Βιοπληροφορική και προσομοίωση Φυσιολογικών Συστημάτων 2021

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 (10/100) - προ επεξεργασία των δεδομένων**:
    1. Διαβάστε τα δεδομένα γονιδιακής έκφρασης από το αρχείο http://139.91.190.186/tei/bioinformatics/assignment.txt
    2. δημιουργείστε ένα heatmap για τα 40 πρώτα γονίδια
- **Data analysis (60/100) – Ανάλυση δεδομένων:**
    1. **Gene expression analysis using means (10/100)**:
        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 40 γονίδια όπου τα 20 πρώτα θα είναι τα γονίδια με τη μέγιστη τιμή στη διαφορά και τα αλλα 20 γονίδια με την ελάχιστη τιμή στη διαφορά
    2. **Gene expression analysis using p-value (10/100)**:
        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 40 γονίδια. Τα 20 πρώτα θα είναι τα γονίδια με τη μέγιστη τιμή στη διαφορά των μέσων όρων και τιμή p-value κάτω από 0.05. Τα άλλα 20 γονίδια με την ελάχιστη τιμή στη διαφορά των μέσων όρων και τιμή p-value κάτω από 0.05.
    3. **Gene expression analysis using Bonferroni corrected p-value (10/100)**:
        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 όλα τα bonferroni γονίδια (είναι λιγότερα από 40). Τα πρώτα θα είναι τα γονίδια με τη μέγιστη τιμή στη διαφορά των μέσων όρων και τιμή κάτω από το Bonferroni corrected p-value.

    4. **Gene expression analysis using q-value (30/100)**:

i. Υλοποιήστε το q-value χωρις τη χρήση έτοιμης συνάρτησης. Μπορείτε να χρησιμοποιήσετε μονο την συνάρτηση ttest_ind για τον υπολογισμό του p-value. 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 40 γονίδια .Τα 20 πρώτα θα είναι τα γονίδια με τη μέγιστη τιμή στη διαφορά των μέσων όρων και τιμή q-value <=0.05. Τα άλλα 20 γονίδια με την ελάχιστη τιμή στη διαφορά των μέσων όρων και τιμή q-value <=0.05.

- **Data Validation (30/100):**
    1. Βρείτε τα 40 γονίδια που εχουν q-value μικρότερο από 0.05 ταξινομημένα με βάση τη διαφορά στους μέσους όρους (βήμα 4 ανάλυσης) και τυπώστε τα.
    2. Διαβάστε το validation dataset από εδώ http://139.91.190.186/tei/bioinformatics/assignment_validate.txt και δημιουργήστε ενα heatmap από τα δεδομένα του validation dataset μονο με τα γονίδια που είχατε επιλέξει από το προηγούμενο βήμα.
    3. Σχολιάστε το heatmap σας σε σχέση με το heatmap του βήματος 4 της ανάλυσης (Gene expression analysis using q-value).

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: 18/04/2021 23:55