Βιοπληροφορική

Ενότητα 12: Αναζήτηση ομοιοτήτων έναντι βάσεων δεδομένων με τη χρήση ευρετικών αλγορίθμων

Αν. καθηγητής Αγγελίδης Παντελής
e-mail: paggelidis@uowm.gr
ΕΕΔΙΠ Μπέλλου Σοφία
e-mail: sbellou@uowm.gr

Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών

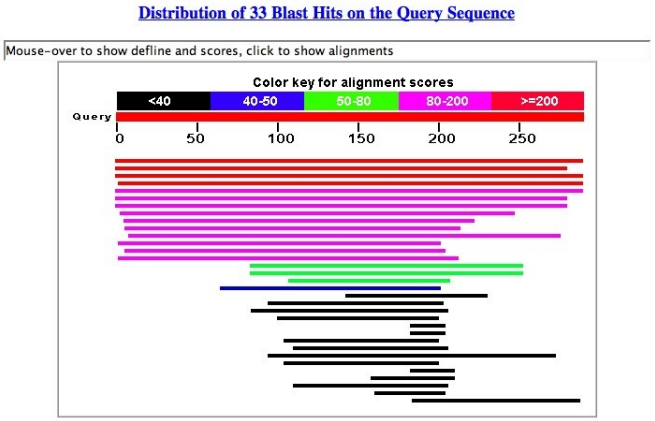
Άδειες Χρήσης

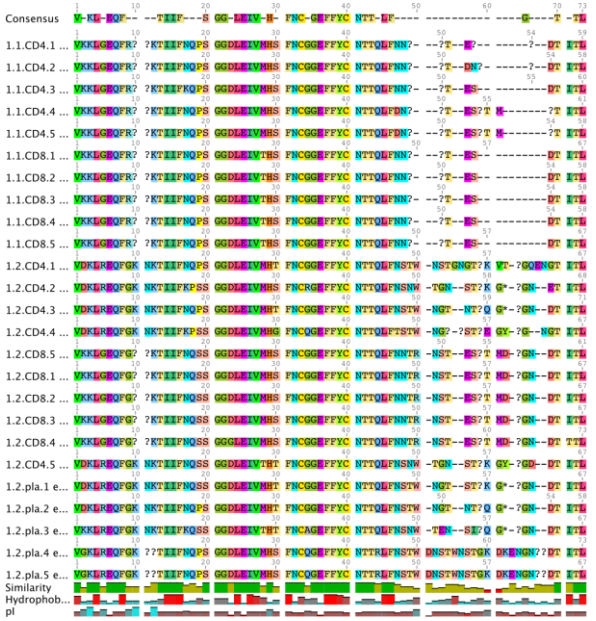
- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ψηφιακά Μαθήματα στο Πανεπιστήμιο Δυτικής Μακεδονίας**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Αναζήτηση ομοιοτήτων έναντι βάσεων δεδομένων με τη χρήση ευρετικών αλγορίθμων





Σκοπός ανάλυσης ομοιότητας αλληλουχιών (1/2)

- Η λειτουργία μίας πρωτεΐνης καθορίζεται από τη δομή της:
  - η οποία καθορίζεται από την αλληλουχία της (αμινοξέα).
    - η οποία καθορίζεται από το γονίδιο που την κωδικοποιεί (DNA).
- Αλλαγές στην αλληλουχία του DNA = μεταλλάξεις:
  - αλλαγές στις πρωτεΐνες που κωδικοποιεί.
    - αλλαγές στη λειτουργία των συγκεκριμένων πρωτεϊνών.
      - αλλαγές στην εξέλιξη του οργανισμού.

Υψηλός βαθμός ομοιότητας είναι ενδεικτικός παρόμοιας λειτουργίας, ενώ χαμηλός βαθμός ομοιότητας υποδηλώνει διαφορετικές λειτουργίες.

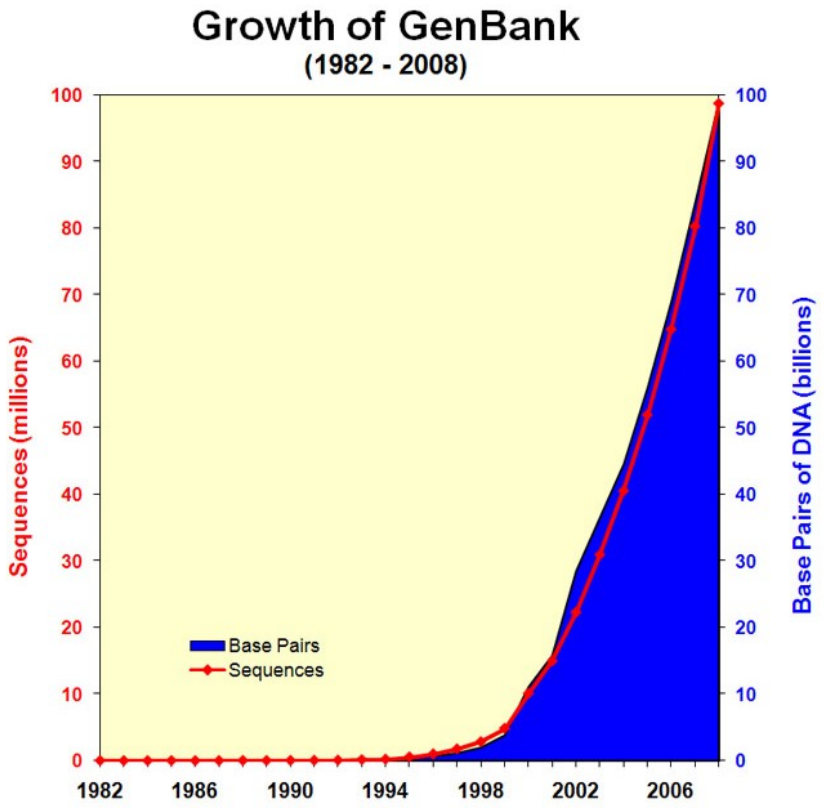
Σκοπός ανάλυσης ομοιότητας αλληλουχιών (2/2)

- **Εξαγωγή συμπερασμάτων για:**
  - Την εξελεγκτική σχέση ή ομολογία δύο γονιδίων.
  - Ποια κομμάτια μιας αλληλουχίας είναι επιρρεπή σε μεταλλαγές.
  - Ποια κομμάτια μιας αλληλουχίας τείνουν να είναι αμετάκλητα.

- Τελικά, για το ποια αμινοξέα σε συγκεκριμένες θέσεις είναι σημαντικά για τη δράση κάθε πρωτεΐνης.

Αναζήτηση ομοιοτήτων έναντι βάσεων δεδομένων

- Προσδιορισμός ολόκληρων γονιδιωμάτων διαφόρων οργανισμών, π.χ. βακτηρίων, μύγας, ψαριού, ανθρώπου.
- **Σκοπός:** Σύγκριση μίας αλληλουχίας με ολόκληρη βάση δεδομένων (πρωτεϊνική ή νουκλεοτιδική).
- **Αποτέλεσμα:** Εξαγωγή συμπερασμάτων για τη λειτουργία αγνώστου γονιδίου ή πρωτεΐνης.



Χρήση δυναμικού προγραμματισμού

- 0.1sec για σύγκριση αλληλουχιών μεγέθους 1kb=1000 βάσεις.
- Η βάση δεδομένων περιέχει 1 εκατομμύριο αλληλουχίες.
- Για να εξεταστεί 1 αλληλουχία έναντι όλων των αλληλουχιών στη βάση δεδομένων θα χρειαστούν.
  - 100,000sec ή 28 ώρες.

Ευρετικοί (πειραματικοί) αλγόριθμοι: Παράγουν λογικά αποτελέσματα, ακόμη και αν δεν είναι αποδεδειγμένα βέλτιστοι ή δεν έχουν εγγύηση απόδοσης.

Χρήση ευρετικών αλγορίθμων (1/2)

- **Παρελθόν:** Όταν η αναζήτηση σε βιολογικές βάσεις ξεκίνησε, η υπολογιστικές δυνατότητες ήταν περιορισμένες.
- **Παρόν:** Οι υπολογιστικές δυνατότητες έχουν βελτιωθεί δραστικά, ωστόσο έχουν αυξηθεί επίσης δραστικά τα βιολογικά δεδομένα.
- Δύο μέθοδοι **50 φορές ταχύτεροι** από τους αλγόριθμους δυναμικού προγραμματισμού.

FASTA (FAST All) BLAST

Χρήση ευρετικών αλγορίθμων (2/2)

- **Heuristic (tried-and-true) methods:** Είναι σχεδόν πάντα αποτελεσματικοί στην εύρεση σχετικών αλληλουχιών μιας βάσης δεδομένων αλλά δεν εγγυώνται ότι αυτή η λύση είναι και βέλτιστη, όπως συμβαίνει με τον δυναμικό προγραμματισμό.
- **FASTA:** Εντοπίζει κοινά μοτίβα μεταξύ της αλληλουχίας και των καταχωρήσεων μίας βιολογικής βάσης και τα ενώνει σε μία στοίχιση.
- **BLAST:** Παρόμοια μέθοδος με την FASTA, αλλά ταχύτερη καθώς αναζητεί ομοιότητες μόνο μεταξύ σημαντικών μοτίβων που εντοπίζονται στην αλληλουχία.

Αλγόριθμοι δυναμικού προγραμματισμού vs. Ευρετικοί αλγόριθμοι

- **Αλγόριθμοι δυναμικού προγραμματισμού:**
  - Μαθηματικά βέλτιστη (βέλτιστες) λύση (λύσεις).
  - Σύμφωνα με συγκεκριμένο σύστημα βαθμονόμησης.
  - Σχετικά αργοί.
  - Υψηλές αποθηκευτικές απαιτήσεις.

- **Ευρετικοί αλγόριθμοι:**
  - Επιτάχυνση της εξεύρεσης λύσης.
  - Δουλεύουν σε συγκεκριμένες περιοχές του πίνακα δυναμικού προγραμματισμού όπου αναμένονται υψηλά σκορ.
  - Μειωμένη ευαισθησία.
  - Βασίζονται στην εύρεση μικρών περιοχών ομοιότητας («λέξεων») και στην επέκτασή τους με χρήση δυναμικού προγραμματισμού.

Αναζήτηση σε βιβλιοθήκες DNA vs. Πρωτεϊνών (1/5)

- Πιο αποτελεσματικό να προσδιορίσεις ομοιότητες μεταξύ πρωτεϊνικών αλληλουχιών παρά νουκλεοτιδικών αλληλουχιών (DNA, RNA).
- DNA αλληλουχίες: 4 διαφορετικές βάσεις A, T, C, G.
- Πρωτεϊνικές αλληλουχίες: 20 διαφορετικά αμινοξέα.
- Παράδειγμα: Αλληλουχία 4 καταλοίπων:
  - DNA: $1/4^4=1/256$ τυχαία στοίχιση.
  - Πρωτεΐνη: $1/20^4=1/160,000$ τυχαία στοίχιση.

Αναζήτηση σε βιβλιοθήκες DNA vs. Πρωτεϊνών (2/5)

- Πλεονασμός στο γενετικό κώδικα:
  - Διαφορετικές τριπλέτες κωδικοποιούν το ίδιο αμινοξύ.
  - Η αλληλουχία DNA μπορεί να διαφέρει και η πρωτεϊνική αλληλουχία μπορεί να ίδια ακριβώς ιδία.

Παράδειγμα:

ATGGAATTAGTTATTAGTGCTTTAATTGTTGAATAA

ATGGAGCTGGTGATCTCAGCGCTGATCGTCGAGTGA

- Στοίχιση χωρίς κενά:

ATGGAATTAGTTATTAGTGCTTTAATTGTTGAATAA

||||| | || || || | || || || | |

ATGGAGCTGGTGATCTCAGCGCTGATCGTCGAGTGA

- 21 ίδια κατάλοιπα (από τα 36): 58% ομοιότητα (identity).

Αναζήτηση σε βιβλιοθήκες DNA vs. Πρωτεϊνών (3/5)

- Μεταφράζουμε κάθε αλληλουχία σε πρωτεΐνη :

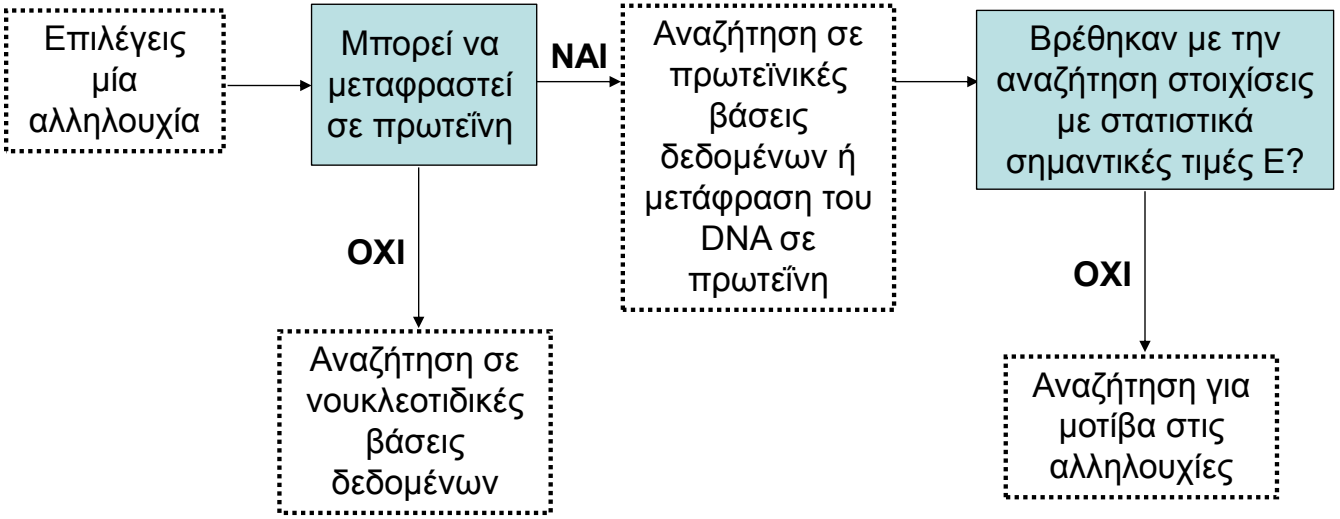
MELVISISALIVE*

MELVISISALIVE*

- 100% ίδια σε πρωτεϊνικό επίπεδο.

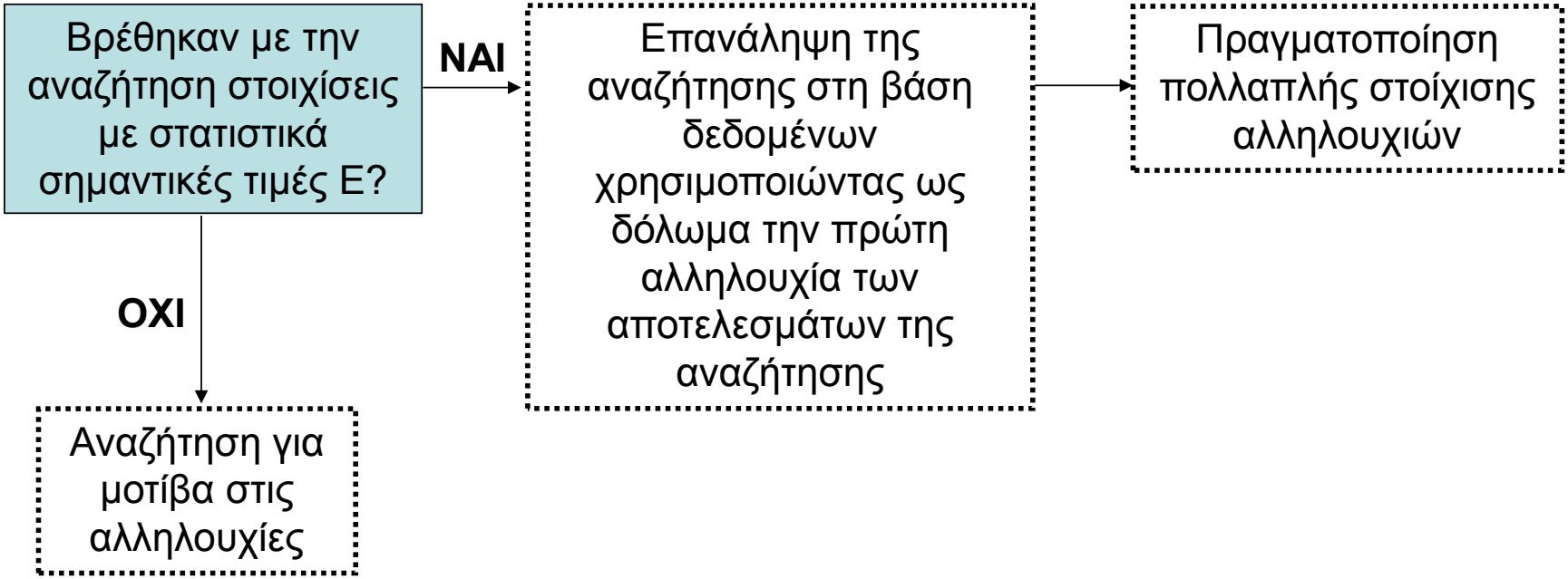
Οι αναζητήσεις με αλληλουχίες DNA (από τις οποίες προκύπτει μία πρωτεΐνη) δίνουν αποτελέσματα με μικρότερη στατιστική σημαντικότητα απ' ότι οι αναζητήσεις με τις αντίστοιχες πρωτεϊνικές αλληλουχίες (Pearson, 2000).

Αναζήτηση σε βιβλιοθήκες DNA vs. Πρωτεϊνών (4/5)



- **Τιμή Ε (E-value):** Η προτεινόμενη αλληλουχία θα πρέπει να έχει Ε με μικρή τιμή και μία καλή στοίχιση με την υπό διερεύνηση αλληλουχία.
- **Τι δείχνει:** Την πιθανότητα το αποτέλεσμα να είναι τυχαίο.
- **Ανώτατο όριο για την τιμή Ε:** 0.01-0.05.
- **Προσοχή:** Η αλληλουχία θα πρέπει να εξεταστεί για επαναλαμβανόμενες περιοχές για την αποφυγή λανθασμένα υψηλής βαθμολόγησης της στοίχισης.

Αναζήτηση σε βιβλιοθήκες DNA vs. Πρωτεϊνών (5/5)



Ο ευρετικός αλγόριθμος FASTA (Fast All)

- Γρήγορη προσέγγιση του αλγόριθμου Smith-Waterman (τοπική στοίχιση).
- **50 φορές πιο γρήγορος** από τους αλγόριθμους δυναμικού προγραμματισμού.
- Δεν εγγυάται την καλύτερη σύγκριση δύο αλληλουχιών – Ευρετικός - πειραματικός αλγόριθμος.

Ο ευρετικός αλγόριθμος FASTA (FASTA3)

- **Ο αλγόριθμος FASTA αναζητεί περιοχές-μοτίβα που να ταιριάζουν μεταξύ της άγνωστης αλληλουχίας και των αλληλουχιών της βάσης δεδομένων.**

- **Αρχή λειτουργίας:** Οι βέλτιστες στοιχίσεις περιέχουν **μικρές περιοχές** όπου οι βαθμολογία στοίχισης είναι μεγαλύτερη από μία τιμή κατωφλίου.

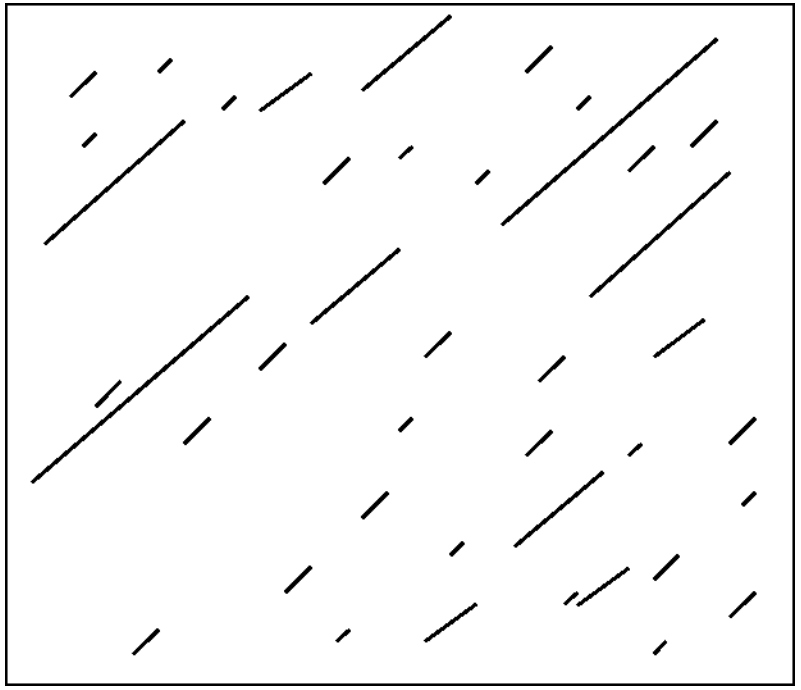
- **Μικρές περιοχές = λέξεις χωρίς κενά = k-tuples:**
  - 2 αμινοξέα στην περίπτωση της πρωτεΐνης.
  - 4-6 νουκλεοτίδια στην περίπτωση DNA.

Ο αλγόριθμος FASTA – Input and output

- **Σκοπός χρήσης:** Σύγκριση μιας άγνωστης αλληλουχίας – input sequence (DNA ή πρωτεΐνη) με όλες τις αλληλουχίες της βάσης δεδομένων.
- **Αποτέλεσμα:** Αναφορά των αλληλουχιών της βάσης με τα περισσότερα ταιριάσματα με την άγνωστη αλληλουχία και οι τοπικές στοιχίσεις των αλληλουχιών της βάσης με την άγνωστη αλληλουχία.
- Άγνωστη αλληλουχία: FASTA format:
  - $1^{η}$ γραμμή: >Πληροφορίες.
  - $2^{η}$ – έως τέλος: Αλληλουχία χωρίς κενά.

Ο αλγόριθμος FASTA –Στάδια ανάλυσης

1. **Κατακερματισμός (Hashing):** Εντοπίζονται μικρές περιοχές χωρίς κενά («λέξεις» συγκεκριμένου μήκους - ktuple), στις οποίες οι δύο αλληλουχίες **ταυτίζονται:**
   - Για πρωτεΐνες, k-tuple= 2.
   - Για νουκλεοτίδια, k-tuple = 4 ή 6.



Μέθοδος κατακερματισμού - Hash method (1/2)

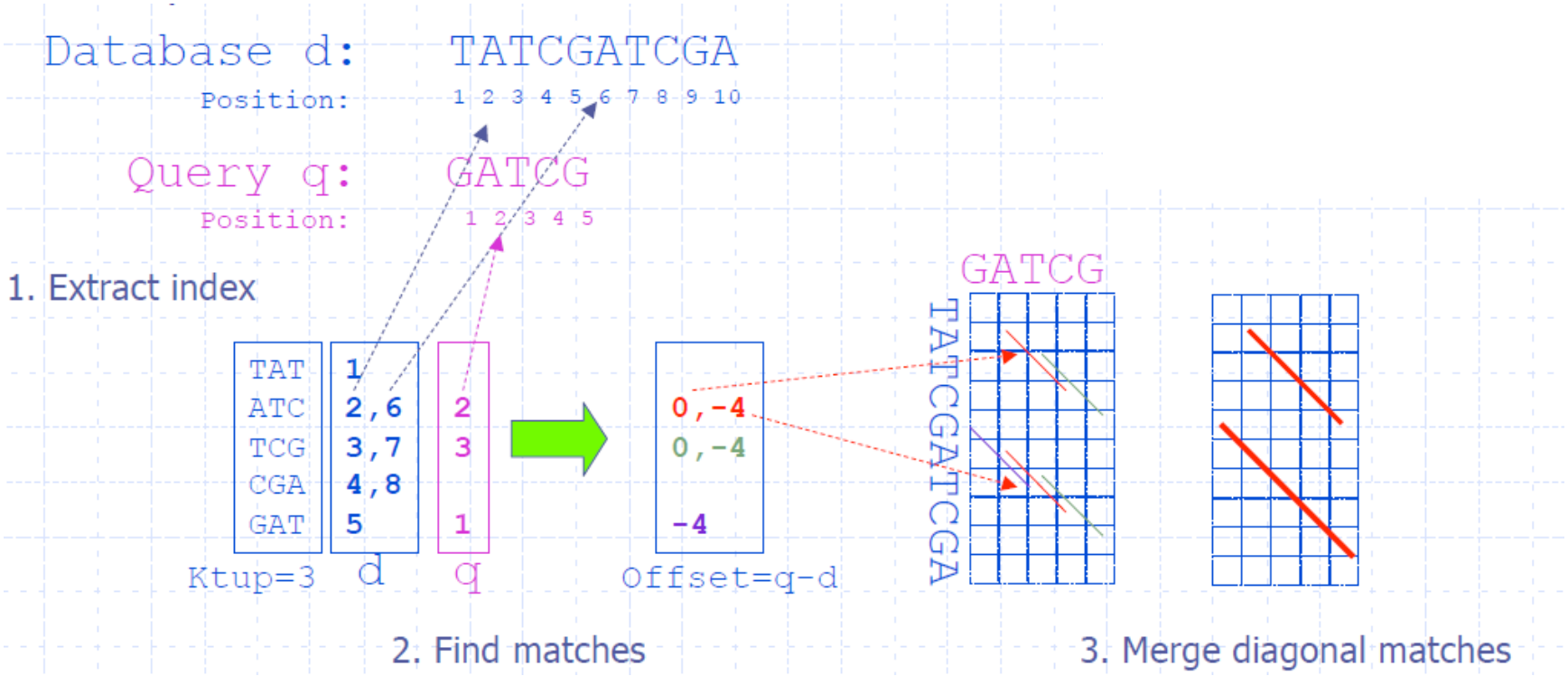
Position	1	2	3	4	5	6	7	8	9	10	11
Sequence 1	**N**	**C**	**S**	**P**	**T**	**A**					
Sequence 2						**A**	**C**	**S**	**P**	**R**	**K**

	Position in		Offset
amino acid	Sequence 1	Sequence 2	posA-posB
A	6	6	0
C	2	7	-5
K	-	11	
N	1	-	
P	4	9	-5
R	-	10	
S	3	8	-5
T	5	-	

- Προσοχή στο κοινό offset των τριών αμινοξέων: C, P, S. Μία κοινή στοίχιση μπορεί να είναι:

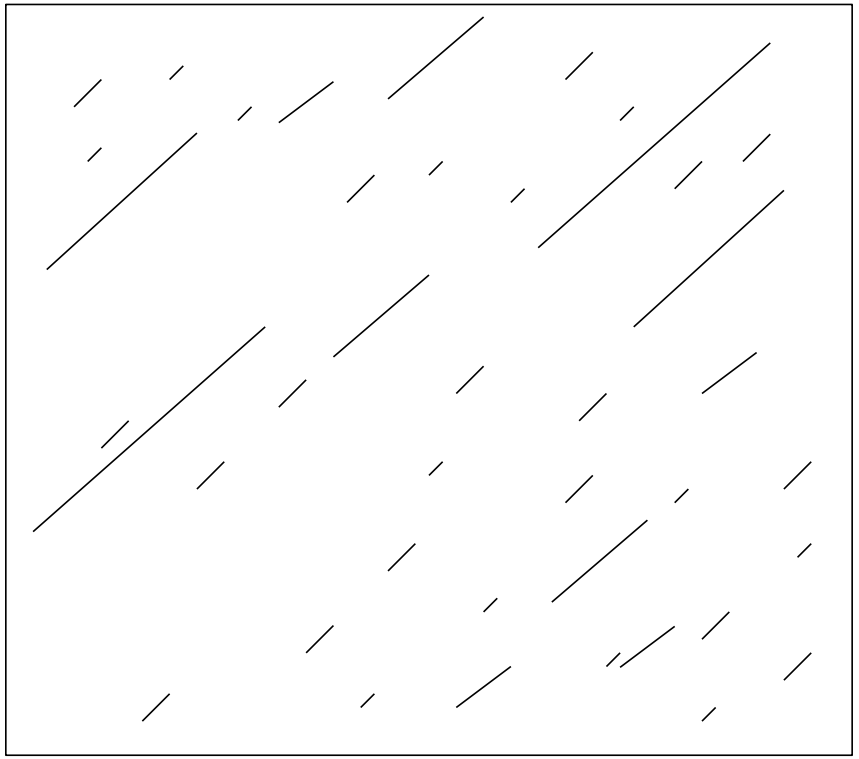
Sequence 1	N	C	S	P	T	A
		I	I	I		
Sequence 2	A	C	S	P	R	K

Μέθοδος κατακερματισμού - Hash method (2/2)



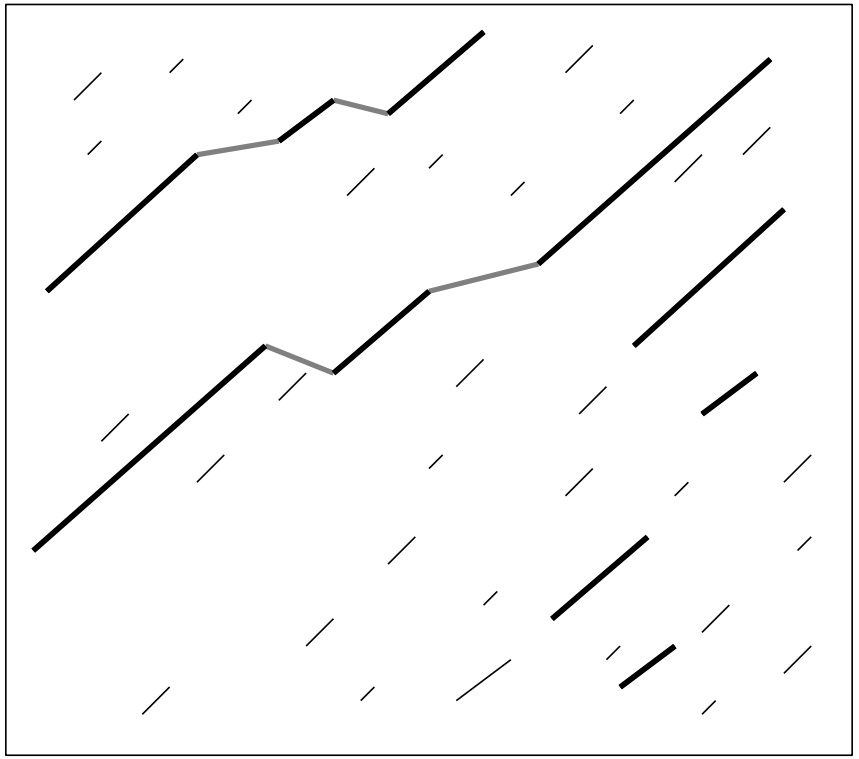
Ο αλγόριθμος FASTA – Στάδια ανάλυσης

1. **Κατακερματισμός (Hashing):** Εντοπίζονται μικρές περιοχές χωρίς κενά («λέξεις» συγκεκριμένου μήκους - ktuple), στις οποίες οι δύο αλληλουχίες **ταυτίζονται:**
   - Για πρωτεΐνες, k-tuple = 2
   - Για νουκλεοτίδια, k-tuple = 4 ή 6.
2. **Βαθμολογία (Scoring):** Εντοπίζονται οι 10 περιοχές με τις περισσότερες k-tuples και τη μεγαλύτερη βαθμολογία (BLOSUM50) – $init_1$ score. Οι περιοχές με τις ομοιότητες κατατάσσονται σε μία λίστα



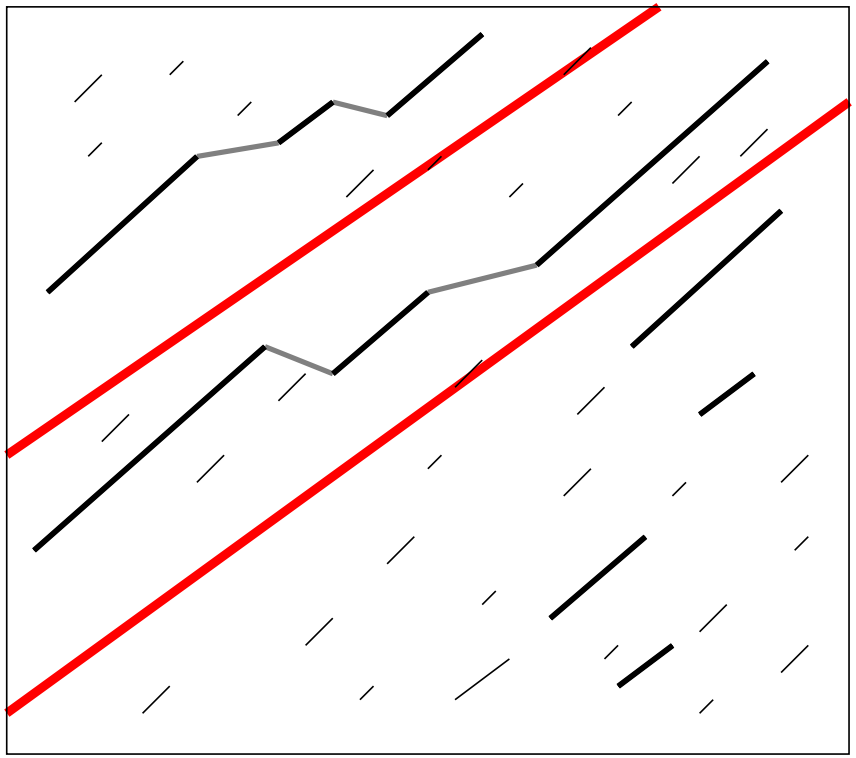
Ο ευρετικός αλγόριθμος FASTA (1/2)

3. **Εισαγωγή κενών (Introduction of gaps):** Δημιουργούνται μεγαλύτερες περιοχές ομοιότητας ενώνοντας διακριτές περιοχές με συγκεκριμένη βαθμολογία.
   - Η νέα βαθμολογία, **init$_n$ score**, προκύπτει από τη βαθμολογία ομοιότητας μείον την ποινή κενών που εισήχθησαν.
   - Ενώνονται οι περιοχές που απέχουν συγκεκριμένα απόσταση (για πρωτεΐνες: 32 όταν k=1 και 16 για k=2).



Ο ευρετικός αλγόριθμος FASTA (2/2)

4. **Στοίχιση (Alignment):** Καθορίζεται το βέλτιστο τμήμα ομοιότητας μεταξύ της αλληλουχίας επερώτησης και της βάσης δεδομένων με τη χρήση του Smith – Waterman. Η βαθμολογία αυτή ονομάζεται opt. score.
5. **Random Sequence Simulation:** Εξετάζεται η στατιστική σημαντικότητα της σύγκρισης. Αποδίδεται ένα Z-score και ένα E-score.



FASTA - Παράδειγμα

```
>>EM_HUM:AF263744; AF263744 Homo sapiens erbb2-interacti  (6409 nt)
 initn: 20580 init1: 20580 opt: 20580  Z-score: 24347.5  bits: 4521.3 E():    0
banded Smith-Waterman score: 20580; 100.0% identity (100.0% similar) in 4116 nt overlap (1-4116:324-4439)

                                          10        20        30
Sequen                                ATGACTACAAAACGAAGTTTGTTTGTGCGG
                                      ::::::::::::::::::::::::::::::
EM_HUM GGAGATAAATTCAACCCAGTGTGTCTAAAAATGACTACAAAACGAAGTTTGTTTGTGCGG
           300       310       320       330       340       350

               40        50        60        70        80        90
Sequen TTGGTACCATGTCGCTGTCTACGAGGGGAAGAGGAGACTGTCACTACTCTTGATTATTCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
EM_HUM TTGGTACCATGTCGCTGTCTACGAGGGGAAGAGGAGACTGTCACTACTCTTGATTATTCT
           360       370       380       390       400       410

              100       110       120       130       140       150
Sequen CATTGCAGCTTAGAACAAGTTCCGAAAGAGATTTTTACTTTTGAAAAAACCTTGGAGGAA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
EM_HUM CATTGCAGCTTAGAACAAGTTCCGAAAGAGATTTTTACTTTTGAAAAAACCTTGGAGGAA
           420       430       440       450       460       470

              160       170       180       190       200       210
Sequen CTCTATTTAGATGCTAATCAGATTGAAGAGCTTCCAAAGCAACTTTTTAACTGTCAGTCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
EM_HUM CTCTATTTAGATGCTAATCAGATTGAAGAGCTTCCAAAGCAACTTTTTAACTGTCAGTCT
           480       490       500       510       520       530
```

Basic Local Alignment Sequence Tool – BLAST (1/2)

- Το υπολογιστικό πακέτο που χρησιμοποιείται και αναφέρεται περισσότερο στη βιοπληροφορική.
- Μεγαλύτερη ταχύτητα / παρόμοια ευαισθησία συγκριτικά με το FASTA.
- Χρησιμοποιείται για την εύρεση **τοπικών** ομοιοτήτων μεταξύ:
  - μιας αλληλουχίας επερώτησης (query sequence).
  - μιας βάσης δεδομένων (πρωτεϊνική ή νουκλεοτιδική).

Basic Local Alignment Sequence Tool – BLAST (2/2)

- Εύρεση μικρών περιοχών («λέξεων» ή k-tuples) όπου η βαθμολογία στοίχισης είναι μεγαλύτερη από ένα κατώφλι.
- **Λέξεις:** Πιθανά σημεία έναρξης μιας καλής τοπικής στοίχισης.

Διαφορά με FASTA:

- Απομακρύνονται οι περιοχές με χαμηλή πολυπλοκότητα, οι οποίες μπορούν να οδηγήσουν σε λάθος συμπεράσματα.
- Ζεύγη υψηλής βαθμολογίας (High Scoring Pairs, HSP).
- Επιλέγονται τα ζεύγη υψηλής βαθμολογίας (High Scoring Pairs, HSP) με στατιστική σημαντικότητα.
- **Βασική υλοποίηση του BLAST: τοπική στοίχιση χωρίς κενά.**

BLAST & FASTA k-tuples

- Ο FASTA αναζητά όλες τις πιθανές λέξεις ίδιου μεγέθους, με k-tuple:
  - 2 αμινοξέα σε περίπτωση πρωτεΐνης.
  - 4-6 νουκλεοτίδια σε περίπτωση DNA.
- Ο BLAST αναζητά τις ομοιότητες μόνο σε σημαντικές περιοχές, με k-tuple:
  - 3 αμινοξέα σε περίπτωση πρωτεΐνης.
  - 11 νουκλεοτίδια σε περίπτωση DNA.
- **ΘΕΩΡΗΤΙΚΑ:** Ο FASTA δίνει περισσότερα αποτελέσματα γιατί ψάχνει όμοιες λέξεις μικρότερου μεγέθους. Δεν είναι όμως πιο ακριβής.

BLAST – Βήματα εφαρμογής, Φιλτράρισμα (1/12)

1. Αρχικά, η αλληλουχία φιλτράρεται για να απομακρυνθούν περιοχές χαμηλής πολυπλοκότητας (Low complexity regions, LCR).

❖ **Παράδειγμα**:

- **Πρωτεΐνη:** PPCDPPPPPKDKKKKDDGPP.
- **DNA:** AAATAAAAAAAATAAAAAAT.

1. **Προσοχή:** Το φιλτράρισμα γίνεται μόνο στην υπό εξέταση αλληλουχία και όχι στις αλληλουχίες της βάσης δεδομένων.

BLAST – Βήματα εφαρμογής, Φιλτράρισμα (2/12)

Η πολυπλοκότητα, K, σε ένα παράθυρο αλληλουχίας μήκους L δίνεται από:

$$K = \frac{1}{L}\log_N(L!/\prod_{alli} n_i!)$$

όπου, **N=4** για DNA και **20** για πρωτεΐνες και n_i: ο αριθμός κάθε καταλοίπου στην αλληλουχία

❖ **Παράδειγμα:** Αλληλουχία **GGGG**

Μήκος, L = 4 και N=4 (DNA)

L! = 1x2x3x4 = 24

n_G=4, n_C=0, n_A=0, n_T=0

$$\prod_{alli} n_i = n_G!n_C!n_A!n_T! = 4!0!0!0! = 24$$

$$K = \frac{1}{4}\log_4(4!/24) = \frac{1}{4}\log_4 1 = 0$$

❖ **Παράδειγμα:** Αλληλουχία **CTGA**

Μήκος, L = 4 και N=4 (DNA)

L! = 1x2x3x4 = 24

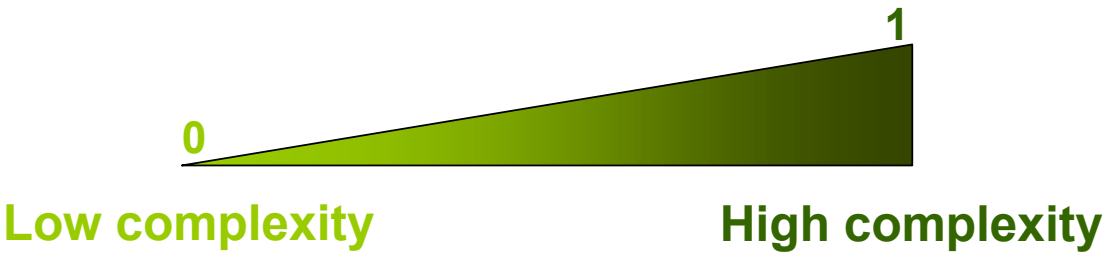
n_G=1, n_C=1, n_A=1, n_T=1

$$\prod_{alli} n_i = n_G!n_C!n_A!n_T! = 1!1!1!1! = 1$$

$$K = \frac{1}{4}\log_4(4!/1) = \frac{1}{4}\log_4 24 = 1.573$$

BLAST – Βήματα εφαρμογής, Φιλτράρισμα (3/12)

$$K = \frac{1}{L}\log_N(L!/\prod_{alli} n_i!)$$



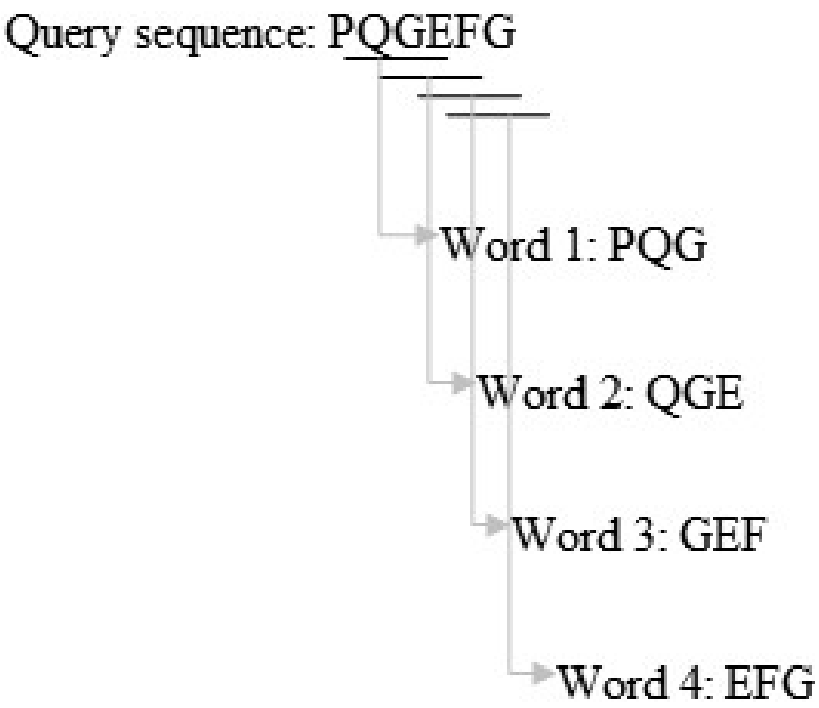
- Ένα παράθυρο 12 καταλοίπων εξετάζεται και υπολογίζεται το K.
- Περιοχές με χαμηλή πολυπλοκότητα (χαμηλό K) δεν υπολογίζονται στην ανάλυση.

BLAST – Βήματα εφαρμογής, Φιλτράρισμα (4/12)

2. Η αλληλουχία επερώτησης τεμαχίζεται σε λέξεις (k-tuples) 3 (πρωτεΐνη) ή 11 (νουκλεοτίδια) γραμμάτων.
3. Χρησιμοποιώντας τον BLOSUM62 υπολογίζεται το σκορ του σωστού ταιριάσματος με κάθε αλληλουχία της βάσης δεδομένων.

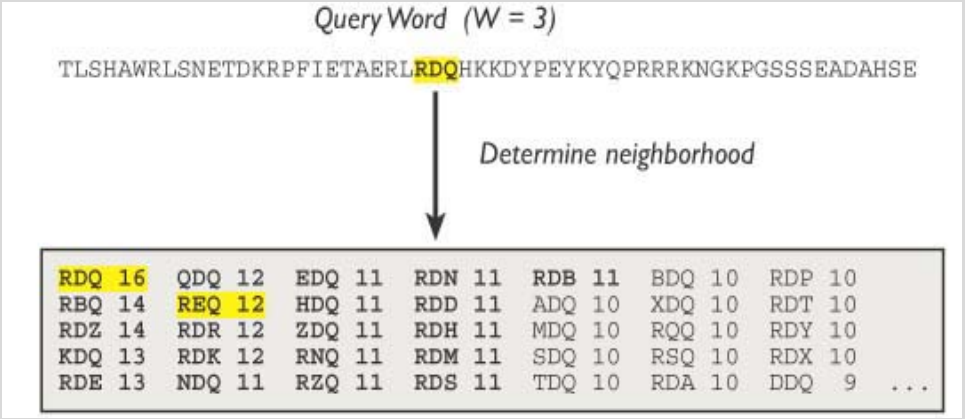
❖ **<u>Παράδειγμα:</u>** PQG.

P-P=7, Q-Q=5, G-G=6, Sum=18.



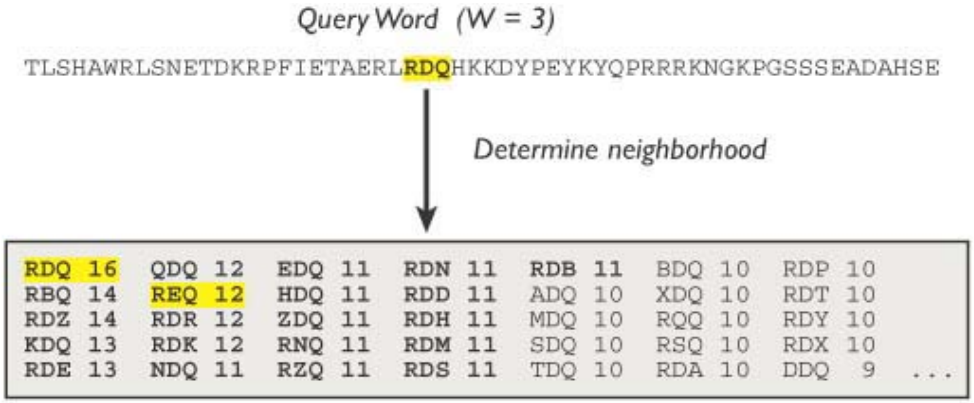
BLAST – Βήματα εφαρμογής, Φιλτράρισμα (5/12)

4. Επίσης, υπολογίζεται και το ταίριασμα με όλους τους πιθανούς συνδυασμούς 3 αμινοξέων (20^2=8000 πιθανοί συνδυασμοί).
   - ❖ **Παράδειγμα 1:** Αλληλουχία επερώτησης: **PQGEFG**, 1η λέξη: PQG
   a. PEG, 7+2+6 = 15
   b. PSG, 7+0+6 = 13
   c. PQA, 7+5+0 = 12…
   - ❖ **Παράδειγμα 2:**



BLAST – Βήματα εφαρμογής, Φιλτράρισμα (6/12)

5. Ορίζεται ένα κατώφλι βαθμολογίας (neighborhood word score threshold, T), για να διατηρηθούν τα πιθανά ταιριάσματα του PQG μόνο στα σημαντικά. Δηλαδή διατηρούνται μόνο τα ταιριάσματα με βαθμολογία μεγαλύτερη του T
   - **Παράδειγμα 1:** Αν T=13 μόνο οι λέξεις με βαθμολογία πάνω από το 13 διατηρούνται, δηλαδή οι a & b
   - Τελικά, από τις 8000 λέξεις διατηρούνται περίπου οι 21.
   - **Παράδειγμα 2:** Για T=11, διατηρούνται μόνο οι 50 πρώτες τριπλέτες που φαίνονται στο πίνακα.



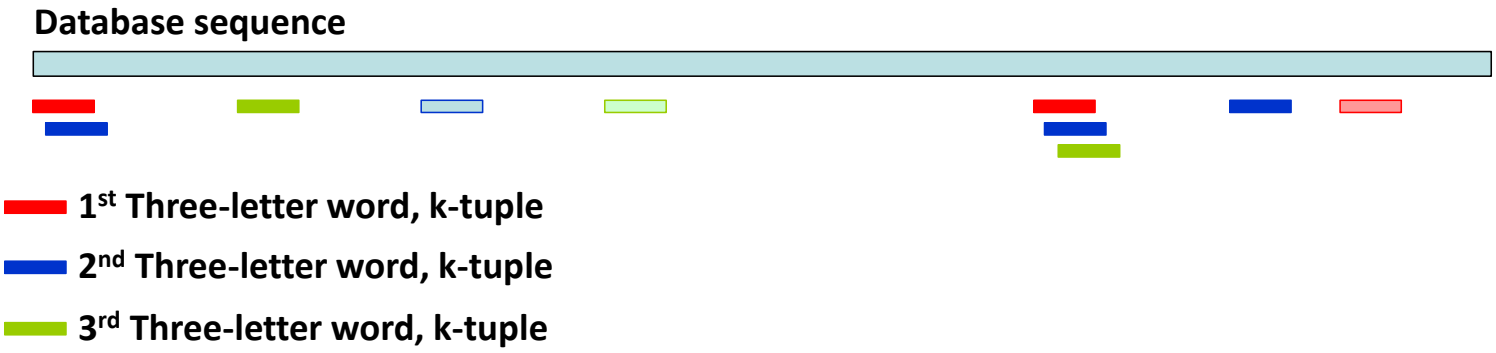
BLAST – Βήματα εφαρμογής, Φιλτράρισμα (7/12)

6. Επαναλαμβάνεται η προηγούμενη διαδικασία για όλες τις τριπλέτες που προκύπτουν από την αλληλουχία.

Προσοχή: Για μία αλληλουχία n βάσεων προκύπτουν περίπου n τριπλέτες, οπότε τελικά ο συνολικός αριθμός λέξεων που έχουμε είναι 50·n.
Παράδειγμα: Για μία αλληλουχία 250 βάσεων προκύπτουν περίπου 250 συνεχόμενες τριπλέτες (για την ακρίβεια προκύπτουν 248). Αν για κάθε τριπλέτες μένουν τελικά 50 λέξεις με μεγαλύτερη βαθμολογία από το κατώφλι, τότε ο συνολικός αριθμός των λέξεων είναι 250x50 = 12,500.

BLAST – Βήματα εφαρμογής, Φιλτράρισμα (8/12)

7. Κάθε αλληλουχία της βάσης δεδομένων σαρώνεται για ακριβή ταύτιση με μία από τις 50 λέξεις που προκύπτουν από την πρώτη τριπλέτα της αλληλουχίας μας, τη δεύτερη τριπλέτα, κτλ.
8. Αν βρεθεί απόλυτα σωστά σύγκριση χρησιμοποιείται για αρχή μιας στοίχισης χωρίς κενά. Αν όχι, πραγματοποιείται η ίδια διαδικασία με τις υπόλοιπες από τις 50 τριπλέτες, οι οποίες δεν έχουν απόλυτη ομοιότητα με την αρχική αλληλουχία.

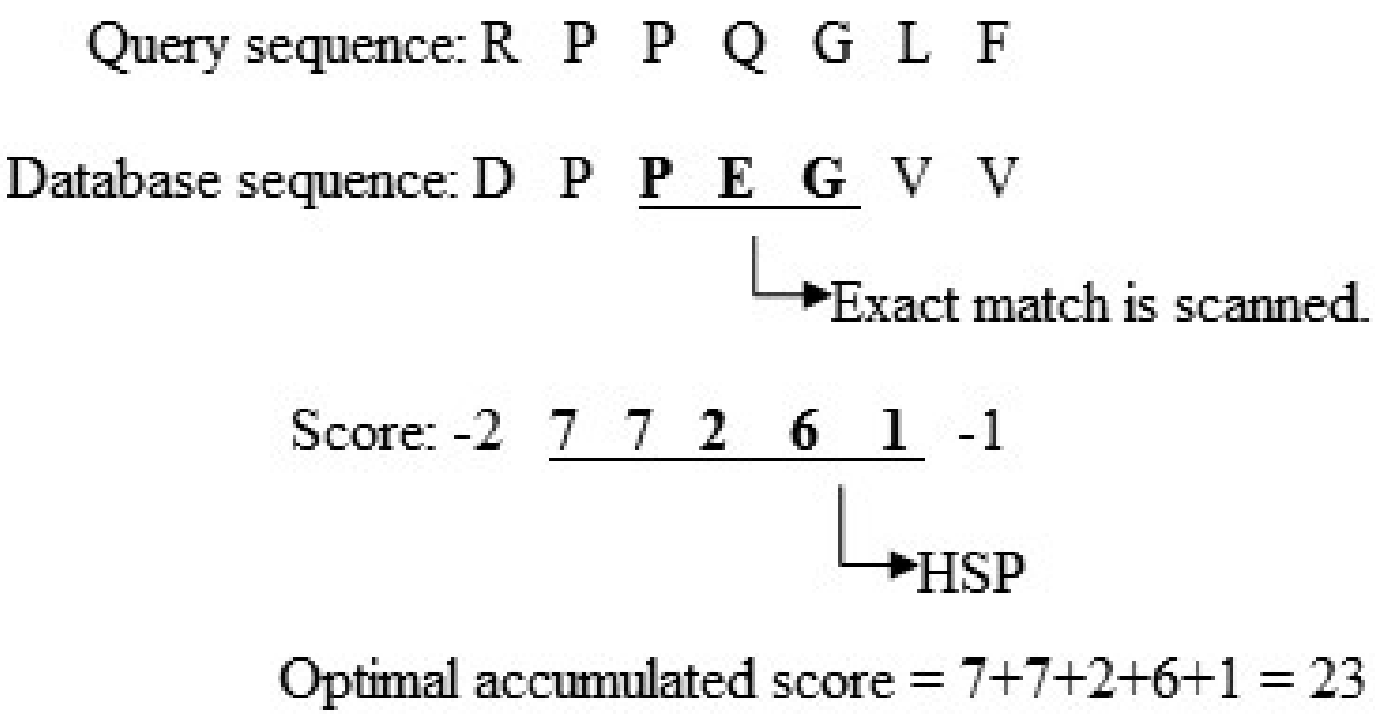


BLAST – Βήματα εφαρμογής, Φιλτράρισμα (9/12)

9. Original BLAST:

- Μόλις βρεθεί, τότε αρχίζει μία στοίχιση με επέκταση και προς τις δύο κατευθύνσεις **χωρίς την εισαγωγή κενών**.
- Η επέκταση συνεχίζεται μόνο στην περίπτωση που **η βαθμολογία αυξάνεται ή παραμένει σταθερή**.
- Τελικά καταλήγουμε στην στοίχιση υψηλής βαθμολογίας – **high scoring segment pair (HSP).**
- **HSP = Local optimal alignment.**

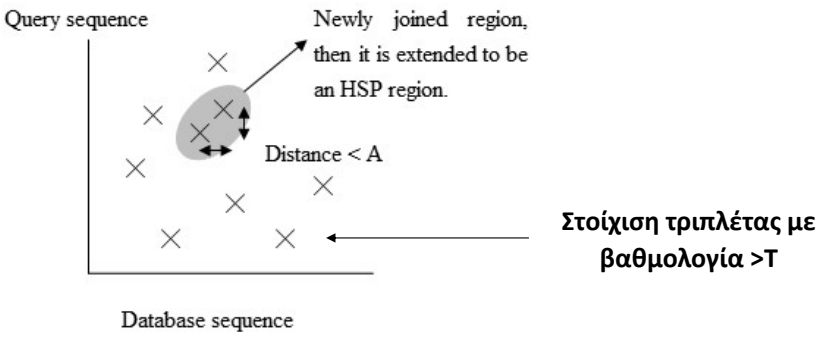
Extending the High Scoring Segment Pair (HSP)



BLAST – Βήματα εφαρμογής, Φιλτράρισμα (10/12)

10. BLAST2 (gapped BLAST):

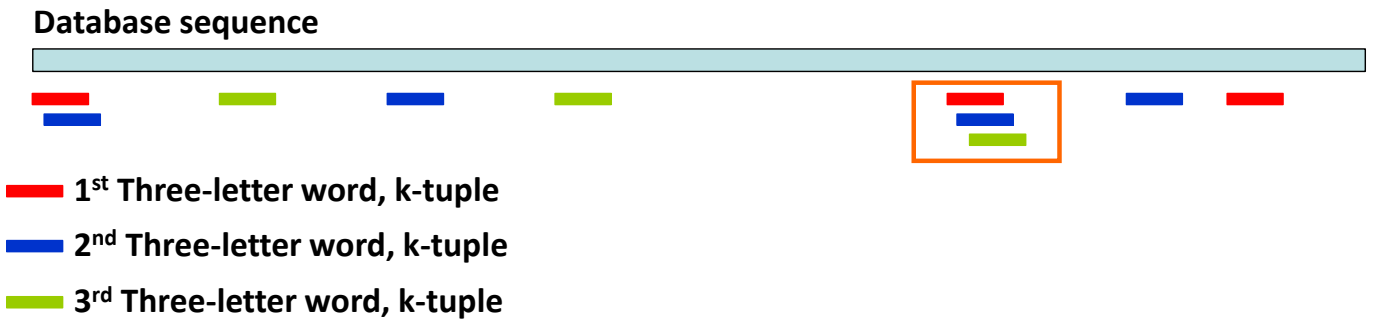
- Το κατώφλι βαθμολογίας για τις λέξεις είναι χαμηλότερο (λιγότερο αυστηρό κριτήριο), π.χ.T=11 και όχι T=13.
- **Αποτέλεσμα:** Μεγαλύτερες λίστες λέξεων για τις οποίες πρέπει να σαρωθεί η κάθε αλληλουχία της βάσης δεδομένων.



BLAST – Βήματα εφαρμογής, Φιλτράρισμα (11/12)

10. BLAST2 (gapped BLAST):

- Ενώνονται οι στοιχίσεις που βρίσκονται στην **ίδια διαγώνιο** και απέχουν **απόσταση μικρότερη από Α.**
- Η πιο μακριά στοίχιση επεκτείνεται όπως προηγουμένως.
- Είναι απαραίτητες τουλάχιστον δύο περιοχές ομοιότητας στην ίδια διαγώνιο για να πραγματοποιηθεί η επέκταση.



BLAST – Βήματα εφαρμογής, Φιλτράρισμα (12/12)

11. Γίνεται το ίδιο για όλες τις λέξεις που προέκυψαν από την αλληλουχία επερώτησης.
12. Αποδίδεται μία βαθμολογία σε κάθε στοίχιση.
13. Διατηρούνται οι στοιχίσεις με βαθμολογία υψηλότερη από ένα κατώφλι S.
14. Καθορίζεται η στατιστική σημαντικότητα των στοιχίσεων.
15. Επαλήθευση με αλγόριθμο δυναμικού προγραμματισμού τοπικής στοίχισης.

BLAST – Βαθμολογία και στατιστική

- Είναι μία στοίχιση «καλή»;
- Το πρόγραμμα BLAST παράγει για κάθε ζεύγος στοίχισης.
  - bit score.
  - expected value (E-value).

BLAST – Bit score vs. BLAST – Expected value (E-value)

- BLAST – Bit score:
  - Ένδειξη της ποιότητας της στοίχισης.
  - Όσο μεγαλύτερο είναι τόσο καλύτερη είναι η στοίχιση.
  - Είναι κανονικοποιημένα, δηλαδή δεν εξαρτώνται από το σύστημα βαθμονόμηση.
  - Μπορούν να συγκριθούν τα bit scores δύο στοιχίσεων.

- BLAST – Expected value (E-value):
  - Ένδειξη της στατιστικής σημαντικότητας της στοίχισης.
  - Δείχνει πόσο πιθανό είναι το αποτέλεσμα να προέκυψε λόγω τυχαιότητας.
  - Όσο πιο χαμηλό, τόσο πιο στατιστικά σημαντική η στοίχιση.
  - π.χ. E-value = 0.05: Η ομοιότητα των αλληλουχιών έχει 5% πιθανότητα να συνέβη τυχαία.

Αναζήτηση ομοιοτήτων έναντι βάσεων δεδομένων

- Η ανάλυση γίνεται σε κεντρικά (remote) υπολογιστικά κέντρα.
- BLAST (http://www.ncbi.nlm.nih.gov/blast/)
  - **BLASTp:** Στοίχιση μιας πρωτεϊνικής αλληλουχίας έναντι βάσης δεδομένων αλληλουχιών πρωτεϊνών.
  - **BLASTn:** Στοίχιση μιας νουκλεοτιδικής αλληλουχίας έναντι βάσης δεδομένων αλληλουχιών DNA.
  - **BLASTx:** Μεταφράζει πρώτα μία αλληλουχία DNA σε πρωτεΐνη και αναζητεί παρόμοιες πρωτεϊνικές αλληλουχίες.
  - **tBLASTn:** Στοίχιση μιας πρωτεϊνικής αλληλουχίας έναντι βάσης δεδομένων μεταφρασμένων νουκλεοτιδικών αλληλουχιών.

BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

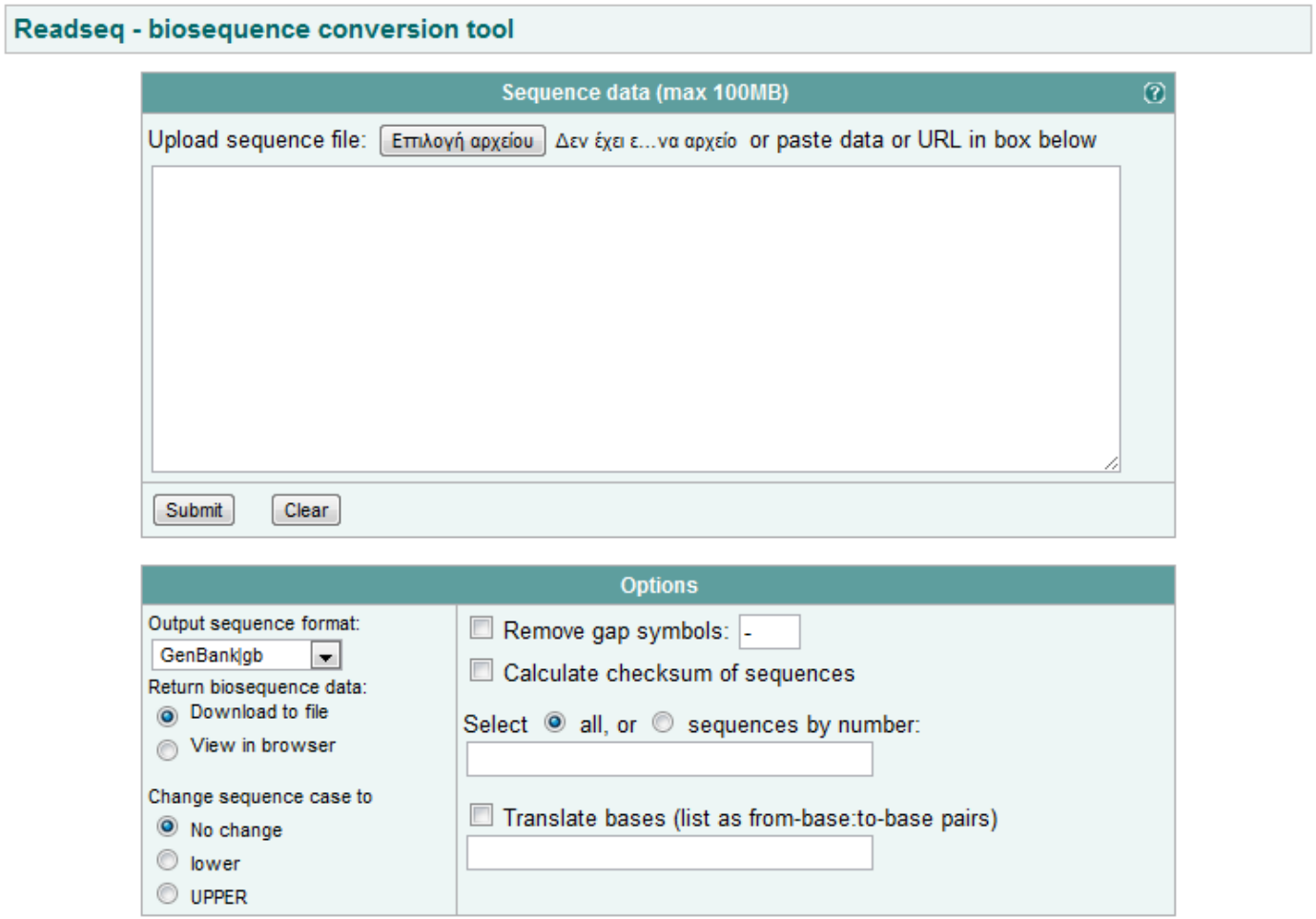
BLAST vs. FASTA

- **BLAST:**
  - Ξεκινάει από περιοχές που είναι απόλυτα όμοιες ή παρόμοιες με την αλληλουχία επερώτησης αρκεί να ξεπερνά η βαθμολογία στοίχισης ένα κατώφλι.
  - Υπολογίζει τη στατιστική σημαντικότητα μόνο ορισμένων στοιχίσεων.
  - Η βασική έκδοση δεν επιτρέπει την εισαγωγή κενών.
  - Πιο γρήγορο από το FASTA.
  - Εφαρμόζει τοπική στοίχιση.
- **FASTA:**
  - Ξεκινάει τη στοίχιση από περιοχές που υπάρχει απόλυτη ταύτιση.
  - Υπολογίζει τη στατιστική σημαντικότητα όλων των στοιχίσεων.
  - Επιτρέπεται η εισαγωγή κενών.
  - Πιο ακριβές από το BLAST.
  - Εφαρμόζει τοπική στοίχιση.

FASTA format

- Η πιο διαδεδομένη μορφή αλληλουχιών.
- 1η γραμμή: >όνομα και πληροφορίες.
- 2η γραμμή: Η αλληλουχία χωρίς κενά.
- Μετατροπή με:
  - Readseq from EBI (http://www.ebi.ac.uk/cgi-bin/readseq.cgi).
  - BCM Launcher (http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html).
- >Name of the protein_length in bps
  MTLRCLEPSGNGGEGTRSQWGTAGSAEEPSPQAARLAKALRELGQTGWYWGSM
  TVNEAKEKLKEAPEGTFLIRDSSHSDYLLTISVKTSAGPTNLRIEYQDGKFRLDSIICVKS
  KLKQFDSVVHLIDYYVQMCKDKRTGPEAPRNGTVHLYLTKPLYTSAPSLQHLCRLTIN
  KCTGAIWGLPLPTRLKDYLEEYKFQV

Readseq


Readseq - biosequence conversion tool
Sequence data (max 100MB)
Upload sequence file:
Επιλογή αρχείου
Δεν έχει ε...να αρχείο
or paste data or URL in box below
Submit
Clear
Options
Output sequence format:
GenBank|gb
Return biosequence data:
Download to file
View in browser
Change sequence case to
No change
lower
UPPER
Remove gap symbols:
-
Calculate checksum of sequences
Select
all, or
sequences by number:
Translate bases (list as from-base:to-base pairs)

BLAST – Εφαρμογή (1/2)

BLAST Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. more...

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. Go

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*

- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*

- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

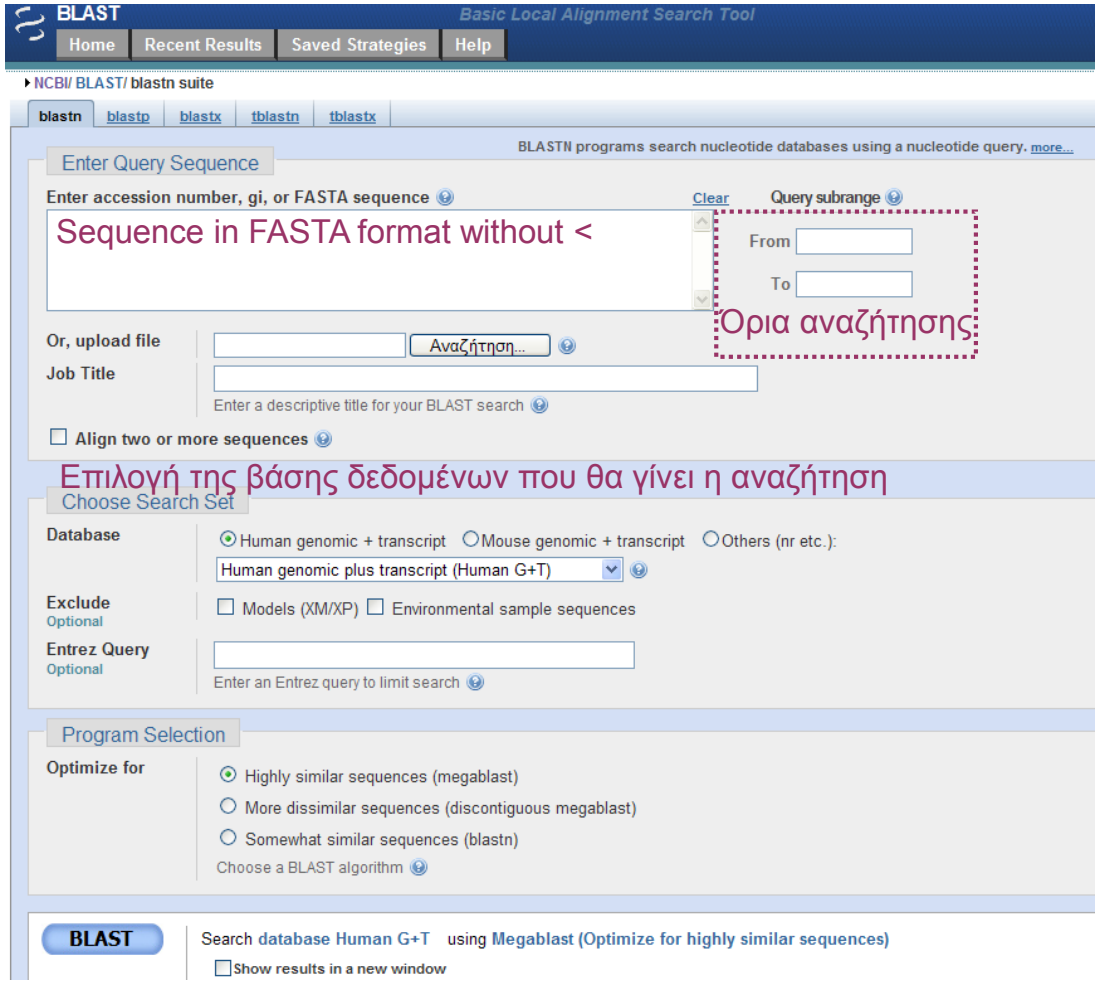
BLAST – Εφαρμογή (2/2)

Basic BLAST

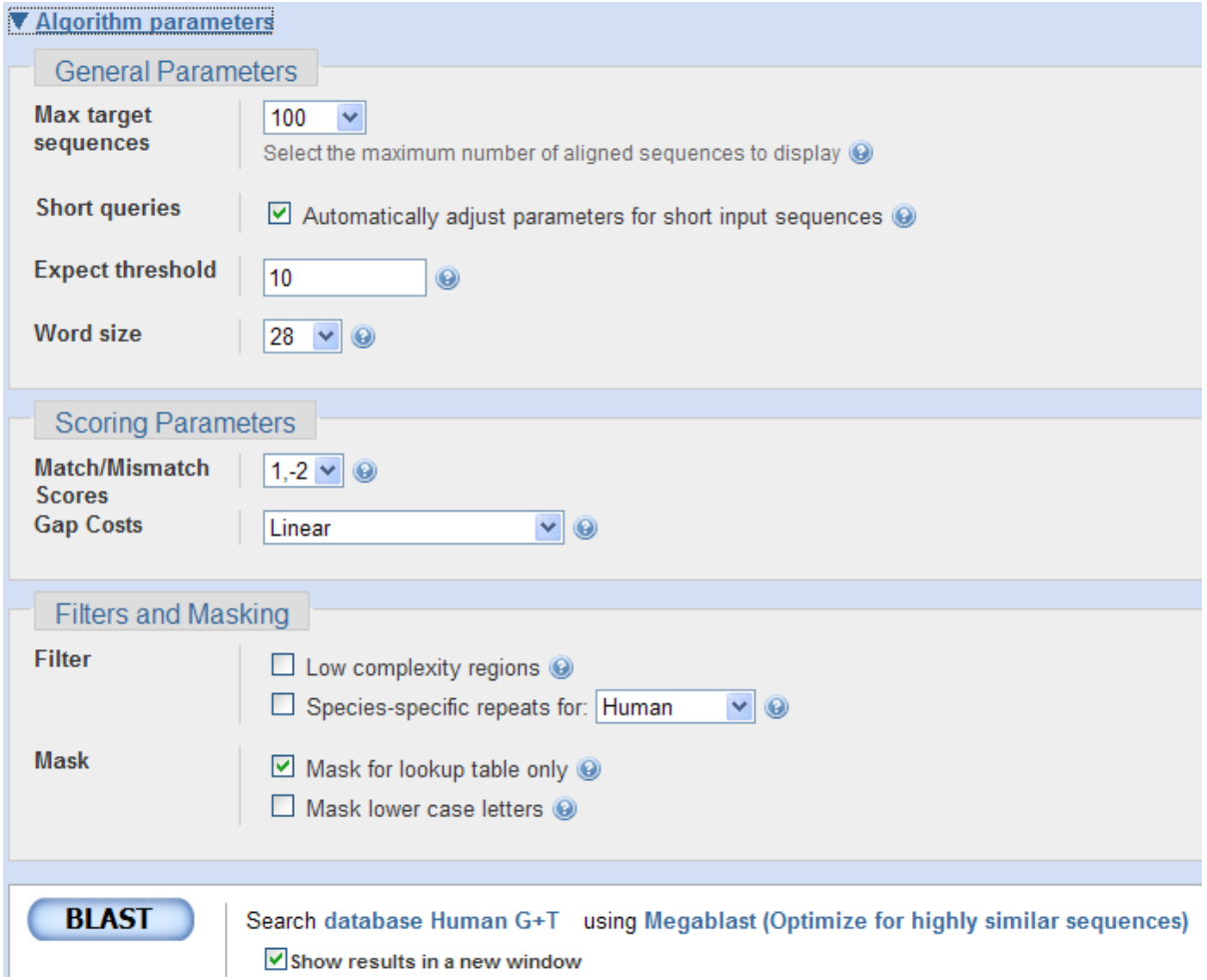
Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

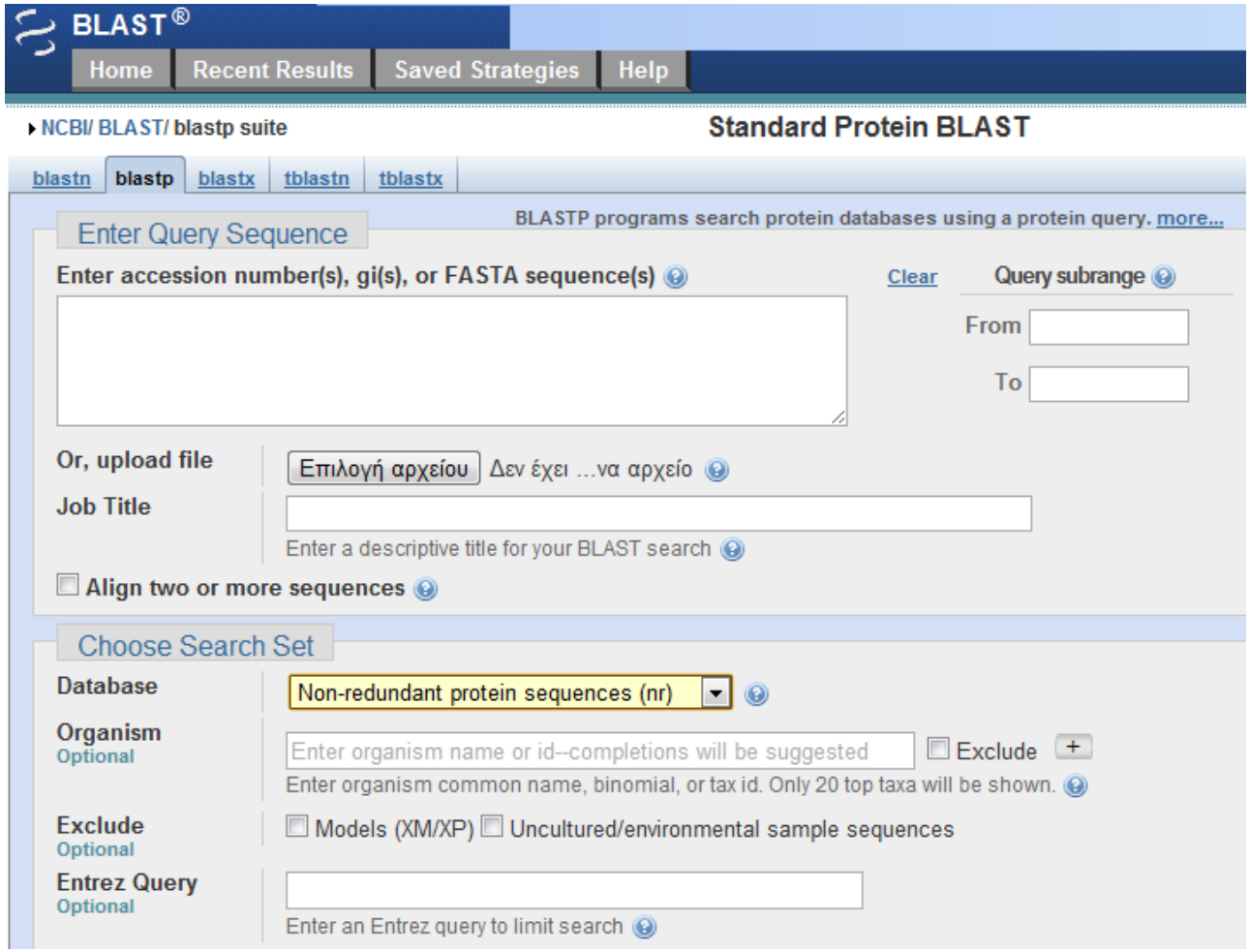
BLAST (1/3)


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastn suite
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
Sequence in FASTA format without <
From
To
Ορια αναζήτησης
Or, upload file
Αναζήτηση...
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Επιλογή της βάσης δεδομένων που θα γίνει η αναζήτηση
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Human genomic plus transcript (Human G+T)
Exclude
Optional
Models (XM/XP)
Environmental sample sequences
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Human G+T using Megablast (Optimize for highly similar sequences)
Show results in a new window

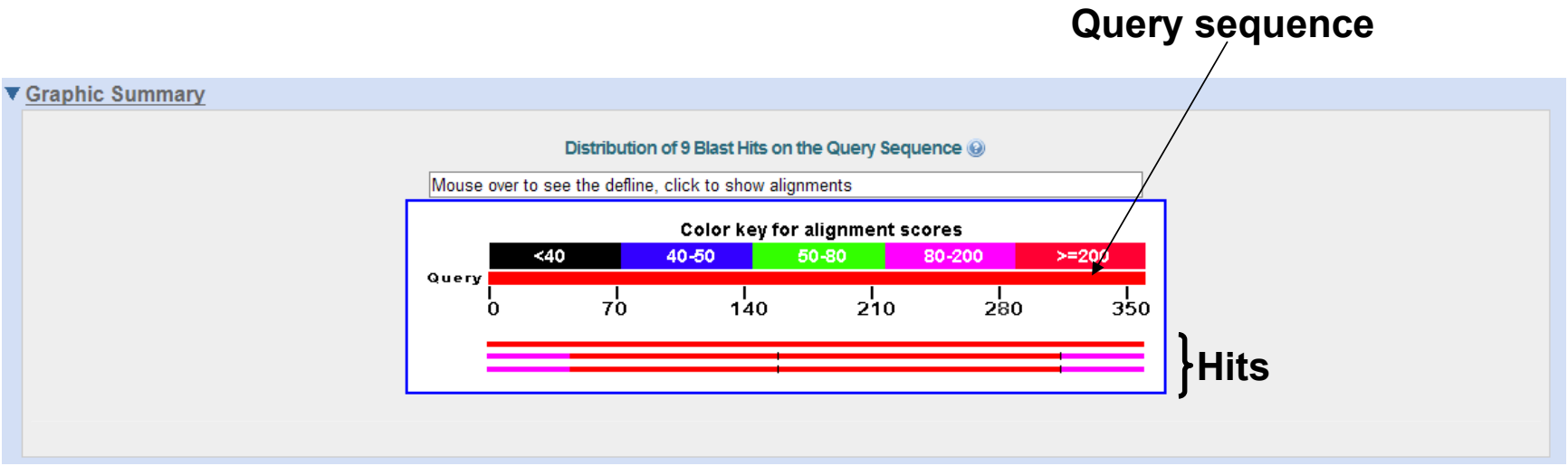
BLAST (2/3)


▼ Algorithm parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
28
Scoring Parameters
Match/Mismatch Scores
1,-2
Gap Costs
Linear
Filters and Masking
Filter
Low complexity regions
Species-specific repeats for:
Human
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Human G+T using Megablast (Optimize for highly similar sequences)
Show results in a new window

BLAST (3/3)

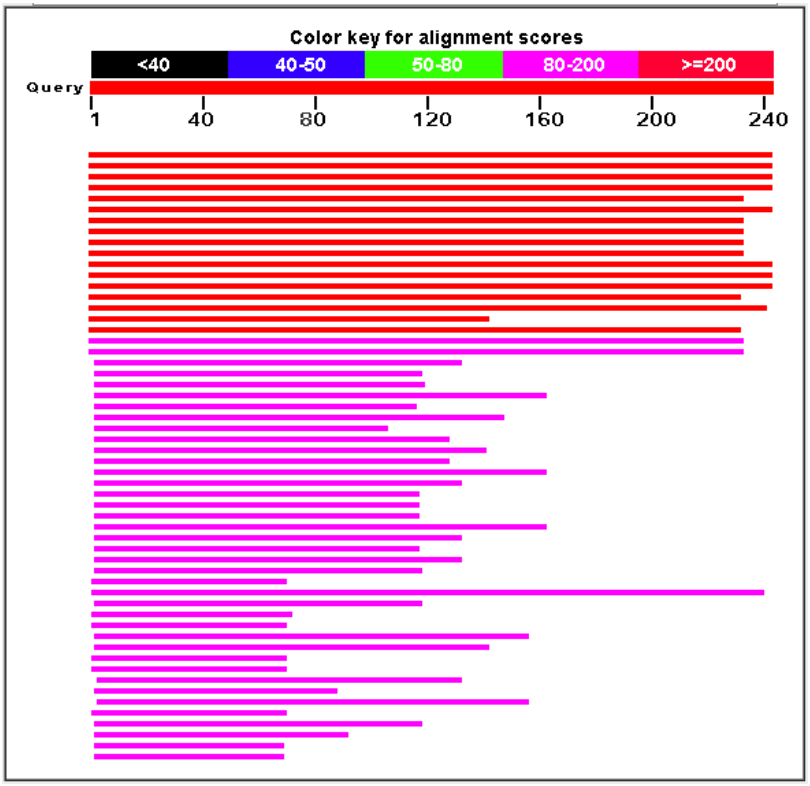

BLAST®
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Επιλογή αρχείου
Δεν έχει ...να αρχείο
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
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

BLAST Results – Graphic Summary



- Εικονική μορφή αποτελεσμάτων.
- Σειρά προτεραιότητας ανάλογα με τη βαθμολόγηση της στοίχισης.

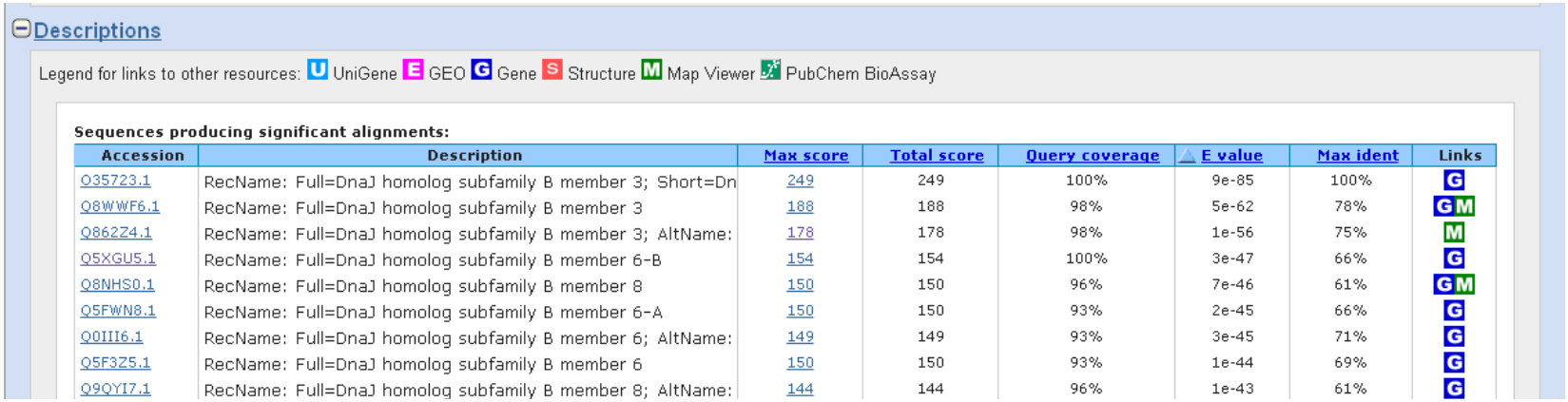
BLAST Results – Domain identification



- **Αποτελέσματα με μικρότερο μήκος:** έχουν κοινό τμήμα (στην αρχή της αλληλουχίας) με την υπό εξέταση αλληλουχία.

BLAST Results – Hit list

- Τα ονόματα των αλληλουχιών που παρουσιάζουν ομοιότητα με την υπό εξέταση αλληλουχία.
- Η σειρά κατάταξης είναι ανάλογη με τον ποσοστό ομοιότητας.



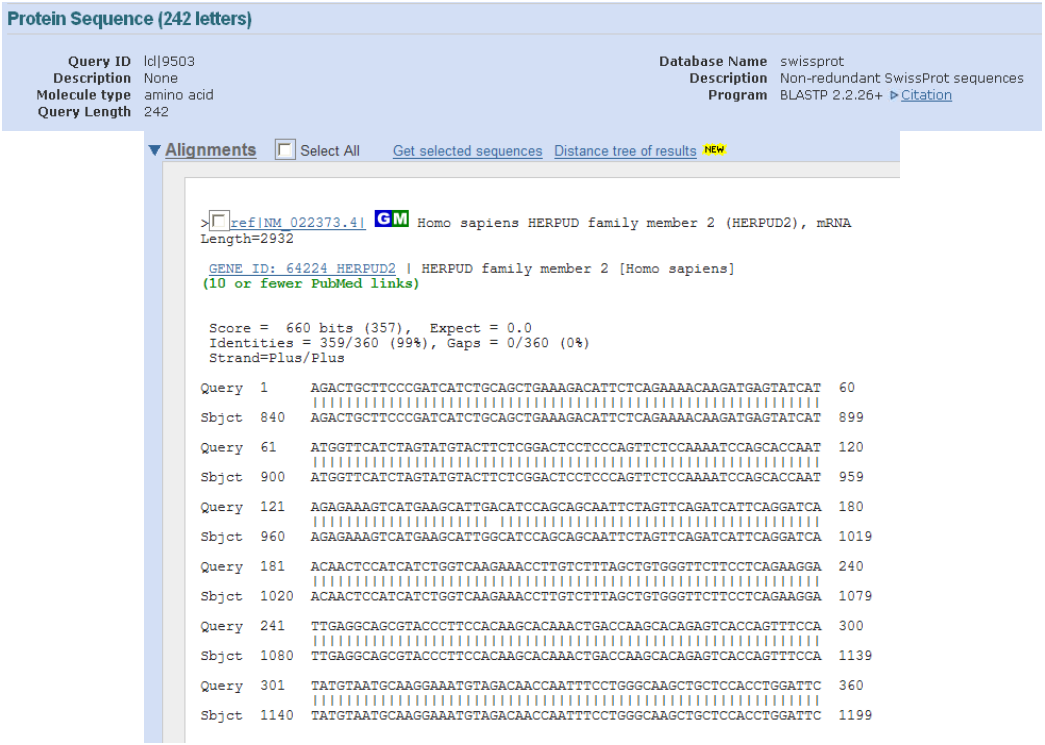
Descriptions

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

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
O35723.1	RecName: Full=DnaJ homolog subfamily B member 3; Short=Dn	249	249	100%	9e-85	100%	G
Q8WWF6.1	RecName: Full=DnaJ homolog subfamily B member 3	188	188	98%	5e-62	78%	GM
Q862Z4.1	RecName: Full=DnaJ homolog subfamily B member 3; AltName:	178	178	98%	1e-56	75%	M
Q5XGU5.1	RecName: Full=DnaJ homolog subfamily B member 6-B	154	154	100%	3e-47	66%	G
Q8NHS0.1	RecName: Full=DnaJ homolog subfamily B member 8	150	150	96%	7e-46	61%	GM
Q5FWN8.1	RecName: Full=DnaJ homolog subfamily B member 6-A	150	150	93%	2e-45	66%	G
Q0III6.1	RecName: Full=DnaJ homolog subfamily B member 6; AltName:	149	149	93%	3e-45	71%	G
Q5F3Z5.1	RecName: Full=DnaJ homolog subfamily B member 6	150	150	93%	1e-44	69%	G
Q9QYI7.1	RecName: Full=DnaJ homolog subfamily B member 8; AltName:	144	144	96%	1e-43	61%	G

BLAST Results – Alignments and parameters



Protein Sequence (242 letters)

Query ID	lcl|9503	**Database Name**	swissprot
Description	None	**Description**	Non-redundant SwissProt sequences
Molecule type	amino acid	**Program**	BLASTP 2.2.26+ ▷Citation
Query Length	242		

▼ **Alignments** Select All Get selected sequences Distance tree of results NEW

>ref|NM_022373.4| GM Homo sapiens HERPUD family member 2 (HERPUD2), mRNA
Length=2932

GENE ID: 64224 HERPUD2 | HERPUD family member 2 [Homo sapiens]
(10 or fewer PubMed links)

Score = 660 bits (357), Expect = 0.0
Identities = 359/360 (99%), Gaps = 0/360 (0%)
Strand=Plus/Plus

```
Query  1     AGACTGCTTCCCGATCATCTGCAGCTGAAAGACATTCTCAGAAAACAAGATGAGTATCAT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  840   AGACTGCTTCCCGATCATCTGCAGCTGAAAGACATTCTCAGAAAACAAGATGAGTATCAT  899

Query  61    ATGGTTCATCTAGTATGTACTTCTCGGACTCCTCCCAGTTCTCCAAAATCCAGCACCAAT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  900   ATGGTTCATCTAGTATGTACTTCTCGGACTCCTCCCAGTTCTCCAAAATCCAGCACCAAT  959

Query  121   AGAGAAAGTCATGAAGCATTGACATCCAGCAGCAATTCTAGTTCAGATCATTCAGGATCA  180
             |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  960   AGAGAAAGTCATGAAGCATTGGCATCCAGCAGCAATTCTAGTTCAGATCATTCAGGATCA  1019

Query  181   ACAACTCCATCATCTGGTCAAGAAACCTTGTCTTTAGCTGTGGGTTCTTCCTCAGAAGGA  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1020  ACAACTCCATCATCTGGTCAAGAAACCTTGTCTTTAGCTGTGGGTTCTTCCTCAGAAGGA  1079

Query  241   TTGAGGCAGCGTACCCTTCCACAAGCACAAACTGACCAAGCACAGAGTCACCAGTTTCCA  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1080  TTGAGGCAGCGTACCCTTCCACAAGCACAAACTGACCAAGCACAGAGTCACCAGTTTCCA  1139

Query  301   TATGTAATGCAAGGAAATGTAGACAACCAATTTCCTGGGCAAGCTGCTCCACCTGGATTC  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1140  TATGTAATGCAAGGAAATGTAGACAACCAATTTCCTGGGCAAGCTGCTCCACCTGGATTC  1199
```

- Όλες οι στοιχίσεις μεταξύ της υπό εξέταση αλληλουχία και των προτεινόμενων αλληλουχιών από τη βάση δεδομένων.
- Μία λίστα από τις παραμέτρους που χρησιμοποιήθηκαν για την αναζήτηση.

BLAST Results – Protein sequence

Alignments

Select All Get selected sequences Distance tree of results Multiple alignment

>sp|O35723.1|DNJB3_MOUSE RecName: Full=DnaJ homolog subfamily B member 3; Short=DnaJ protein homolog 3; AltName: Full=Heat shock protein J3; Short=HSJ-3;
AltName: Full=MSJ-1
Length=242

GENE ID: 15504 Dnajb3 | DnaJ (Hsp40) homolog, subfamily B, member 3
[Mus musculus] **(Over 10 PubMed links)**

Score = 249 bits (637), Expect = 9e-85, Method: Compositional matrix adjust.
Identities = 120/120 (100%), Positives = 120/120 (100%), Gaps = 0/120 (0%)

```
Query  1    MVDYYEVLGVPRQASAEAIRKAYRKLALKWHPDKNPEHKEEAERRFKQVAQAYEVLSDVR  60
            MVDYYEVLGVPRQASAEAIRKAYRKLALKWHPDKNPEHKEEAERRFKQVAQAYEVLSDVR
Sbjct  1    MVDYYEVLGVPRQASAEAIRKAYRKLALKWHPDKNPEHKEEAERRFKQVAQAYEVLSDVR  60

Query  61   KREVYDRCGEVGEVGGGGAAGSPFHDAFQYVFSFRDPAEVFREFFGGHDPFSFDFFGGDP  120
            KREVYDRCGEVGEVGGGGAAGSPFHDAFQYVFSFRDPAEVFREFFGGHDPFSFDFFGGDP
Sbjct  61   KREVYDRCGEVGEVGGGGAAGSPFHDAFQYVFSFRDPAEVFREFFGGHDPFSFDFFGGDP  120
```

>sp|Q8WWF6.1|DNJB3_HUMAN RecName: Full=DnaJ homolog subfamily B member 3
Length=145

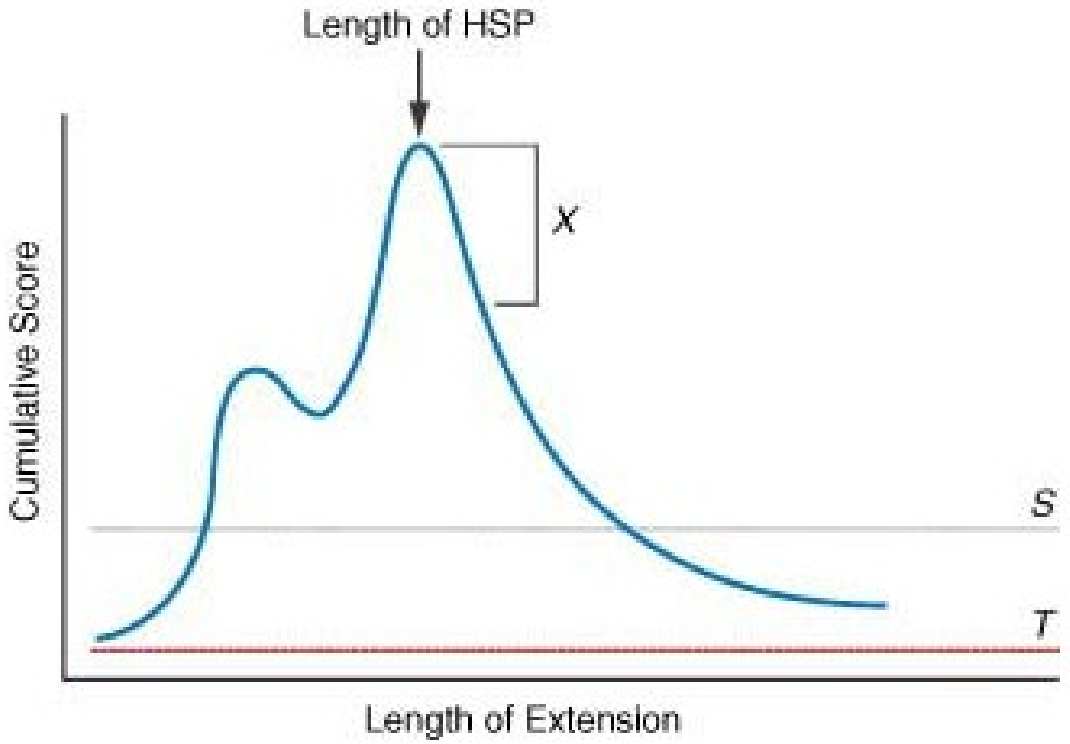
GENE ID: 414061 DNAJB3 | DnaJ (Hsp40) homolog, subfamily B, member 3
[Homo sapiens] **(10 or fewer PubMed links)**

Score = 188 bits (478), Expect = 5e-62, Method: Compositional matrix adjust.
Identities = 92/118 (78%), Positives = 101/118 (86%), Gaps = 1/118 (1%)

```
Query  1    MVDYYEVLGVPRQASAEAIRKAYRKLALKWHPDKNPEHKEEAERRFKQVAQAYEVLSDVR  60
            MVDYYEVL VPRQAS+EAI+KAYRKLALKWHPDKNPE+KEEAERRFKQVA+AYEVLSD +
Sbjct  1    MVDYYEVLDVPRQASSEAIKKAYRKLALKWHPDKNPENKEEAERRFKQVAEAYEVLSDAK  60

Query  61   KREVYDRCGEVGEVGGGGAAGSPFHDAFQYVFSFRDPAEVFREFFGGHDPFSFDFFGG  118
            KR++YDR GE G   GG   G PF D F+YVFSFRDPA+VFREFFGG DPFSFD  G
Sbjct  61   KRDIYDRYGEAG-AEGGCTGGRPFEDPFEYVFSFRDPADVFREFFGGQDPFSFDLLGN  117
```

Extending the High Scoring Segment Pair (HSP)



Neighborhood Score Threshold (T)

Minimum Score (S)

Τέλος Ενότητας

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημείωμα Αναφοράς

- Copyright Πανεπιστήμιο Δυτικής Μακεδονίας, Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών, Αγγελίδης Παντελής. «**Βιοπληροφορική**». Έκδοση: 1.0. Κοζάνη 2015. Διαθέσιμο από τη δικτυακή διεύθυνση:

  https: //eclass.uowm.gr/courses/ICTE102/

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.