Ειδικά Θέματα Εφαρμοσμένης Στατιστικής: Statistical Genetics – Bioinformatics

(Topics in Applied Statistics: Statistical Genetics – Bioinformatics)

ΔΙΔΑΣΚΩΝ: Π.ΠΑΠΑΣΤΑΜΟΥΛΗΣ

Γενικά Στοιχεία Μαθήματος

Κωδικός: 61234
Τύπος: Επιλογής Ομάδας Μαθημάτων 1
Επίπεδο: Μεταπτυχιακό
Έτος σπουδών: Α'
Εξάμηνο σπουδών: 2ο
ECTS: 3
Γλώσσα διδασκαλίας: Αγγλική

Περιεχόμενο Μαθήματος

Η σύγχρονη Βιολογία είναι μία επιστήμη που παράγει μεγάλο όγκο δεδομένων. Το μάθημα θα φέρει σε επαφή τους φοιτητές με βιολογικά δεδομένα από μεθόδους αλληλούχισης «επόμενης γενιάς» (όπως μικροσυστοιχίες γονιδίων, αλληλούχισης RNA) και θα παρουσιάσει τα κύρια εργαλεία συμπερασματολογίας καθώς και τις στατιστικές προκλήσεις που απαντώνται σε αυτά, όπως ο έλεγχος του ρυθμού των ψευδώς θετικών αποτελεσμάτων η μοντελοποίηση βιολογικών δεδομένων απαρίθμησης μέσω γενικευμένων γραμμικών μοντέλων έλεγχοι για τη διαφοροποίηση της γονιδιακής έκφρασης μοντελοποίηση της ετερογένειας μέσω μοντέλων μείξεων κατανομών μοντέλα κλασικής ή Μπεϋζιανής Στατιστικής για την εκτίμηση του βαθμού έκφρασης γονιδίων/ισομορφών.

Προαπαιτούμενα

Το μάθημα απευθύνεται σε κοινό με καλό υπόβαθρο στη Στατιστική. Πιο συγκεκριμένα, απαιτείται:

- Καλή γνώση:
  - Εκτιμητικής και Ελέγχων Υποθέσεων
  - Γενικευμένων Γραμμικών Μοντέλων
- Βασικές γνώσεις:
  - Υπολογιστικής Στατιστικής
  - Μπεϋζιανής Συμπερασματολογίας
  - Προγραμματισμού σε **R**
- Σχετικά προσφερόμενα μαθήματα (δεν είναι προαπαιτούμενα):
  - Μοντέλα Bayes στη Στατιστική
  - Στατιστική Μάθηση
  - Στατιστική για μεγάλο όγκο δεδομένων

Επιδιωκόμενα Μαθησιακά Αποτελέσματα

Μετά την επιτυχή ολοκλήρωση του μαθήματος, οι φοιτητές θα είναι σε θέση να:

- γνωρίζουν τις εφαρμογές της Στατιστικής στη Γενετική και Βιοπληροφορική
- αντιμετωπίζουν με τον σωστό τρόπο προβλήματα ελέγχου πολλαπλών υποθέσεων
- μάθουν νέες Στατιστικές μεθοδολογίες που αναπτύχθηκαν τα τελευταία 20 χρόνια, όπως το πάντρεμα μεθόδων κλασικής και Μπεϋζιανής Στατιστικής
- έρθουν σε επαφή με σύνολα δεδομένων όπως μικροσυστοιχίες, αλληλούχισης RNA, και μονοκυτταρικά δεδομένα
- υλοποιούν κατάλληλους αλγορίθμους στην **R** και **Bioconductor**.

Συνιστώμενη Βιβλιογραφία

- Holmes, Susan and Wolfgang Huber. Modern Statistics for Modern Biology. Cambridge University Press, 2019
- Efron, Bradley. Large scale inference: Empirical Bayes Methods for Estimation, Testing and Prediction. Cambridge University Press, 2010
- Gentleman, Robert, et al., eds. Bioinformatics and computational biology solutions using R and Bioconductor. Springer Science & Business Media, 2006
- McLachlan, Geoffrey and David Peel. Finite Mixture Models. Wiley Series in Probability and Statistics, 2000
- Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B, 1995
- Dudoit, Sandrine and Shaffer, Juliet Popper and Boldrick, Jennifer C. Multiple hypothesis testing in microarray experiments. Statistical Science, 2003
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expres- sion analysis of digital gene expression data. Bioinformatics, 2010
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 2014
- Li, B., Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 2011
- Glaus, P, Honkela, A, Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics, 2012
- Hensman, J, Papastamoulis, P, Glaus, P, Honkela, A, Rattray, M. Fast and accurate approximate inference of transcript expression from RNA-seq data. Bioinformatics, 2015
- Lönnstedt, Ingrid and Speed, Terry. Replicated Microarray data. Statistica sinica, 2002
- Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology, 2004

Διδακτικές και Μαθησιακές Μέθοδοι

Για την υλοποίηση των κατάλληλων αλγορίθμων θα χρησιμοποιηθεί το λογισμικό ανοικτού κώδικα R, το οποίο είναι διαθέσιμο σε όλα τα δημοφιλή λειτουργικά συστήματα (Windows, Mac, Linux). Πιο συγκεκριμένα θα χρησιμοποιηθούν εξειδικευμένες βιβλιοθήκες που είναι διαθέσιμες στο αποθετήριο https://www.bioconductor.org, όπως limma, DeSeq2, edgeR, BitSeq, rsem-EBSeq.

Μέθοδοι Αξιολόγησης και Βαθμολόγησης

Θα πρέπει να παραδοθούν 2 εργασίες οι οποίες αντιστοιχούν στο ≈ 50% του τελικού βαθμού. Το υπόλοιπο ≈ 50% αντιστοιχεί στην τελική εξέταση.