ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

Βιοπληροφορική

Ενότητα 8:

Αναζήτηση Ομοιοτήτων σε Βάσεις Δεδομένων Ακολουθιών, 2 ΔΩ

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

- Κατανόηση της αναγκαιότητας των ευριστικών αλγορίθμων στην αναζήτηση ομοιοτήτων σε βάσεις δεδομένων ακολουθιών.
- Επεξήγηση των μεθόδων FASTA και BLAST.

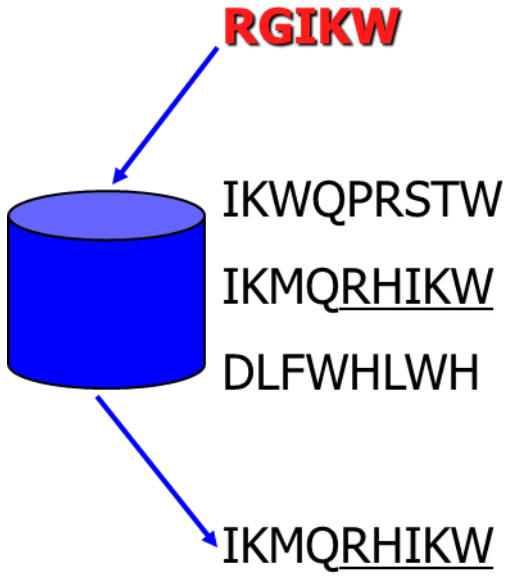
Λέξεις Κλειδιά

- Λέξεις κλειδιά: Αναζήτηση ομοιοτήτων σε βάσεις δεδομένων ακολουθιών, Εφαρμογή φίλτρων.

- Key words: Sequence database similarity searching, FASTA, BLAST, Two-hit BLAST, Gapped BLAST, Filtering.

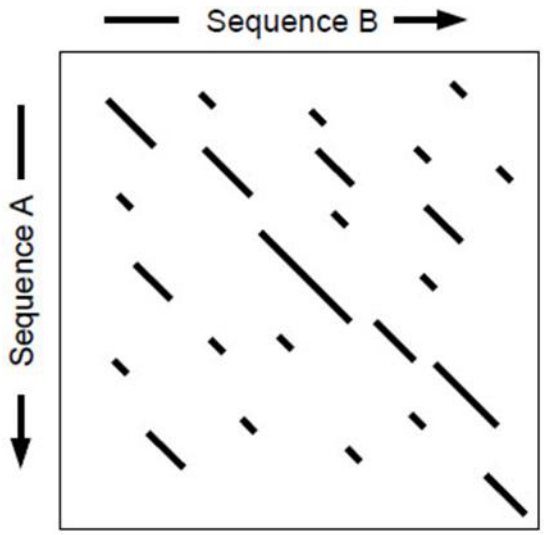
Αναζήτηση Ομοιοτήτων

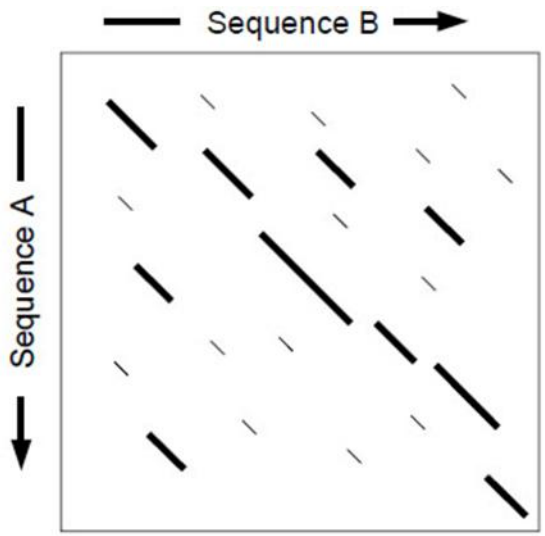
- Δεδομένα.
  - Ακολουθία επερώτησης (query sequence).
  - Ακολουθίες στη Βάση Δεδομένων (subject sequences).
- Αναζήτηση.
  - Μέθοδοι δυναμικού προγραμματισμού:
    - Πρακτικοί μόνο για "μικρά" προβλήματα.
  - Ευριστικοί Αλγόριθμοι:
    - Γρήγορη αναζήτηση.
    - Δεν εγγυώνται την βέλτιστη στοίχιση.
    - FASTA (Lipman and Pearson, 1985).
    - BLAST (Altschul et al, 1990).



FASTA 1/3

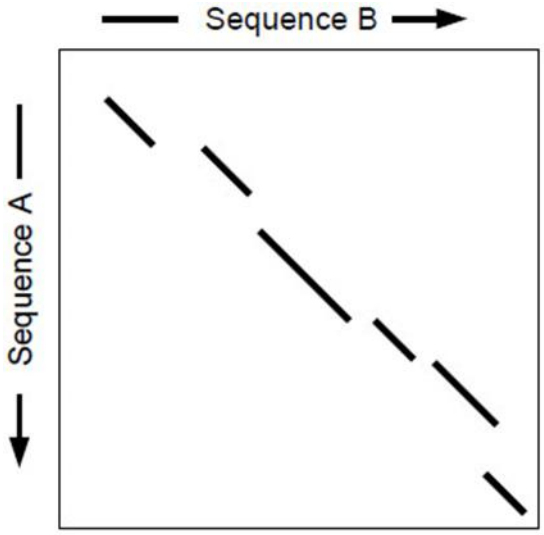
- **K-tuples (K-tup)**: λέξεις μεγέθους k.
- Για κάθε ακολουθία της ΒΔ:
  - αναγνώριση διαγωνίων ταύτισης.
  - βαθμολόγηση με τη χρήση πίνακα αντικατάστασης και επιλογή των διαγωνίων με το μεγαλύτερο score.

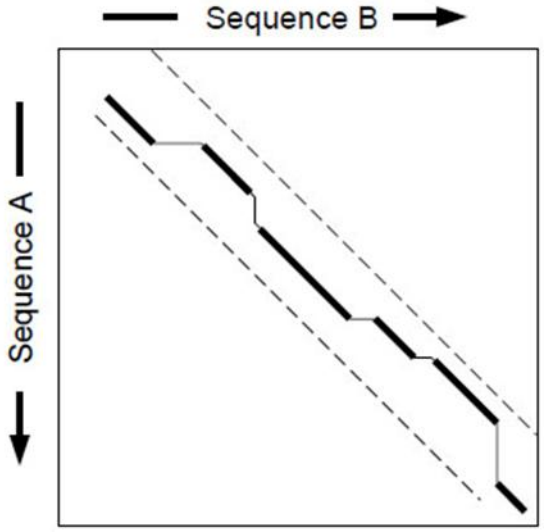




FASTA 2/3

- Για κάθε ακολουθία της ΒΔ:
  - Ένωση των διαγωνίων με score μεγαλύτερο ενός κατωφλίου.
  - Χρήση δυναμικού προγραμματισμού (banded Smith-Waterman) για τη βελτιστοποίηση της στοίχισης.





FASTA 3/3

- K-tuples (K-tup):
  - πρωτεΐνες: 2-tuples.
  - DNA: 6-tuples.

- η **τιμή του k** επηρεάζει την **ευαισθησία** και την **ταχύτητα** της αναζήτησης.
  - μεγάλη τιμή k:
    - μείωση λέξεων που ταυτίζονται αλλά δεν αντιστοιχούν σε πραγματικές στοιχίσεις (background word hits).
    - αύξηση ταχύτητας αλλά μείωση ευαισθησίας (sensitivity).

BLAST 1/12

Query: NKCKT**PQG**QRLVN, W=3, T=13

neighborhood words

word	score
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
…	…

BLAST 2/12


Expanded Word List
HEA
HDA
HQA
HKA
HES
HEC
HET
HEG
HEV
...
Sequence Database
>seq1
GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
>seq2
FQAKLFGIATAVASEHLLGAGTKCCEAPMCQFLAGKPEQVASVRIILEHGTRETLN
>seq3
QAKLFGFQAKLFGIAHECASEVAGKPVASRETLEAPMCQGKPVASVRIILETRETLNEQIILEHG
>seq4
VASVRGAGTKCCEAPMCQFLAGKPVASVRIILETRETLNEQIILEHGHLLFLAGKPVTL
>seq5
HGHLLFLAGKPVASRETLEAGAGTKCCEHGHLLFLAGKPVASVRIILETAGKPVASVRI
>seq6
ILETRETLNGKPVASRETLEAPMCQEQIIHEVVAGKPVASRETLEAPMCQG
>seq7
LETHGHLLFLAGPMCQFLVRIIRIILETAGKPVASVRIKPVASVHGHLLFLAGKPETLEARET
...

BLAST 3/12

```
>query
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
>seq1
GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Align at seed word.

```
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
                                    H+A
    GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Extend sequence to identify HSP.

```
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
            KL Q+LA KPEHP+ R+ L+A H+A +I++    L +
    GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Stop extension of HSP when quality of the alignment reaches a threshold value. Calculate significance.

```
KLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTE
KL Q+LA KPEHP+ R+ L+A H+A +I++    L +
KLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQ
```

```
Score = 42.0 bits (97), Expect = 0.004
Identities = 17/36 (47%), Positives = 26/36 (72%)
```

BLAST 4/12

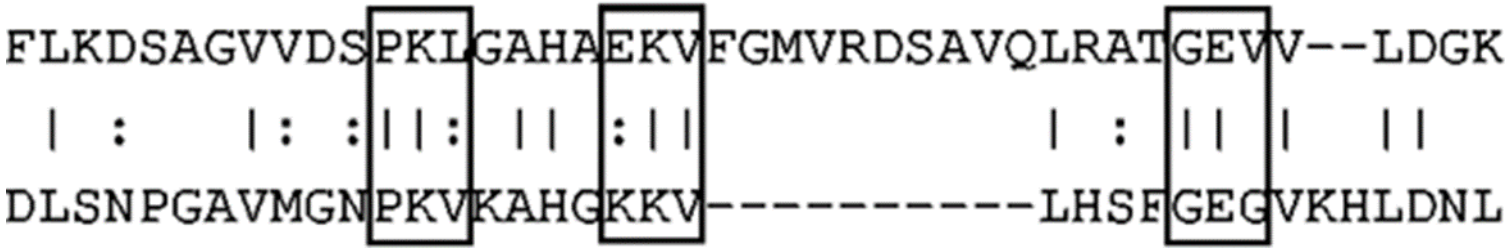
- **Words**: όλες οι δυνατές λέξεις μεγέθους W:
  - πρωτεΐνες: 3 κατάλοιπα.
  - DNA: 11 κατάλοιπα.
- Δημιουργία ευρετηρίου με όλες τις λέξεις στις ακολουθίες της ΒΔ.
- Για κάθε λέξη της ακολουθίας επερώτησης
  - εύρεση όλων των λέξεων (neighborhood words) που στοιχίζονται μαζί της με score μεγαλύτερο ενός κατωφλίου (**neighborhood score threshold, T**) χρησιμοποιώντας έναν πίνακα αντικατάστασης.

BLAST 5/12

- Αναζήτηση των neighborhood words έναντι του προκατασκευασμένου πίνακα με όλες τις λέξεις των ακολουθιών της ΒΔ.
- Επέκταση της στοίχισης μεταξύ της ακολουθίας επερώτησης και των ακολουθιών της ΒΔ για την αναγνώριση ενός **High-scoring Segment Pair (HSP)**.
- Τερματισμός της επέκτασης όταν το score γίνει μικρότερο ενός προκαθορισμένου κατωφλίου.
- **Μεγαλύτερο T**:
  - μείωση των προσπαθειών εύρεσης HSPs.
  - **αύξηση ταχύτητας αλλά μείωση ευαισθησίας**.

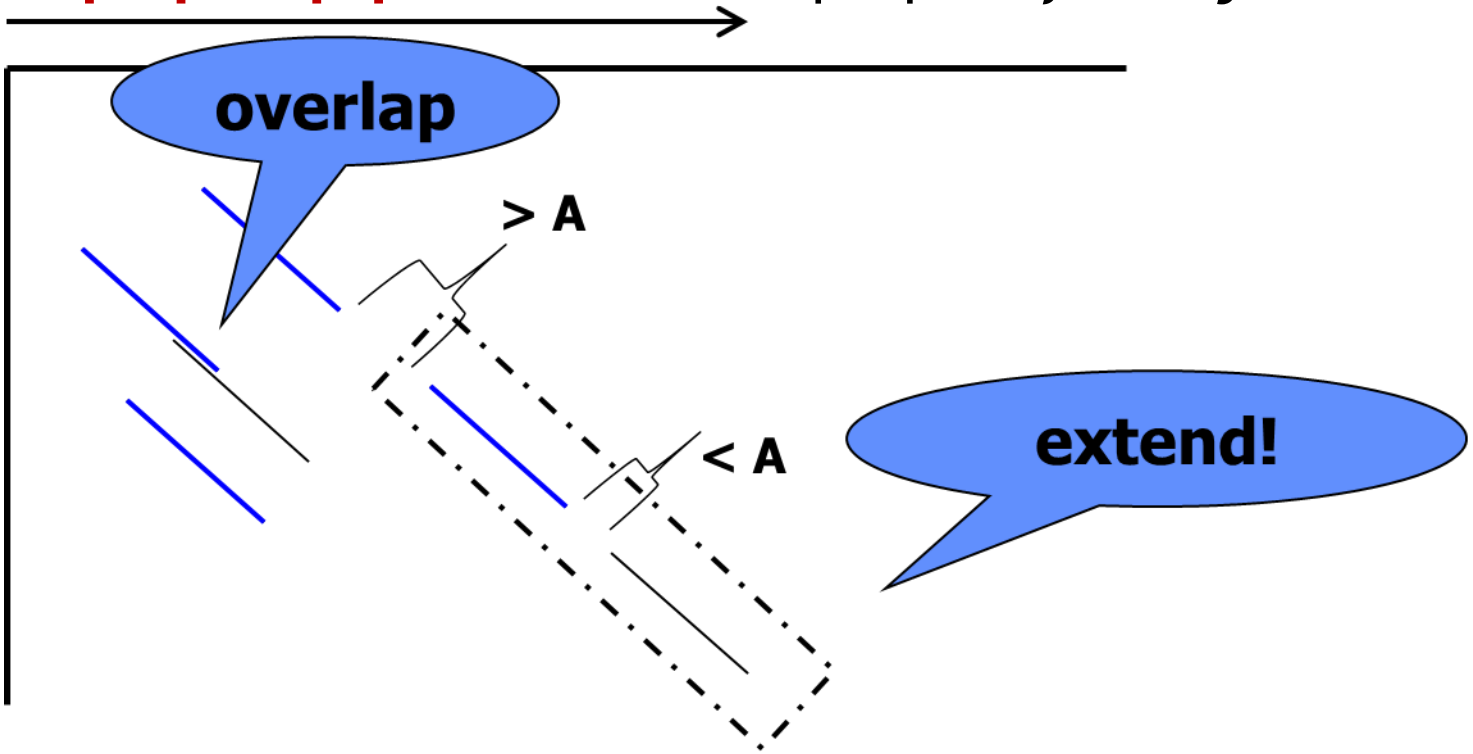
BLAST 6/12

- **two-hit BLAST**:
  - Παρατηρήσεις:
    - Η επέκταση γύρω από τα seed words αποτελεί το 90% του χρόνου εκτέλεσης του BLAST.
    - Τα HSP έχουν μεγαλύτερο μήκος και περιέχουν πολλές λέξεις μήκους W.



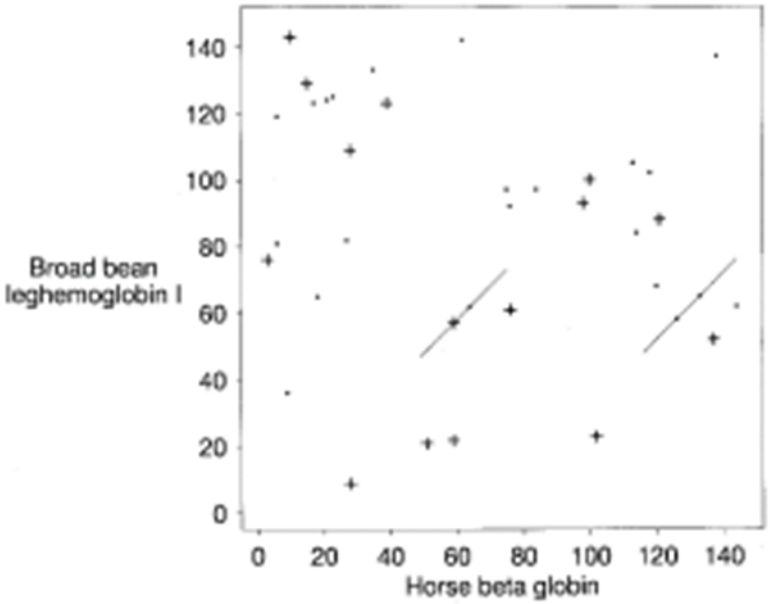
BLAST 7/12

- **two-hit BLAST**:
  - Επέκταση μόνο όταν υπάρχουν **δύο μη επικαλυπτόμενα** ζεύγη λέξεων σε **απόσταση μικρότερη** από ένα κατώφλι μεταξύ τους.



BLAST 8/12

- **two-hit BLAST**:
  - για να διατηρηθεί η ευαισθησία της αναζήτησης, απαιτείται **μικρότερο T** (neighborhood score threshold).
    - περισσότερα hits.
    - αλλά μόνο ένα μικρό ποσοστό από αυτά σχετίζονται με δεύτερο hit.
  - αύξηση της ταχύτητας.

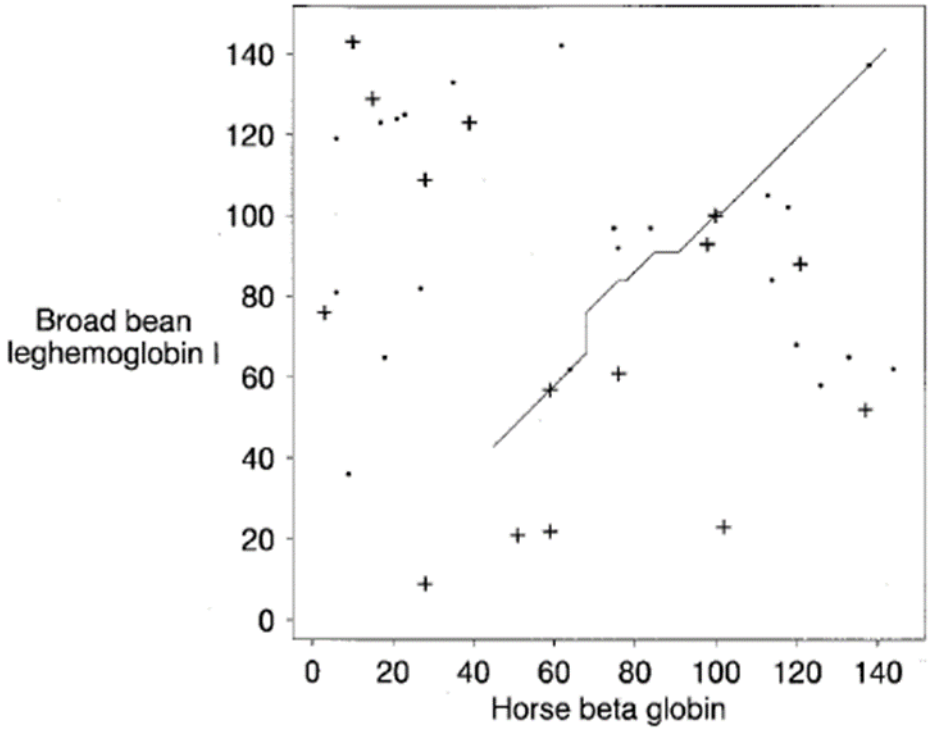


BLAST 9/12

- **Gapped BLAST**:
  - two-hit BLAST για τη δημιουργία ενός HSP.
  - επέκταση του HSP με κενά, αν έχει score μεγαλύτερο ενός κατωφλίου S_g.

BLAST 10/12

- **Gapped BLAST**.




```
Leghemoglobin   43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                   F  L +    V+ +PK+ AH +KV           L + GE V  LD   G+
Beta globin     45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90

Leghemoglobin   91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                   +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin     91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

BLAST 11/12

- **Εφαρμογή φίλτρων**.

```
Length = 810
Score = 91.3 bits (223), Expect = 2e-18
Identities=53/137(38%), Positives=70/137 (50%), Gaps=17/137 (12%)

Q: 1   RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR 60
       +RRR   R  RRRR R R R R R RR    R  +R +  R+RR + RRRR + +R R R
S: 132 QRRREHEREERRRRERERERERGRGRRDENERDPKREQEERQRREQERRRREQEQRERER 191

Q: 61  RRRR-------------RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR----RR 103
       R  R              RR + RR + RRRR + + RR R+RR  R       +R    RR
S: 192 RGERDEEDDENQRDPDWRREQERREQERRRREQEQERRERQRRGGRDDEDENQRDPDWRR 251

Q: 104 RRRRRRRRRRRRRRRR 120   "QUERY SEQUENCE"
        ++RR + + RR R RR      "IDENTICAL/SIMILAR RESIDUES"
S: 252 EQKRREQEQERRERERR 268  "DATABASE TARGET"
```

BLAST 12/12

- Εφαρμογή φίλτρων:
  - επαναλήψεις.
  - περιοχές χαμηλής πολυπλοκότητας.
  - δημιουργία στατιστικώς σημαντικών στοιχίσεων, αλλά χωρίς βιολογικό νόημα.
  - **Soft filtering**:
    - φιλτράρισμα μόνο στη φάση αναζήτησης.
  - **Hard filtering**:
    - φιλτράρισμα στη φάση αναζήτησης και στη φάση τελικής στοίχισης.

Servers

- FASTA:
  - http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
  - http://www.ebi.ac.uk/Tools/sss/fasta/
- BLAST:
  - http://blast.ncbi.nlm.nih.gov/
  - http://www.ebi.ac.uk/Tools/sss/ncbiblast/
- Sequence Similarity Searching:
  - http://www.ebi.ac.uk/Tools/sss/

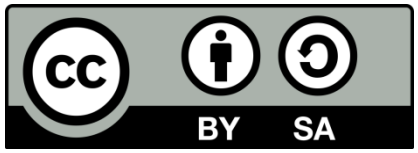
Βιβλιογραφία

- 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.