ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

# Βιοπληροφορική

## Ενότητα 11:

Πολλαπλή Στοίχιση Ακολουθιών, 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

- κατανόηση της έννοιας και των εφαρμογών της πολλαπλής στοίχισης ακολουθιών.
- επιλογή των ακολουθιών προς στοίχιση.
- παρουσίαση των διαφορετικών μορφοποιήσεων της πολλαπλής στοίχισης.
- κατανόηση των εναλλακτικών μεθόδων βαθμολόγησης της πολλαπλής στοίχισης ακολουθιών.

# Λέξεις Κλειδιά

- Λέξεις κλειδιά: πολλαπλή στοίχιση ακολουθιών.
- Key words: Multiple Sequence Alignment, Multiple alignment editors, Minimum Entropy, Sum-of-Pairs Score.

# Πολλαπλή Στοίχιση 1/3

- αποκαλύπτει συντηρημένες περιοχές
- αντιστοίχιση καταλοίπων με κριτήρια ομοιότητας σε επίπεδο
  - δομής
  - εξέλιξης
  - λειτουργίας
  - ακολουθίας

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

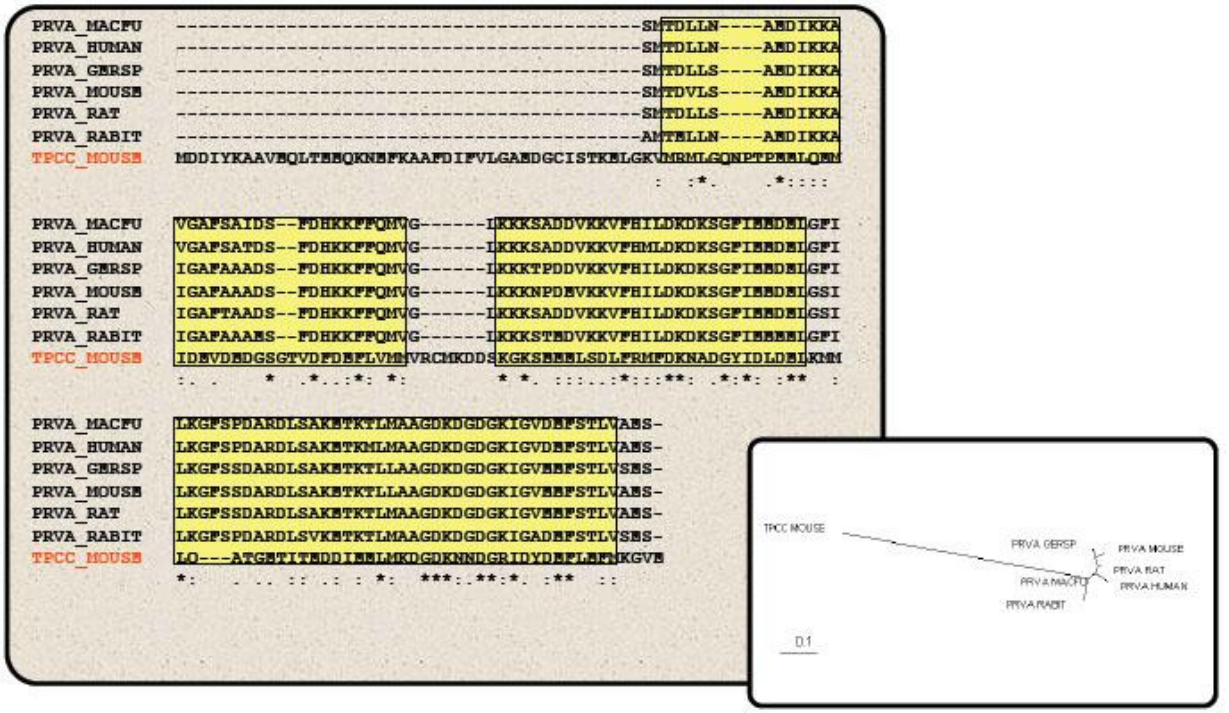
# Πολλαπλή Στοίχιση 2/3

- Εύρεση απομακρυσμένων ομόλογων πρωτεϊνών
- Δημιουργία profile που περιγράφει ένα domain
- Φυλογενετική ανάλυση
- Εύρεση συντηρημένων περιοχών που αντιστοιχούν σε υποκινητές, καταλυτικά κέντρα κ.α.
- Μελέτη της επίδρασης SNPs στη δομή και λειτουργία
- Πρόβλεψη δευτεροταγούς και τριτοταγούς δομής πρωτεϊνών
- Σχεδιασμός εκκινητών PCR

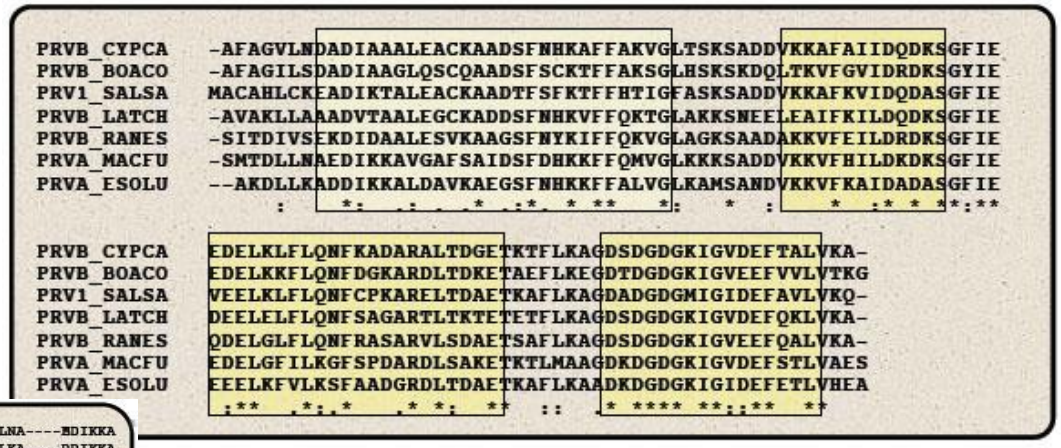
# Πολλαπλή Στοίχιση 3/3

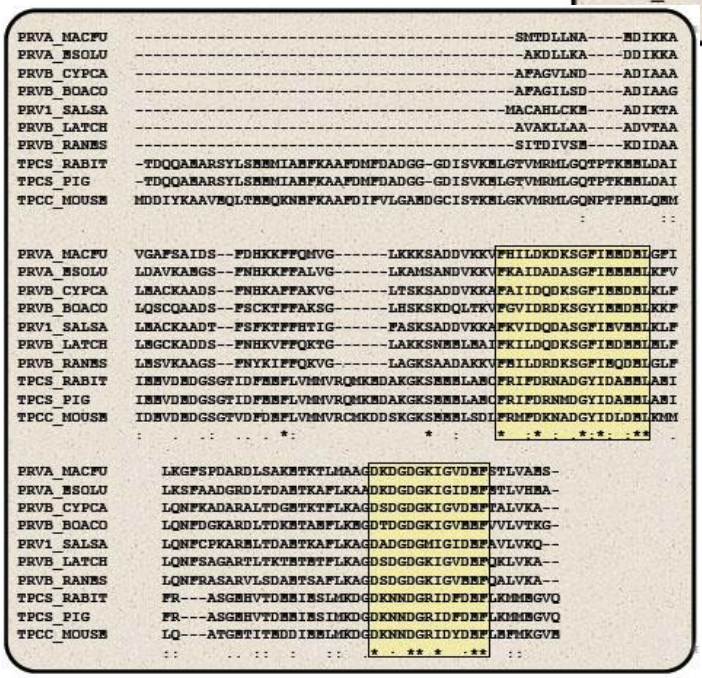

# Επιλογή Ακολουθιών προς Στοίχιση 1/3

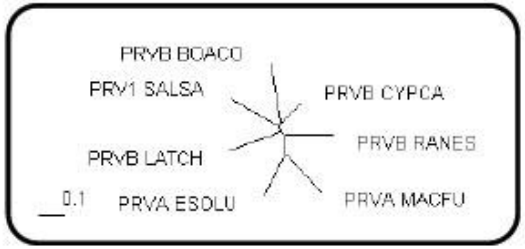
- Η χρήση πανομοιότυπων ακολουθιών στην πολλαπλή στοίχιση δεν προσφέρει πληροφορίες για τη σχέση απομακρυσμένων ομόλογων αλληλουχιών.



# Επιλογή Ακολουθιών προς Στοίχιση 2/3

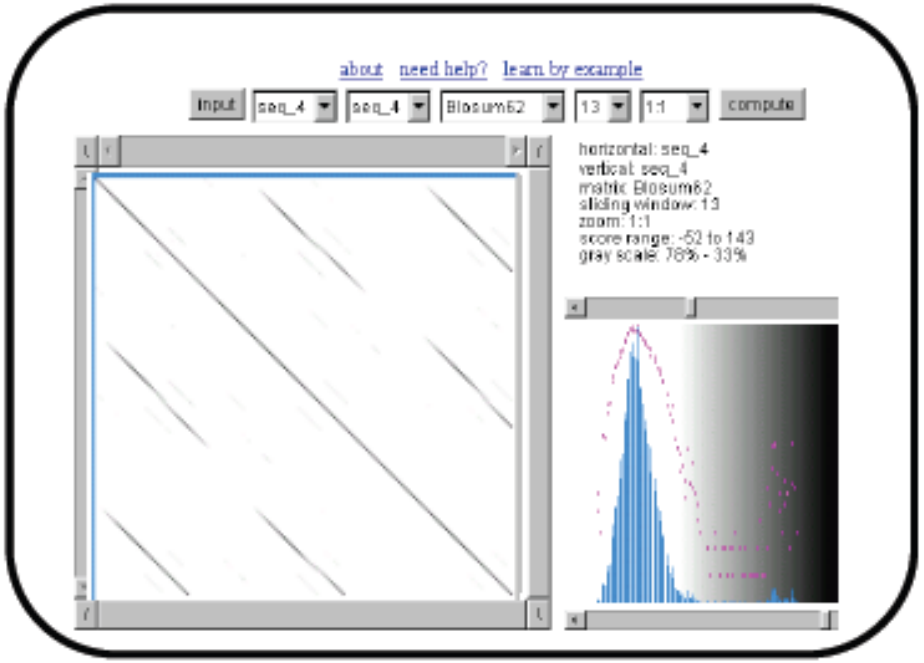


# Επιλογή Ακολουθιών προς Στοίχιση 3/3

- ακολουθίες με διαφορετικό αριθμό επαναλήψεων
  - αναγνώριση των επαναλήψεων (π.χ. dotplot)
  - διαχωρισμός της στοίχισής τους




# Μορφοποίηση Πολλαπλής Στοίχισης 1/2

- **Readseq** http://www.ebi.ac.uk/cgi-bin/readseq.cgi
  - Μετατροπή από το ένα format στο άλλο

- **FASTA** (.fa ή .fasta ή .fst)

```
>IXI_234
TSPASIRPPAGPSSRPAMVSSRRTRPSPPGPRRPTGRPCCSAAPRRPQATGGWKTCSGTC
TTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSRSAGSRPNRFAPTLMSSCITSTTG
PPAWAGDRSHE
>IXI_235
TSPASIRPPAGPSSR---------RPSPPGPRRPTGRPCCSAAPRRPQATGGWKTCSGTC
TTSTSTRHRGRSGW----------RASRKSMRAACSRSAGSRPNRFAPTLMSSCITSTTG
PPAWAGDRSHE
>IXI_236
TSPASIRPPAGPSSRPAMVSSR--RPSPPPPRRPPGRPCCSAAPPRPQATGGWKTCSGTC
TTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSR--GSRPPRFAPPLMSSCITSTTG
PPPPAGDRSHE
>IXI_237
TSPASLRPPAGPSSRPAMVSSRR-RPSPPGPRRPT----CSAAPRRPQATGGYKTCSGTC
TTSTSTRHRGRSGYSARTTTAACLRASRKSMRAACSR--GSRPNRFAPTLMSSCLTSTTG
PPAYAGDRSHE
```

# Μορφοποίηση Πολλαπλής Στοίχισης 2/2

- **Phylip** (.phy ή .phylip)

```
 4 131
IXI_234      TSPASIRPPA GPSSRPAMVS SRRTRPSPPG PRRPTGRPCC SAAPRRPQAT
IXI_235      TSPASIRPPA GPSSR----- ----RPSPPG PRRPTGRPCC SAAPRRPQAT
IXI_236      TSPASIRPPA GPSSRPAMVS SR--RPSPPP PRRPPGRPCC SAAPPRPQAT
IXI_237      TSPASLRPPA GPSSRPAMVS SRR-RPSPPG PRRPT----C SAAPRRPQAT

             GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRKS MRAACSRSAG
             GGWKTCSGTC TTSTSTRHRG RSGW------ ----RASRKS MRAACSRSAG
             GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRKS MRAACSR--G
             GGYKTCSGTC TTSTSTRHRG RSGYSARTTT AACLRASRKS MRAACSR--G

             SRPNRFAPTL MSSCITSTTG PPAWAGDRSH E
             SRPNRFAPTL MSSCITSTTG PPAWAGDRSH E
             SRPPRFAPPL MSSCITSTTG PPPPAGDRSH E
             SRPNRFAPTL MSSCLTSTTG PPAYAGDRSH E
```

- **Clustal** (.aln)

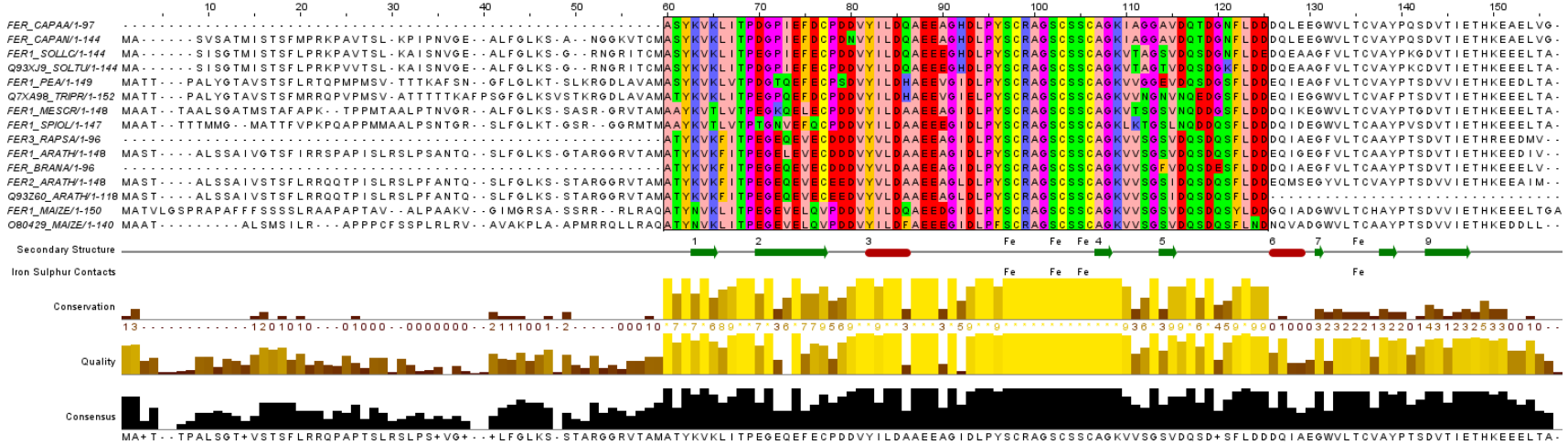
```
IXI_234      TSPASIRPPAGPSSRPAMVSSRRTRPSPPGPRRPTGRPCCSAAPRRPQATGGWKTCSGTC
IXI_235      TSPASIRPPAGPSSR---------RPSPPGPRRPTGRPCCSAAPRRPQATGGWKTCSGTC
IXI_236      TSPASIRPPAGPSSRPAMVSSR--RPSPPPPRRPPGRPCCSAAPPRPQATGGWKTCSGTC
IXI_237      TSPASLRPPAGPSSRPAMVSSRR-RPSPPGPRR----PTCSAAPRRPQATGGYKTCSGTC
             *****:*********         ***** ***    * ***** *******:*******

IXI_234      TTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSRSAGSRPNRFAPTLMSSCITSTTG
IXI_235      TTSTSTRHRGRSGWRA----------SRKSMRAACSRSAGSRPNRFAPTLMSSCITSTTG
IXI_236      TTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSR--GSRPPRFAPPLMSSCITSTTG
IXI_237      TTSTSTRHRGRSGYSARTTTAACLRASRKSMRAACSR--GSRPNRFAPTLMSSCLTSTTG
             *************: *          ***********  **** **** *****:*****

IXI_234      PPAWAGDRSHE
IXI_235      PPAWAGDRSHE
IXI_236      PPPPAGDRSHE
IXI_237      PPAYAGDRSHE
             **  *******
```

# Multiple alignment editors

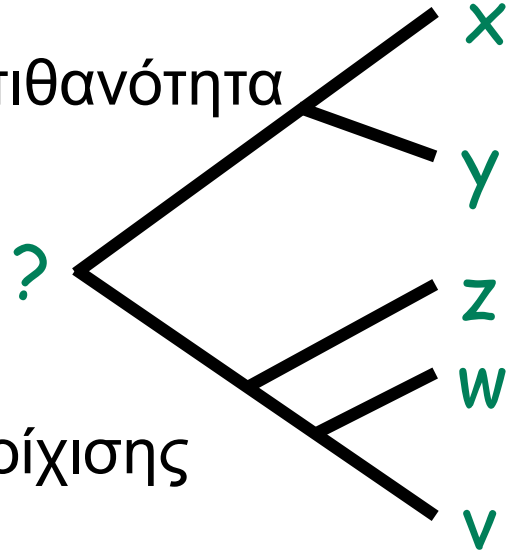
- Έλεγχος και πιθανή τροποποίηση πολλαπλών στοιχίσεων
  - Jalview http://www.jalview.org/
  - SeaView http://pbil.univ-lyon1.fr/software/seaview.html




- Δημιουργία sequence logos
  - WebLogo http://weblogo.berkeley.edu/

# Βαθμολόγηση Πολλαπλής Στοίχισης 1/6

- Ορισμένες θέσεις είναι καλύτερα συντηρημένες από άλλες.
- Οι ακολουθίες δεν είναι ανεξάρτητες αλλά σχετίζονται με κάποιο φυλογενετικό δέντρο.
  - Τα κατάλοιπα μιας στήλης έχουν προέλθει από ένα "αρχέγονο" κατάλοιπο.
  - Η στοίχιση πρέπει να μεγιστοποιεί την πιθανότητα εύρεσης κοινού προγόνου.
- Απλοποιήσεις:
  - Αγνοούμε φυλογενετικά δένδρα
  - Στατιστικά ανεξάρτητες οι στήλες της στοίχισης



# Βαθμολόγηση Πολλαπλής Στοίχισης 2/6

στοίχιση m

AC-GCGG-C

m = AC-GC-GAG

GCACC-GAG

- $m_i^j$ = κατάλοιπο στη στήλη i στην ακολουθία j
  - π.χ. $m_4^2$ = G
- $c_{ia}$ = παρατηρούμενος αριθμός καταλοίπων a στη στήλη i
  - π.χ. $c_{1A}$=2, $c_{1C}$=0, $c_{1G}$=1, $c_{1T}$=0, $C_{1-}$=0
- $p_{ia}$ = πιθανότητα του καταλοίπου a στη στήλη i

# Βαθμολόγηση Πολλαπλής Στοίχισης 3/6

- Minimum Entropy
  - $S(m_i) = -\sum_a c_{ia} \log_2 p_{ia}$
    - $S(m_i)$ το score της i στήλης της στοίχισης m
  - μέτρο της μεταβλητότητας που παρατηρείται στη στήλη i
  - καλή στοίχιση
    - ελαχιστοποιεί την συνολική εντροπία $\sum_i S(m_i)$

# Βαθμολόγηση Πολλαπλής Στοίχισης 4/6

- Sum-of-Pairs Score
  - άθροισμα των scores όλων των "επαγόμενων" ανά δύο στοιχίσεων
  - $S(m) = \sum_i S(m_i)$
  - $S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$
    - όπου $s(m_i^k, m_i^l)$ το score της στοίχισης των καταλοίπων στις ακολουθίες k και l, όπως προκύπτει από ένα πίνακα αντικατάστασης

# Βαθμολόγηση Πολλαπλής Στοίχισης 5/6

- Sum-of-Pairs Score

επαγόμενες ανά δύο στοιχίσεις

- **Seq1: ALLE**
- **Seq2: GLLD**
- **Seq3: WLGD**

```
Seq1: ALLE
Seq2: GLLD

Seq1: ALLE
Seq3: WLGD

Seq2: GLLD
Seq3: WLGD
```

- Blosum50
  - s(L-L) = 5
  - s(L-G) = -4
- SP = SP(1) + SP(2) + SP(3) + SP(4)
  - SP(2) = s(L-L) + s(L-L) + s(L-L) = 15
  - SP(3) = s(L-L) + s(L-G) + s(L-G) = -3

# Βαθμολόγηση Πολλαπλής Στοίχισης 6/6

- **Weighted Sum-of-Pairs Score**
  - $S(m_i) = \sum_{k<l} w_{kl}\, s(m_i^k, m_i^l)$

| Ακολουθία | στήλη A | στήλη B | στήλη C |
|---|---|---|---|
| 1 | N | N | N |
| 2 | N | N | N |
| 3 | N | N | N |
| 4 | N | N | C |
| 5 | N | C | C |
| **Sum-of-pairs score:** | **60** | **24** | **6** |

# Βιβλιογραφία


- David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press; 2nd edition (March 12, 2013).

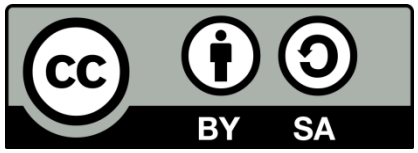
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2nd edition (May 4, 2009).

- 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.