



Βιοπληροφορική

Ενότητα 9:

Αναζήτηση Ομοιοτήτων σε
ΒΔ Ακολουθιών - Στατιστική
Σημαντικότητα, 1 ΔΩ

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

- Παρουσίαση των εφαρμογών της αναζήτησης ομοιότητας με αμινοξικές και νουκλεοτιδικές ακολουθίες.
- Επεξήγηση των στατιστικών μεθόδων για την αξιολόγηση της σημαντικότητας μιας στοίχισης.



Λέξεις Κλειδιά

- Λέξεις κλειδιά: Στατιστική σημαντικότητα.
- Key words: Statistical significance, Z-score, E-value.



Αναζήτηση Ομοιοτήτων 1/2

● Αναζήτηση πρωτεϊνών:

– Περισσότερο ευαίσθητη για την εύρεση απομακρυσμένων ομόλογων ακολουθιών.

- Εκφυλισμός γενετικού κώδικα.
- Συντηρητικές αντικαταστάσεις.
- Αλφάβητο 20 γραμμάτων έναντι 4 για το DNA (στατιστική σημαντικότητα στοιχίσεων)

– Πιο γρήγορη.

		Second Letter					
		T	C	A	G		
First Letter	T	TTT } Phe TTC } TTA } Leu TTG }	TCT } Ser TCC } TCA } TCG }	TAT } Tyr TAC } TAA Stop TAG Stop	TGT } Cys TGC } TGA Stop TGG Trp	T C A G	
	C	CTT } Leu CTC } CTA } CTG }	CCT } Pro CCC } CCA } CCG }	CAT } His CAC } CAA } Gln CAG }	CGT } Arg CGC } CGA } CGG }	T C A G	
	A	ATT } Ile ATC } ATA } Met ATG }	ACT } Thr ACC } ACA } ACG }	AAT } Asn AAC } AAA } Lys AAG }	AGT } Ser AGC } AGA } Arg AGG }	T C A G	
	G	GTT } Val GTC } GTA } GTG }	GCT } Ala GCC } GCA } GCG }	GAT } Asp GAC } GAA } Glu GAG }	GGT } Gly GGC } GGA } GGG }	T C A G	



Αναζήτηση ομοιοτήτων 2/2

● Αναζήτηση DNA:

- Περισσότερο ευαίσθητη όταν συγκρίνονται κοντινές εξελικτικά ακολουθίες.
- Για μη κωδικοποιούσες ακολουθίες.
- Εύρεση μεταλλάξεων πλαισίου ανάγνωσης (frameshift mutations).



Στατιστική Σημαντικότητα 1/7

- Έστω η στοίχιση δύο ακολουθιών με score s .
- Η στοίχιση είναι **τυχαία** ή έχει **βιολογικό νόημα**;



Στατιστική Σημαντικότητα 2/7

● Ολική Στοίχιση:

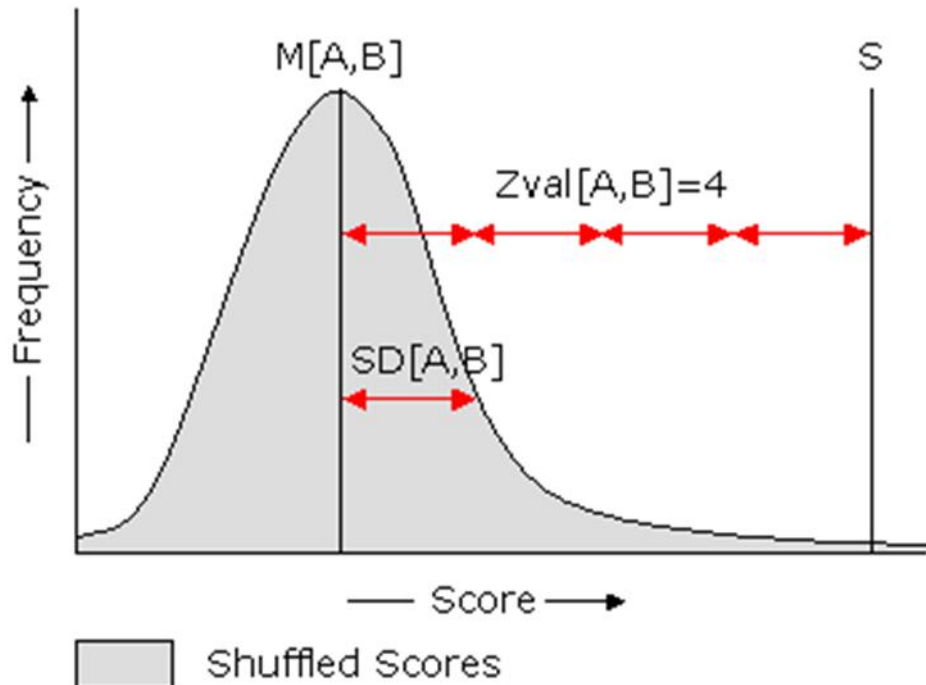
- Δεν είναι γνωστή η κατανομή των scores της στοίχισης τυχαία επιλεγμένων αλληλουχιών.
- Δημιουργία πολλών **τυχαίων ακολουθιών ίδιου μήκους και αμινοξικής σύστασης**.
- Υπολογισμός των scores s' της στοίχισής τους.
- Υπολογισμός του **Z-score**.
 - $Z\text{-score} = (s - \bar{s})/sd$.
 - \bar{s} = μέση τιμή s' .
 - sd = τυπική απόκλιση.



Στατιστική Σημαντικότητα 3/7

● Ολική Στοίχιση:

- Αν το Z-score είναι μικρό, η στοίχιση δεν είναι στατιστικώς σημαντική.

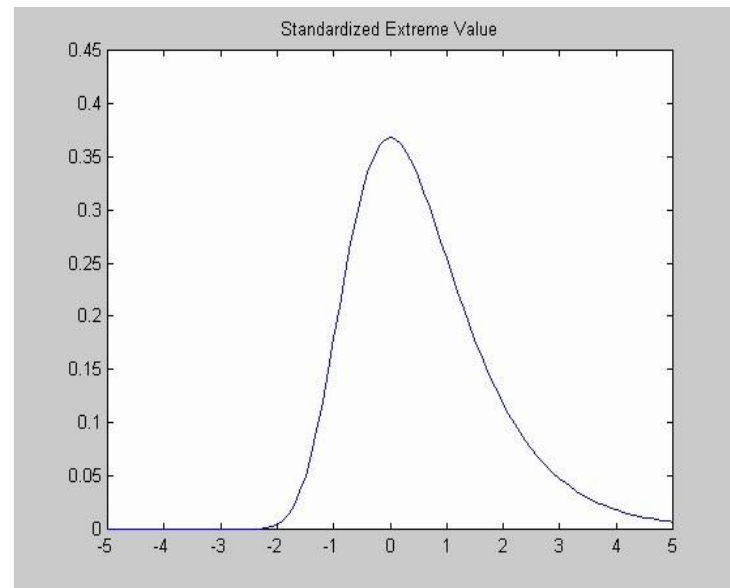




Στατιστική Σημαντικότητα 4/7

● Τοπική στοίχιση χωρίς κενά:

- Η βαθμολογία S των τυχαίων στοιχίσεων ακολουθεί την **κατανομή ακραίων τιμών** (Extreme value distribution ή Gumbel).





Στατιστική Σημαντικότητα 5/7

● Τοπική στοίχιση χωρίς κενά:

- P-value = $P(S \geq s) = 1 - e^{-Kmne^{\lambda s}}$.
- **P-value** πιθανότητα να προκύψει τυχαία στοίχιση με βαθμολογία μεγαλύτερη ή ίση του s .
- K, λ παράμετροι κατανομής.
 - εξαρτώνται από:
 - σύστημα βαθμολόγησης.
 - συχνότητες υποβάθρου.
- m, n μήκη αλληλουχιών.



Στατιστική Σημαντικότητα 6/7

● Τοπική στοίχιση χωρίς κενά:

– $E(S \geq s) = Kmne^{-\lambda s}$.

- **$E(S \geq s)$** πλήθος τυχαίων στοιχίσεων με βαθμολογία μεγαλύτερη ή ίση του s .

–
$$S_{bit} = \frac{\lambda s - \ln K}{\ln 2}$$

● **S_{bit} κανονικοποιημένο score:**

- Συγκρίσιμα αποτελέσματα που έχουν προκύψει από διαφορετικά συστήματα βαθμολογίας.



Στατιστική Σημαντικότητα 7/7

● Τοπική στοίχιση χωρίς κενά:

– E-value = $E(S_{\text{bit}} \geq s_{\text{bit}}) = mn2^{-S_{\text{bit}}}$.

- **E-value** πλήθος τυχαίων στοιχίσεων με βαθμολογία μεγαλύτερη ή ίση του s_{bit} .
- "**Πραγματική**" στοίχιση: **E-value** $\rightarrow 0$.
- Για ακολουθίες μήκους > 100 κατάλοιπα.

Τύπος Ακολουθίας	E-value	Ταυτότητα Καταλοίπων
Νουκλεοτιδική	$< 10^{-6}$	$> 70\%$
Αμινοξική	$< 10^{-4}$	$> 25\%$



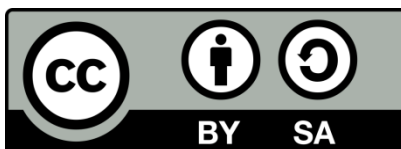
Βιβλιογραφία

- 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



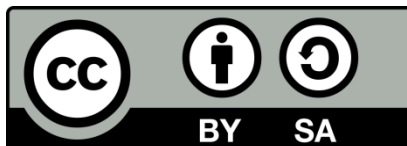
Σημείωμα Αναφοράς

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.