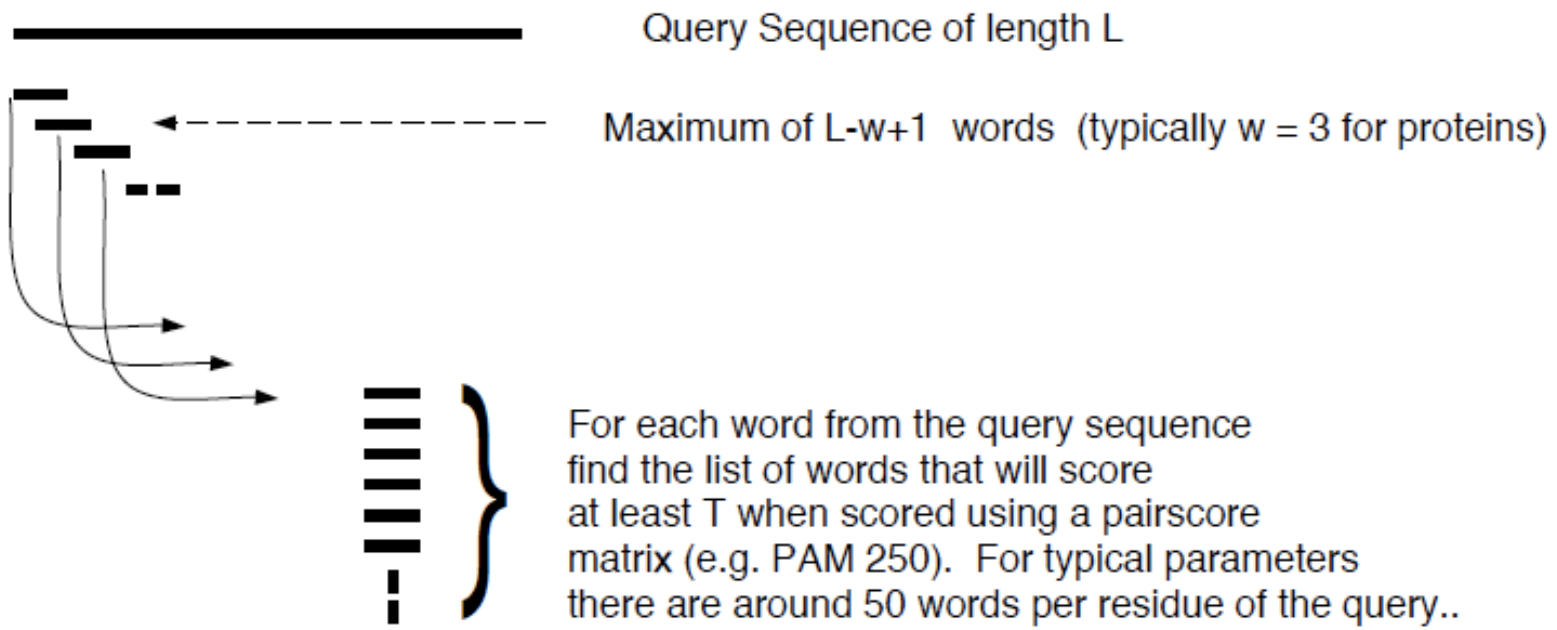
- Πρωτεΐνες: w=3

- 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

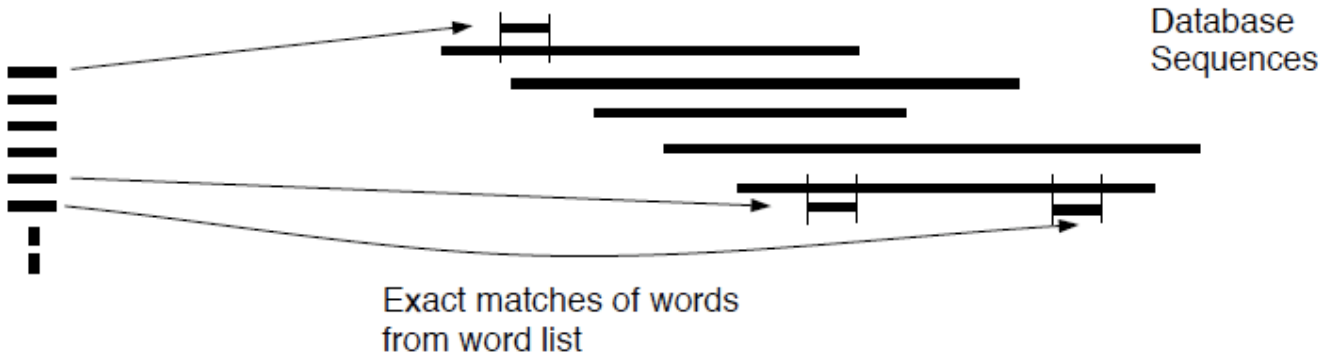
(1) For the query find the list of high scoring words of length w.



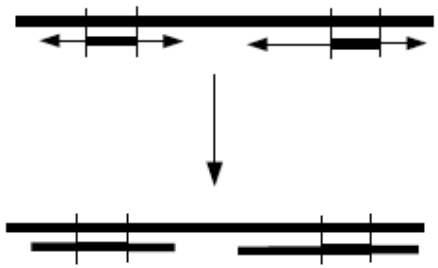
- PQG
- 20 X 20 X 20 = 8.000 words
- PQG X 8.000 words
- PQG X PEG = 7 + 2 + 6=15
- Όριο τιμής T

Αλγόριθμος BLAST

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



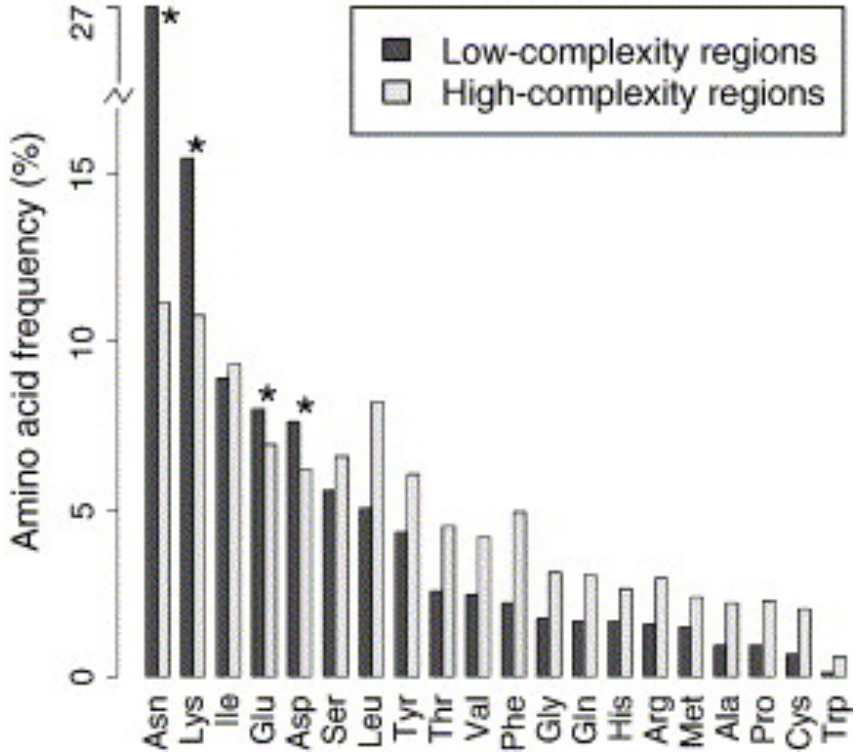
(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.



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.

```
PEGADINDAKK LINEDQPR
        ↓
PEGADINDAKKKKKKKKKKKKKK LINEDQPR
```

- Μεταλλάξεις στην ακολουθία 2.

```
DSAKL IMTCK PIMQEYGA
        ↓
DSAKL IMTCKKKKKKKKKKKKPIMQEYGA
```

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

```
PEGADINDAKKKKKKKKKKKKKK LINEDQPR
      |  ||||||||||||||  |
DSAKL IMTCKKKKKKKKKKKK - - PIMQEYGA
```

Φιλτράρισμα περιοχών χαμηλής πολυπλοκότητας

- Φιλτράρισμα (masking)
- Και για BLAST και για FASTA.
- Φιλτράρεται η ακολουθία επερώτησης μόνο.

```
PEGADINDAKKKKKKKKKKKKKK LINEDQPR
      |    ||||||||||||||    |
DSAKL IMTCKKKKKKKKKKKKK - - PIMQEYGA
```

- Χ για πρωτεΐνες και Ν για νουκλεϊκά οξέα (ή μικρά γράμματα)

```
PEGADINDAXXXXXXXXXXXXXXX LINEDQPR

DSAKL IMTCXXXXXXXXXXXXX - - PIMQEYGA
```

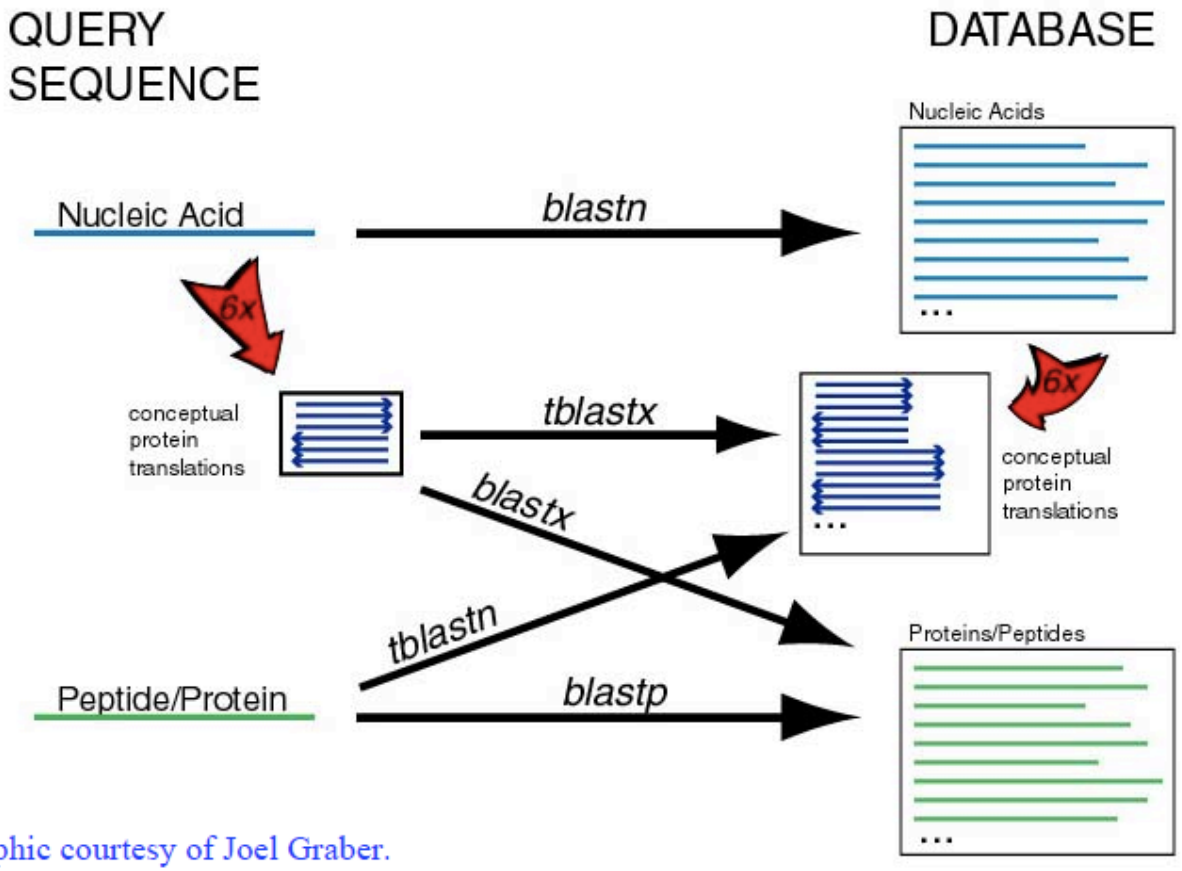
- Φίλτρα του Blast:
  - Dust: νουκλεοτίδια
  - Seg: πρωτεΐνες
- Άλλες ακολουθίες που μπορεί να φιλτράρονται:
  - Επαναλήψεις Alu
  - Φορείς κλωνοποίησης
  - Διαμεμβρανικές περιοχές
  - Coiled-coils

Blast

Blast

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

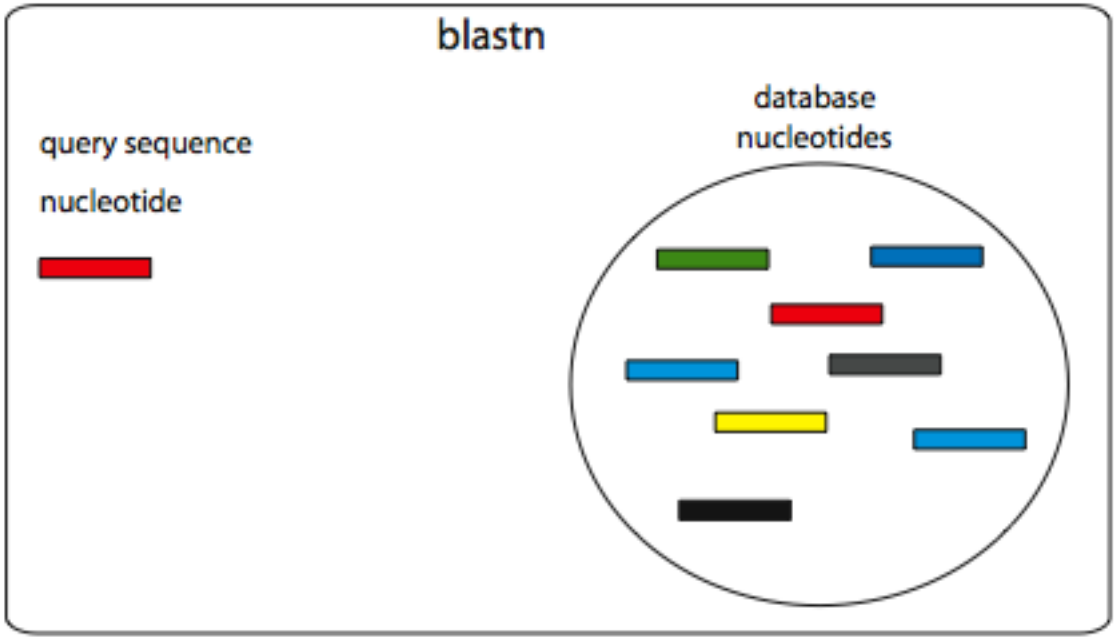
Blast



Graphic courtesy of Joel Graber.

Blastn / MegaBlast

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



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

MegaBlast

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

Blastn

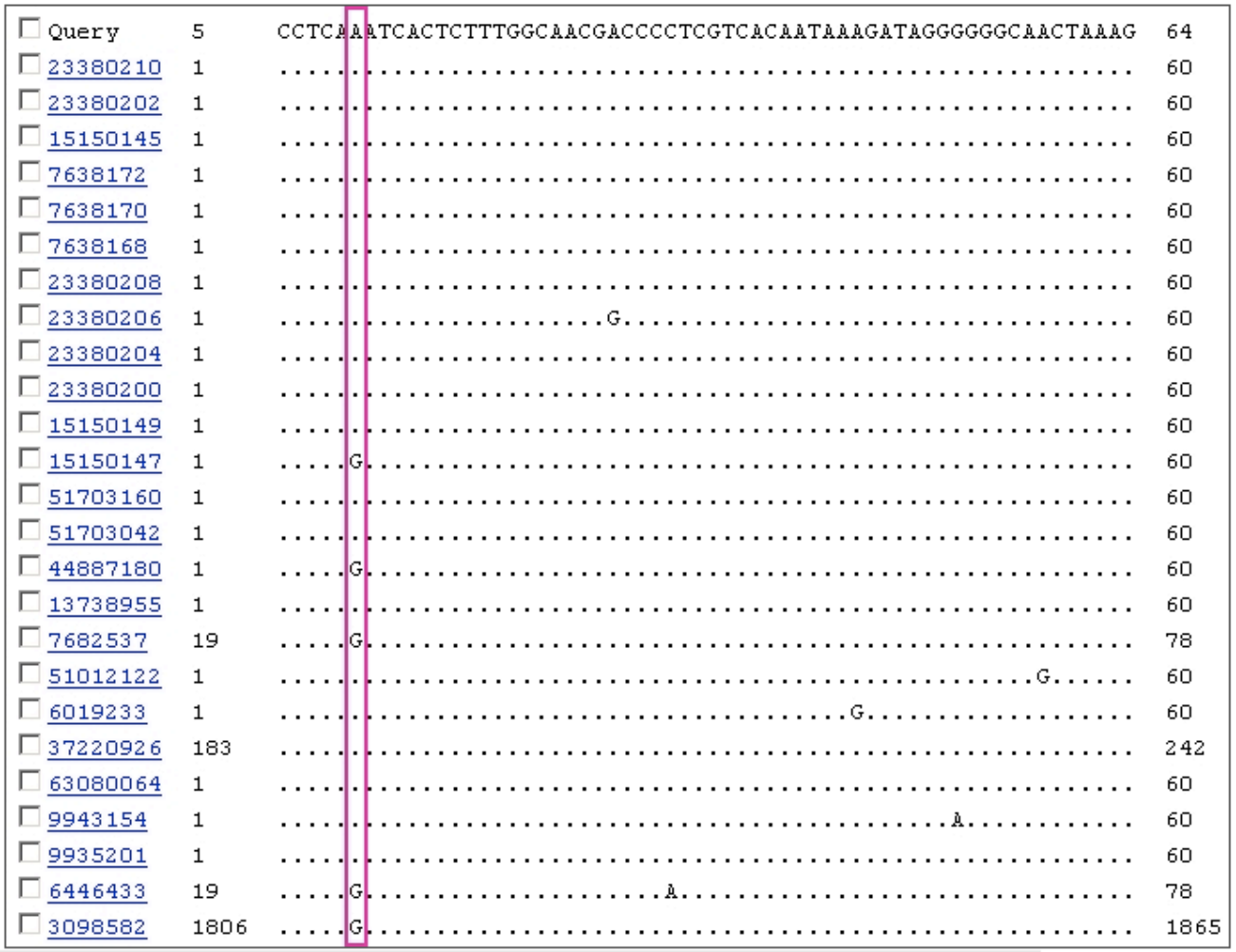
Παράδειγμα: Έλεγχος εξειδίκευσης ζεύγους εκκινητών (primers)

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

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

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

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


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

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

Blastn

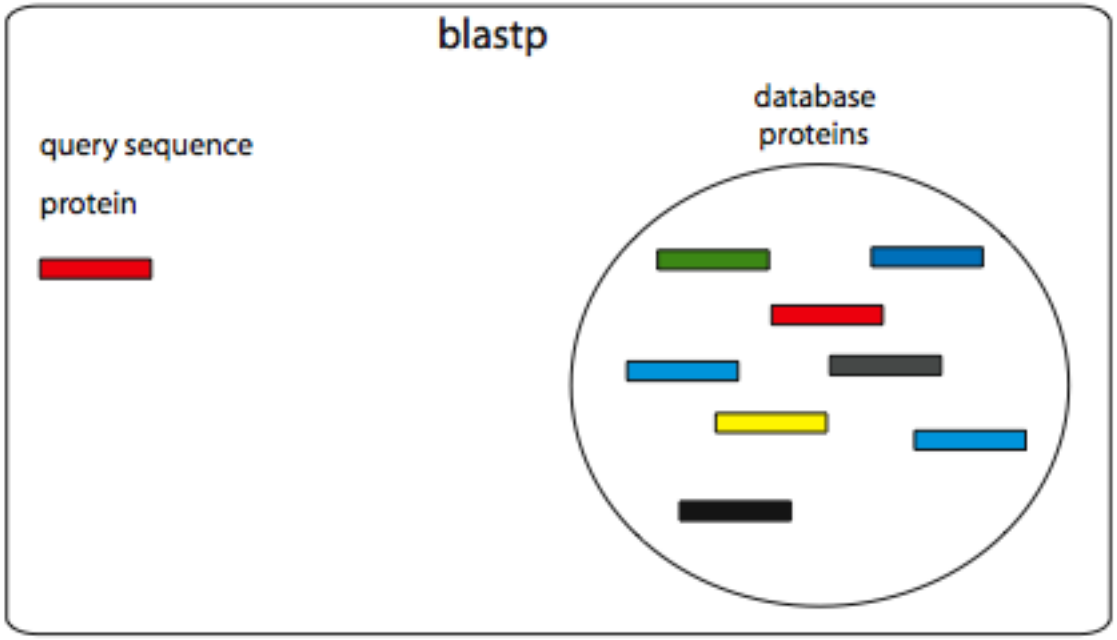
Παράδειγμα: Εντοπισμός SNPs σε ακολουθίες του ιού HIV-1 για ανθεκτικότητα σε φάρμακα



Blastp

- Πρωτεΐνη Χ πρωτεΐνες

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



Translated Blast

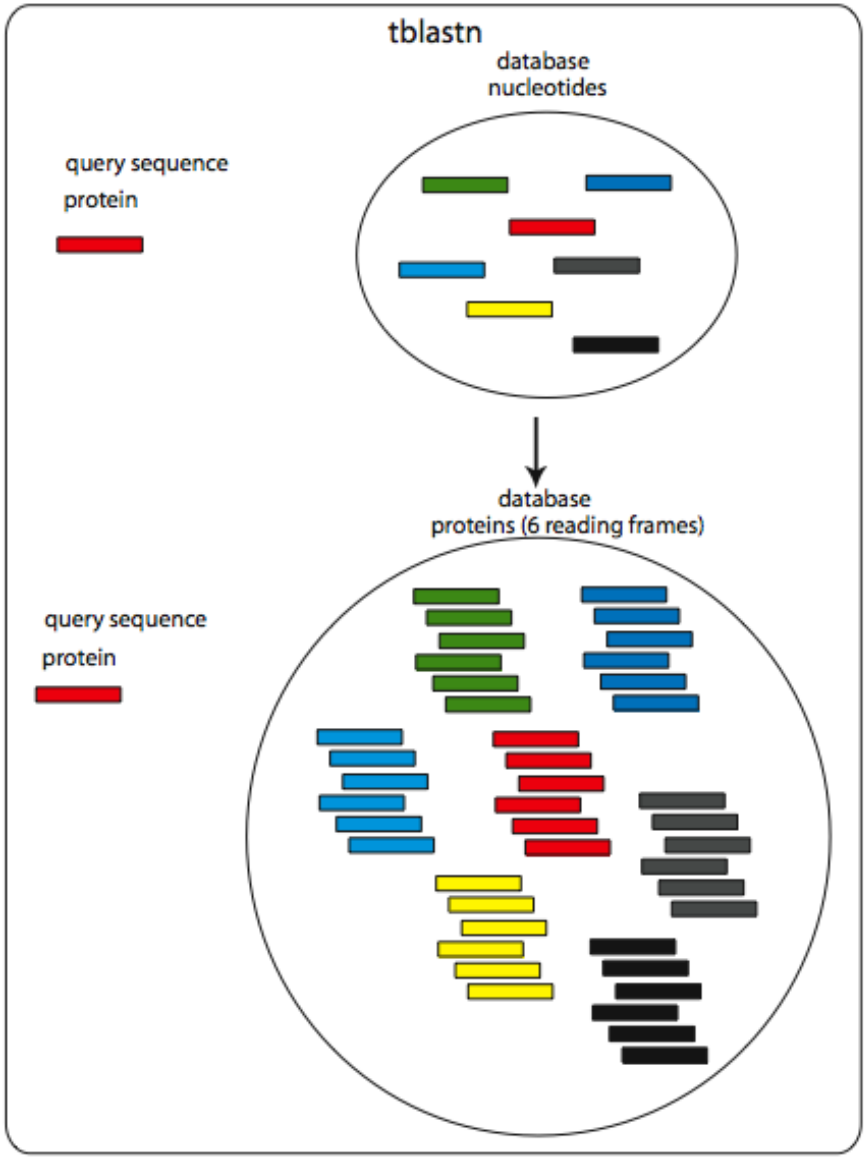
- Η νουκλεοτιδική ακολουθία ενός γονιδίου εμφανίζεται λιγότερο συντηρημένη από την αμινοξική ακολουθία της πρωτεΐνης του.

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

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

tblastn

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



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

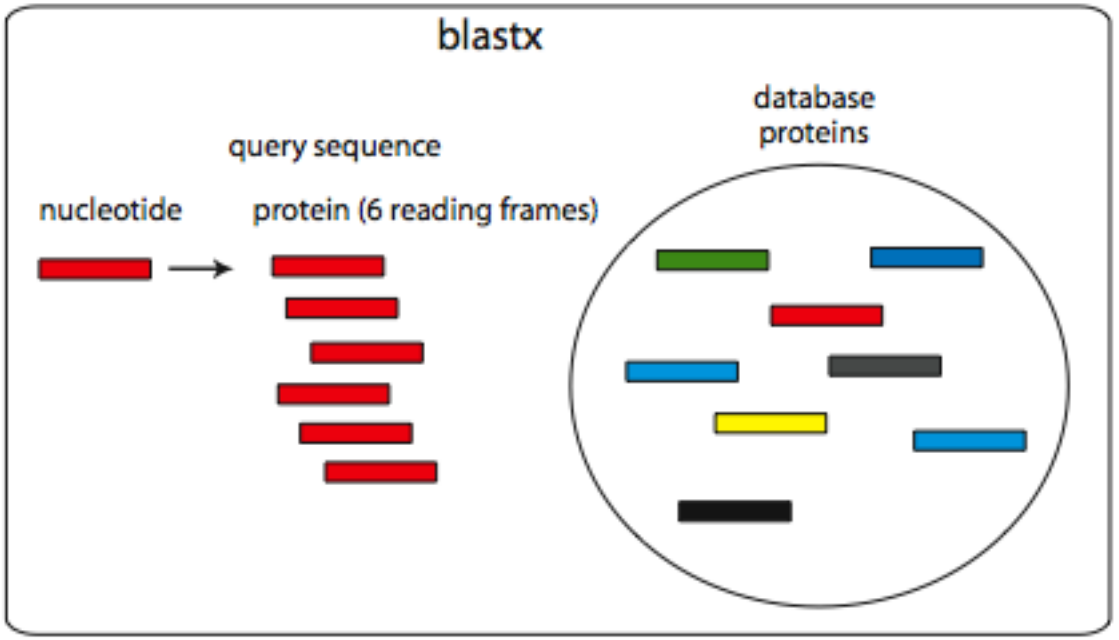
tblastn

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

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

Blastx

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



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

Blastx

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

Alignments

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

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

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

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

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

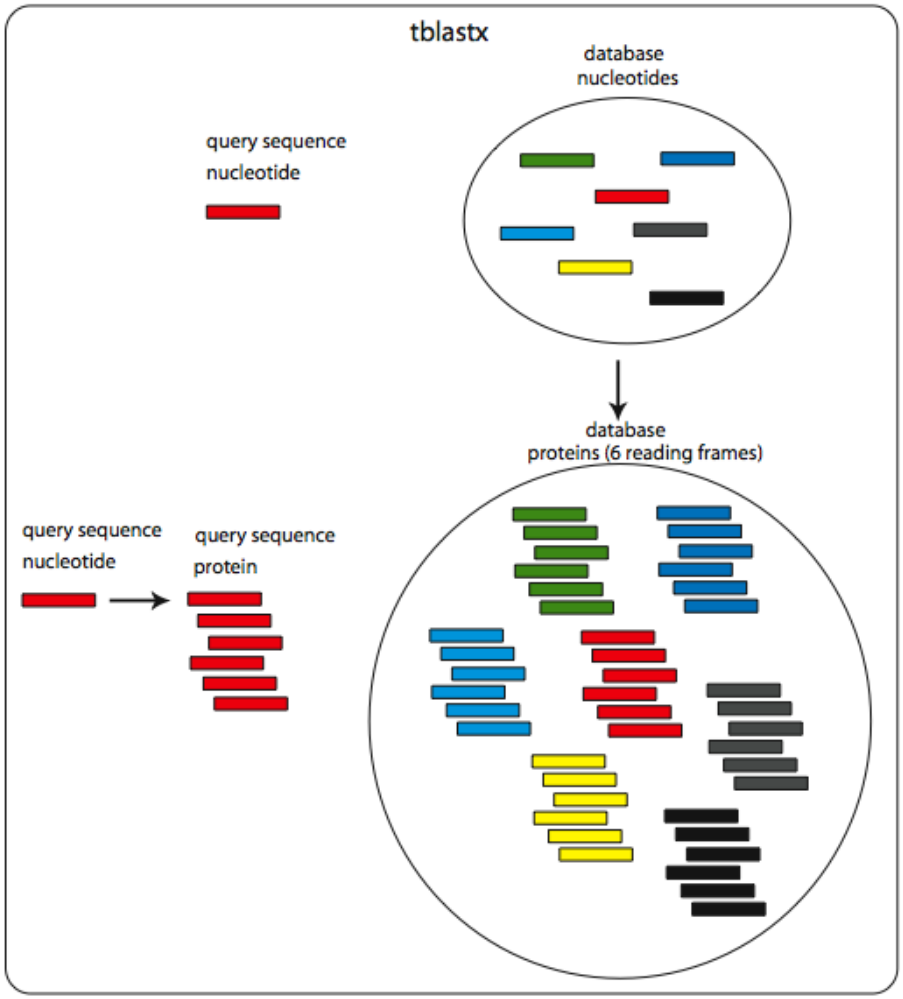
```
Query  2    EEDIVIRSENFTNNAKTIIVQLKESIKINCTRPNNNTRKSIPIATGGAIYATGDIIGDIR  181
            EEDIVIRSENFTNNAKTIIVQLKESIKINCTRPNNNTRKSIPIATGGAIYATGDIIGDIR
Sbjct  1    EEDIVIRSENFTNNAKTIIVQLKESIKINCTRPNNNTRKSIPIATGGAIYATGDIIGDIR  60

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

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

tblastx

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

- 6X6 blastp



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

tblastx

- Αναζήτηση (διαειδική) για άγνωστα μέχρι σήμερα γονίδια.

Blast και φυλογένεση

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

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

The Closest BLAST Hit Is Often Not the Nearest Neighbor

Liisa B. Koski, G. Brian Golding

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

Received: 23 January 2001 / Accepted: 20 February 2001

Επαλήθευση ομολογίας μέσω ενδιάμεσων ακολουθιών

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

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

Επαλήθευση ομολογίας μέσω ενδιάμεσων ακολουθιών

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