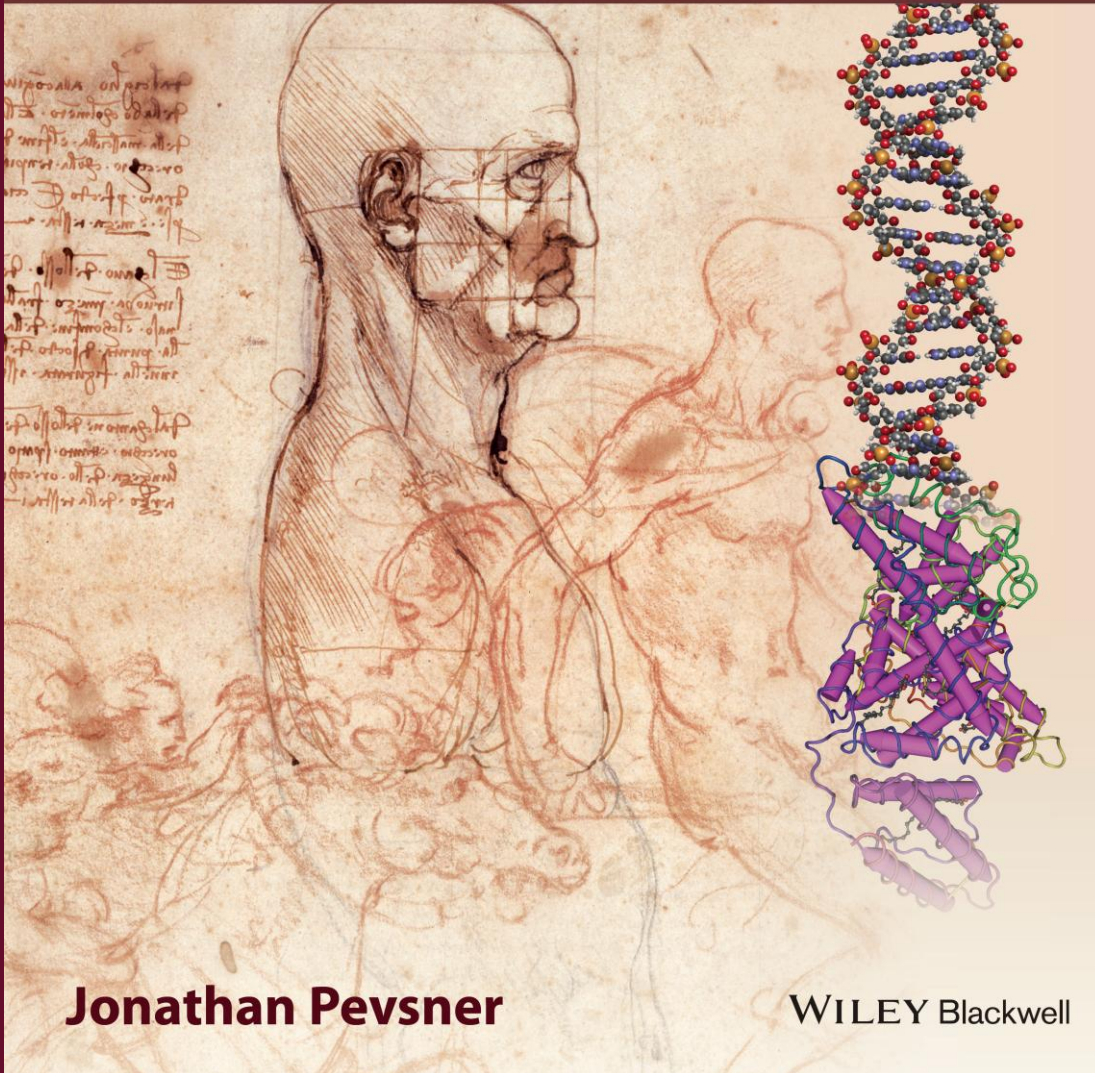


BIOINFORMATICS AND FUNCTIONAL GENOMICS

third edition



Jonathan Pevsner

WILEY Blackwell

Κεφάλαιο 4

Το Βασικό Εργαλείο Αναζήτησης Τοπικής Στοίχισης (BLAST)

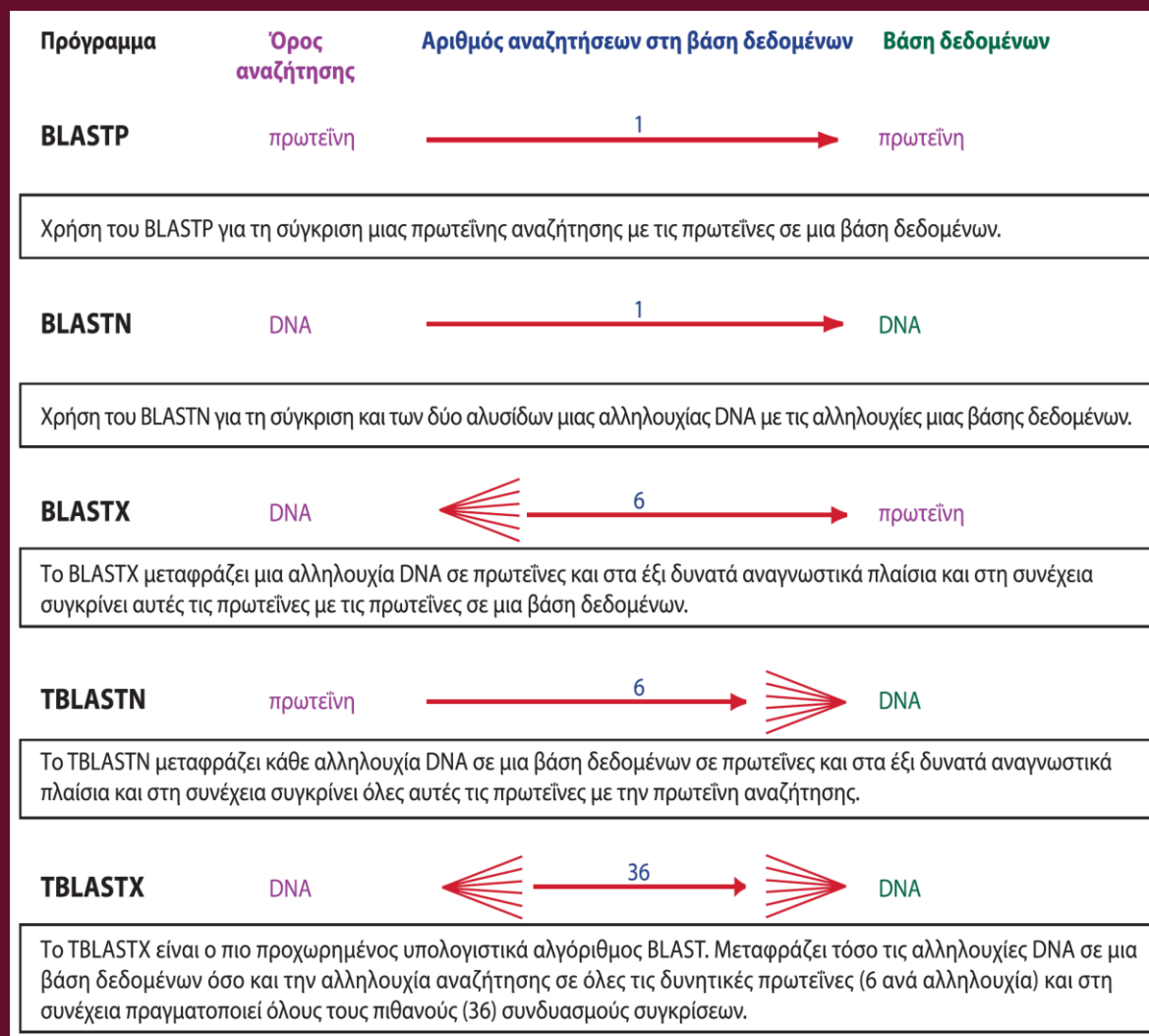
Ακαδημαϊκές
Εκδόσεις



The image shows the NCBI BLAST Standard Protein BLAST interface. It includes a navigation bar with 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. A user is logged in as 'pevsner'. The main section is titled 'Standard Protein BLAST' and contains several input fields and options. Numbered arrows point to specific parts of the interface:

- 1** points to the 'Enter Query Sequence' section, where a protein sequence is entered: `>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]`.
- 2** points to the 'Choose Search Set' section, where the 'Database' is set to 'Reference proteins (refseq_protein)'.
- 3** points to the 'Entrez Query' field, which contains the query 'perutz m[Author]'.
- 4** points to the 'Program Selection' section, where the 'Algorithm' is set to 'blastp (protein-protein BLAST)'.
- 5** points to the 'BLAST' button and the 'Algorithm parameters' section at the bottom.

Εικόνα 4.1 Η αρχική σελίδα αναζήτησης BLASTP στο NCBI. Η αλληλουχία αναζήτησης μπορεί να εισαχθεί ως κωδικός καταχώρισης, αναγνωριστικό GI ή αλληλουχία σε μορφή FASTA όπως φαίνεται εδώ (βέλος 1). Η βάση δεδομένων πρέπει να επιλεγεί (βέλος 2), εφόσον δεν έχει επιλεγεί η προεπιλεγμένη βάση δεδομένων (όπως εδώ, όπου έχει επιλεγεί η βάση δεδομένων πρωτεϊνών RefSeq). Η επιλογή επισημαίνεται με κίτρινο χρώμα. Η αναζήτηση μπορεί να περιοριστεί σε έναν συγκεκριμένο οργανισμό ή ταξινομική ομάδα και μπορεί να περιοριστεί περαιτέρω με βάση συγκεκριμένες καταχωρίσεις Entrez (βέλος 3). Εδώ περιορίζουμε την αναζήτηση σε καταχωρίσεις που περιλαμβάνουν τον συγγραφέα Max Perutz. Σε αυτό το κεφάλαιο αναλύουμε τον αλγόριθμο BLASTP (βέλος 4), ενώ οι αλγόριθμοι PSI-BLAST, PHI-BLAST και DELTA-BLAST περιγράφονται στο Κεφάλαιο 5. Πολλές από τις παραμέτρους αναζήτησης είναι δυνατόν να τροποποιηθούν (βέλος 5).



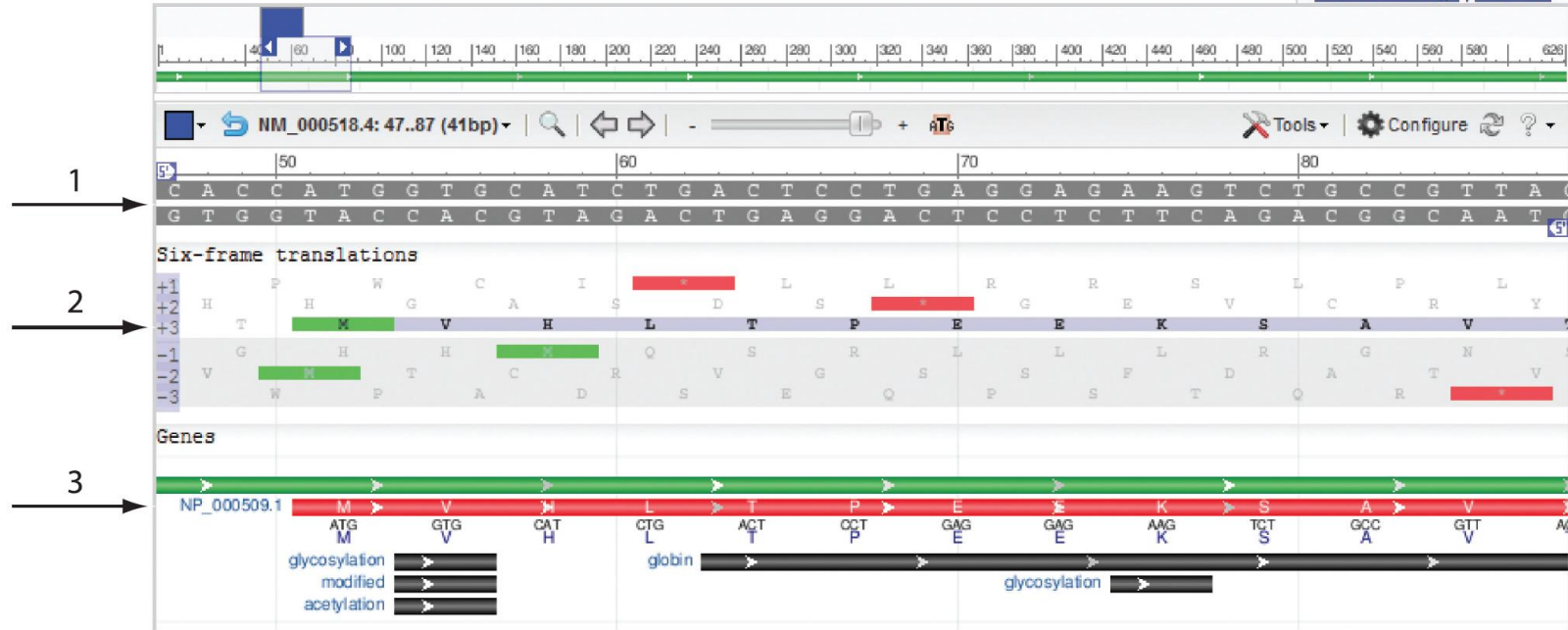
Εικόνα 4.2 Επισκόπηση των πέντε βασικών αλγορίθμων του BLAST. Σημειώστε ότι η κατάληξη P αναφέρεται σε πρωτεΐνες (όπως στο BLASTP), η κατάληξη N σε νουκλεοτίδια και η κατάληξη X σε DNA που μεταφράζεται δυνητικά σε έξι πρωτεϊνικές αλληλουχίες. Η κατάληξη T αναφέρεται στη μετάφραση (Translation), κατά την οποία όλες οι νουκλεοτιδικές αλληλουχίες σε μια βάση δεδομένων DNA μεταφράζονται δυνητικά σε πρωτεΐνες σε έξι αναγνωστικά πλαίσια.

Homo sapiens hemoglobin, beta (HBB), mRNA

NCBI Reference Sequence: NM_000518.4

[GenBank](#) [FASTA](#)

[Link To This Page](#) [Feedback](#)



Εικόνα 4.3 Το DNA είναι δυνατόν να κωδικοποιεί πρωτεΐνες σε έξι διαφορετικά αναγνωστικά πλαίσια. Για να το κατανοήσουμε αυτό καλύτερα, εξετάζουμε τη νουκλεοτιδική καταχώριση του HBB (του γονιδίου της ανθρώπινης β-σφαιρίνης) στο NCBI και επιλέγουμε την προβολή «graphics» (γραφικά). Στην προβολή αυτή απεικονίζονται οι δύο συμπληρωματικές αλυσίδες (αλληλουχίες) του DNA (βέλος 1). Σε αυτή την προβολή σε μεγέθυνση, απεικονίζεται μόνο ένα τμήμα της αλληλουχίας του HBB. Από τον πάνω κλώνο κωδικοποιούνται τρεις πιθανές αμινοξικές αλληλουχίες (αναγνωστικά πλαίσια +1, +2, +3) οι οποίες παρατίθενται με γκρι χρώμα με βάση την ονοματολογία ενός γράμματος. Σε αυτή την περίπτωση, το πλαίσιο +3 αντιστοιχεί στο πλαίσιο που χρησιμοποιείται πραγματικά για τη μετάφραση στα κύτταρα (βέλος 2). Σημειώστε ότι τα πλαίσια +1 και +2, καθώς και το πλαίσιο -3 περιλαμβάνουν κωδικόνια τερματισμού (συμβολίζονται με αστερίσκους κόκκινου χρώματος). Στο κάτωτο τμήμα της εικόνας παρουσιάζεται η αλληλουχία αμινοξέων της αντίστοιχης πρωτεΐνης (βέλος 3) και τα αντίστοιχα νουκλεοτίδια (αντίστοιχο πλαίσιο +3). Οι περιοχές με μαύρη σκίαση αντιστοιχούν σε θέσεις μετα-μεταφραστικών τροποποιήσεων (π.χ. γλυκοζυλίωση), καθώς και στη χαρακτηριστική επικράτεια της σφαιρίνης.

Πίνακας 4.1 Βάσεις δεδομένων πρωτεϊνών στο NCBI στις οποίες μπορούν να πραγματοποιηθούν αναζητήσεις BLAST. Η δίσωση (#) υποδηλώνει, κατά προσέγγιση, τον αριθμό των αλληλουχιών στη βάση δεδομένων. Προσαρμοσμένος από BLAST, NCBI, <http://BLAST.ncbi.nlm.nih.gov/>.

Βάση δεδομένων	Περιγραφή	# αλληλουχιών
nr	Μη πλεονάζουσες αλληλουχίες (δηλαδή μόνο ένα αντίγραφο κάθε αλληλουχίας) από όλες τις παρακάτω βάσεις: Genebank (μεταφράσεις των κωδικών περιοχών, CDS) + PDB + SwissProt + PIR + PRF, εκτός από αλληλουχίες από προγράμματα δειγμάτων WGS που συλλέχθηκαν από το περιβάλλον	65 εκατομμύρια
Πρωτεΐνες αναφοράς	Αλληλουχίες πρωτεϊνών αναφοράς στο NCBI	50 εκατομμύρια
UniProtKB/SwissProt	Μη πλεονάζουσες αλληλουχίες UniProtKB/SwissProt	450.000
Πατενταρισμένες πρωτεϊνικές αλληλουχίες	Πρωτεϊνικές αλληλουχίες από το τμήμα Πατεντών της GenBank	1,3 εκατομμύρια
PDB (Protein Data Bank)	Βάση δεδομένων PDB	77.000
Μεταγονιδιωματικές πρωτεΐνες	Πρωτεΐνες από τα μεταγονιδιωματικά προγράμματα WGS (env_nr)	6,5 εκατομμύρια
Μεταγράψιμα	Αλληλουχίες μεταγράφων από τυχαία προσπέλαση μεταγραφώματος (TSA)	770.000

Πίνακας 4.2 Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών στο NCBI στις οποίες μπορούν να πραγματοποιηθούν αναζητήσεις BLAST. Η δίσση (#) υποδηλώνει, κατά προσέγγιση, τον αριθμό των αλληλουχιών στη βάση δεδομένων. Προσαρμοσμένος από BLAST, NCBI, <http://BLAST.ncbi.nlm.nih.gov/>.

Βάση δεδομένων	Περιγραφή	# αλληλουχιών
Γονιδιωματικές αλληλουχίες ανθρώπου + μετάγραφα (Human Genomic + Transcript)	Υπομνηματισμός αλληλουχιών RNA του ανθρώπου, έκδοση 104 του NCBI, σύνολο συναρμολογημάτων του γονιδιώματος του ανθρώπου	55.000
Γονιδιωματικές αλληλουχίες ποντικού + μετάγραφα (Mouse Genomic + Transcript)	Υπομνηματισμός αλληλουχιών RNA του ποντικού (<i>Mus musculus</i>) του NCBI, σύνολο συναρμολογημάτων του γονιδιώματος του ποντικού	Μη διαθέσιμη πληροφορία
nr/nt	Όλες οι αλληλουχίες GenBank+EMBL+DDBJ+PDB+RefSeq, δεν περιλαμβάνονται οι EST, STS, GSS, WGS, TSA, ούτε οι αλληλουχίες πατεντών και οι HTGS των φάσεων 0, 1 και 2	25 εκατομμύρια
refseq_rna	Αλληλουχίες αναφοράς μεταγράφων στο NCBI	3,5 εκατομμύρια
refseq_genomic	Γονιδιωματικές αλληλουχίες αναφοράς στο NCBI	2,7 εκατομμύρια
NCBI Genomes (Γονιδιώματα)	Αλληλουχίες των χρωμοσωμάτων στο NCBI	28.000
Ετικέτες εκφραζόμενης αλληλουχίας (EST, <u>E</u> xpressed <u>S</u> equences <u>T</u> ags)	Ετικέτες εκφραζόμενης αλληλουχίας από τα αντίστοιχα τμήματα των βάσεων δεδομένων GenBank+EMBL+DDBJ	75 εκατομμύρια
Αλληλουχίες διερεύνησης γονιδιώματος (GSS, <u>G</u> enome <u>S</u> urvey <u>S</u> equences)	Αλληλουχίες διερεύνησης γονιδιώματος στις οποίες περιλαμβάνονται γονιδιωματικές αλληλουχίες μονής ανάγωσης, αλληλουχίες παγίδευσης εξονίων και αλληλουχίες Alu από PCR	36 εκατομμύρια
Γονιδιωματικές αλληλουχίες από μαζική αλληλούχιση (HTGS, <u>H</u> igh- <u>T</u> hroughput <u>G</u> enomic <u>S</u> equences)	Γονιδιωματικές αλληλουχίες από μαζική αλληλούχιση γονιδιωμάτων που όμως δεν έχουν ολοκληρωθεί (αλληλουχίες από τις φάσεις 0, 1 και 2)	153.000
Αλληλουχίες πατεντών	Νουκλεοτιδικές αλληλουχίες από το τμήμα Πατεντών της GenBank	21 εκατομμύρια
PDB (<u>P</u> rotein <u>D</u> ata <u>B</u> ank)	Νουκλεοτιδικές αλληλουχίες της βάσης PDB	8.000
alu	Επαναλαμβανόμενες αλληλουχίες <i>Alu</i> του ανθρώπου	325
Ετικέτες θέσης αλληλουχίας (STS, <u>S</u> equences <u>T</u> agged <u>S</u> ites)	Αλληλουχίες από τα τμήματα Ετικετών θέσης αλληλουχίας των βάσεων δεδομένων GenBank+EMBL+DDBJ	1,3 εκατομμύρια
Τυχαία προσπέλαση ολικού γονιδιώματος (WGS, <u>W</u> hole- <u>G</u> enome <u>S</u> hotgun)	Συναρμολογήματα αλληλουχιών από τυχαία προσπέλαση ολικού γονιδιώματος	116 εκατομμύρια
Αλληλουχίες μεταγράφων από τυχαία προσπέλαση μεταγραφώματος (TSA, <u>T</u> ranscriptome <u>S</u> hotgun <u>A</u> ssembly sequences)	Συναρμολογήματα αλληλουχιών μεταγράφων από τυχαία προσπέλαση μεταγραφώματος (TSA)	15 εκατομμύρια
Αλληλουχίες 16S ριβοσωμικού RNA (Βακτήρια και Αρχαία)	Αλληλουχίες 16S ριβοσωμικού RNA (από Βακτήρια και Αρχαία)	7.500

Algorithm parameters

General Parameters

1 → **Max target sequences** Select the maximum number of aligned sequences to display ?

2 → **Short queries** ☒ Automatically adjust parameters for short input sequences ?

3 → **Expect threshold** ?

4 → **Word size** ?

5 → **Max matches in a query range** ?

Scoring Parameters

6 → **Matrix** ?

7 → **Gap Costs** ?

8 → **Compositional adjustments** ?

Filters and Masking

9 → **Filter** ☐ Low complexity regions ?

10 → **Mask** ☐ Mask for lookup table only ?
☐ Mask lower case letters ?

BLAST Search database **Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**
☐ Show results in a new window

Εικόνα 4.4 Προαιρετικές παράμετροι του BLASTP. Τα αριθμημένα βέλη αναφέρονται στη συζήτηση στο κείμενο.

(α) Προεπιλογή: προσαρμογή του πίνακα βαθμολόγησης με βάση τη σύσταση

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: [ref|NP_648360.2|](#) Length: 120 Number of Matches: 1

Range 1: 32 to 114 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
31.6 bits(70)	0.050	Compositional matrix adjust.	21/88(24%)	40/88(45%)	12/88(13%)
Query 29	HLCSHSLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--	87	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q		
Sbjct 32	KLCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQM	86			
Query 88	-----KRGIVEQCCTSIICSLYQLENYC	109	+ G+ ++CC C++ ++ YC		
Sbjct 87	LKTRRLRDGVFDECCLKSCTMDEVLYYC	114			

(β) Χωρίς προσαρμογή για τη σύσταση (μόνο με φιλτράρισμα περιοχών χαμηλής πολυπλοκότητας από προεπιλογή)

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: [ref|NP_648360.2|](#) Length: 120 Number of Matches: 1

Range 1: 33 to 114 [GenPept](#) [Graphics](#)

Score	Expect	Identities	Positives	Gaps
33.5 bits(75)	0.009	21/87(24%)	40/87(45%)	12/87(13%)
Query 30	LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--	87	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q	
Sbjct 33	LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML	87		
Query 88	-----KRGIVEQCCTSIICSLYQLENYC	109	+ G+ ++CC C++ ++ YC	
Sbjct 88	KTRRLRDGVFDECCLKSCTMDEVLYYC	114		

(γ) Προσαρμογή χρησιμοποιώντας στατιστική με βάση τη σύσταση

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: [ref|NP_648360.2|](#) Length: 120 Number of Matches: 1

Range 1: 33 to 114 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
30.4 bits(67)	1e-04	Composition-based stats.	21/87(24%)	40/87(45%)	12/87(13%)
Query 30	LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--	87	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q		
Sbjct 33	LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML	87			
Query 88	-----KRGIVEQCCTSIICSLYQLENYC	109	+ G+ ++CC C++ ++ YC		
Sbjct 88	KTRRLRDGVFDECCLKSCTMDEVLYYC	114			

Εικόνα 4.5 Επίδραση των αλλαγών των πινάκων βαθμολόγησης και των παραμέτρων αναζήτησης στις κατά ζεύγη στοιχίσεις σε αναζητήσεις BLASTP. Εδώ χρησιμοποιήθηκε ως όρος αναζήτησης η ανθρωπίνη ινσουλίνη (NP_000198.1) για αναζήτηση με το BLASTP στη βάση δεδομένων πρωτεϊνών RefSeq με περιορισμό οργανισμού στην *Drosophila*. (α) Από την αναζήτηση με τις προεπιλεγμένες ρυθμίσεις προκύπτει στοίχιση με μια πρωτεΐνη ινσουλίνης της *Drosophila* με σκορ ίσο με 31,6 bits και τιμή *E* ίση με 0,05. Τα αποτελέσματα παρουσιάζονται (β) χωρίς προσαρμογή για τη σύσταση και (γ) μετά από προσαρμογή χρησιμοποιώντας στατιστική ανάλυση με βάση τη σύσταση. Οι αναμενόμενες τιμές γι' αυτές τις τρεις αναζητήσεις σημειώνονται σε πλαίσια με κόκκινο περίγραμμα.

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI Welcome pevsner. [Sign Out]

NCBI/ BLAST/ blastp suite/ Formatting Results - U4X4JS8B014

ⓘ Your search is limited to records matching entrez query: txid6656 [ORGN].

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [YouTube](#) [How to read this page](#) [Blast report description](#)

gi|4504349|ref|NP_000509.1| hemoglobin subunit...

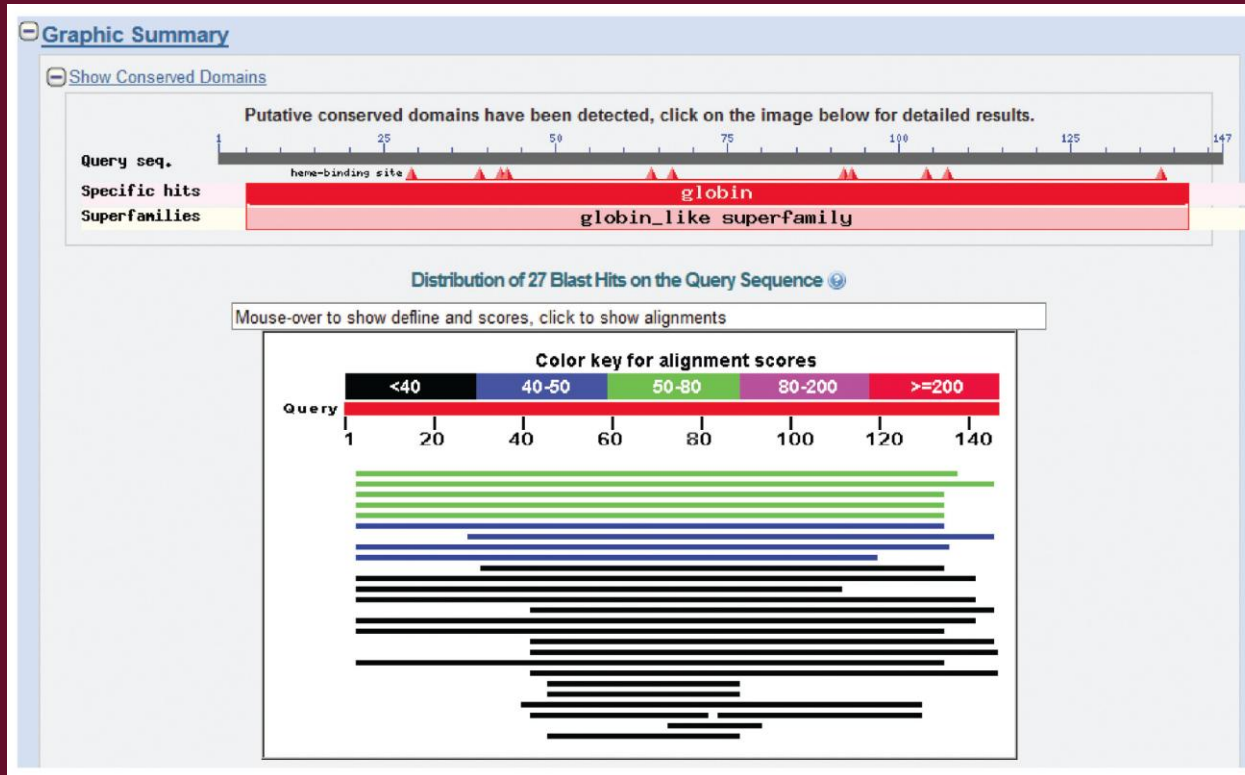
Query ID	Id 51620	Database Name	refseq_protein
Description	gi 4504349 ref NP_000509.1 hemoglobin subunit beta [Homo sapiens]	Description	NCBI Protein Reference Sequences
Molecule type	amino acid	Program	BLASTP 2.2.28+ Citation
Query Length	147		

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

Εικόνα 4.6 Στο άνω μέρος της αναφοράς αποτελεσμάτων του BLAST περιγράφεται η αναζήτηση που πραγματοποιήθηκε και αναφέρονται ο όρος αναζήτησης (βέλος 1), το μήκος του όρου αναζήτησης (βέλος 2), η βάση δεδομένων (βέλος 3) και το πρόγραμμα που χρησιμοποιήθηκε (BLASTP 2.2. 28 σε αυτή την περίπτωση, βέλος 4). Στο κάτω μέρος, πρόσθετοι σύνδεσμοι παρέχουν μια σύνοψη της αναζήτησης με λεπτομέρειες των στατιστικών της στοιχείων (βέλος 5) και ταξινομικές αναφορές των αποτελεσμάτων (βέλος 6).

Search Parameters		
Program	blastp	
Word size	3	
Expect value	10 ← 1	
Hitlist size	100	
Gapcosts	11,1	
Matrix	BLOSUM62 ← 2	
Filter string	F	
Genetic Code	1	
Window Size	40	
Threshold	11 ← 3	
Composition-based stats	2	
Database		
Posted date	Jun 12, 2013 10:46 AM	
Number of letters	6,910,040,539 ← 4	
Number of sequences	19,996,853	
Entrez query	txid10090 [ORGN]	
Karlin-Altschul statistics		
Lambda	0.320339	0.267
K	0.136843	0.041
H	0.422367	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Εικόνα 4.7 Σύνοψη μιας αναζήτησης BLAST. Στο άνω μέρος εμφανίζονται οι παράμετροι αναζήτησης [π.χ. το πρόγραμμα που χρησιμοποιήθηκε, η αναμενόμενη τιμή (βέλος 1), ο πίνακας βαθμολόγησης (βέλος 2), τα φίλτρα που εφαρμόστηκαν και το κατώφλι (βέλος 3)]. Στο μέσο του πίνακα περιγράφεται η βάση δεδομένων. Σε αυτό το παράδειγμα η βάση δεδομένων περιέχει περίπου επτά δισεκατομμύρια αμινοξικά κατάλοιπα (βέλος 4) και τα αποτελέσματα έχουν περιοριστεί στο txid10090 (δηλαδή στον ποντικό). Στο κάτω μέρος παρατίθενται οι τιμές των στατιστικών παραμέτρων των Karlin-Altschul (λ , K και H).



Εικόνα 4.8 Στη γραφική αναφορά των αποτελεσμάτων του BLAST περιλαμβάνονται η απεικόνιση των συντηρημένων επικρατειών (εδώ φαίνεται μια στοίχιση με μια πρωτεΐνη της οικογένειας των σφαιρινών) και μια παρουσίαση της έκτασης των στοίχισεων που προέκυψαν από την αναζήτηση χρησιμοποιώντας έναν χρωματικό κώδικα. Ο οριζόντιος άξονας αντιστοιχεί στο μήκος της αλληλουχίας αναζήτησης (147 αμινοξέων της β-σφαιρίνης). Κάθε στοίχιση στη βάση δεδομένων έχει χρώμα που εκφράζει το σκορ της (π.χ. οι πέντε στοίχισεις με πράσινο χρώμα έχουν σκορ 50-80) και μήκος ανάλογο με την έκταση της συντηρημένης αλληλουχίας (μία από τις πέντε στοίχισεις πράσινου χρώματος στη βάση δεδομένων περιλαμβάνει μια περιοχή που εκτείνεται πλήρως μέχρι και το καρβοξυτελικό άκρο της HBB, ενώ οι άλλες τέσσερις έχουν μερική κάλυψη της έκτασης της β-σφαιρίνης). Αυτό το είδος γραφικής παρουσίασης των αποτελεσμάτων είναι χρήσιμο επειδή εμφανίζει συνοπτικά τις περιοχές των αλληλουχιών από τη βάση δεδομένων που στοιχίζονται με τον όρο αναζήτησης.

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:2

Alignments Download GenPept Graphics Distance tree of results Multiple alignment							
	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396833.1 PREDIC	59.7	59.7	91%	1e-10	29%	XP_003396832.1
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref XP_003494220.1 PREDI	58.5	58.5	97%	3e-10	28%	XP_003494219.1
<input type="checkbox"/>	PREDICTED: globin-like [Megachile rotundata]	57.8	57.8	89%	6e-10	29%	XP_003707185.1
<input type="checkbox"/>	PREDICTED: globin-like [Apis florea]	53.9	53.9	89%	1e-08	30%	XP_003690810.1
<input type="checkbox"/>	globin 1 [Apis mellifera]	52.8	52.8	89%	4e-08	30%	NP_001071291.1
<input type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396831.1 PREDIC	45.1	45.1	89%	2e-05	26%	XP_003396830.1
<input type="checkbox"/>	PREDICTED: neuroglobin-like, partial [Acyrtosiphon pisum]	42.4	42.4	80%	2e-04	23%	XP_001946608.2
<input type="checkbox"/>	globin, putative [Ixodes scapularis]	42.7	42.7	90%	2e-04	25%	XP_002414906.1

Εικόνα 4.9 Σε μια τυπική αναφορά αποτελεσμάτων του BLASTP περιλαμβάνεται μια λίστα αλληλουχιών της βάσης δεδομένων που στοιχίζονται με τον όρο αναζήτησης. Επίσης, παρέχονται σύνδεσμοι γι' αυτές τις αλληλουχίες (δηλαδή προς τις αντίστοιχες καταχωρίσεις στη βάση NCBI Protein) και για τις κατά ζεύγη στοιχίσεις τους με την αλληλουχία αναζήτησης. Τέλος, στην αναφορά αποτελεσμάτων παρατίθεται το bit σκορ και η τιμή E για κάθε στοίχιση. Σημειώστε ότι οι καλύτερες στοιχίσεις, που βρίσκονται στην κορυφή της λίστας, έχουν υψηλά bit σκορ και μικρές τιμές E.

COBALT

Constraint-based Multiple Alignment Tool

My NCBI

Welcome pevsner. [Sign Out](#)

[Home](#)
[Recent Results](#)
[Help](#)

[Phylogenetic Tree](#)
[Edit and Resubmit](#)
[Back to Blast Results](#)
[Download](#)

Multiple Alignment Results - gi|4504349|ref|NP_000509.1| hemoglobin subunit... - Cobalt RID U57PC4Y5211 (8 seqs)

Descriptions
☒ Select All
 [Re-align](#)
[Alignment parameters](#)

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

Accession	Description	Links
<input checked="" type="checkbox"/> XP_003396832.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396833.1 PREDICTED: cytoglobin	GM
<input checked="" type="checkbox"/> XP_003494219.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref XP_003494220.1 PREDICTED: cytoglobin	GM
<input checked="" type="checkbox"/> XP_003707185.1	PREDICTED: globin-like [Megachile rotundata]	G
<input checked="" type="checkbox"/> XP_003690810.1	PREDICTED: globin-like [Apis florea]	G
<input checked="" type="checkbox"/> NP_001071291.1	globin 1 [Apis mellifera] >emb CAJ43389.1 globin 1 [Apis mellifera] >emb CAJ43388.1 globin 1 [Apis mellifera]	UGM
<input checked="" type="checkbox"/> XP_003396830.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396831.1 PREDICTED: cytoglobin	GM
<input checked="" type="checkbox"/> XP_001946608.2	PREDICTED: neuroglobin-like, partial [Acyrtosiphon pisum]	GM
<input checked="" type="checkbox"/> XP_002414906.1	globin, putative [Ixodes scapularis] >gb EEC18571.1 globin, putative [Ixodes scapularis]	G

Alignments
☒ Select All
 [Re-align](#)

Mouse over the sequence identifier for sequence title

View Format: [Compact](#)
 Conservation Setting: [2 Bits](#)

<input checked="" type="checkbox"/> XP_003396832	1	MGTFILRFFGFSSDDNRIDEATGLTEKQKGLVQNTWAVIRKDEVASGI	AVMTIFFKTYPEYQRYFSAFADVFPDELPA	NK	80
<input checked="" type="checkbox"/> XP_003494219	1	MGTFILRFFGISSDDNRIDEATGLTEKQKGLVQNTWAVIRKDEVASGI	AVMTIFFKTYPEYQRYFSAFADVFPDELPA	NK	80
<input checked="" type="checkbox"/> XP_003707185	1	MDSFLRLGISS--DNRIDQATGLTEKQKGLVQNTWSIIRKDEVGAG	VLVMCAFFKKYPSYVQYFEAFKDIPLDQLPDNK		79
<input checked="" type="checkbox"/> XP_003690810	1	MGTFILRFLGISSDDNRIDQATGLTERQKGLVQNTWAVVRKDEVASGI	AVMTAFFKKYPEYQRYFTAFMDTFLNELPA	NK	80
<input checked="" type="checkbox"/> NP_001071291	1	MGTFILRFLGISSDDNRIDQATGLTERQKGLVQNTWAVVRKDEVASGI	AVMTAFFKKYPEYQRYFTAFMDTFLNELPA	NK	80
<input checked="" type="checkbox"/> XP_003396830	1	MGSVLTYF--LGNPDDVVDPKGLINKKRIIRETWGLRANSVKVG	VDIMISYFKRPQHHRAFPFFKDIADDLLDNK		79
<input checked="" type="checkbox"/> XP_001946608	1	-----SCDLTR-----FI	FLFLYRLFEHQELLQLFTKFGELKTRDAQNS		42
<input checked="" type="checkbox"/> XP_002414906	1	MSW---LFGSAS---ADMPSTKGLTTSKCAIKDTWTMFRRETRTNAL	SLFVALFSRYPEYQRMFPNFADVALKDMMQCP		75

Εικόνα 4.10 Στο κάτω μέρος της αναφοράς αποτελεσμάτων μιας αναζήτησης BLASTP (ή άλλης αναζήτησης BLAST) περιλαμβάνεται μια σειρά κατά ζεύγη στοιχίσεων όπως αυτές στην Εικόνα 4.5. Χρησιμοποιώντας την επιλογή για αλλαγή μορφοποίησης, τα αποτελέσματα μπορούν να εμφανιστούν ως πολλαπλή στοίχιση, όπως στη συγκεκριμένη περίπτωση μιας ομάδας αλληλουχιών σφαιρινών. Υπάρχουν και άλλες επιλογές μορφοποίησης των αποτελεσμάτων που επιτρέπουν στον χρήστη να επιθεωρήσει περιοχές ομοιότητας και απόκλισης σε μια οικογένεια πρωτεϊνών.

Homo sapiens hemoglobin, epsilon 1 (HBE1), mRNA

Sequence ID: [reflNM_005330.3](#) Length: 816 Number of Matches: 1Range 1: 203 to 705 [GenBank](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
410 bits(454)	5e-113	393/503(78%)	3/503(0%)	Plus/Plus
CDS:hemoglobin subun	1			M V H
Query	3	ATTGCTTCTGACACAACCTGTGTTCACTAGCAACCTCAAA---CAGACACCATGGTGCAT		
Sbjct	203	ATCTGCTTCCGACACAGCTGCAATCACTAGCAAGCTCTCAGGCCTGGCATCATGGTGCAT		
CDS:hemoglobin subun	1			M V H
CDS:hemoglobin subun	4	L T P E E K S A V T A L W G K V N V D E		
Query	60	CTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA		
Sbjct	263	TTTACTGCTGAGGAGAAGGCTGCCGTTACTAGCCTGTGGAGCAAGATGAATGTGGAAGAG		
CDS:hemoglobin subun	4	F T A E E K A A V T S L W S K M N V E E		
CDS:hemoglobin subun	24	V G G E A L G R L L V V Y P W T Q R F F		
Query	120	GTGGTGGTGGAGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGACCCAGAGGTTCTTT		
Sbjct	323	GCTGGAGGTGAAGCCTTGGGCAGACTCCTCGTTGTTTACCCCTGGACCCAGAGATTTTTT		
CDS:hemoglobin subun	24	A G G E A L G R L L V V Y P W T Q R F F		
CDS:hemoglobin subun	44	E S F G D L S T P D A V M G N P K V K A		
Query	180	GAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCT		
Sbjct	383	GACAGCTTTGAAACCTGTCTGCTCCCTCTGCCATCCTGGGCAACCCCAAGGTCAAGGCC		
CDS:hemoglobin subun	44	D S F G N L S S P S A I L G N P K V K A		
CDS:hemoglobin subun	64	H G K K V L G A F S D G L A H L D N L K		
Query	240	CATGGCAAGAAAGTGCTCGGTGCCCTTTAGTGATGGCCCTGGCTCACCTGGACAACCTCAAG		
Sbjct	443	CATGGCAAGAAAGTGCTGACTTCCTTTGGAGATGCTATTAAAAACATGGACAACCTCAAG		
CDS:hemoglobin subun	64	H G K K V L T S F G D A I K N M D N L K		
CDS:hemoglobin subun	84	G T F A T L S E L H C D K L H V D P E N		
Query	300	GGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAAC		
Sbjct	503	CCCGCCTTTGCTAAGCTGAGTGAGCTGCACTGTGACAAGCTGCATGTGGATCCTGAGAAC		
CDS:hemoglobin subun	84	P A F A K L S E L H C D K L H V D P E N		
CDS:hemoglobin subun	104	F R L L G N V L V C V L A H H F G K E F		
Query	360	TTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCATCACTTTGGCAAGAATTTC		
Sbjct	563	TTCAAGCTCCTGGGTAACGTGATGGTGAATTATTCTGGCTACTCACTTTGGCAAGGAGTTC		
CDS:hemoglobin subun	104	F K L L G N V M V I I L A T H F G K E F		
CDS:hemoglobin subun	124	T P P V Q A A Y Q K V V A G V A N A L A		
Query	420	ACCCACCAAGTGAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCC		
Sbjct	623	ACCCCTGAAGTGAGGCTGCCTGGCAGAGCTGGTGTCTGCTGTGCCATTGCCCTGGCC		
CDS:hemoglobin subun	124	T P E V Q A A W Q K L V S A V A I A L A		
CDS:hemoglobin subun	144	H K Y H		
Query	480	CACAAGTATCACTAAGCTCGCTT 502		
Sbjct	683	CATAAGTACCACCTGAGTTCTCTT 705		
CDS:hemoglobin subun	144	H K Y H		

Εικόνα 4.11 Στην περίπτωση αναζητήσεων BLASTN, η επιλογή κωδικής αλληλουχίας (CDS) στη σελίδα αλλαγής μορφοποίησης επιτρέπει την εμφάνιση των αμινοξικών αλληλουχιών που αντιστοιχούν στις κωδικές περιοχές του όρου αναζήτησης και των αλλη-λουχιών από τη βάση δεδομένων με τις οποίες στοιχίζεται η αλληλουχία αναζήτησης. Στη συγκεκριμένη περίπτωση χρη-σιμοποιήθηκε ως όρος το DNA της β-σφαιρίνης του ανθρώπου (NM_000518) και παρουσιάζ-εται μια στοίχιση με τη στενά σχετιζόμενη ε1-σφαιρίνη. Παρα-τίθενται επίσης οι αντίστοιχες αμινοξικές αλληλουχίες. Οι θέσεις που δεν είναι συντη-ρημένες (mismatches) εμφανίζο-νται με μοβ γραμματοσειρά.

Φάση 1: Ρυθμίσεις: Δημιουργία μιας λίστας λέξεων ($w=3$) με σκορ πάνω από το κατώφλι T

- Αλληλουχία αναζήτησης: β-σφαιρίνη ανθρώπου NP_000509.1 (περιλαμβάνει ...VTALWGKVNVD...).
- Ανάγνωση αλληλουχίας, απόρριψη περιοχών χαμηλής πολυπλοκότητας ή άλλο είδος φίλτρου, κατασκευή του πίνακα καταγραφής (lookup table).
- Λέξεις που προκύπτουν από την αλληλουχία αναζήτησης (HBB):

VTALWGKVNVD

- Δημιουργία μιας λίστας λέξεων που ταυτίζονται με τον όρο αναζήτησης (πάνω και κάτω από το κατώφλι).
Εξετάστε τη λέξη LWG στην αλληλουχία αναζήτησης και τα σκορ (εξάγονται από έναν πίνακα BLOSUM62) για διάφορες άλλες λέξεις.

LWG 4+11+6=21

IWG 2+11+6=19

MWG 2+11+6=19

VWG 1+11+6=18

FWG 0+11+6=17

AWG 0+11+6=17

LWS 4+11+0=15

LWN 4+11+0=15

LWA 4+11+0=15

LYG 4+ 2+6=12

LFG 4+ 1+6=11

FWS 0+11+0=11

AWS -1+11+0=10

CWS -1+11+0=10

IWC 2+11-3=10

κατώφλι (=12)

παραδείγματα
λέξεων πάνω
από το κατώφλι
ή ίσων με αυτό

παραδείγματα
λέξεων κάτω
από το κατώφλι

Εικόνα 4.12 Παρουσίαση της λειτουργίας του αλγορίθμου BLAST. Στην αρχική φάση, μια αλληλουχία αναζήτησης (όπως η ανθρώπινη β-σφαιρίνη) αναλύεται σε λέξεις καθορισμένου μεγέθους (π.χ. $w = 3$) και καταρτίζεται ένας κατάλογος λέξεων με σκορ κατωφλιού (π.χ. $T = 11$). Ένας κατάλογος με αρκετές πιθανές λέξεις από την αλληλουχία αναζήτησης παρατίθεται στην εικόνα (από LWG έως IWC). Σε μια αναζήτηση BLAST υπάρχουν 8.000 λέξεις για $w = 3$. Για μια δεδομένη λέξη, όπως για παράδειγμα η αλληλουχία LWG, καταρτίζεται ένας κατάλογος λέξεων με σκορ μεγαλύτερο ή ίσο με κάποιο κατώφλι T (π.χ. 12). Σε αυτό το παράδειγμα εμφανίζονται 15 λέξεις μαζί με τις βαθμολογίες τους από τον πίνακα BLOSUM62. Δέκα από αυτές είναι πάνω από το κατώφλι και πέντε κάτω από αυτό.

Φάση 2: Σάρωση και επέκταση

- Επιλογή λέξεων πάνω από το κατώφλι T (LWG, IWG, MWG, VWG, FWG, AWG, LWS, LWN, LWA, LYG).
- Σάρωση της βάσης δεδομένων για καταχωρίσεις («επιτυχίες») που ταιριάζουν με τη λίστα λέξεων.
- Δημιουργία ενός πίνακα κατακερματισμού (hash table) με τις θέσεις όλων των επιτυχιών για κάθε λέξη.
- Εκτέλεση στοίχισης χωρίς κενά.
- Εκτέλεση στοίχισεων με εισαγωγή κενών.

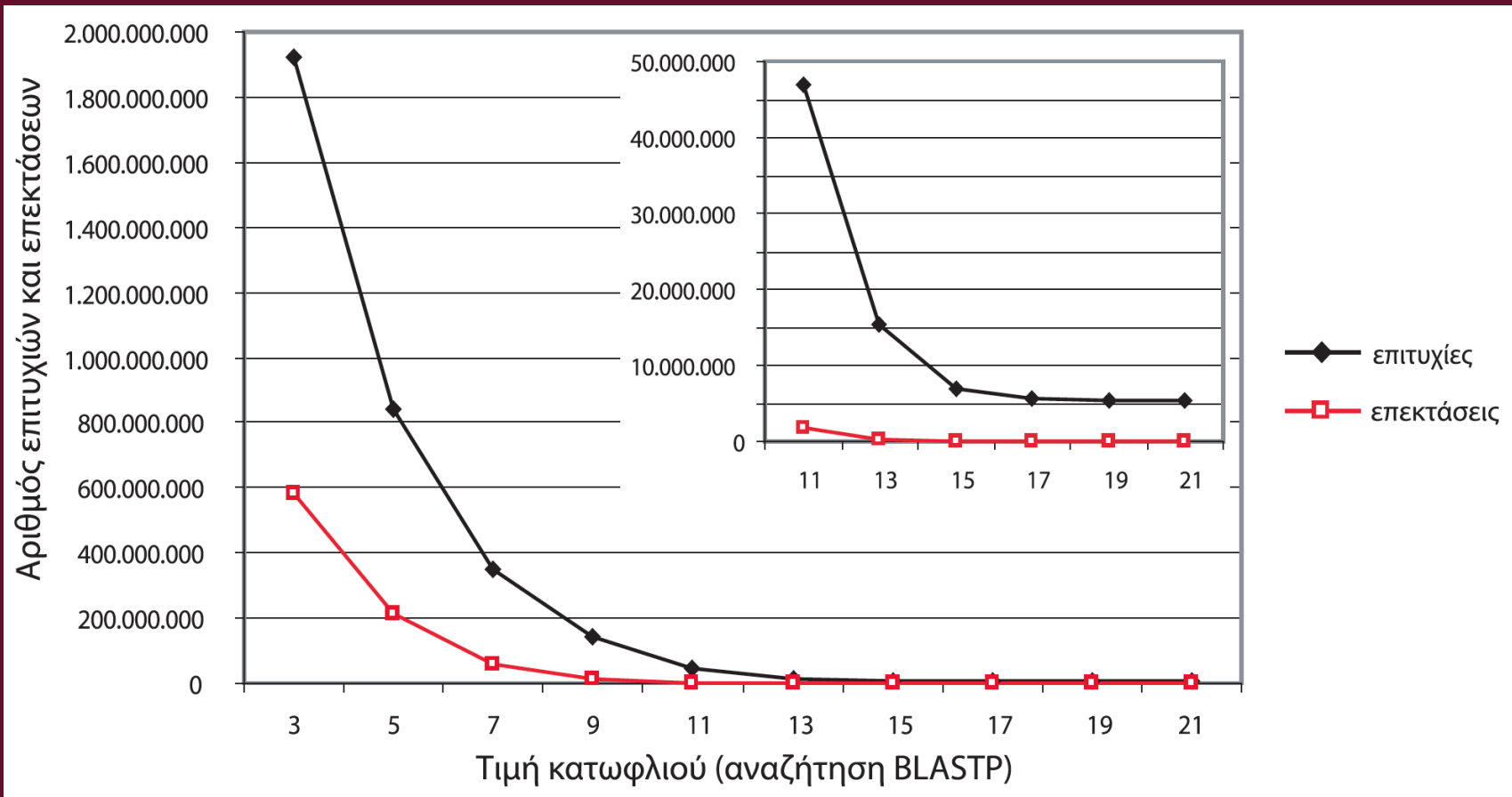
```
LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV HBB
L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F D G+ +V
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-----DLSHGSAQV HBA
```

← επέκταση επέκταση →
ένα ζευγάρι λέξης με την αλληλουχία
της α-σφαιρίνης στην πρώτη φάση αναζήτησης
το οποίο εκκινεί την επέκταση

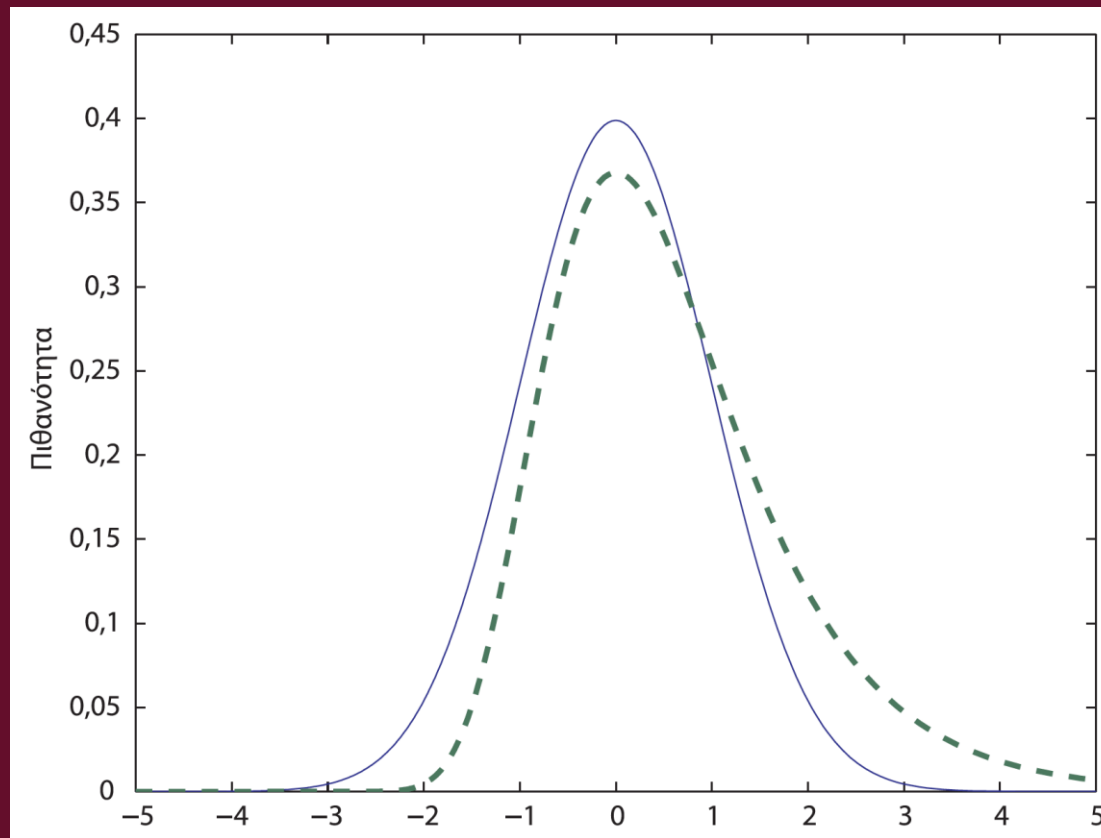
Φάση 3: Έλεγχος επιβεβαίωσης

- Υπολογισμός των προσθηκών, απαλοιφών και αναντιστοιχιών (για στοίχισης που προέκυψαν από τη φάση 2).
- Εφαρμογή στατιστικής ανάλυσης με βάση τη σύσταση (για BLASTP, TBLASTN).
- Δημιουργία στοίχισης με εισαγωγή κενών.

Εικόνα 4.12 Στη φάση 2 γίνεται σάρωση μιας βάσης δεδομένων για να βρεθούν καταχωρίσεις που ταιριάζουν με τη λίστα των λέξεων. Οι στοίχισεις που προκύπτουν επεκτείνονται και προς τις δύο κατευθύνσεις με ή χωρίς εισαγωγή κενών, αλλά οι θέσεις στοίχισης δεν αποθηκεύονται (για να μη μειώνεται η απόδοση). Οι επιτυχείς στοίχισεις στη βάση δεδομένων επεκτείνονται και προς τις δύο κατευθύνσεις για να ανιχνευθούν ζεύγη τμημάτων υψηλού σκορ (HSP). Αν το σκορ ενός HSP υπερβαίνει ένα συγκεκριμένο κατώφλι S , τότε αυτό περιλαμβάνεται στα αποτελέσματα του BLAST. Στη φάση 3 πραγματοποιείται αναδρομικός έλεγχος και καταγράφονται οι θέσεις των προσθηκών, των απαλοιφών και των αναντιστοιχιών. Παρατηρήστε ότι στο συγκεκριμένο παράδειγμα το ζεύγος λέξεων που εκκινεί το στάδιο επέκτασης δεν περιέχει μια ακριβή ταύτιση (δηλαδή μόνο δύο κατάλοιπα στο πλαίσιο LWG στοιχίζονται με την αλληλουχία AWG). Η βασική αρχή της χρήσης ενός κατωφλίου T για τις αναζητήσεις πρωτεϊνών είναι ότι επιτρέπεται η εκκίνηση της επέκτασης τόσο από ακριβείς ταυτίσεις όσο και από σχετικές ομοιότητες. Για τις αναζητήσεις νουκλεοτιδίων BLASTN απαιτούνται απόλυτες ταυτίσεις και όχι λέξεις που έχουν σκορ πάνω από ένα όριο.



Εικόνα 4.13 Η επίδραση της μεταβολής της τιμής του κατωφλίου (άξονας x) στον αριθμό των αποτελεσμάτων από την αναζήτηση στη βάση δεδομένων (μαύρη γραμμή) και του πλήθους των επεκτάσεων (κόκκινη γραμμή). Οι αναζητήσεις BLASTP εκτελέστηκαν χρησιμοποιώντας ως αλληλουχία αναζήτησης την ανθρώπινη β-σφαιρίνη.



Εικόνα 4.14 Σύγκριση της κανονικής κατανομής (συνεχής γραμμή) με την κατανομή ακραίων τιμών (διακεκομμένη γραμμή). Από τη σύγκριση μιας αλληλουχίας αναζήτησης με ένα σύνολο τυχαίων αλληλουχιών ίδιου μήκους προκύπτουν σκορ που ακολουθούν την κατανομή ακραίων τιμών (και όχι την κανονική κατανομή). Το εμβαδόν κάτω από την καμπύλη ισούται με 1. Για την κανονική κατανομή, η μέση τιμή (μ) εντοπίζεται στο μηδέν, ακριβώς στο κέντρο της καμπύλης, και η πιθανότητα Z να επιτευχθεί ένα σκορ x δίνεται σε μονάδες τυπικής απόκλισης (σ) του x από τη μέση τιμή: $Z = (x - \mu)/\sigma$. Σε αντίθεση με την κανονική κατανομή, η κατανομή ακραίων τιμών είναι μη συμμετρική και κλίνει προς τα δεξιά (θετική λοξότητα). Η κατανομή ακραίων τιμών ακολουθεί τη συνάρτηση $f(x) = \exp(-\exp(-x))$. Το σχήμα της κατανομής ακραίων τιμών καθορίζεται από τη χαρακτηριστική τιμή u και τη σταθερά απομείωσης λ ($u = 0$ και $\lambda = 1$)

Πίνακας 4.3 Σχέση μεταξύ των τιμών E και p στο πρόγραμμα BLAST με βάση την εξίσωση (4.8). Οι μικρές τιμές E (0,05 ή μικρότερες) σχετίζονται στενά με τις τιμές p .

E	p
10	0,99995460
5	0,99326205
2	0,86466472
1	0,63212056
0,1	0,09516258
0,05	0,04877058
0,001	0,00099950
0,0001	0,0001000

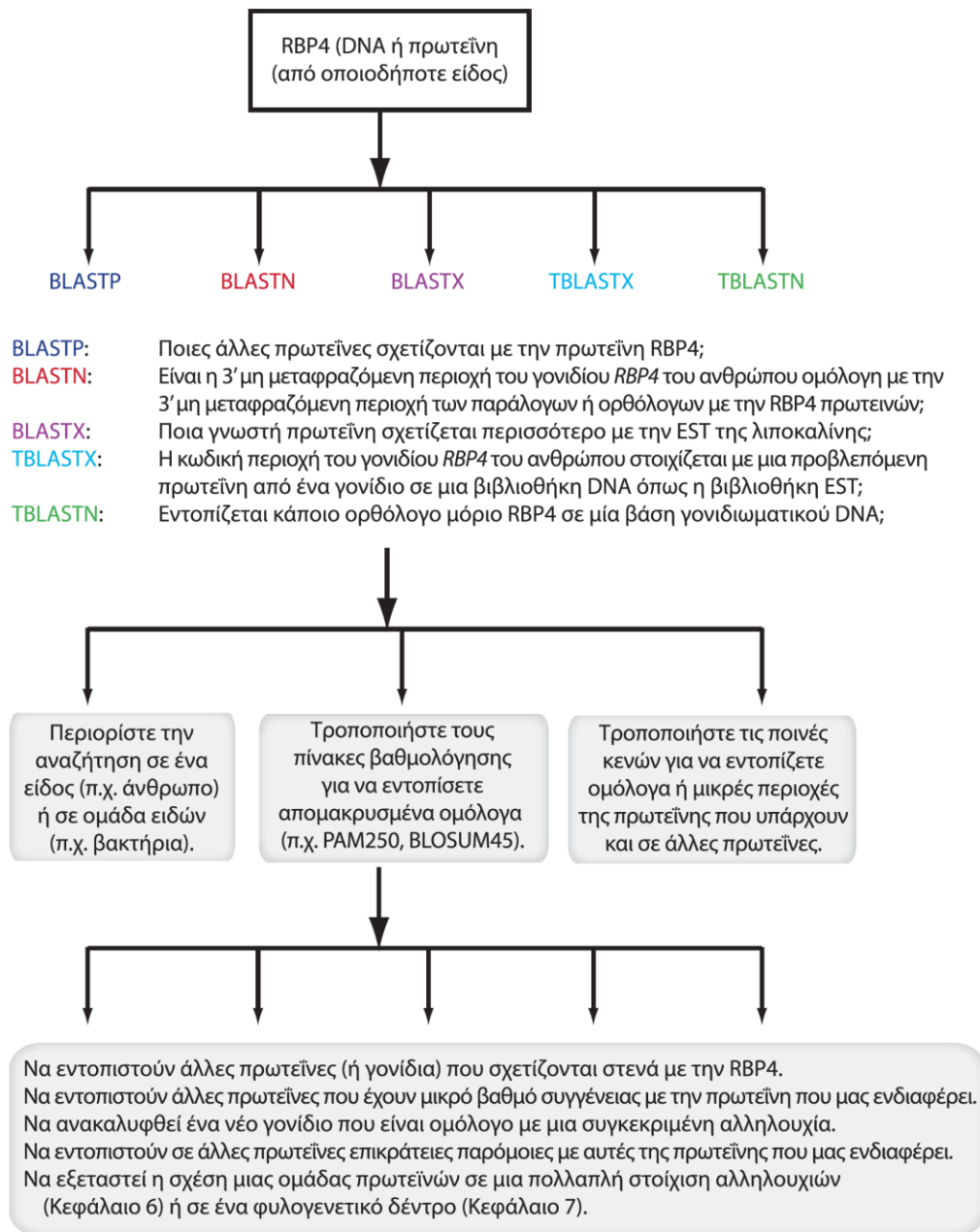
Σημείο εκκίνησης:
μια μοριακή
αλληλουχία

Στρατηγικές
αναζήτησης

Χαρακτηριστικά
ερωτήματα

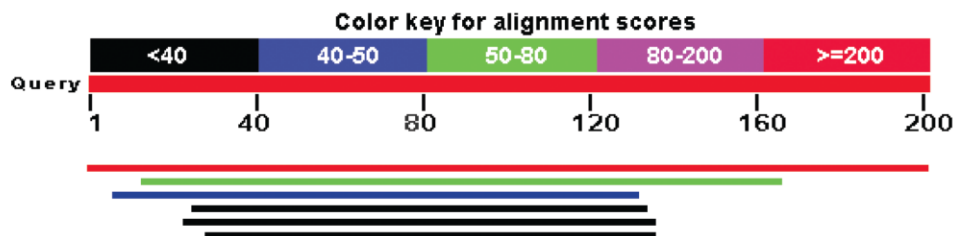
Ρυθμιζόμενες
παράμετροι
αναζήτησης

Στόχοι:
τα αποτελέσματα
της αναζήτησης
BLAST



Εικόνα 4.15 Επισκό-πηση των στρατηγικών αναζήτησης BLAST. Πάρα πολλά είδη προβλημάτων μπορούν να απαντηθούν με τα διά-φορα είδη αναζητήσεων BLAST, από τον χαρα-κτηρισμό του γονιδιώ-ματος ενός οργανισμού έως την αξιολόγηση των παραλλαγών αλληλου-χίας ενός γονιδίου.

(α) Συνοπτική γραφική απεικόνιση των αποτελεσμάτων



(β) Λίστα στοιχίσεων

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 6

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input checked="" type="checkbox"/>	retinol-binding protein 4 precursor [Homo sapiens]	420	420	100%	1e-150	100%	NP_006735.2
<input checked="" type="checkbox"/>	apolipoprotein D precursor [Homo sapiens]	55.5	55.5	76%	1e-09	28%	NP_001638.1
<input checked="" type="checkbox"/>	glycodelin precursor [Homo sapiens] >reflNP_002562.2 glycodelin precursor [Homo s	40.0	40.0	62%	5e-04	26%	NP_001018059.1
<input checked="" type="checkbox"/>	protein AMBP preproprotein [Homo sapiens]	35.0	35.0	54%	0.034	23%	NP_001624.1
<input checked="" type="checkbox"/>	complement component C8 gamma chain precursor [Homo sapiens]	32.3	32.3	56%	0.18	25%	NP_000597.2
<input checked="" type="checkbox"/>	lipocalin-15 precursor [Homo sapiens]	28.5	28.5	53%	3.4	23%	NP_976222.1

(γ) Στοιχισμός κατά ζεύγη της RBP4 και της C8G

complement component C8 gamma chain precursor [Homo sapiens]

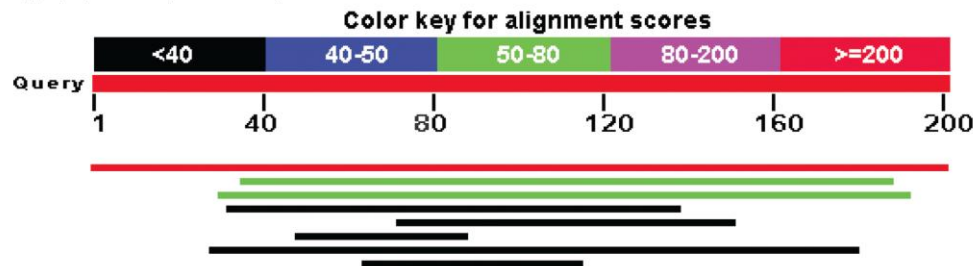
Sequence ID: [reflNP_000597.2](#) Length: 202 Number of Matches: 1

Range 1: 33 to 139 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
32.3 bits(72)	0.18	Compositional matrix adjust.	28/114(25%)	49/114(42%)	8/114(7%)
Query 24	VSSFRVKNFDKARFSSTWYAMAKKDPEGLFLQDNIVAEFSVDETG-QMSATAKGRVRL				82
Sbjct 33	ISTIQPKANFDAQQFAGTNLLVAVGSACRFLQEQGHRAEATTLHVAPQGTAMAVSTFRKL				92
Query 83	NNWDVCAADMVGTFTDTEPAKFKMKYWGVSFLQKGNDDHWIVDTIDYDIYAVQY				136
Sbjct 93	DG--ICWQVRQLYGDIGVLGRFLLQARDA-----RGAVHVVVAETDYQSFVLY				139

Εικόνα 4.16 Αποτελέσματα αναζήτησης BLASTP χρησιμοποιώντας ως όρο αναζήτησης την ανθρώπινη πρωτεΐνη RBP στη βάση δεδομένων nr και με περιορισμό στις πρωτεΐνες RefSeq του ανθρώπου. (α) Στη γραφική παρουσίαση των αποτελεσμάτων περιλαμβάνονται 6 επιτυχίες, αλλά μόνο η μία (η ίδια η RBP4) έχει υψηλό σκορ (κόκκινη ράβδος) και εκτείνεται σε όλο το μήκος της αλληλουχίας αναζήτησης. (β) Στα αποτελέσματα του BLASTP περιλαμβάνεται μια λίστα στοιχίσεων. Από την εξέταση των τιμών *E* προκύπτει περαιτέρω ότι, εκτός από την ίδια την RBP, μπορεί να έχουν εντοπιστεί αρκετά αυθεντικά παράλογα πρωτεϊνικά μόρια. Σε αυτά περιλαμβάνεται ο παράγοντας συμπληρώματος 8γ (C8G), με τιμή *E* ίση με 0,18. Είναι άραγε η πρωτεΐνη C8G ομόλογη με την RBP4; (γ) Από τη στοίχιση κατά ζεύγη των RBP4 και C8G, που περιλαμβάνεται στα αποτελέσματα του BLASTP, προκύπτει 25% ταύτιση αμινοξέων και στοίχιση ενός μοτίβου GXW (κόκκινο ορθογώνιο) που είναι σταθερά συντηρημένο σε όλες τις λιποκαλίνες (π.χ. στην RBP4).

(α) Συνοπτική γραφική απεικόνιση των αποτελεσμάτων



(β) Λίστα στοιχίσεων

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input type="checkbox"/>	complement component C8 gamma chain precursor [Homo sapiens]	412	412	100%	3e-147	100%	NP_000597.2
<input type="checkbox"/>	lipocalin-15 precursor [Homo sapiens]	69.7	69.7	76%	1e-14	34%	NP_976222.1
<input type="checkbox"/>	protein AMBP preproprotein [Homo sapiens]	68.9	68.9	80%	1e-13	25%	NP_001624.1
<input type="checkbox"/>	retinol-binding protein 4 precursor [Homo sapiens]	33.1	33.1	52%	0.12	25%	NP_006735.2
<input type="checkbox"/>	tenascin-X isoform 1 precursor [Homo sapiens] ← Not homologous	30.0	30.0	39%	1.5	31%	NP_061978.6
<input type="checkbox"/>	neuroblastoma-amplified sequence [Homo sapiens] ← Not homologous	29.6	29.6	20%	2.1	44%	NP_056993.2
<input type="checkbox"/>	neutrophil gelatinase-associated lipocalin precursor [Homo sapiens]	28.9	28.9	75%	2.9	21%	NP_005555.2
<input type="checkbox"/>	HBS1-like protein isoform 1 [Homo sapiens] ← Not homologous	28.5	28.5	25%	5.4	33%	NP_006611.1

Εικόνα 4.17 Αποτελέσματα αναζήτησης BLASTP στη βάση μη επαναλαμβανόμενων πρωτεϊνικών αλληλουχιών με περιορισμό στις ανθρώπινες πρωτεΐνες και με αλληλουχία αναζήτησης τον ανθρώπινο παράγοντα συμπληρώματος 8γ (C8G). (α) Στη γραφική απεικόνιση των αποτελεσμάτων περιλαμβάνονται 8 στοιχίσεις. Μία από αυτές αφορά την ίδια την πρωτεΐνη C8G (κόκκινη ράβδος, υψηλό σκορ), ενώ υπάρχουν αρκετές στοιχίσεις με χαμηλό σκορ (μαύρες ράβδοι) που εκτείνονται σε μικρές περιοχές της αμινοξικής αλληλουχίας. (β) Στη λίστα των στοιχίσεων περιλαμβάνεται η RBP4 και άλλα μέλη της οικογένειας των λιποκαλινών. Αυτή η «αντίστροφη» αναζήτηση υποστηρίζει την υπόθεση ότι ο παράγοντας C8G, που ταυτοποιήθηκε ως ομόλογη πρωτεΐνη με την RBP4 σε προηγούμενη αναζήτηση στην οποία ο όρος αναζήτησης ήταν η RBP4, είναι αυθεντικό ομόλογό της. Στην παρούσα αναζήτηση, τρεις στοιχίσεις δεν αντιστοιχούν σε ομόλογες αλληλουχίες (βέλη). Οι τιμές *E* γι' αυτές τις στοιχίσεις είναι πολύ υψηλές και οι πρωτεΐνες αυτές ανήκουν σε άλλες πρωτεϊνικές οικογένειες και όχι στις λιποκαλίνες (όπως μπορεί να επιβεβαιωθεί από ξεχωριστές αναζητήσεις BLAST).

(γ) Αποτελέσματα στοίχισης με μη ομόλογες πρωτεΐνες

Download
GenPept
Graphics

tenascin-X isoform 1 precursor [Homo sapiens]

Sequence ID: [ref|NP_061978.6|](#) Length: 4242 Number of Matches: 1

Range 1: 3255 to 3330 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
30.0 bits(66)	1.5	Compositional matrix adjust.	25/81(31%)	36/81(44%)	6/81(7%)
Query 73	TTLHVAPQGTAMAVSTFRKLD-GICWQVRQLYGDTGVLGRFLLQARDARGAVHVVVAETD	131			
Sbjct 3255	TPLPVEPRLGELAVAAVTSDSVGLSWTVAQ-----GPFDSFLVQYRDAQGQPQAVFVSGD	3309			
Query 132	YQSFAVLYLERAGQLSVKLYA	152			
Sbjct 3310	LRAVAVSGLDPARKYKFLFLFG	3330			

Download
GenPept
Graphics

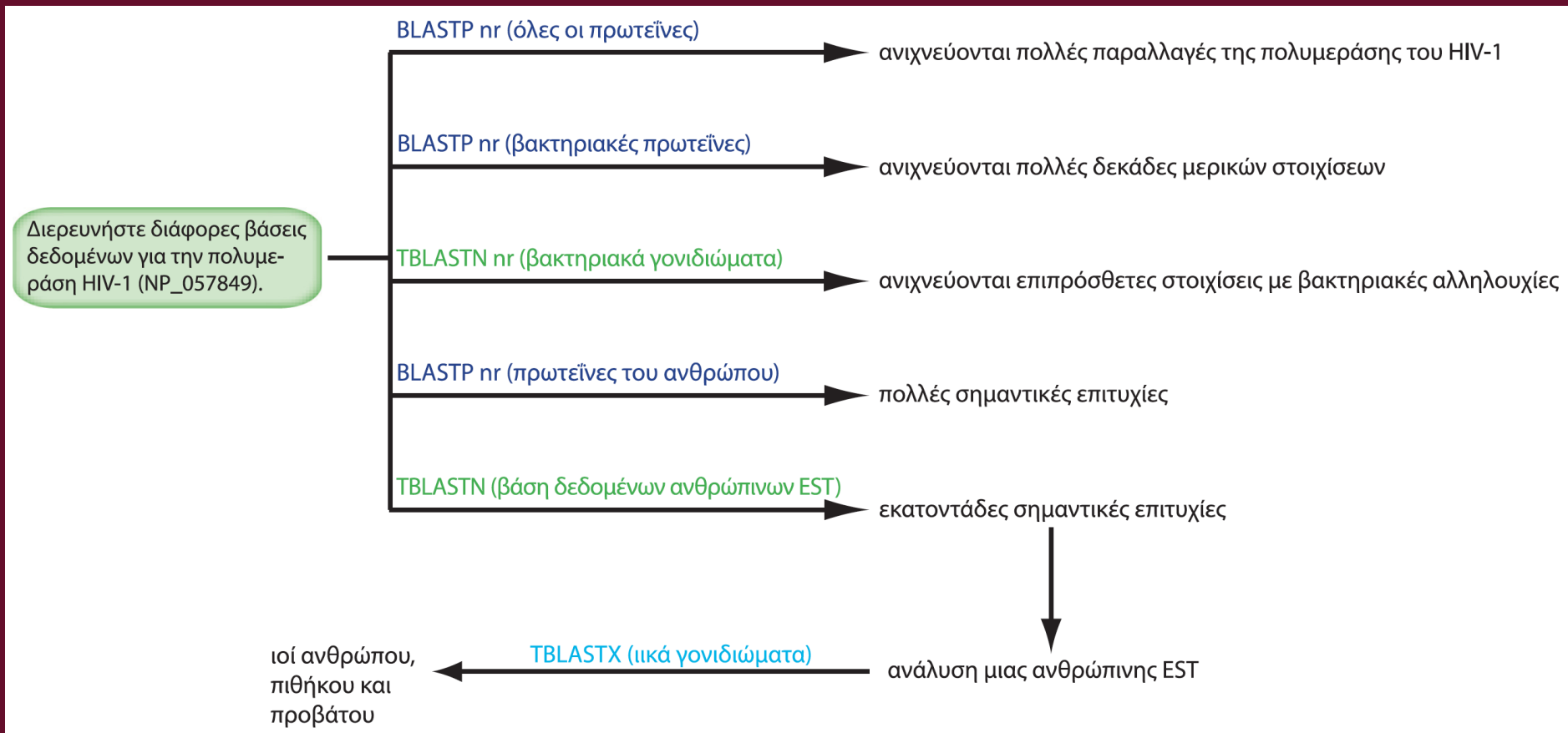
neuroblastoma-amplified sequence [Homo sapiens]

Sequence ID: [ref|NP_056993.2|](#) Length: 2371 Number of Matches: 1

Range 1: 2323 to 2360 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

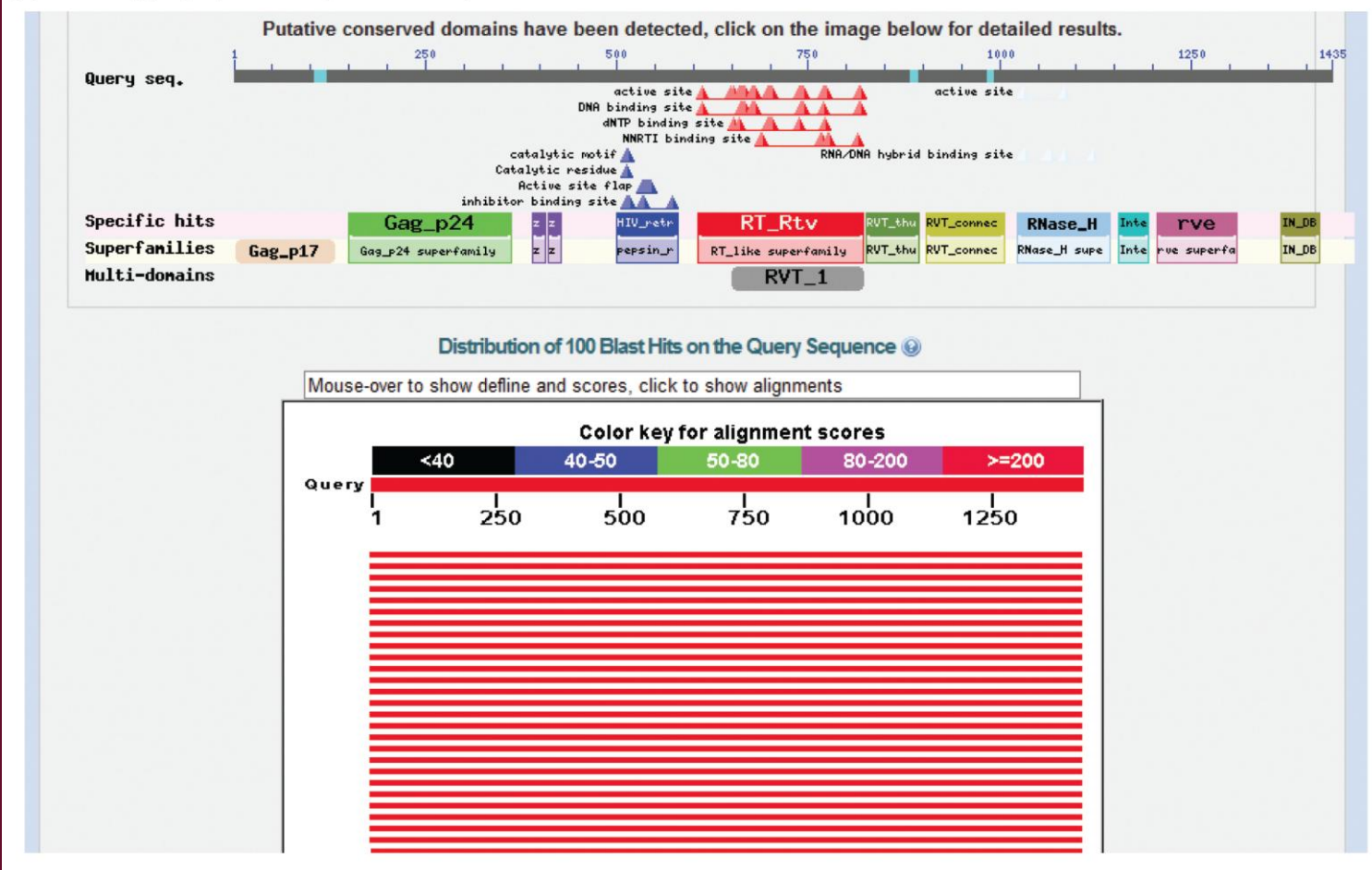
Score	Expect	Method	Identities	Positives	Gaps
29.6 bits(65)	2.1	Compositional matrix adjust.	18/41(44%)	23/41(56%)	3/41(7%)
Query 49	GTWLLVAVGSACRFLQEQGHRAEATTTLHVAPQGTAMAVSTF	89			
Sbjct 2323	GRWDAEELG---RHLREAGHEAEAGSLLLAVRGTHQAFRTF	2360			

Εικόνα 4.17 (γ) Από την επισκόπηση των στοιχίσεων κατά ζεύγη μεταξύ του C8G και δύο υποθετικών μη ομόλογων πρωτεϊνών προκύπτει ότι αυτές οι πρωτεΐνες είναι πολύ μεγαλύτερες από τις τυπικές λιποκαλίνες (4.242 και 2.371 κατάλοιπα αμινοξέων). Η ισομορφή 1 της τενασκίνης X (tenascin X isoform 1) δε στοιχίζεται με το εξαιρετικά συντηρημένο μοτίβο GXW. Η ενισχυόμενη στο νευροβλάστωμα αλληλουχία (neuroblastoma-amplified sequence) επίσης δε στοιχίζεται με το μοτίβο GXW και η περιοχή στοίχισης εκτείνεται σε μόνο 41 κατάλοιπα αμινοξέων. Τα παραπάνω αποτελέσματα υπογραμμίζουν την ανάγκη να εξετάζονται οι κατά ζεύγη στοιχίσεις μετά από μια αναζήτηση BLAST. Η τιμή E παρέχει ένα στατιστικό επιχείρημα για την αξιολόγηση της πιθανής ομολογίας, αλλά θα πρέπει να συμπληρώνεται από την εξέταση των βιολογικών ιδιοτήτων των πρωτεϊνών. Σε αυτό το παράδειγμα, η RBP4, ο παράγοντας C8G και οι άλλες λιποκαλίνες είναι υδατοδιαλυτές, υδρόφιλες πρωτεΐνες που συναντώνται σε υψηλές συγκεντρώσεις και πιθανώς επιτελούν παρόμοιες λειτουργίες ως πρωτεΐνες-μεταφορείς. Οι πρωτεΐνες αυτές εμφανίζουν επίσης παρόμοιες τρισδιάστατες δομές (βλ. Κεφάλαιο 13).



Εικόνα 4.18 Επισκόπηση αναζητήσεων BLAST με κοινό όρο αναζήτησης την πρωτεΐνη Pol του ιού HIV-1. Συχνά διεξάγουμε μια σειρά αναζητήσεων BLAST για να απαντήσουμε σε ζητήματα που αφορούν ένα συγκεκριμένο γονίδιο, μια πρωτεΐνη ή έναν οργανισμό. Ο αριθμός των στοιχίσεων με αλληλουχίες της βάσης δεδομένων από μια αναζήτηση BLAST ποικίλλει ευρέως, από καμία έως χιλιάδες στοιχίσεις, και εξαρτάται από τη φύση του όρου αναζήτησης, τη βάση δεδομένων και τις παραμέτρους αναζήτησης.


(α) Συνοπτική γραφική απεικόνιση των αποτελεσμάτων






Εικόνα 4.19 Αναζήτηση BLASTP για την πρωτεΐνη Pol του HIV-1 (NP_057849). (α) Στη γραφική απεικόνιση των αποτελεσμάτων διακρίνονται οι συντηρημένες επικράτειες της πρωτεΐνης. Οι ράβδοι των συντηρημένων επικρατειών παρέχουν συνδέσμους παραπομπής στη βάση Conserved Domain Database του NCBI (Κεφάλαια 5 και 6). Οι σύνδεσμοι αντιστοιχούν στις πρωτεϊνικές επικράτειες (Gag_p17, Gag_p24) και οι συντομογραφίες περιλαμβάνουν την rnr (ρετροϊκή ασπαρτική πρωτεάση), την rnt (αντίστροφη μεταγραφάση ή RNA εξαρτώμενη DNA πολυμεράση), την RNaseH (ριβονουκλεάση H) και την rne (κεντρική επικράτεια ιντεγκράσης). Οι κόκκινες οριζόντιες ράβδοι αντιστοιχούν σε περιοχές στενών στοιχίσεων με ιικές πρωτεΐνες.



(β) Λίστα στοιχίσεων (με αγκύρωση στην αλληλουχία αναζήτησης – οι τελείες υποδηλώνουν ταυτίσεις)


Query	1	MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQI	60
NP_057849	1	60
P0C6F2	1K.....	60
P03366	1	60
P03367	1	60
P04587	1	60
AAD03191	1Q.R.....	60
P35963	1A...K.....Q.R.....D.....	60
P12497	1K.....Q.....	60
P20875	1R.....S.....	60
AAD03200	1R...R...Q.....S.....	60
P20892	1K.....Q.....I.....	60
Q73368	1S.....	60
BAB85751	1Q.....M.....	60
AFB39387	1Q.....R.....A.....	60
P03369	1K.....	60
P05959	1K.K.....R.R.....S.A.....	60
AAG30116	1I.....K.....R..L.....Q..I.....A.....	60
AAD03217	1I.....Q.....	60


R


R,K

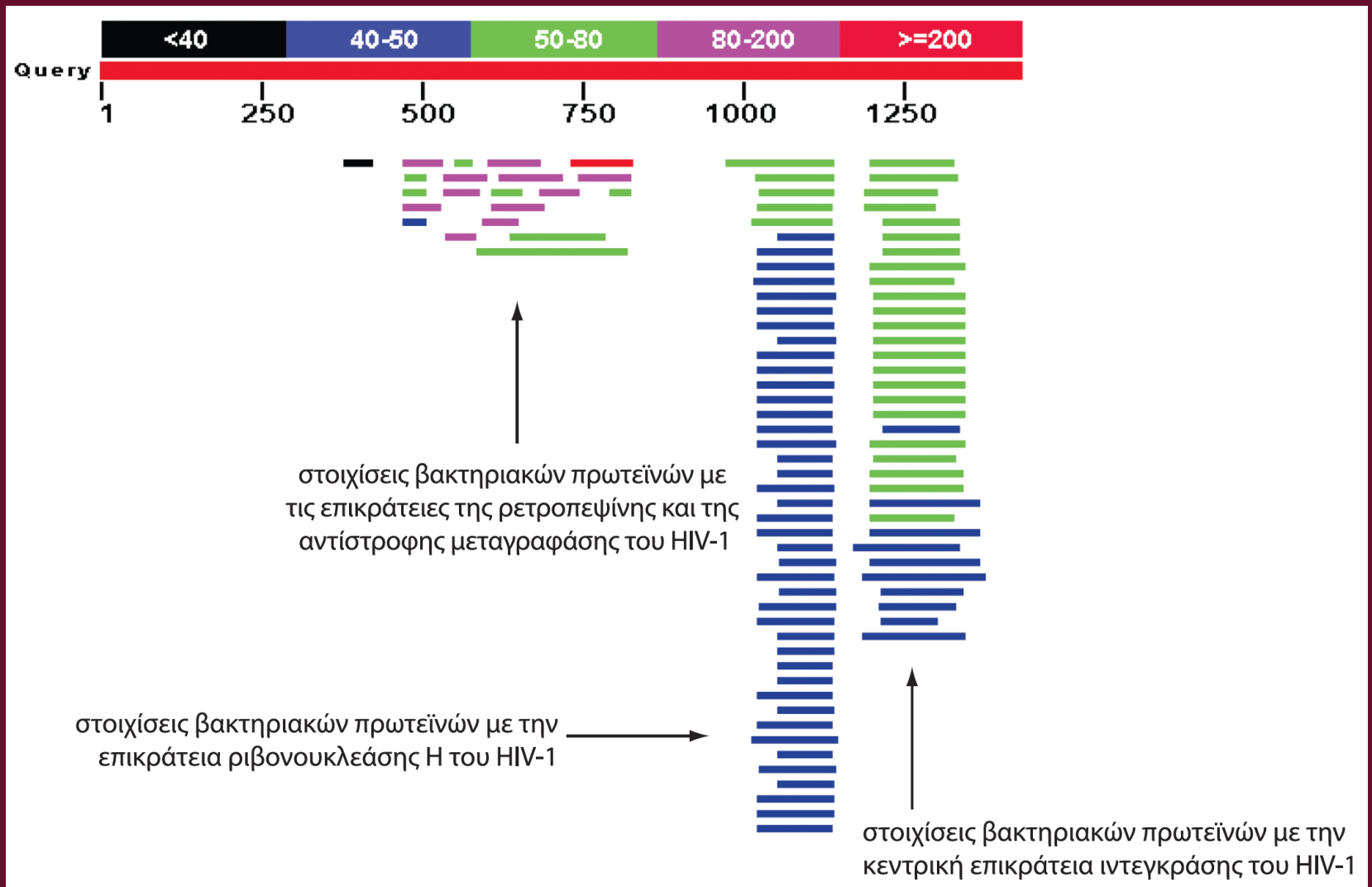


R R,Q


R

Εικόνα 4.19 (β) Οι επιλογές του BLAST επιτρέπουν διάφορες μορφές παρουσίασης της στοίχισης, όπως είναι η αγκυρωμένη στον όρο αναζήτησης διαμόρφωση, στην οποία οι κουκκίδες αντιστοιχούν σε συντηρημένες αμινοξικές θέσεις που στοιχίζονται με την αλληλουχία αναζήτησης. Σε αυτή τη μορφή παρουσίασης τονίζονται οι διαφορές των αλληλουχιών των πρωτεϊνών. Τα βέλη δείχνουν κατάλοιπα αργινίνης (R) στην αναζήτηση που είναι απολύτως συντηρημένα ή που αντικαθίστανται από λυσίνη (K) ή από γλουταμίνη (Q).

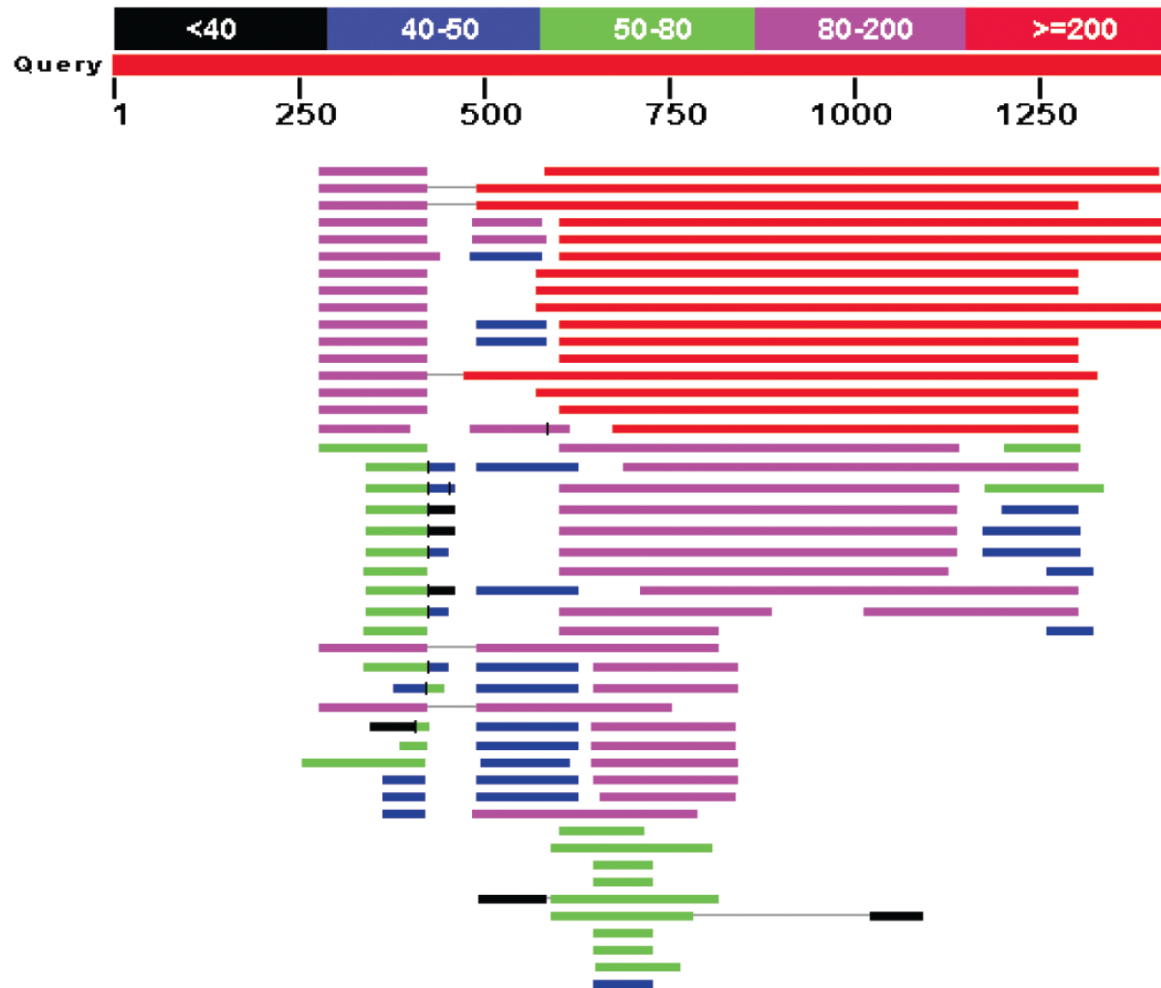
Human immunodeficiency virus 1 [viruses] taxid 11676		
ref NP_057849.4	Gag-Pol [Human immunodeficiency virus 1]	2971 0.0
ref NP_789740.1	Pol [Human immunodeficiency virus 1]	2052 0.0
ref NP_705927.1	reverse transcriptase [Human immunodeficiency virus 1]	1149 0.0
ref YP_001856242.1	reverse transcriptase [Human immunodeficiency virus 1]	1149 0.0
ref NP_789739.1	reverse transcriptase p51 subunit [Human immunodeficiency virus 1]	912 0.0
ref NP_057850.1	Pr55(Gag) [Human immunodeficiency virus 1]	908 0.0
ref NP_705928.1	integrase [Human immunodeficiency virus 1]	602 0.0
ref YP_001856243.1	integrase [Human immunodeficiency virus 1]	602 0.0
ref NP_579880.1	capsid [Human immunodeficiency virus 1]	481 4e-156
ref NP_579876.2	matrix [Human immunodeficiency virus 1]	271 7e-81
ref NP_705926.1	retropepsin [Human immunodeficiency virus 1]	204 2e-57
ref YP_001856241.1	retropepsin [Human immunodeficiency virus 1]	204 2e-57
ref NP_579881.1	nucleocapsid [Human immunodeficiency virus 1]	130 5e-32
ref NP_787043.1	Gag-Pol Transframe peptide [Human immunodeficiency virus 1]	119 4e-28
Simian immunodeficiency virus [viruses] taxid 11723		
ref NP_687035.1	Gag-Pol [Simian immunodeficiency virus]	1687 0.0
ref NP_054369.1	gag protein [Simian immunodeficiency virus]	502 1e-159
Human immunodeficiency virus 2 [viruses] taxid 11709		
ref NP_663784.1	gag-pol fusion polyprotein [Human immunodeficiency virus 2]	1675 0.0
ref NP_056837.1	gag polyprotein [Human immunodeficiency virus 2]	523 3e-167
Simian immunodeficiency virus SIV-mnd 2 [viruses] taxid 159122		
ref NP_758887.1	pol protein [Simian immunodeficiency virus SIV-mnd 2]	1377 0.0
ref NP_758886.1	gag protein [Simian immunodeficiency virus SIV-mnd 2]	486 2e-153
Feline immunodeficiency virus [viruses] taxid 11673		
ref NP_040973.1	pol polyprotein [Feline immunodeficiency virus]	489 2e-148
ref NP_040972.1	gag protein [Feline immunodeficiency virus]	158 8e-38
Equine infectious anemia virus [viruses] taxid 11665		
ref NP_056902.1	pol polyprotein [Equine infectious anemia virus]	424 1e-123
ref NP_056901.1	gag protein [Equine infectious anemia virus]	154 2e-36
Candida albicans SC5314 [ascomycetes] taxid 237561		
ref XP_888860.1	hypothetical protein Ca019_6468 [Candida albicans SC5314]	90 2e-15
ref XP_721310.1	hypothetical protein Ca019.6468 [Candida albicans SC5314]	86 1e-14
Sus scrofa (wild boar, ...) [even-toed ungulates] taxid 9823		
ref XP_003482346.1	PREDICTED: hypothetical protein LOC100348234	90 2e-15
Tribolium castaneum (rust-red flour beetle) [beetles] taxid 7070		
ref XP_001815322.1	PREDICTED: similar to orf [Tribolium castaneum]	89 5e-15
ref XP_001808495.1	PREDICTED: similar to orf [Tribolium castaneum]	88 8e-15
Candida dubliniensis CD36 [ascomycetes] taxid 573826		
ref XP_002421195.1	retrovirus-related Pol polyprotein from Candida dubliniensis	88 6e-15
Moniliophthora perniciosa FA553 [basidiomycetes] taxid 554373		
ref XP_002387985.1	hypothetical protein MPER_13056 [Moniliophthora perniciosa]	88 7e-15

Εικόνα 4.20 Η σελίδα ταξινομικής αναφοράς για μια αναζήτηση BLASTP μας παρέχει πληροφορίες για τα διάφορα είδη οργανισμών που φέρουν ομόλογες πρωτεΐνες με την πρωτεΐνη αναζήτησης από τον HIV-1. Οι περισσότερες στοιχίσεις αφορούν άλλες ιικές πρωτεΐνες, αλλά κάποιες αφορούν πρωτεΐνες άλλων ειδών, όπως του κουνελιού, των μυκήτων, του χοίρου και των εντόμων. Τα σύμβολα /// υποδηλώνουν ότι υπάρχει μια σειρά από άλλες στοιχίσεις που δεν απεικονίζονται.



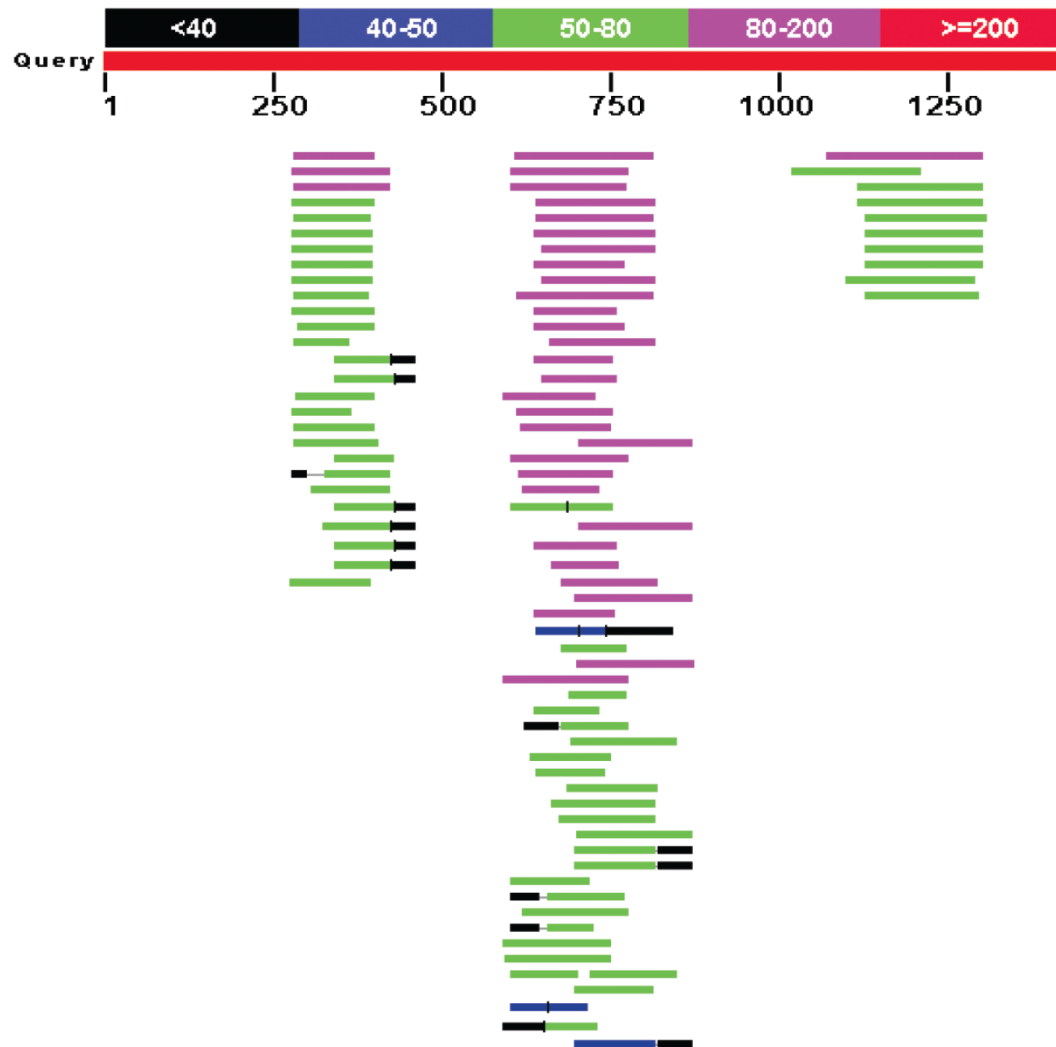
Εικόνα 4.21 Αποτέλεσμα αναζήτησης BLASTP με όρο αναζήτησης την Pol του HIV-1 και με περιορισμό των αποτελεσμάτων στα βακτήρια. Η γραφική απεικόνιση των αποτελεσμάτων επιτρέπει την ταυτοποίηση των επικρατειών του HIV-1 που έχουν ομόλογες βακτηριακές αλληλουχίες. Φαίνονται επίσης τα μήκη της έκτασης των στοιχίσεων και το πλήθος των στοιχισμένων βακτηριακών αλληλουχιών.

(α) Αναζήτηση BLASTP για την Pol του HIV-1 στη βάση μη επαναλαμβανόμενων αλληλουχιών πρωτεϊνών του ανθρώπου

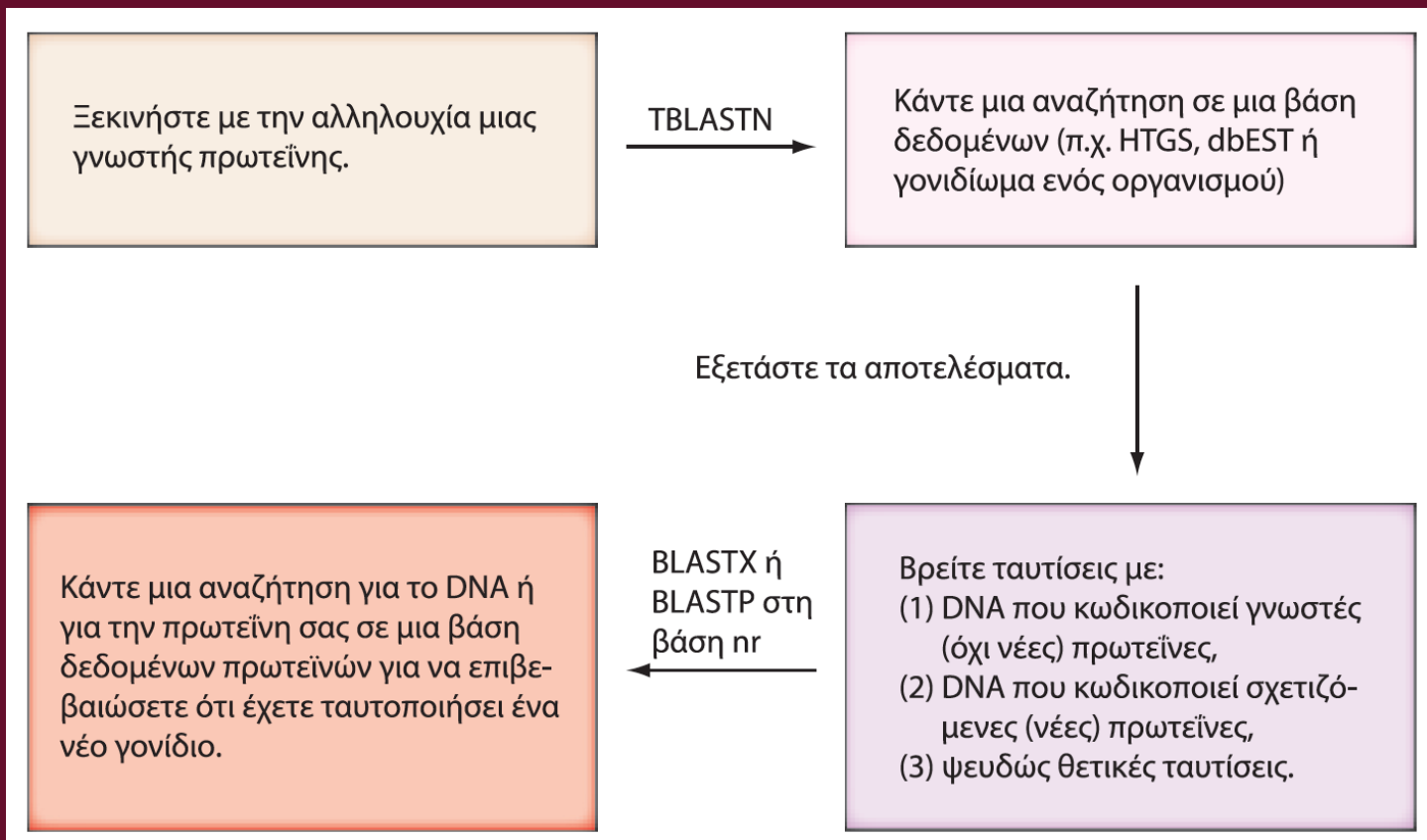


Εικόνα 4.22 (α) Γραφική απεικόνιση των αποτελεσμάτων της αναζήτησης BLASTP με όρο την πρωτεΐνη Pol. Η αναζήτηση στοιχίσεων περιορίστηκε σε ανθρώπινες πρωτεΐνες. Ορισμένες επιτυχίες έχουν πολύ υψηλό σκορ.

(β) Αναζήτηση TBLASTN για την Pol του HIV-1 στη βάση ανθρώπινων ετικετών εκφραζόμενης αλληλουχίας



Εικόνα 4.22 (β) Εκφράζονται άραγε στον άνθρωπο μετάγραφα που κωδικοποιούν πρωτεΐνες ομόλογες με την πρωτεΐνη Pol του HIV-1; Παρουσιάζονται τα αποτελέσματα αναζήτησης TBLASTN για την ιική πρωτεΐνη Pol με περιορισμό της αναζήτησης σε EST του ανθρώπου. Πολλά ανθρώπινα γονίδια μεταγράφονται ενεργά και από τα μετάγραφά τους προκύπτουν πρωτεΐνες που είναι δυνητικά ομόλογες με την Pol του HIV-1.



Εικόνα 4.23 Πώς να ανακαλύψετε ένα νέο γονίδιο χρησιμοποιώντας αναζητήσεις BLAST. Αρχίστε με την αλληλουχία μιας γνωστής πρωτεΐνης όπως η ανθρώπινη β-σφαιρίνη. Πραγματοποιήστε μια αναζήτηση TBLASTN σε μια βάση δεδομένων DNA. Είναι απίθανο να υπάρχουν πολλά «νέα» γονίδια στα καλά χαρακτηρισμένα γονιδιώματα οργανισμών όπως του ανθρώπου, του σακχαρομύκητα και της *E. coli*. Μπορεί επομένως να είναι χρήσιμο να κάνετε αναζητήσεις σε βάσεις δεδομένων για οργανισμούς των οποίων τα γονιδιώματα δεν είναι καλά χαρακτηρισμένα, ούτε πλήρως υπομνηματισμένα. Με τις αναζητήσεις TBLASTN μπορεί να απαντηθούν τρία σημαντικά ζητήματα: (1) Αντιστοιχεί άραγε η αλληλουχία αναζήτησης σε γνωστές πρωτεΐνες που έχουν ήδη υπομνηματιστεί; (2) Αντιστοιχεί ο όρος αναζήτησης σε ομόλογες πρωτεΐνες που δεν έχουν ακόμη υπομνηματιστεί (δηλαδή σε «νέα» γονίδια); (3) Με ποιον τρόπο η αλληλουχία DNA που αντιστοιχεί στο υποτιθέμενο νέο γονίδιο μπορεί να διερευνηθεί χρησιμοποιώντας τον αλγόριθμο BLASTX στη βάση μη επαναλαμβανόμενων αλληλουχιών DNA; Με τα αποτελέσματα της αναζήτησης αυτής είναι δυνατόν να επιβεβαιωθεί ότι το DNA κωδικοποιεί πράγματι μια πρωτεΐνη που δεν ταυτίζεται απόλυτα με οποιαδήποτε άλλη γνωστή, χαρακτηρισμένη πρωτεΐνη.

(α) Αποτελέσματα αναζήτησης TBLASTN με όρο την ανθρώπινη β-σφαιρίνη σε EST νηματοδών

Ac_EH1r_01A07_M13 Adult *Anguillicola crassus* *Anguillicola crassus* cDNA clone Ac_EH1r_01A07

Sequence ID: [gb|JK511422.1|](#) Length: 559 Number of Matches: 1

Range 1: 40 to 483				GenBank	Graphics	▼ Next Match	▲ Previous Match	
Score	Expect	Method			Identities	Positives	Gaps	Frame
149 bits(375)	6e-44	Compositional matrix adjust.			69/148(47%)	97/148(65%)	1/148(0%)	+1
Query 1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVPWTQRFESFGDLSTPDAVMGNPK						60	
Sbjct 40	MV I E +A+ +LW K+NV+E+G +A+ RLL+V PWTQR F +FG+LST A+M N K						219	
Query 61	VKAHGKKVLGAFSDGLAHLDNLKGTFTALSELHCDKLHVDPENFRLLGNVLVCVLAHHFG						120	
Sbjct 220	VAKHGTTVMGGDLRAIQNMDDIKNAYRELSVMHSEKLVDPDNFRLLSEHITLCMAAKFG						399	
Query 121	-KEFTPPVQAAAYQKVVAGVANALAHKYH	147						
Sbjct 400	EFT VQ A+QK + V +AL +YH							
	PTFTADVQEAWQKFLMAVTSALGRQYH	483						

(β) Αποτελέσματα αναζήτησης BLASTX για μια EST νηματοδών, στα οποία φαίνεται η καλύτερη στοίχισή της με μια πρωτεΐνη σπονδυλωτών

RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chain

Sequence ID: [sp|P80946.1|HBBA_ANGAN](#) Length: 147 Number of Matches: 1

Range 1: 1 to 147				GenPept	Graphics	▼ Next Match ▲ Previous Match	
Score	Expect	Method	Identities		Positives	Gaps	Frame
290 bits(742)	2e-97	Compositional matrix adjust.	136/147(93%)		141/147(95%)	0/147(0%)	+1
Query 43	VEWTD	AEHTAILSLWKKINVEEIGPQAMRRLIVCPWTQRHFANFGNLSTAAAIMNNEKV				222	
Sbjct 1	VEWI+ E TAI S W KIN+EEIGPQAMRRLIVCPWTQRHFANFGNLSTAAAIMNN+KV					60	
Query 223	AKHGTTVMGGDLRAIQNMDDIKNAYRELSVMHSEKLVDPDNFRLLSEHITLCMAAKFGP					402	
Sbjct 61	AKHGTTVMGGDLRAIQNMDDIKNAYR+LSVMHSEKLVDPDNFRLL+EHITLCMAAKFGP					120	
Query 403	TEFTADVQEAWQKFLMAVTSALGRQYH	483					
Sbjct 121	TEFTADVQEAWQKFLMAVTSAL RQYH	147					

Εικόνα 4.24 Η διαδικασία αναζήτησης ενός νέου γονιδίου παρουσιάστηκε χρησιμοποιώντας ως όρο αναζήτησης την ανθρώπινη β-σφαιρίνη (NP_000509) και διεξάγοντας μια αναζήτηση στη βάση δεδομένων ετικετών εκφραζόμενης αλληλουχίας (EST) με περιορισμό των αποτελεσμάτων στους νηματοδείς. (α) Στις στοιχίσεις περιλαμβάνεται μία EST από τον μύκητα *Anguillicola crassus* (κωδικός καταχώρισης GenBank JK511422.1). (β) Από την αναζήτηση BLASTX στη βάση μη επαναλαμβανόμενων αλληλουχιών DNA με όρο αυτόν τον κωδικό καταχώρισης προέκυψαν στοιχίσεις με γνωστές β-σφαιρίνες. Η καλύτερη στοίχιση ήταν με μια σφαιρίνη σπονδυλωτών. Εφόσον από την αναζήτηση αυτή δεν προέκυψε στοίχιση με κάποια σφαιρίνη του *A. crassus*, συμπεραίνουμε ότι από την εργασία μας προέκυψε μια αλληλουχία DNA που κωδικοποιεί μια σφαιρίνη νηματοδών η οποία δεν είχε περιγραφεί προηγουμένως. Αυτή η νέα σφαιρίνη μπορεί στη συνέχεια να χαρακτηριστεί με ταυτοποίηση της πλήρους αλληλουχίας της, των ομολόγων της, της εξέλιξής της, της δομής και της λειτουργίας της.