



Gene set-based clustering of gene expression data

Vlad Popovici

- Gene expression profiling is a way of studying activity of genes of different organisms, organs or cell lines etc and what molecular pathways are active. Usually, it is performed using microarray or RNA sequencing technology.
- Clustering of gene expression profiles is used to determine if there are any similar and diverse groups of samples.
- When clustering expression profiles various similarity measures can be considered, depending on the objective of the analysis. One interesting perspective is brought by the so-called pathway activation scores, or similar parameters, that try to capture the activity of a given pathway or gene set.
- However, when clustering samples (expression profiles) one needs a similarity score between the said samples. Here we propose to use a gene-set measure, e.g. $\text{Similarity}(x1, x2) = \text{abs}(\text{Score}(x1, \text{gene_set}) - \text{Score}(x2, \text{gene_set}))$, to produce groupings of samples relative to a reference gene set.

Aims of the project

- The project will implement (in R) a series of functions (eventually as a package) allowing such clustering with various activation scores (Z-score, Kolmogorov-Smirnoff, GSEA etc) and a number of alternative similarity functions (e.g. absolute difference, log-ratio etc). For testing purposes, MSigDB (<http://www.gsea-msigdb.org/gsea/msigdb/>) will be used for gene set selection and publicly available tissue-specific expression data sets from GTEx (<https://www.gtexportal.org>)
- Main steps:
 - Study a tutorial on data clustering in R
 - Study some examples of R packages
 - Get familiar with data representation (within existing R packages)
 - Study some methods for gene set scoring (R packages – look at **singscore** package!)
 - Implement the new similarity function
 - Compare with correlation-based similarity
 - Identify colon-tissue data from GTEx and test the functions/package on it – use the TPM data (normalized RNA-Seq)