Joana Da Silva Alves, Céline Godfroid and Clara Heiman

# Wiki

WGCNA is a weighted gene co-expression network analysis that extracts network and modules from gene expression data (amount of mRNA transcribed). This method allows us to find the strongest connections and to remove the least important connections in a network. It is based on the idea that a network in biology is based on a scale-free topology.

**Cleaning and preprocessing of the data**

The first step we took was to clean the data and preprocess it: remove the outliers and the non-attributed values (NA). On this data we did the Pearson correlation to see how the genes are correlated together. With the values of the Pearson correlation we built two matrices: one weighted and one un-weighted.
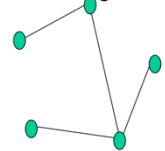
[i]**Un-weighted and weighted matrices**

The difference between these two matrices is that in an un-weighted matrix the genes are either connected or not connected and these connections have equal strengths, whereas in a weighted matrix all genes are connected but these connections have various strengths. WGCNA is based on the weighted matrix.



Weighted Network View
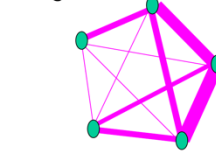
•All genes are connected
•Connection Widths=Connection strenghts

Unweighted View

Some genes are connected
All connections are equal

**Adjacency matrix and edge list**

The weighted matrix can also be called an adjacency matrix. Based on this adjacency matrix, we can extract an edge list, which gives us the simple connections between the genes. From this edge list we can build a network of the gene connections.

**Similarity and dissimilarity matrix**

To build the similarity matrix we took our initial data. This matrix gives us the numerical distance between the genes. We can also construct a dissimilarity matrix, which is built from 1-the similarity matrix.

**Topological Overlap Matrix (TOM)**

This matrix is also a similarity matrix but this one is built on the amount of genes that two genes have in common, therefore the overlap of neighbors between two genes.



*Figure 1: The more neighbors' two genes have in common the redder the result is, on the contrary the least genes two genes is common have the more yellow the results are.*

**Hierarchical clustering**

It is a method to categorize our genes on their dissimilarity, therefore the closest genes, the ones with the lowest dissimilarity, will be put together.



*Figure 2: : We first make one group per gene, in this figure we have 5 genes and therefore 5 groups at the beginning. When the use the information in our dissimilarity matrix and group the genes according to their distances. In our example, we can group gene 1 and 2, and 3 and 5 since the distance between these genes are the lowest. We continue this process to finally have one group with all the genes in it.*

**Defining the modules**

From our hierarchical clustering we can find the various modules that build our network with two different methods. The first one is based on where we cut our dendrogram, if we cut high in our tree we have few modules. On the other hand, if we cut low we have more modules. We

chose the second method, which is defining the minimum number of genes in each module. We chose the random value of 20 so we didn't have too many or too little modules.

## Gene enrichment

### Gene ontology

It is based on three defined biological categories: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). In these categories, there are sub-categories that define specific biological functions such as nucleic acid binding (MF), mitotic division (BP), extracellular space (CC)

The gene enrichment method compares the modules we found with the GO categories. To do this, we used a hypergeometric test which calculated if the p-value of the overlap between our modules and the GO categories is significant. If it is, the modules are enriched. This test verifies that the overlap is not random but truly significant.
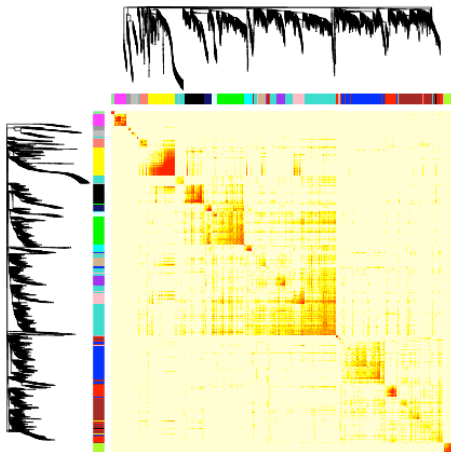


*Figure 3: The band color corresponds to the modules. This band helps to visualize the modules on the hierarchical clustering. The modules can be identify visually.*

## Final results

This is our result for the gene enrichment, we choose the most significant gene enriched in each module. Most "termName" are enriched for BP, a few for CC.

Then we compare if there was some "termName" that appears in two or more modules. In our example it is for extracellular space. Maybe these two modules are involved in the same function, they may be work together.

| Module | termName | Gene ontology | Module | termName | Gene ontology |
|--------|----------|---------------|--------|----------|---------------|
| 1 | Nucleic acid binding | MF | 16 | Growth hormone secretion | BP |
| 2 | Membrane budding | BP | 17 | Integral component of plasma membrane | CC |
| 3 | nucleus | CC | 18 | Cellular response to stress | BP |
| 4 | Blood vessel morphogenesis | BP | 19 | Transferase complex | CC |
| 5 | Immune system process | BP | 20 | Co-factor metabolic process | BP |
| 6 | Immune response | BP | 21 | Brush border | CC |
| 7 | ribosome | CC | 22 | Extracellular space | CC |
| 8 | Calcium ion-dependent exocytosis of neurotransmitter | BP | 23 | Beta-catenin binding | MF |
| 9 | Extracellular matrix | CC | 24 | Epidermis development | BP |
| 10 | Phospholipid translocation | BP | 25 | Extracellular space | CC |
| 11 | myofibril | CC | 26 | Inner ear morphogenesis | BP |
| 12 | Chromosome | CC | 27 | Mitochondrial respiratory chain | CC |
| 13 | Mitotic nuclear division | BP | 28 | Formation of translation preinitiation complex | BP |
| 14 | Catalytic activity | MF | 29 | Sterol biosynthetic process | BP |
| 15 | Olfactive receptor activity | MF | --- | --- | --- |

*Figure 4: Result for the enrichment.*

**Applications**

WGCNA is used to analyze genetic data. It has many applications in neurobiology and other biological research such as DNA methylation.

---

[i] https://labs.genetics.ucla.edu/horvath/GeneralFramework/NetworkConstruction.pdf, 27 mai 2016