Βιοπληροφορική και προσομοίωση Φυσιολογικών Συστημάτων 2020

Άσκηση 1

(Ατομική)

Στην παρούσα άσκηση καλείστε να δημιουργήστε τη δική σας γονιδιακή υπογραφή η οποία θα εστιάζεται στη κατηγοριοποίηση δειγμάτων Cancer & Normal. Για την υλοποίηση πρέπει να ακολουθήσετε τα παρακάτω βήματα:

- **Data pre-processing (20/100) - προ επεξεργασία των δεδομένων**:
    1. Διαβάστε τα δεδομένα γονιδιακής έκφρασης από το αρχείο http://139.91.190.186/tei/bioinformatics/assignment1.txt
    2. Ταξινομήσετε τα δεδομένα σας με βάση τη κλάση (columns) και να τα φέρετε σε μια μορφή οπου πρώτα θα εμφανίζονται τα Cancer samples και μετα τα Normal samples
    3. δημιουργείστε ένα heatmap για τα 100 πρώτα γονίδια
- **Data analysis (60/100) – Ανάλυση δεδομένων:**
    1. **Gene expression analysis using means (20/100)**:
        i. Για κάθε γονίδιο βρειτε τη μέση τιμή ανα κλάση (Cancer, Normal) και τη διαφορά τους.
        ii. Ταξινομείστε τα γονιδια με βάση τη διαφορά στους μέσους όρους
        iii. δημιουργείστε ένα heatmap για 100 γονίδια όπου τα 50 πρώτα θα είναι τα γονίδια με τη μέγιστη τιμή στη διαφορά και τα αλλα 50 γονίδια με την ελάχιστη τιμή στη διαφορά

    2. **Gene expression analysis using p-value (20/100)**:
        i. Για κάθε γονίδιο βρείτε το p-value (μπορείτε να χρησιμοποιήσετε την συνάρτηση ttest_ind).
        ii. Ταξινομείστε τα γονίδια με βάση τη διαφορά ***στους μέσους όρους***.
        iii. Δημιουργείστε ένα heatmap για 100 γονίδια. Τα 50 πρώτα θα είναι τα γονίδια με τη μέγιστη τιμή στη διαφορά των μέσων όρων και τιμή p-value κάτω από 0.0001. Τα άλλα 50 γονίδια με την ελάχιστη τιμή στη διαφορά των μέσων όρων και τιμή p-value κάτω από 0.0001.
    3. **Gene expression analysis using q-value (20/100)**:
        i. Για κάθε γονίδιο βρείτε το q-value με βάση το p-value. Ο τύπος για το q-value είναι *q-value = p-value * n/(n-k)* οπου *n = number of genes*, *k = rank in gene list* όπως περιγράφεται εδώ www.nonlinear.com/progenesis/qi/v2.4/faq/pq-values.aspx.
        ii. Ταξινομείστε τα γονίδια με βάση τη διαφορά ***στους μέσους όρους***
        iii. Δημιουργείστε ένα heatmap όλα για 100 γονίδια .Τα 50 πρώτα θα είναι τα γονίδια με τη μέγιστη τιμή στη διαφορά των μέσων όρων και τιμή q-value <=0.05.

Τα άλλα 50 γονίδια με την ελάχιστη τιμή στη διαφορά των μέσων όρων και τιμή q-value <=0.05.

- **Results interpretation (20/100):**
    1. Στο αρχείο http://139.91.190.186/tei/bioinformatics/h.all.v7.1.entrez.gmt θα βρείτε τα hallmark genes για τον καρκίνο, δηλαδή γονίδια που είναι γνωστό ότι επηρεάζουν με κάποιο τρόπο τη νόσο. Κάθε γραμμή είναι μια λίστα από hallmark genes και μπορεί ένα γονίδιο να εμφανίζεται σε πάνω από μια λίστες. Διαβάστε το αρχείο και αποθηκεύστε το σε μια δομή.
    2. Βρείτε για κάθε μια από τις παραπάνω μεθοδολογίες
        i. Πόσα από τα γονίδια που επιλέξατε είναι μέσα στα hallmark genes (αφαιρέστε τυχών διπλές εγγραφές του ίδιου γονίδιου).
        ii. Σε ποια hallmark gene list (πρώτη κολώνα στα δεδομένα του h.all.v7.1.entrez.gmt αρχείου) έχετε τα περισσότερα γονίδια.
        iii. Συγκρίνετε τις 3 μεθοδολογίες και περιγράψτε ποια θεωρείτε καλύτερη και για ποιο λόγο.

Όλα τα αρχεία είναι tab delimited files.

Τρόπος παράδοσης: Eclass

Ανεβάστε ενα αρχείο με όνομα τον αριθμό μητρώου που θα περιέχει το ipynb file <αριθμός_ μητρώου>_exercise1.ipynb (αν το σύστημα δεν σας επιτρέπει να ανεβάσετε το αρχείο λόγο κατάληξης, προσθέστε το .txt ή συμπιέστε το και ανεβάστε το).

Deadline: 23/04/2020 23:55