School of Electronics and Computer Science
University of Southampton

COMP3212(2023/24): Computational Biology Lab One

Issue	29 January 2024
Deadline	09 February 2024

Objective

To learn how to handle biological data

Preliminaries

For exercises in this module, we will use `Python` programming language in a `Jupyter` notebook environment. If you are unfamiliar with this, or need help getting started, please refer to *Lab 0*

Part 1

Here we introduce concentration measurements of mRNA (taken by microarray) and protein (taken by LC-MS) derived from 500 genes. These genes are expressed at different concentrations dependent upon the phases (G1, G2, S and M) within the cell cycle using the dataset provided, alongside some terms that indicate biological process (GOBP), molecular function (GOMF) and where they are located in the cell/component (GOCC). As computational biologists, we are very interested in modelling the concentration of protein in particular, as proteins influence the majority of cell behaviour.
We will attempt to answer the question "Can the concentration of any particular protein be reliably inferred from it's respective mRNA level?"

1. Import the cell cycle dataset excel spreadsheet (using Pandas). You may need to do some tidying of the data such as dropping rows with missing NaN values.

2. Do some exploratory data analysis by:

    - Calculate the variance and mean of the protein and mRNA concentrations
    - Generate a histogram of one of the cell cycle stages of the RNA and protein distribution.
    - State what you notice about these, and how this might affect inferring Protein from mRNA concentrations.
    - Generate a scatterplot of the RNA vs Protein concentrations for each stage of the cell cycle.
    - Fit a linear model (use sklearn or your own direct solve of linear regression)
    - How accurate would predicitons of Protein concentration be using just RNA concentration? Can you quantify this?

Part 2

Exploring more deeply what we can learn from this data.
Tasks

1. Find all genes that contain ‘cell cycle’ in their GOBP term and plot them as a scatterplot (with different colour) overlaid across all genes for each cell cycle phase.
   - Calculate the correlations.
   - Comment on how these compare, link this to your understanding of the Cell Cycle.
2. Find all genes that contain ‘ribosome’ in their GOCC term and plot them as a scatterplot (with different colour) overlaid across all genes for each cell cycle phase.
   - Calculate the correlations.
   - Comment on how these compare, link this to your understanding of the Cell Cycle.
3. Count the number of occurrences of every GOBP term across all genes, what are some of the difficulties that arise when using these terms?
4. Calculate the *change* in mRNA/protein levels across the cell cycle by taking the difference at each stage (G1-S, S-G2, G2-G1), and standardize the differences by mean-centering and variance scaling.
5. What do we notice about changes in the cell cycle? Is there any apparent clustering of GO terms?
6. Continue the exploration using other terms you have met in the lectures, try to find clusters or correlations

Report and marking

You can ask a demonstrator or myself to mark your work in the lab, or you can submit a pdf output from your work.