BIO310 Introduction to Bioinformatics
Lab 7 & HW 3
Spring 2019

April 6, 2020

Instructions:

- We expect you to start working on this assignment in the recitation and submit the progress you've made to `Lab7` assignment. Your overall effort will be graded out of 10 and it will eventually contribute to your lab grade. This grade will be assigned as a number from 1-10; no effort being 1 and full effort being 10.
- The complete assignment is to be submitted as the `HW3` assignment by the due date (to be announced). This will be graded out of 100 and will contribute to your homework grade.
- Submit a PDF document for the answers of the write-up questions. The plots should be appropriately labeled, figures should have captions and should be appropriately cited within the main text.
- Upload the code online on SuCourse. The code you submit should be in a format that is ready to run. In submitting the code on SuCourse, compress it as a ZIP file with the name `BIO310-HWX-YourName.zip` where you substitute in your first and last names into the file name in place of 'YourName' and X with the current homework number.
- Please follow the submission instructions, not adhering the submission standards will lead to point deduction.

Generate string compositions and de Bruijn graphs

1. Write a function to generate a k-mer composition of a given string. See Rosalind for more details.
2. Write a function to construct the de Bruijn graph of a collection of k-mers. See Rosalind for more details.
3. Using the above, construct the de Bruijn graphs `DeBruijn2(TAATGCCATGGGATGTT)`, `DeBruijn3(TAATGCCATGGGATGTT)`, and `DeBruijn4(TAATGCCATGGGATGTT)`, What do you notice?
4. How does the graph `DeBruijn3(TAATGCCATGGGATGTT)` compare to `DeBruijn3(TAATGGGATGCCATGTT)`?

This Colab Notebook might be useful.

Study of Yeast Gene Expression

In this homework you will work on analyzing a gene expression dataset. This data is from (**?**), which measured the gene expression of almost all genes in yeast during the metabolic shift from fermentation to respiration. Expression levels were measured at seven time points during the diauxic shift. The original data is preprocessed to remove artifacts and genes with many missing values.
You will use the provided data in the file `yeastExpression.txt`. The first column lists the gene names, the remaining columns contain the log2 transformed gene expression values measured at 7 different time points. The data is tab limited. You may convert this data to any format as you wish, for your subsequent analysis.

1. To understand the data, pick any two genes and plot their expression values as a single graph. Use a line graph to see the trend across the time. The x-axis should be the time points, the y-axis is the log2 gene expression values. Please properly label your graph. Explain with one sentence what happens to the expression over time for these genes.
2. Calculate the Spearman correlation between the expression of the genes you chose. Is there any correlation? If there is is it positive or negative?
3. If a gene's expression is not varying during the measured time points, it would be hard to identify its role in the process. Therefore, we will filter the genes whose expression do not change over time. Calculate the variance of expression for each gene. (In Python `num.var` function). Rank the genes based on their variance in descending order (the highest variable gene should be at the top). Which gene varies the most? Which gene varies the least?
4. In a typical analysis genes with low variance are removed. Here we will filter even more to obtain a manageable list. For this, select the top 800 most varying genes. Report the lowest variance in this 800 gene list.
5. Cluster the genes in the filtered dataset using hierarchical clustering with average linkage function and with a distance function of your choice. Display your clustering results with a heatmap. You may do this by using Python or you may use this tool `http://www2.heatmapper.ca/expression/`. The accepted file formats are specified in the link `http://www2.heatmapper.ca/about/instructions/`: Are the genes clustered well ? Do you observe any interesting result? State with one sentence.