Integration of PloGO2 and WGCNA for proteomics

J.Wu and D.Pascovici

October 27, 2020

Abstract

This vignette describes a tailored workflow of using WGCNA for protein network analysis. A publicly available protein ratio dataset was used to demonstrate the major workflow steps and its outputs Wu et al. (2016). It also demonstrates that the WGCNA output can be seamlessly input into PloGO2 and further characterised functionally using PloGO2.

1 WGCNA workflow for proteomics

WGCNA Langfelder and Horvath (2008) is a popular correlation network functional analysis tool which can be used to generate a potentially large number of related data subsets via clustering. The WGCNA was proposed and mostly used for genomic data correlation analysis. We tailored the WGCNA workflow to be more suitable for proteomic data analysis, and include this wrapper in the PloGO2 package for completeness.

Our in-house tailored WGNCA workflow starts from a protein abundance or ratio dataset. It is assumed that the data has been properly normalised. The protein expression will be log-transformed before the analysis starts. Firstly, a soft power was selected for constructing the weighted correlation matrix by using the approximate scale-free network criteria. A scale free network is one where the topology is dominated by a few highly connected nodes, which link the rest of the less connected nodes to the system; it is assumed that most biologically relevant networks should satisfy this property. By raising the correlation to the selected soft power, the correlation network becomes scale-free. The cut-off of the parameter RsquaredCut was set as 0.85. Signed networks, instead of unsigned for gene network analysis, are constructed using the soft power selected. Then TOM (topology overlap metrics) distance was calculated from the network adjacency. Hierarchical clustering was performed based on the TOM distance. A set of clusters were obtained by using the dynamic tree cutting method with the parameter value of minClusterSize as 20. An automatic cluster merging function was invoked to merge the closely correlated clusters from the dynamic tree cutting.

The script for an example WGNCA workflow can be found in the script folder of this package. An example of the input of WGCNA is as follows.

```
> library(PloGO2)
> library(xtable)
> path <- system.file("files", package = "PloGO2")
> print(xtable(read.csv(file.path(path,"rice.csv"))[1:6,c(1:2,5:7)]))
```

	Accession	Uniprot	R1.X127_N.126	R2X127_N.126	R1.X127_C.126
1	350295	Q0DI31	0.96	1.01	0.94
2	537402	Q94JF2	1.14	0.99	1.19
3	5777627	Q7XSU8	1.01	1.02	1.02
4	6002788	Q9LWY6	0.91	0.93	0.98
5	9828445	Q6EUD7	0.95	0.98	0.98
6	11990470	Q9FRU0	0.95	1.01	0.73

The included WGNCA workflow can be executed using the following command

```
> source(file.path(system.file("script", package = "PloGO2"), "WGCNA_proteomics.R"))
```

A number of plots will be produced to visualise the overall weighted correlation network, cluster profiles and eigenprotein boxplots. Some examples are as follows.

Fig 1 shows the cluster dendrogram for the initial and merged clusters.

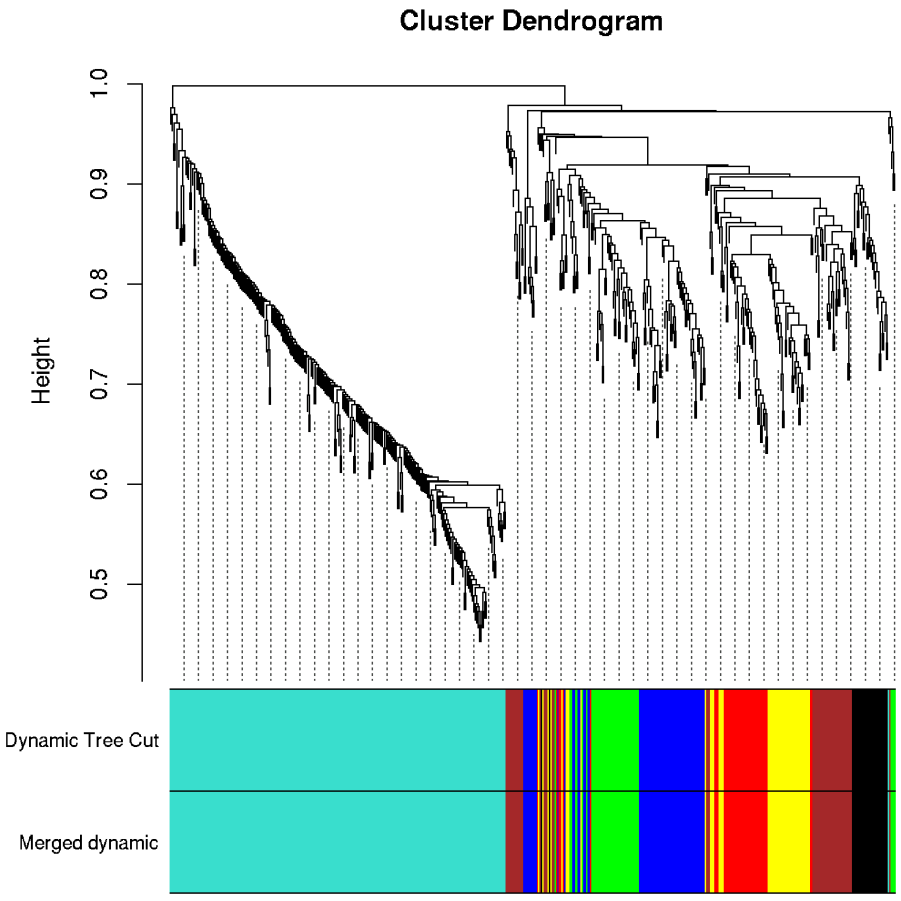

Figure 1: Cluster dendrogram

Fig 2 shows the boxplots for the top 6 hub proteins for the red cluster.

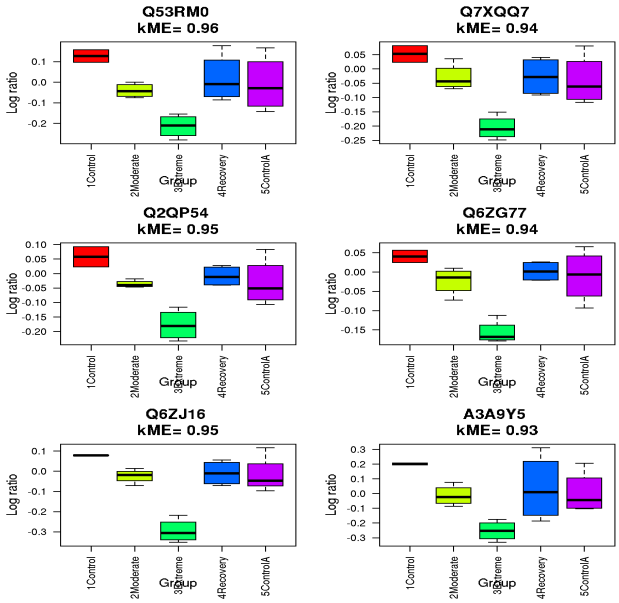

Figure 2: Boxplot for hub proteins - red

Fig 3 shows the cluster profile for all clusters.

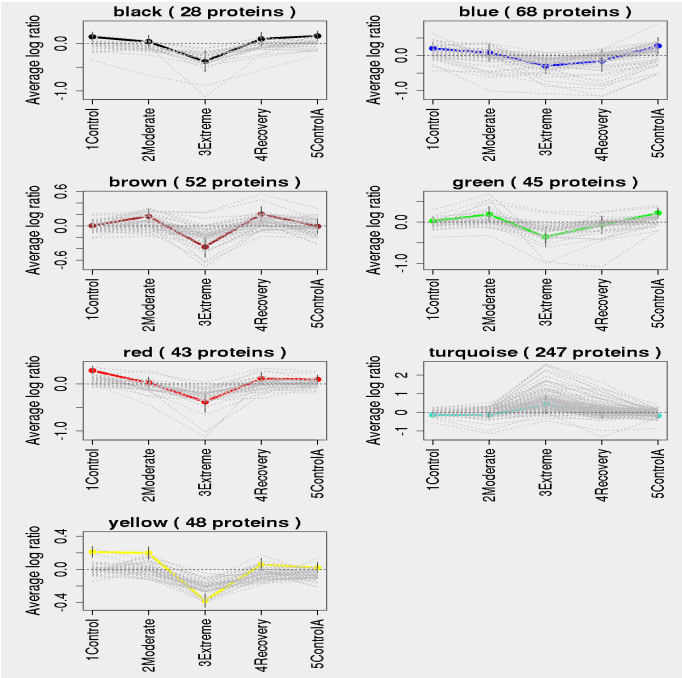

Figure 3: Cluster profile

The output of the WGCNA analysis can be summarised in a multi-tab Excel file which include the overall data and proteins in every cluster ordered by their kME (fuzzy module membership) values. Hub proteins will be easily identified The proteins with the highest kMEs can be viewed as the hub proteins of that cluster. Therefore, the hub proteins are those on the top of the list. An example of the result for one cluster, "red", is as follows.

```
> print(xtable(readWorkbook("ResultsWGCNA.xlsx", "red")[1:6,c(1:2,5:7,23:24)]))
```

	Accession	Uniprot	R1.X127_N.126	R2X127_N.126	R1.X127_C.126	moduleColors	kMEred
1	115453785.00	Q53RM0	1.17	1.10	0.93	red	0.96
2	115488938.00	Q2QP54	1.10	1.02	0.96	red	0.95
3	42407940.00	Q6ZJ16	1.08	1.08	0.93	red	0.95
4	125549171.00	Q7XQQ7	1.02	1.08	0.95	red	0.94
5	115445929.00	Q6ZG77	1.06	1.02	0.99	red	0.94
6	125583193.00	A3A9Y5	1.22	1.23	0.92	red	0.93

2 Using input from WGCNA workflow

The output of the WGCNA can be directly fed into PloGO2 to perform the KEGG pathway analysis, with pre-processed KEGG pathway DB and optional abundance file (refer to the PloGO2 vignette for details).

```
> annot.folder <- genAnnotationFiles("ResultsWGCNA.xlsx", DB.name = file.path(path, "pathwayD
> res <- PloPathway(reference="AllData", filesPath=annot.folder,
+         data.file.name = file.path(path, "Abundance_data.csv"),
+         datafile.ignore.cols = 1)
```

References

Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

Yunqi Wu, Mehdi Mirzaei, Dana Pascovici, Joel M Chick, Brian J Atwell, and Paul A Haynes. Quantitative proteomic analysis of two different rice varieties reveals that drought tolerance is correlated with reduced abundance of photosynthetic machinery and increased abundance of clpd1 protease. *Journal of proteomics*, 143:73–82, 2016.