ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

Βιοπληροφορική

Ενότητα 10:

Αναζήτηση Ομοιοτήτων σε ΒΔ Ακολουθιών - Blast, (2/2) 1ΔΩ

Τμήμα: Βιοτεχνολογίας

Όνομα καθηγητή: **Τ. Θηραίου**

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Μαθησιακοί Στόχοι

- Αναφορά στις παραλλαγές του BLAST.
- Εξοικείωση με τη διεπαφή του BLAST.

Λέξεις Κλειδιά

- Λέξεις κλειδιά: Βάσεις δεδομένων BLAST, Παραλλαγές BLAST, Ανταποδοτικό BLAST.

- Key words: BLAST databases, BLAST family of programs, PSI-BLAST, PHI-BLAST, Reciprocal BLAST, BLAST interface.

Παραλλαγές του BLAST 11/12

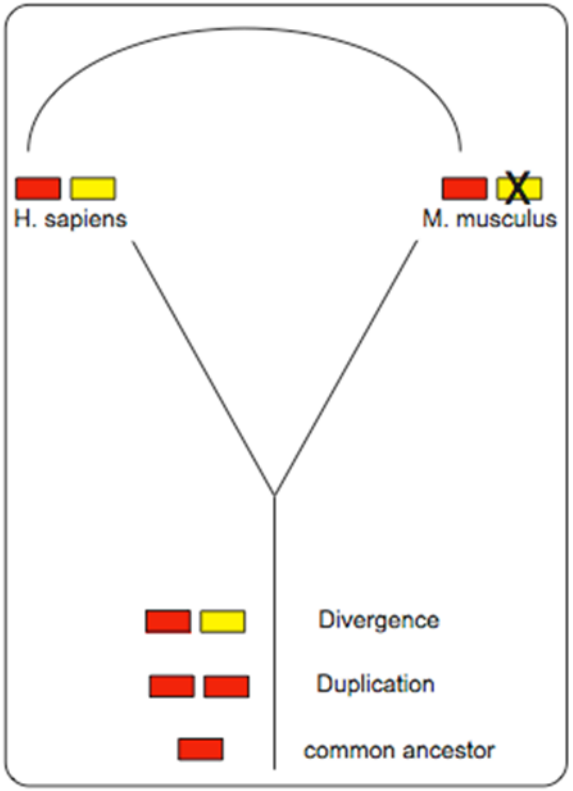
- **PHI-BLAST**:
  - **μοτίβο (pattern)**.
    - χαρακτηρίζει μια οικογένεια πρωτεϊνών.
    - π.χ. [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].

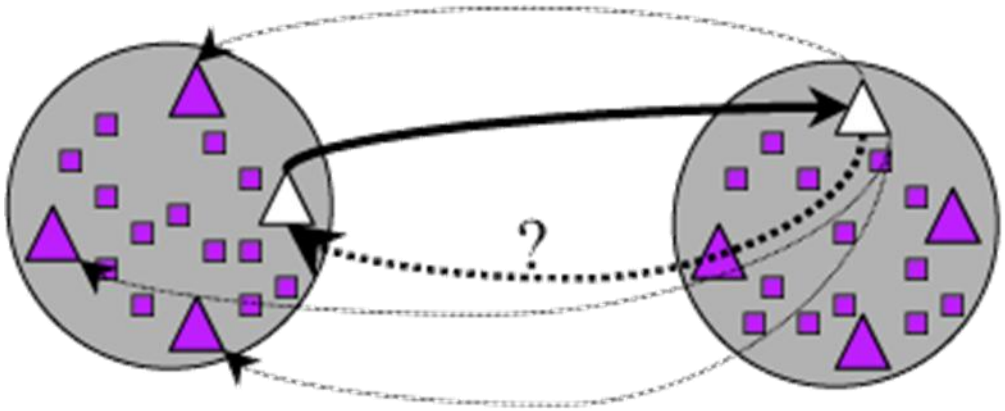
Παραλλαγές του BLAST 12/12

- **PHI-BLAST**:
  - Δεδομένα εισόδου για την αναζήτηση:
    - ακολουθία επερώτησης.
    - μοτίβο (regular expression) που υπάρχει στην ακολουθία επερώτησης.
  - Αναζήτηση ακολουθιών που **περιέχουν το μοτίβο** και έχουν **ομοιότητα** με την ακολουθία επερώτησης στη γειτονική περιοχή του μοτίβου.
  - **Μείωση των hits** που **δεν έχουν πραγματική ομολογία** με την ακολουθία επερώτησης.

Ανταποδοτικό BLAST 1/3

- **Best Reciprocal (BLAST) Hit**:





Ανταποδοτικό BLAST 2/3

- Εύρεση ορθόλογων γονιδίων / πρωτεϊνών.
- Αναζήτηση BLAST με την ακολουθία α του οργανισμού Α στις ακολουθίες του οργανισμού Β.
  - καλύτερο hit η ακολουθία β.
- Αναζήτηση BLAST με την ακολουθία β του οργανισμού Β στις ακολουθίες του οργανισμού Α.
  - καλύτερο hit η ακολουθία α.
- Οι ακολουθίες α και β είναι ορθόλογες.

Ανταποδοτικό BLAST 3/3

- Κρίσιμες παράμετροι:
  - είδος φίλτρου: soft filtering vs hard filtering.
  - αλγόριθμος τελικής στοίχισης: BLAST vs Smith-Waterman.
  - τιμές κατωφλίου: E-value ή bit-score, μήκος στοίχισης.
- Σφάλματα:
  - πρόσφατος εκτεταμένος γονιδιακός διπλασιασμός.
  - γονιδιακή σύντηξη.
  - domain rearrangements.

Διεπαφή BLAST 1/6

Basic BLAST

Choose a BLAST program to run.

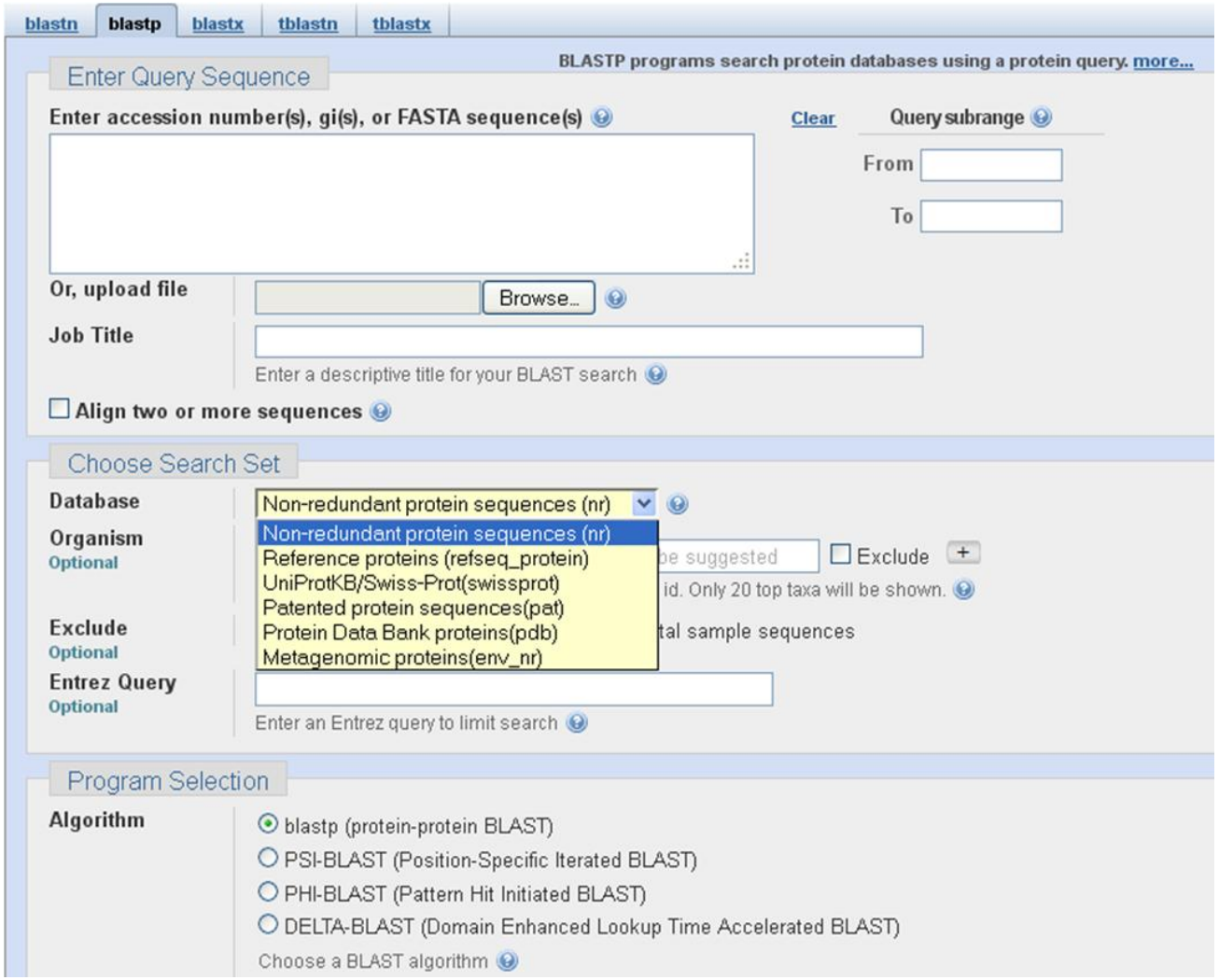
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using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

Specialized BLAST

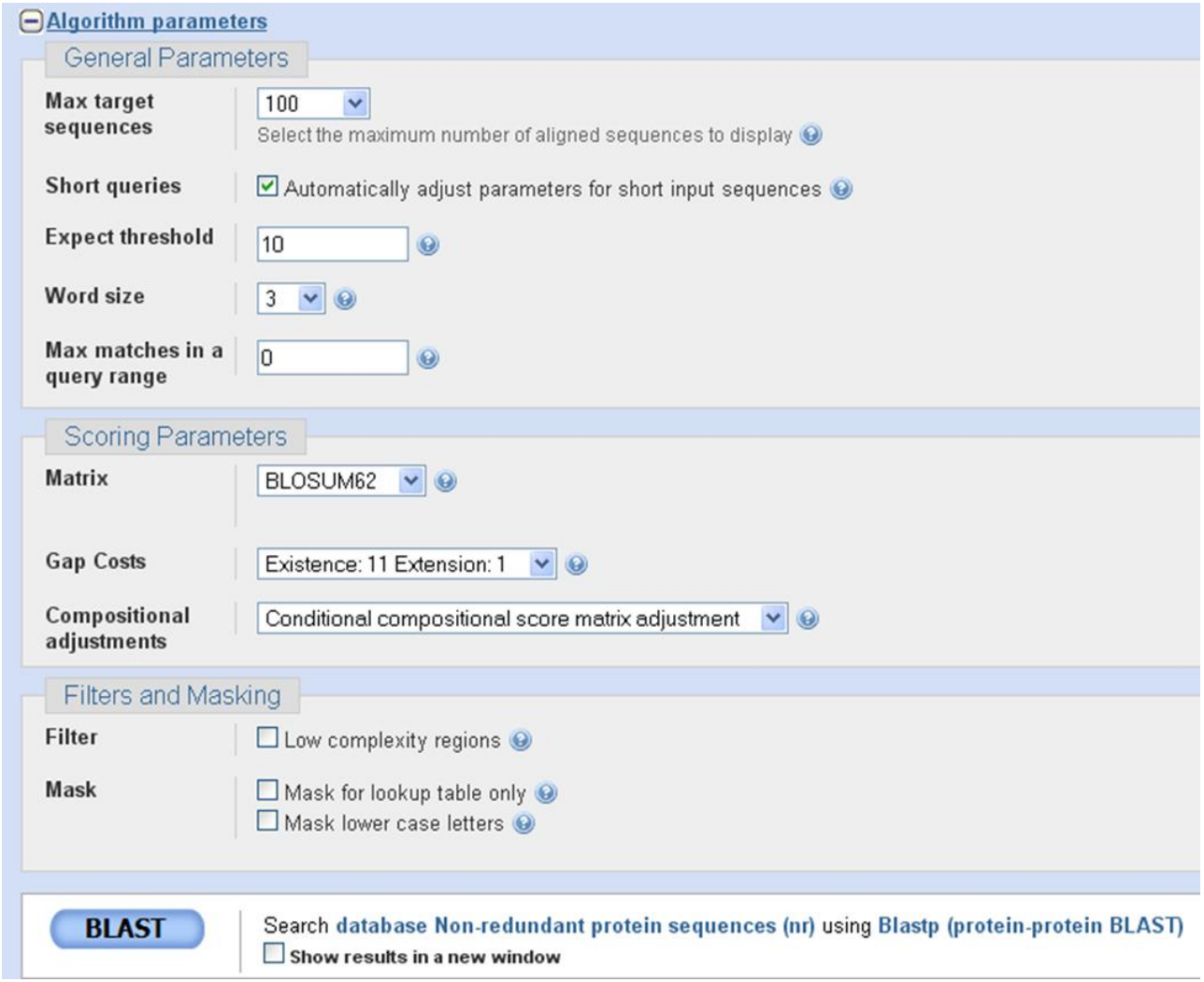
Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene

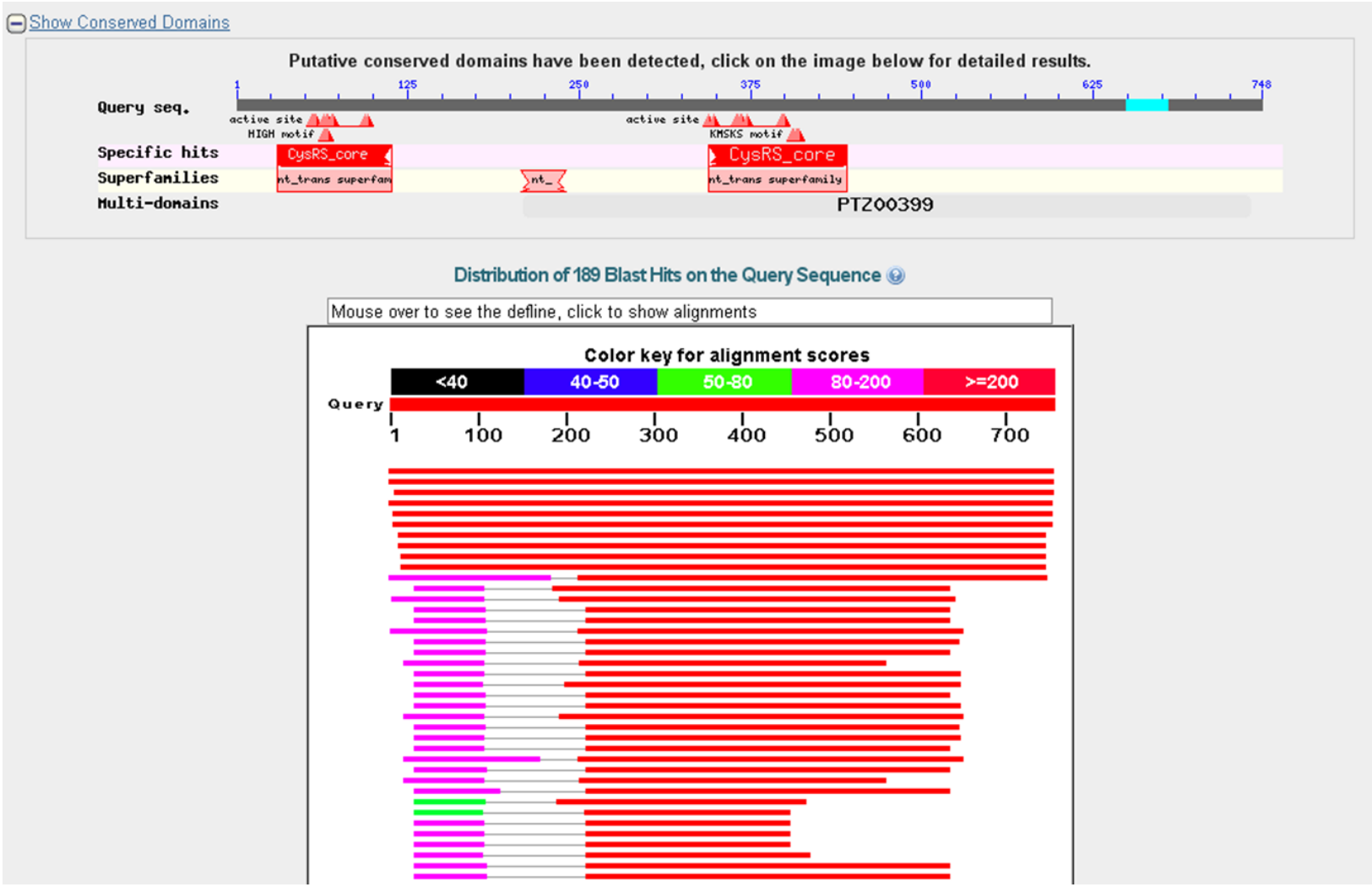
Διεπαφή BLAST 2/6



Διεπαφή BLAST 3/6



Διεπαφή BLAST 4/6



Διεπαφή BLAST 5/6

Descriptions

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

NEW - alignment score below the threshold on the previous iteration

- alignment was checked on the previous iteration

Run PSI-Blast iteration 2 with max 500 Go

Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NEW P49589.3	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	1489	1489	100%	0.0	100%	G M
NEW Q4R550.1	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	1447	1447	100%	0.0	97%	
NEW Q9ER72.2	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	1360	1360	99%	0.0	90%	G M
NEW Q5F408.1	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	1246	1246	99%	0.0	82%	G M
NEW Q5M7N8.1	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	1158	1158	99%	0.0	77%	G
NEW Q7ZWR2.1	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	1153	1153	99%	0.0	77%	G
NEW Q7KN90.1	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	821	821	97%	0.0	59%	G M
NEW Q291L4.1	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	821	821	97%	0.0	59%	G
NEW P53852.1	RecName: Full=Cysteine--tRNA ligase; AltName: Full=Cysteinyl-tRNA	589	589	97%	0.0	45%	G
NEW Q09860.1	RecName: Full=Probable cysteine--tRNA ligase; AltName: Full=Cysteir	561	561	97%	0.0	44%	G
NEW Q54KR1.1	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	473	580	94%	5e-155	47%	

Διεπαφή BLAST 6/6

>sp|Q54KR1.1|SYCC_DICDI RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cysteinyl-tRNA synthetase; Short=CysRS
Length=660

Sort alignments for this subject sequence by:
E value Score Percent identity
Query start position Subject start position

Score = 473 bits (1216), Expect = 5e-155, Method: Compositional matrix adjust.
Identities = 250/532 (47%), Positives = 350/532 (66%), Gaps = 17/532 (3%)

```
Query  213  SKLPKFWEGDFHRDMEALNVLPPDVLTRVSEYVPEIVNFVQKIVDNGYGYVSNGSVYFDT  272
            S L K WE  F  DM+ LNVLPPD LTRV+EYVP+IV +V+KI+ NG+ Y SNGSVYFDT
Sbjct  124  SDLSKKWETAFFEDMKLLNVLPPDALTRVTEYVPQIVEYVEKIISNGFAYESNGSVYFDT  183

Query  273  AKFASSEKHSYGKLVPEAVGDQKALQEGEGDLSISADRLSEKRSPNDFALWKASKPGEPS  332
              F  S+ H YGKL P +VG++K   EGEG L+ ++  +SEKRS  DFALWK SKPGEP
Sbjct  184  VAF--SKAHDYGKLEPNSVGNEKLAAEGEGSLTATS-AVSEKRSGFDFALWKKSKPGEPV  240

Query  333  WPCPWGKGRPGWHIECSAMAGTLLGASMDIHGGGFDLRFPHHDNELAQSEAYFENDCWVR  392
            W  PWG+GRPGWHIECSAMA  LLG ++DIH GG DL+FPHHDNELAQSEA++ N  W+
Sbjct  241  WNSPWGEGRPGWHIECSAMASDLLGGNIDIHSGGSDLKFPHHDNELAQSEAFYGNRQWIN  300

Query  393  YFLHTGHLTIAGCKMSKSLKNFITIKDALKKHSARQLRLAFLMHSWKDTLDYSSNTMESA  452
            YF+H+GHL I G KMSKSLKNFITIK AL+K+++RQ+R+ F++H +   ++YS  +M  A
Sbjct  301  YFVHSGHLLIDGLKMSKSLKNFITIKQALEKYTSRQMRMFFILHKYDKAMNYSPESMGYA  360

Query  453  LQYEKFLNEFFLNVKDILR-APVDITGQFEKWGEEEAELNKNFYDKKTAIHKALCDNVDT  511
            ++ EK   EFF   K ILR +P+ +  QF  W + E +LNK+  +    +H+ + DN +T
Sbjct  361  IEMEKTFVEFFHTAKQILRDSPLSLP-QF--WTQAEKDLNKHLQNANDQVHQFILDNFNT  417

Query  512  RTVMEEMRALVSQCNLYMAARKAVRKRPNQALLENIALYLTHMLKIFGAVEEDSS---LG  568
              ++ +  LV++ N+Y+ +    +  P   L+  IA Y+T++  +FG  E  ++   +G
Sbjct  418  SDALKTLSDLVNKTNVYIRSCAEQKTNPRLNLISAIAEYITYIFSVFGLTESSTASSMIG  477

Query  569  FPVGGPGTSLSLEATVMPYLQVLSEFREGVRKIAREQKVPEILQLSDALRDNILPELGVR  628
            F   G G   ++E  + P L  L++FR  VR  A  +    IL+  D LRD +LP LGV+
Sbjct  478  FGSAGKG---NIEEEMTPILNALTQFRSEVRASAIAKDTTSILKTCDNLRDEVLPLLGVK  534

Query  629  FEDHEGLPTVVKLVDRNTLLkereekrrveeekrkkkeeaarrkqeqeaakLAKMKIPPS  688
             +D      + K  D+ TL    +++   ++E  KKK+     K+++   K  K KIPP
Sbjct  535  IDDKSATTAMWKFEDKETL----KKEIEQKKEIEKKKQADKEEKEKKLKEKFEKSKIPPQ  590

Query  689  EMFLSETDKYSKFDENGLPTHDMEGKELSKGQAKKLKKLFEAQEKLYKEYLQ  740
            ++F++ETDKYSKF+E G+PTHD EG E++K Q KKL+K ++ Q K +  YL+
Sbjct  591  QLFINETDKYSKFNELGMPTHDKEGVEITKSQLKKLQKEYDNQTKEHNNYLK  642
```

Έλεγχος Αποτελεσμάτων

- Ομοιότητα σε επαρκές **μήκος** των ακολουθιών.
- Υψηλό ποσοστό **ταυτόσημων** καταλοίπων.
- Εμφάνιση χαρακτηριστικών **δομικών / λειτουργικών μοτίβων**.
- Ποιότητα δεδομένων στις βάσεις.
- Υποθετικά γονίδια.

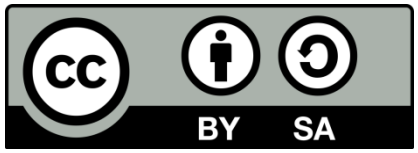
Βιβλιογραφία

- David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press; 2nd edition (March 12, 2013).
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2nd edition (May 4, 2009).
- Andreas D. Baxevanis, B. F. Francis Ouellette, "Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins", Wiley-Interscience; 3rd edition (October 29, 2004).

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.

- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδεια χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα Γεωπονικού Πανεπιστημίου Αθηνών**» έχει χρηματοδοτήσει μόνο την 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.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημείωμα Αναφοράς

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.