ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

Βιοπληροφορική

Ενότητα 9:

Αναζήτηση Ομοιοτήτων σε ΒΔ Ακολουθιών - Στατιστική Σημαντικότητα, 1 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

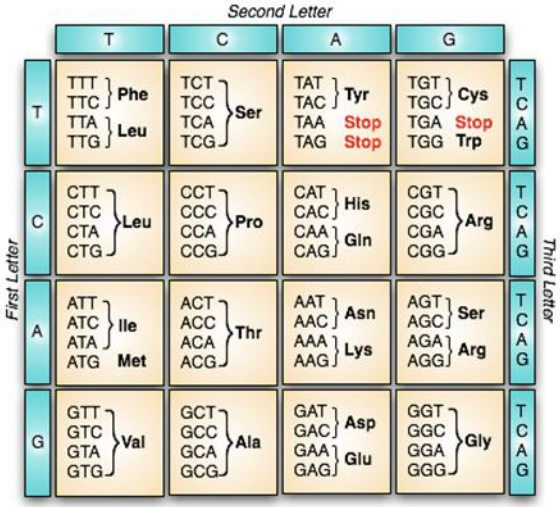
- Παρουσίαση των εφαρμογών της αναζήτησης ομοιότητας με αμινοξικές και νουκλεοτιδικές ακολουθίες.
- Επεξήγηση των στατιστικών μεθόδων για την αξιολόγηση της σημαντικότητας μιας στοίχισης.

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.



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

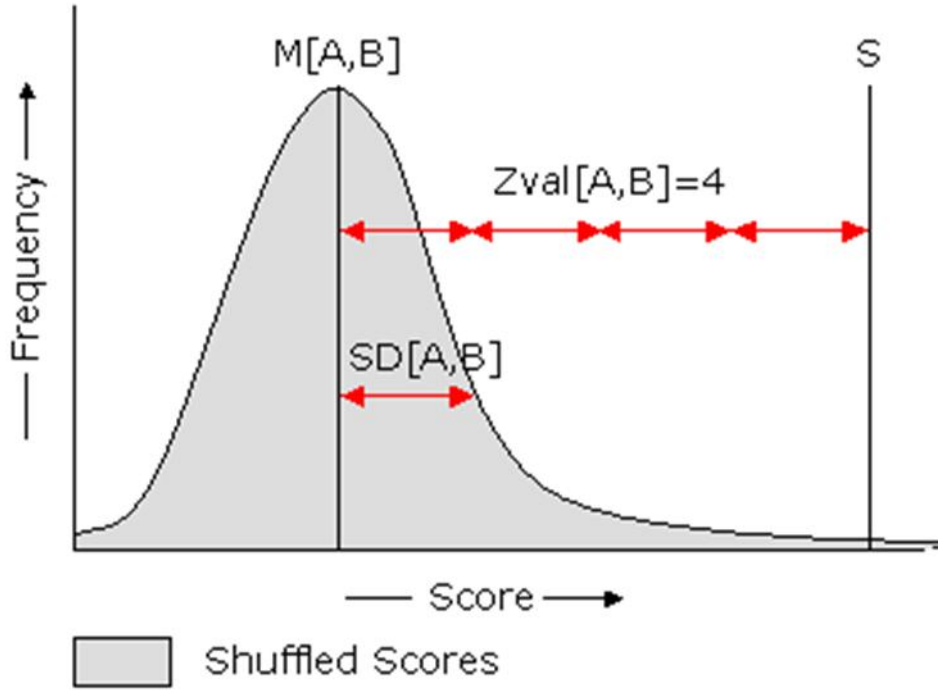
- Έστω η στοίχιση δύο ακολουθιών με score s.
- Η στοίχιση είναι **τυχαία** ή έχει **βιολογικό νόημα**;

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-score = (s- ṡ)/sd.
    - ṡ = μέση τιμή s′.
    - sd = τυπική απόκλιση.

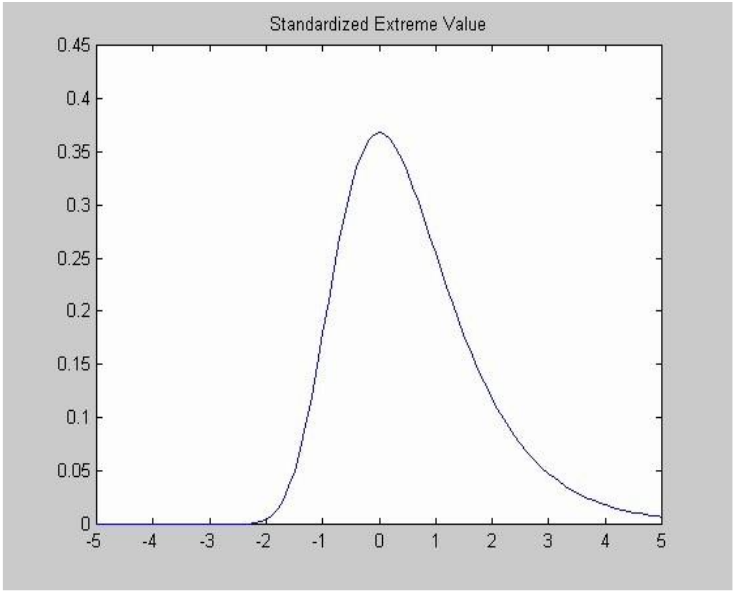
Στατιστική Σημαντικότητα 3/7

- **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≥s)** πλήθος τυχαίων στοιχίσεων με βαθμολογία μεγαλύτερη ή ίση του s.
  - $$S_{bit} = \frac{\lambda s - \ln K}{\ln 2}$$
    - **Sbit κανονικοποιημένο score**:
      - Συγκρίσιμα αποτελέσματα που έχουν προκύψει από διαφορετικά συστήματα βαθμολογίας.

Στατιστική Σημαντικότητα 7/7

- **Τοπική στοίχιση χωρίς κενά**:
  - E-value = $E(S_{bit} \geq s_{bit}) = mn2^{-S_{bit}}$.
    - **E-value** πλήθος τυχαίων στοιχίσεων με βαθμολογία μεγαλύτερη ή ίση του s_{bit}.
    - **"Πραγματική" στοίχιση: E-value → 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).

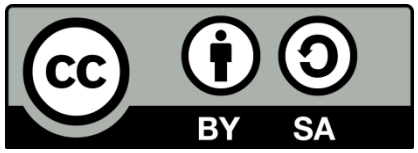
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2nd edition (May 4, 2009).

- 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
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.