Modular analysis of gene expression data using ISA

Introduction:

The aim of this project was to find clusters of genes whose expression is related for a subset of samples. We also wanted to correlate the expression of these newly created clusters to particular phenotypes.
The data we used for this project came from the CoLaus cohort study which groups people aged between 35 to 75 years old. The data we actually used the most was composed of expression levels for 19'903 protein coding genes for 555 people that we called samples. The genetic material has been preserved with lymphoblastic cell lines.
For this project we used the ISA (Iterative Signature Algorithm) to create the clusters of sample according to gene expression levels. Thank to this method, human bias is mostly avoided and further analysis would help to guess function and correlation of/between unknown genes.
The first step of this analysis was to make the raw data ready to be used by ISA algorithm. They were therefore cleared of the NAs values, log 2 transformed and z-scored gene then sample wise.

Methods:

ISA : Iterative Signature Algorithm. ISA is a biclustering method and is a soft clustering algorithm, which means modules can overlap. The input is our matrix of data (gene expression data of 19'903 genes for 555 samples) and we have a set of biclusters as an output.

Use of `isa` function and its parameters using the "isa2 package" on R :

```
isa(data,     thr.row=seq(1,6,by=0.5),thr.col=seq(1,6,by=0.5),
no.seeds=100, direction=c("updown", "updown"), cor.limit=0.5)
```

- `data` → matrix of data
- `thr.row` and `thr.col` → thresholds of the rows and the columns respectively.
- `no.seeds` → number of random seed vectors to generate.
- `direction` → keep values that are significantly higher than the mean (direction="up"), or the ones that are significantly lower than the mean (direction="down"); or both (direction="updown").
- `cor.limit` → specifies the correlation limit above which two modules are considered to be the same

Enrichment analysis:

As the result of the first ISA run gave way too much modules, a selection for using a cor.limit= 0.5 was used to reduce this amount to 65 modules.

Then an enrichment analysis was performed using the following databases: Disease_Signatures_from_GEO_down_2014, Disease_Signatures_from_GEO_up_2014, GO_Biological_Process_2015, GO_Cellular_Component_2015, GO_Molecular_Function_2015, HMDB_Metabolites, KEGG_2016, OMIM_Disease, WikiPathways_2016. This enrichment analysis was performed using a python package called “enrichr”.We looked at the 65 modules we got individually and in each database.

First we search for the most relevant modules in each database separately. In order to do so, p-values (obtained using a Fisher’s exact test or an hypergeometric test) and combined score (combination of p-value and z-score(“modification to Fisher's exact test in which we compute a z-score for deviation from an expected rank”(package informations))) were considered. The highest the combine score and the smallest the p-value, the better pour our analysis. This choice of the relevant modules was however quite arbitrary as we chose the ones that seems to be the best ones among the whole database without any particular criteria.Then comparisons databases for one module were made and then we search for links among all the databases. Last, the biological pathways, disease or metabolites that came out of this analysis were put in relation with the actual phenotypes of the corresponding people.

Analysis: Biological VS phenotypical

Our phenotype analysis consisted to use a code a the R logiciel to run a t-test for every phenotype of each module with the remaining sample. The code was able to point us the exact module and for which phenotype it differs significantly from the remaining samples. Because we run t-tests for 39 phenotypes times 65 modules we end up with a huge number of statistic tests (more than 2500). As a result we had to deal with a Multi hypothesis testing and we used a correction to lower our threshold. We wanted at first to use the Bonferroni correction which end up giving us no significant results at all. The problem come from the fact that our variables (modules and phenotypes) are not all independent from one another. We arbitrarily decided to use the number of modules for our correction factor which gave us a threshold low enough to select some interesting results and high enough to observe results (p-value = 7.7*10^-4).

We come across several problems in our phenotypic analysis. We got some incorrect significant results due to the low number of samples in some modules. We also got non-significant results that seems interesting from the the enrichment analysis. We understood that the close range of phenotypes won’t let us find more all the possible results from our enrichment analysis. At the end we found some modules who were interesting in terms of significance and size in the module.

Results:

An interesting module we found was the 82nd. It put in relation leukocyte transendothelial migration and cell adhesion molecules with complement activation.

These are all related to the concept of inflammation. Related phenotypes of uricemia and creatinine levels were found. Uricemia was the phenotype linking both the metabolic syndrome aspect and the inflammation aspect together.
We are however not in the capacity to get much more information than this link with the informations at our disposal. Further analysis would be need to have a more precise idea of this connection between the genetic aspect of inflammation and the measurable molecular data.

Challenges and further analysis:

We agreed that for this project our biggest challenge was not to have a specific research question to start from. Starting from a data driven approach didn't give us any target phenotype and sample which result in non-significatif phenotype analysis that could have been interesting for non-measured phenotypes. If we had to go forward with our research we would start back our analysis from a link that seems interesting (for example the module 82). From this link we would then ask a specific research question and try to replicate our significant results with a more targeted and larger cohort. If we were able to replicate our results from those analysis we would then experimentally validate it trough laboratory experimentations including genetic tools. Hopefully from this pattern we would be able to end up with a scientific discovery.

Conclusion:

Although this project did not lead to any significant results yet, if further investigations were made, this could be a powerful tool to helps find new relation between disease and genes, to find molecules implicated in a particular pathway, etc.
However more precise and rigorous methods should be defined, in particular regarding the enrichment analysis part.