Algorithms in Bioinformatics – “Final assignment”

*Lovisa Franzén, 2019*

**Project idea**

During the course, I was intrigued by the use and application of graphs and networks, and therefore I would like to involve it in my final project somehow. In my daily work, I mainly work with RNA-seq data and I would like to make use of that kind of data for this project as well to make it more applicable and relevant to my work.

A normal workflow in bulk RNA-seq analysis is to first identify differentially expressed genes (DEGs) between two (or more) groups, and thereafter study the most highly and significant DEGs for instance using pathway enrichment methods. Pathways can in turn be analysed and visualised as networks, since many cellular pathways may overlap or belong to the same group of pathways – pathways such as “macrophage polarisation” and “t cell receptor signalling” could for instance belong to the same broader “inflammation” category and could therefore form its own subnetwork.

My idea for a course project would thus be to perform DE and pathway analysis, and thereafter do pathway visualisation as a graph (e.g. using *networkX*), using a public dataset that is of relevance for my PhD project. The project would also involve visualising the intermediate results, and finally, interpreting the generated pathway network.

**Further details**

I have identified an interesting RNA-seq dataset, available at Gene Expression Omnibus, accession number [GSE83888](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83888), that I would like to work with. The data consists of three groups, with three samples in each group, which corresponds to 1) precultured (“PC”) human foetal lungs, 2) human foetal lung explants after four days culture with budesonide (“Bud”) treatment, or 3) without (“Way”) budesonide treatment.

All the programming/data analysis for the project will be done in *python*, as opposed to R which is what I am normally used to program in for these kinds of analyses. I will try to simplify the workflow up until the generation of a network in order for the scope of the project to better suit the given timeframe.

The workflow as I see it will consist of the following steps:

1. Merge data to form one set
   * Remove the “PC” group for simplicity, and focus on the “Bud” and “Way” groups
2. Perform DE analysis between the treated, “Bud”, and untreated, “Way”, groups
3. Run pathway analysis on significant DEGs
   * Identify a reasonable fold-change and q-value cut-off
   * Identify a good way to perform pathway analysis in python with provides data a format that can be applied to subsequent network analysis
4. Explore possibilities to analyse and visualize pathways as a network using NetworkX in python

If time allows, I will also like to prepare the whole workflow as a python pipeline.

The results with be summarised in a Jupyter Notebook.