Tutorial 1

Week 1

Short Videos

Here are three short videos to give you more understanding of biology

- Cell structure
- From DNA to Protein
- Inner life of cell

Exploring Biological Data

This week we are going to reinforce concepts learnt during lectures about the cell structure, and the central dogma of molecular biology regarding the process of information flow from DNA to protein.

Here we introduce concentration measurements of mRNA (taken by microarray) and protein (taken by LC-MS) of 500 genes taken at different timesteps within the cell cycle wikipedia within this dataset Cell-Cycle-Set.xlsx, 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. It is thought that the concentration of any particular protein can be reliably inferred from it's respective mRNA level, we will be exploring that concept in this tutorial.

During these tutorials we highly recommend you use Python as the primary programming language, and all of the examples will use Python; however you may choose to use MATLAB/Java.

Tasks:

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. Perform exploratory analysis of the data, thus:
    1. Generate a histogram of one of the cell cycle stages of the RNA and protein distribution. Do you notice anything interesting with regards to the mean/variance of the distribution?
    2. Look at the pairwise correlations between each of the RNA/protein columns (this can be achieved using the corr() function). Does the change in timestep have much effect on the relationship(s) between RNA and protein?
    3. Generate a scatterplot of the RNA versus. protein for each cell cycle stage. Fit a linear model to the data, can we infer protein concentration from RNA concentration?

Week 2

Short Videos

One of the major revolutions in science in the last 10-20 years has been in sequences. Here are some videos describing the major sequencing technologies.

- Sanger sequencing
- Illumina sequencing
- Nanopore Technology
- The 100,000 Genomes Project
- This is rather longer, but gives a picture of what is happening now in the space of sequencing

Continuing with the cell cycle, let's go into some deeper analysis of the functional role of the proteins and see if there are comparisons between RNA and protein.

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. Is there a stronger/weaker correlation?
2. Repeat task 1 by finding genes that contain 'ribosome' in their GOCC term.

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 level 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. Repeat tasks 1 and 2 by plotting the changes in levels with GOBP/GOCC labelling. What do we notice about changes in the cell cycle? Is there any apparent clustering of GO terms?

We encourage you to use other terms in GOBP, GOMF or GOCC and try to find clustering/correlations by using them.

Here is a jupyter notebook with some solutions jupyter and raw python python

Intranet: Student Resources | Systems & Support