**Project Plan**

Dimitri Wirjowerdojo

**Working Title**

Cluster …

**Background**

The organisation of proteins is usually done in terms of their involvement in a specific function or set of processes. Databases such as KEGG, Reactome, and BRENDA, are available online, though the pathways listed may contain bias towards a specific research question. The greater coverage of the whole -omic approach can be made use to recreate such pathways with less bias or even lead to the discovery of new ones.

**Aim**

Use high-throughput transcriptomics data to derive groups of proteins with a coherent function.

**Objectives**

* Encode the transcriptome of each tissue into a vector.
* Create distance matrix (e.g. Euclidean distance, Manhattan, correlation, etc.) based on the vector created previously. It may be important to set a cut-off threshold.
* Apply clustering (e.g. MCL) to such matrix to find coherently expressed genes.

**Time Plan**

V4-5 : Sort out PDC account, start creating script to convert expression profiles into vector (~20k long for protein-coding genes only).

V5-6 : Make a script for creating distance matrix, try subset the data to try different distance metrics. Test thresholding. If needed, then use PDC resources.

V6-7 : Create script for clustering, and run on subset.

**Repository**

Scripts will be made available at <http://github.com/dmtr13/cluster/>

*Some interesting reads:*

<http://people.revoledu.com/kardi/tutorial/Similarity/MutivariateDistance.html>

<http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/slides/Clustering%20microarray%20data.pdf>

<http://math.usu.edu/jrstevens/stat5570/2.2.Distances.pdf>

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S2-S2>

<http://homes.di.unimi.it/valenti/SlideCorsi/MB0910/HierarchicalClustering.pdf>

<http://www.rna-seqblog.com/assessing-dissimilarity-measures-for-sample-based-hierarchical-clustering-of-rna-sequencing-data-using-plasmode-datasets/>

<https://www.bioconductor.org/help/course-materials/2002/Summer02Course/Distance/distance.pdf>

<http://qpcrupdate.info/haeseleer-bioinf-2005.pdf>

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559>