



Βιοπληροφορική

Ενότητα 11:

Πολλαπλή Στοίχιση Ακολουθιών, 1ΔΩ

Τμήμα: **Βιοτεχνολογίας**

Όνομα καθηγητή: **Τ. Θηραίου**



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο



ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

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  ---ADKPKRPLSAYMLWLN SARES IKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSSE
trybr  KKDSNAPKRAMTSMFFSSDFRS----KHSDLS-IVEMSKAAGA AWKELGP
mouse  -----KPKRFRSAYNIYVSESFQ----EAKDDS-AQGK LKLVNEAWKNLSP
          ***. ::: . . . . : . . . * . *: *

chite  AATAKQNYIRALQ EYERNGG-
wheat  ANKLGGEYNKAI AAYNKGESA
trybr  AEKDKERYKREM-----
mouse  AKDDRIRYDNEMKSWEEQMAE
          * : .* . :

```

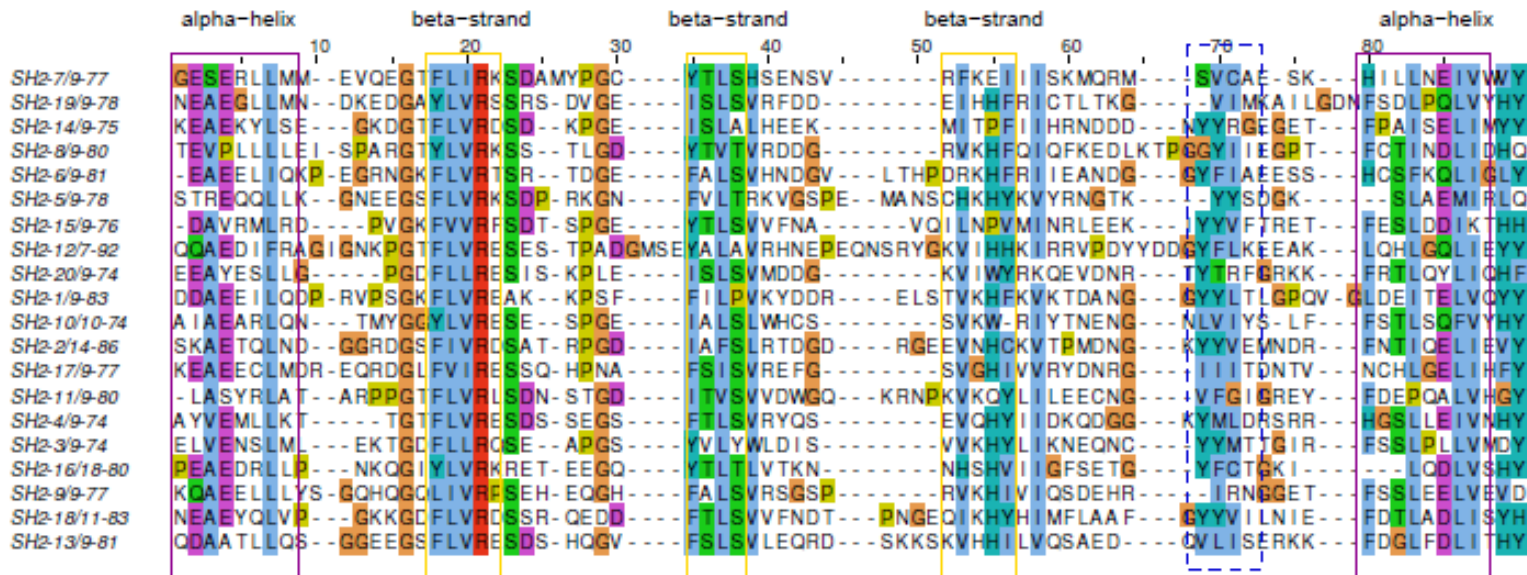


Πολλαπλή Στοίχιση 2/3

- 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 Στοιχείση 1/3

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

```

PRVA_MOUSE -----SMTDLLN---AMDIKKA
PRVA_HUMAN -----SMTDLLN---AMDIKKA
PRVA_GERSP -----SMTDLLS---AMDIKKA
PRVA_MOUSE -----SMTDVLN---AMDIKKA
PRVA_RAT -----SMTDLLS---AMDIKKA
PRVA_RABIT -----AMTELLN---AMDIKKA
TPCC_MOUSE MDDIYKAAVBQLTEBQKNEFKAAFDIFVLGANDGCISTKELGKVMRMLGQNPTPEELQNY
: : * . : * : : :

PRVA_MOUSE VGFASADS--FDHKKFFQMVG----IKKKSADDVKKVPHILDKDKSGFIEBDELGFI
PRVA_HUMAN VGFASATDS--FDHKKFFQMVG----IKKKSADDVKKVPHILDKDKSGFIEBDELGFI
PRVA_GERSP IGAFAAADS--FDHKKFFQMVG----IKKKTPODVKKVPHILDKDKSGFIEBDELGFI
PRVA_MOUSE IGAFAAADS--FDHKKFFQMVG----IKKKNPDEVKKVPHILDKDKSGFIEBDELGSI
PRVA_RAT IGAFAAADS--FDHKKFFQMVG----IKKKSADDVKKVPHILDKDKSGFIEBDELGSI
PRVA_RABIT IGAFAAAS--FDHKKFFQMVG----IKKKSADDVKKVPHILDKDKSGFIEBDELGFI
TPCC_MOUSE IDEVDDEDGSGTVDFFDFLVMIVRCMKDDSKGKSBEELSDFRMFDKNADGYIDLDLAKM
: . . * . : : * : : * * . : : : : * : : : : * : : :

PRVA_MOUSE LKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDNPFSTLVANS-
PRVA_HUMAN LKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDNPFSTLVANS-
PRVA_GERSP LKGFSSDARDLSAKETKTLMAAGDKDGDGKIGVNEPFSTLVANS-
PRVA_MOUSE LKGFSSDARDLSAKETKTLMAAGDKDGDGKIGVNEPFSTLVANS-
PRVA_RAT LKGFSSDARDLSAKETKTLMAAGDKDGDGKIGVNEPFSTLVANS-
PRVA_RABIT LKGFSPDARDLSVKTETKTLMAAGDKDGDGKIGADNPFSTLVANS-
TPCC_MOUSE LQ---ATGHTITEDDIEELMKDGDKNNDGRIDYDFELFELKGV
* : . . . . . : : * : : * : : * : : * : : * : :
  
```

Diagram illustrating the relationship between sequences:

```

TPCC_MOUSE ---|
                |
                |--- PRVA_GERSP --- PRVA_MOUSE
                |--- PRVA_RAT --- PRVA_HUMAN
                |--- PRVA_MACFU
                |--- PRVA_RABIT
  
```

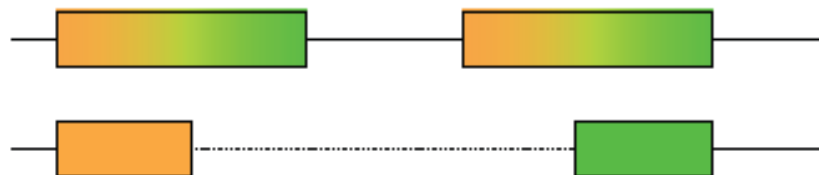
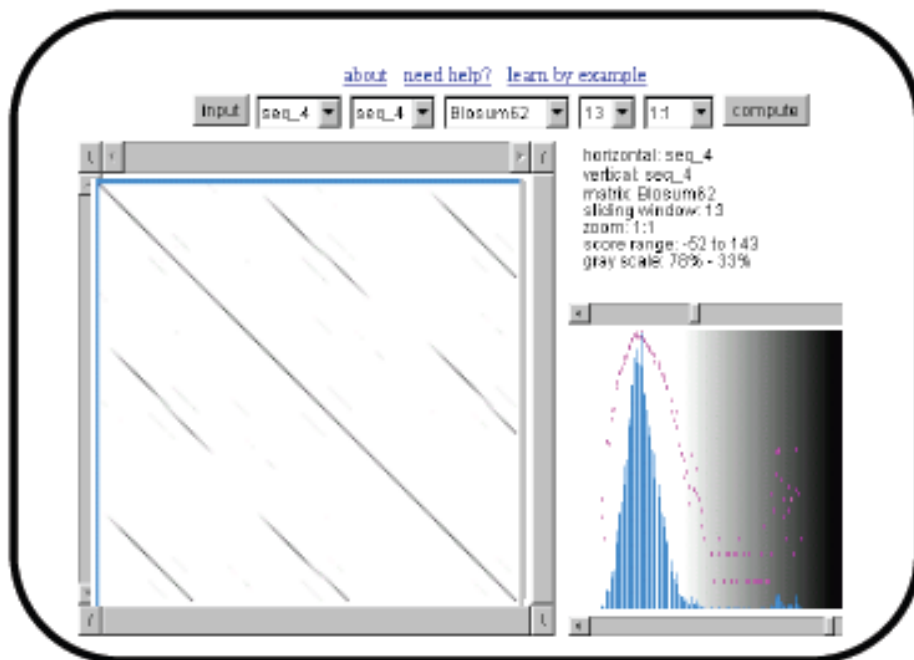
D1



Επιλογή Ακολουθιών προς Στοιχίση 3/3

- ακολουθίες με διαφορετικό αριθμό επαναλήψεων

- αναγνώριση των επαναλήψεων (π.χ. dotplot)
- διαχωρισμός της στοιχισής τους





Μορφοποίηση Πολλαπλής Στοιχείσης 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
TSPASIRPPAGPSSRPAMVSSRRTRPSPPGRRRPTGRPCCSAAPRRPQATGGWKTCSGTC
TTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACRSRSAGSRPNRFAPITLMSSCITSTTG
PPAWAGDRSHE
>IXI_235
TSPASIRPPAGPSSR-----RSPPGRRRPTGRPCCSAAPRRPQATGGWKTCSGTC
TTTSTSTRHRGRSGW-----RASRKSMRAACRSRSAGSRPNRFAPITLMSSCITSTTG
PPAWAGDRSHE
>IXI_236
TSPASIRPPAGPSSRPAMVSSR--RSPPPRRRPPGRPCCSAAPRRPQATGGWKTCSGTC
TTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSR--GSRPPRFAPITLMSSCITSTTG
PPPPAGDRSHE
>IXI_237
TSPASLRPPAGPSSRPAMVSSRR-RSPPGRRRPT----CSAAPRRPQATGGYKTCSGTC
TTTSTSTRHRGRSGYSARTTTAACLRASRKSMRAACSR--GSRPNRFAPITLMSSCLTSTTG
PPAYAGDRSHE
```



Μορφοποίηση Πολλαπλής Στοιχείσης 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--RSPPPP  PRRPPGRPCC  SAAPRRPQAT
IXI_237      TSPASLRPPA  GPSSRPAMVS  SRR-RPSPPG  PRRPT----C  SAAPRRPQAT

GGWKTCSGTC  TTSTSTRHRG  RSGWSARTTT  AACLRASRKS  MRAACRSRAG
GGWKTCSGTC  TTSTSTRHRG  RSGW-----  ----RASRKS  MRAACRSRAG
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--RSPPPPPRRPPGRPCCSAAPRRPQATGGWKTCSGTC
IXI_237      TSPASLRPPAGPSSRPAMVSSRR-RPSPPGPRR----PTCSAAPRRPQATGGYKTCSGTC
*****:*****          ***** **      * ***** *****:*****

IXI_234      TTSTSTRHRGRSGWSARTTTAACLRASRKS MRAACRSRAGSRPNRFAPTLMSSCITSTTG
IXI_235      TTSTSTRHRGRSGWRA-----SRKSMRAACRSRAGSRPNRFAPTLMSSCITSTTG
IXI_236      TTSTSTRHRGRSGWSARTTTAACLRASRKS MRAACSR--GSRPPRFAPPLMSSCITSTTG
IXI_237      TTSTSTRHRGRSGYSARTTTAACLRASRKS MRAACSR--GSRPNRFAPTLMSSCLTSTTG
*****: *          ***** ***** *****:*****

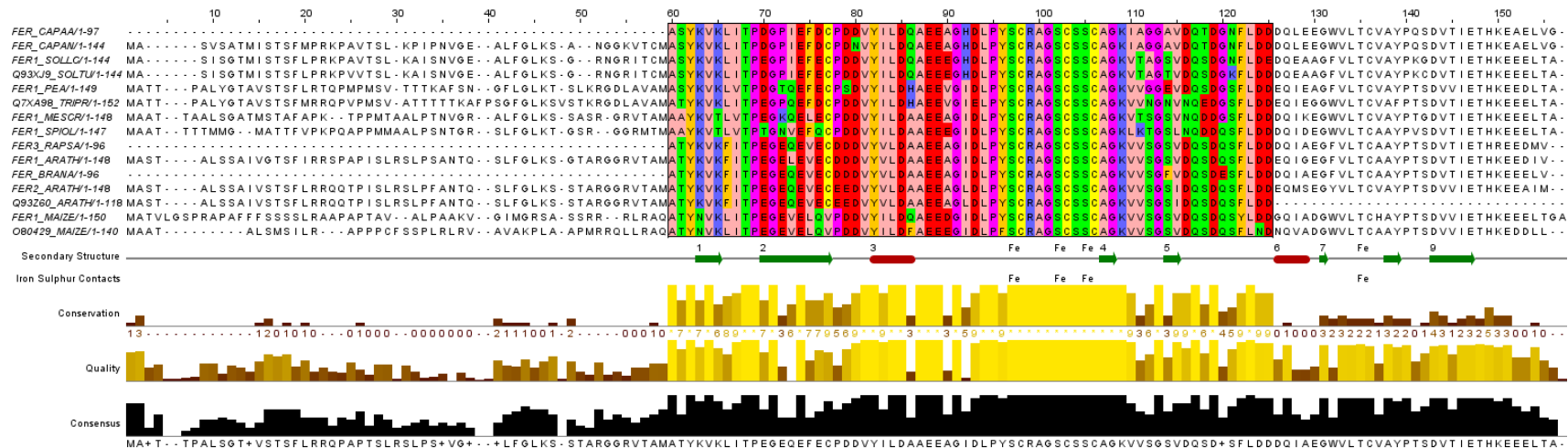
IXI_234      PPAWAGDRSHE
IXI_235      PPAWAGDRSHE
IXI_236      PPPPAGDRSHE
IXI_237      PPAYAGDRSHE
** *****

```



Multiple alignment editors

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



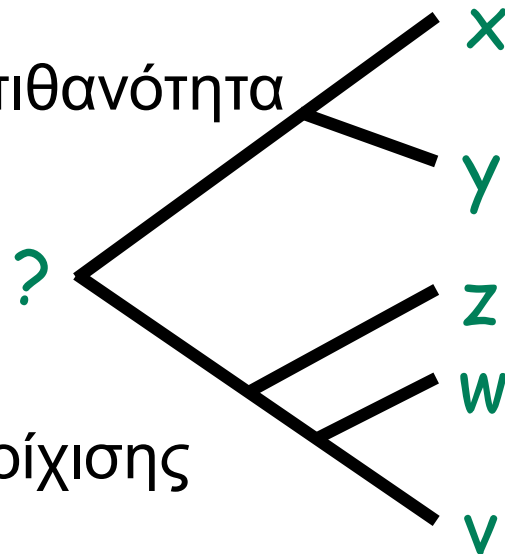
- Δημιουργία sequence logos

- WebLogo <http://weblogo.berkeley.edu/>



Βαθμολόγηση Πολλαπλής Στοιίχισης 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 στοιίχιση πρέπει να μεγιστοποιεί την πιθανότητα εύρεσης κοινού προγόνου.
- **Απλοποιήσεις:**
 - Αγνοούμε φυλογενετικά δένδρα
 - Στατιστικά ανεξάρτητες οι στήλες της στοιίχισης





Βαθμολόγηση Πολλαπλής Στοίχισης 2/6

στοίχιση m

m = AC-GCGG-C
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 Στοιχείσης 3/6

● Minimum Entropy

– $S(m_i) = -\sum_a c_{ia} \log_2 p_{ia}$

- $S(m_i)$ το score της i στήλης της στοιχείσης m
- μέτρο της **μεταβλητότητας** που παρατηρείται στη στήλη i
- **καλή στοιχείση**
 - **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 στοίχισεων
- $S(m) = \sum_i S(m_i)$
- $S(m_i) = \sum_{k < l} s(m_i^k, m_i^l)$
 - 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 Στοιχίσης 5/6

Seq1: ALLE

Seq2: GLLD

- Sum-of-Pairs Score

επαγόμενες ανά δύο στοιχίσεις

- Seq1: ALLE

Seq1: ALLE

Seq3: WLGD

- Seq2: GLLD

Seq2: GLLD

- Seq3: WLGD

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 Στοιχείσης 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
1N.....N.....N.....
2N.....N.....N.....
3N.....N.....N.....
4N.....N.....C.....
5N.....C.....C.....
Sum-of-pairs score:	60	24	6



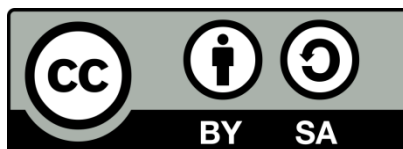
Βιβλιογραφία

- 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

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ



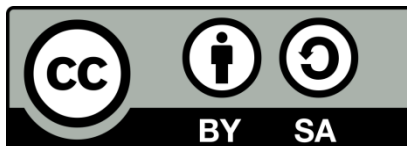
Σημείωμα Αναφοράς

Copyright Γεωπονικό Πανεπιστήμιο Αθηνών 2015. Τμήμα Βιοτεχνολογίας, Θηραίου Τριάς. «Βιοπληροφορική». Έκδοση: 1.0. Αθήνα 2015. Διαθέσιμο από τη δικτυακή διεύθυνση:
<https://mediasrv.aua.gr/eclass/courses/OCDB100/>



Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά, Παρόμοια Διανομή 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων, π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».



Η άδεια αυτή ανήκει στις άδειες που ακολουθούν τις προδιαγραφές του Ορισμού Ανοικτής Γνώσης [2], είναι ανοικτό πολιτιστικό έργο [3] και για το λόγο αυτό αποτελεί ανοικτό περιεχόμενο [4].

[1] <http://creativecommons.org/licenses/by-sa/4.0/>

[2] <http://opendefinition.org/okd/ellinika/>

[3] <http://freedomdefined.org/Definition/EI>

[4] <http://opendefinition.org/buttons/>



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.