WGCNA revisited: Module identification

Shaomin Yan, and Guang Wu*
National Engineering Research Center for Non-Food Biorefinery, State Key Laboratory of Non-Food Biomass and Enzyme Technology, Guangxi Biomass Engineering Technology Research Center, Guangxi Key Laboratory of Biorefinery, Guangxi Academy of Sciences, 98 Daling Road, Nanning, 530007, Guangxi, China
*Corresponding author’s e-mail: hongguanglishibahao@gxas.cn

Abstract. WGCNA is a very popular R language software package used in biomedical field. It is mainly designed to deal with gene microarray samples to find clusters (modules), which include highly correlated genes, in biomedical studies. Theoretically, it results in the hierarchical modules, which are different from the results based on the edge density inside module and outside modules. In this study, we address this difference and indicate that caution should be exercised when using these two methods to interpret the implication of studies.

1. Introduction

WGCNA (Weighted Gene Co-expression Network Analysis) is a very popular R language software package used in biomedical field [1-3]. WGCNA is mainly designed to deal with gene microarray samples, especially the transcriptomic data from Gene Expression Omnibus (GEO) [4, 5], in order to build a correlation network [1-3]. Thus, it is a power tool to find clusters (modules), which include highly correlated genes, in biomedical studies. These modules then serve to screen genes to identify potential biomarkers as well as therapeutic targets. In its first release, the WGCNA developers had used it in studies on pairwise relationships among gene transcripts [6-10], carcinoma in brain [11], Saccharomyces cerevisiae cycle [12], mouse genes [13-15], brain tissue in human and primate [16-18], type 2 diabetes [19], chronic fatigue syndrome [20] and phenotypes in plants [21].

Figure 1. Publications from 2007 to 2020 using WGCNA R language software package found in PubMed.gov with additional 342 publications from January 2021 to May 2, 2021.
Since then, its application is increasing significantly with 1765 articles documented in PubMed.gov on May 2, 2021 with keyword “WGCNA” including software developers’ own articles (Figure 1) [22]. Therefore, WGCNA is really very popular in biomedical research community.

Theoretically, WGCNA detects its modules according to a weighed network adjacency, which is derived from correlations between different gene profiles. However, when applying WGCNA to study the distributions of PM$_{2.5}$ [23] and SO$_2$ [24] in China, we found that the modules detected by WGCNA could be geographically irrelevant places although the PM$_{2.5}$ and SO$_2$ profiles are highly-correlated. This suggests that the modules detected by WGCNA are somewhat different from other network analysis [25].

Highly likely, most researchers in biomedical community are unaware of this difference. Hence we attempt to elaborate this difference in this study.

2. Where is the Difference?
In order to find out the difference in network construction and module detection, we need to examine how WGCNA works according to its tutorials [26].

2.1. Data pre-processing
This step is almost the same for all data analysis.

2.2. Choosing the soft-thresholding power
This should be a particular contribution of WGCNA for correlation network, because it uses the following commands to find the weighted network adjacency.

```r
# Choose a set of soft-thresholding powers
powers = c(c(1:10), seq(from = 12, to=20, by=2))
# Call the network topology analysis function
sft = pickSoftThreshold(datExpr, powerVector = powers, verbose = 5)
```

2.3. One-step network construction and module detection
This step automatically works to generate the modules with the following command.

```r
net = blockwiseModules(datExpr, power = 6,
                        TOMType = "unsigned", minModuleSize = 30,
                        reassignThreshold = 0, mergeCutHeight = 0.25,
                        numericLabels = TRUE, pamRespectsDendro = FALSE,
                        saveTOMs = TRUE,
                        saveTOMFileBase = "femaleMouseTOM",
                        verbose = 3)
```

Because this command is a function call, it is not clear how this function does network construction and module detection. So we still cannot find the difference in this step.

2.4. Step-by-step construction of the gene network and identification of modules
This step-by-step procedure provides the opportunity to understand how WGCNA works to construct a gene network and to identify modules, which has the following commands.

For co-expression similarity and adjacency,

```r
softPower = 6;
adjacency = adjacency(datExpr, power = softPower);
```
For Topological Overlap Matrix (TOM)

```r
# Turn adjacency into topological overlap
TOM = TOMsimilarity(adjacency);
dissTOM = 1 - TOM
```

For Clustering using TOM

```r
# Call the hierarchical clustering function
geneTree = hclust(as.dist(dissTOM), method = "average");
# Plot the resulting clustering tree (dendrogram)
sizeGrWindow(12,9)
plot(geneTree, xlab="", sub="", main = "Gene clustering on TOM-based dissimilarity", labels = FALSE, hang = 0.04);
```

Here, we can find that WGCNA uses the standard R function hclust to detect the module, where branches of the hierarchical clustering dendrogram correspond to modules [2]. Thereafter, the following command is used for module identification.

```r
# We like large modules, so we set the minimum module size relatively high:
minModuleSize = 30;
# Module identification using dynamic tree cut:
dynamicMods = cutreeDynamic(dendro = geneTree, distM = dissTOM, deepSplit = 2, pamRespectsDendro = FALSE, minClusterSize = minModuleSize);
table(dynamicMods)
# Convert numeric labels into colors
dynamicColors = labels2colors(dynamicMods)
table(dynamicColors)
# Plot the dendrogram and colors underneath
sizeGrWindow(8,6)
plotDendroAndColors(geneTree, dynamicColors, "Dynamic Tree Cut", dendroLabels = FALSE, hang = 0.03, addGuide = TRUE, guideHang = 0.05, main = "Gene dendrogram and module colors")
```

3. How this difference works

Theoretically, the standard R function hclust, though flashClust was implemented as a fast version of hclust [3], is a hierarchical cluster analysis on a set of dissimilarities using Euclidean distance, which works: "Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance–Williams dissimilarity update formula according to the particular clustering method being used." [27].

On the one hand, the graph theory says: "Many networks consist of modules which are densely connected themselves but sparsely connected to other modules." [28]. On the other hand, it says: "Branches of the dendrogram group together densely interconnected, highly co-expressed genes. Module identification amounts to the identification of individual branches ("cutting the branches off the dendrogram")." [26]. However, difference on concept still exists between hierarchical clusters and graph theory clusters.

Figure 2 shows 19 modules generated by WGCNA. Graphically, Figure 2 demonstrates that a gene (branch) usually does not directly connect with any genes (branches) in other modules, unless through
the gene on the top of dendrogram. The numbers of genes in modules from 0 to 18 are 99, 609, 460, 409, 316, 312, 221, 211, 157, 123, 106, 100, 94, 91, 77, 76, 58, 47, and 34, respectively.

According to the commands in 2.4, we can see that there are two starting points for comparison with the other method.

3.1. First starting point for comparison
The adjacency in command is the first starting point for comparison because adjacency is a matrix generated by soft-thresholding powers.

adjacency = adjacency(datExpr, power = softPower)

Applying adjacency to igraph [28], the left panel in Figure 3 shows 45 modules, of which 31 modules are single gene modules. The numbers of genes in the modules are 2, 16, 26, 35, 73, 91, 97, 99, 122, 363, 448, 592, 748, 857, and 1 for 31 single gene modules.

3.2. Second starting point for comparison
The dissTOM in command is the second starting point for comparison because dissTOM is a matrix that will be used in command hclust.

dissTOM = 1-TOM

For Clustering using TOM

# Call the hierarchical clustering function
geneTree = hclust(as.dist(dissTOM), method = "average");

Applying dissTOM to igraph [28], the right panel in Figure 3 shows 106 modules, of which 102 modules are single gene modules. The numbers of genes in modules are 397, 735, 883, 1483, and 1 for 102 single gene modules. However, the command, minModuleSize = 30, in WGCNA makes it impossible to have single gene modules.

The difference between Figures 2 and 3 is the difference between hierarchical modules and the modules based on the edge density inside module and outside modules. According to graph theory, edge has no length, whereas hierarchical cluster uses Euclidean distance. It is hard to say which module is more suitable for genetic studies unless a very detailed study is conducted to compare genes with clinical traits. Clearly more studies are needed to clarify this difference in future.
4. Conclusion

In this study, we draw attention to the difference between hierarchical modules operated in WGCNA and the modules based on the edge density inside and outside modules operated in network software package. These two methods generate different numbers of modules and different elements in modules, thus caution should be exercised when using these two methods to interpret the implication of studies.

Acknowledgments

The work is partially supported by Scientific Development Fund of Guangxi Academy of Sciences (2021YFJ1203).

References

[1] Zhang, B., Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol., 4: Article 17.
[2] Langfelder, P., Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 29: 559.
[3] Langfelder, P., Horvath, S. (2012) Fast R functions for robust correlations and hierarchical clustering. J. Stat. Softw., 46: 1–17.
[4] Edgar, R., Domrachev, M., Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res., 30: 207-210.
[5] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A. (2013) NCBI GEO: Archive for functional genomics data sets--update. Nucleic Acids Res., 41: D991-D995.
[6] Zhou, X., Kao, M.C., Wong, W. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. Proc. Natl. Acad. Sci. U.S.A., 99: 12783–12788.
[7] Steffen, M., Petti, A., Aach, J., D'haeseleer, P., Church, G. (2002) Automated modelling of signal transduction networks. BMC Bioinformatics., 3: 34.
[8] Stuart, J.M., Segal, E., Koller, D., Kim, S.K. (2003) A Gene-coexpression network for global discovery of conserved genetic modules. Science., 302: 249–255.
[9] Carey, V.J., Gentry, J., Whalen, E., Gentleman, R. (2005) Network structures and algorithms in Bioconductor. Bioinformatics., 21: 135–136.
[10] Schaefer, J., Strimmer, K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics., 21: 754–764.
[11] Chuang, C.L., Jen, C.H., Chen, C.M., Shieh, G.S. (2008) A pattern recognition approach to infer time-lagged genetic interactions. Bioinformatics., 24: 1183–1190.

[12] Cokus, S., Rose, S., Haynor, D., Gronbech-Jensen, N., Pellegrini, M. (2006) Modelling the network of cell cycle transcription factors in the yeast *Saccharomyces cerevisiae*. BMC Bioinformatics., 7: 381.

[13] Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A.C., Liu, L.M., Wu, H., Geschwind, D.H., Febbo, P.G., Kornblum, H.I., Cloughesy, T.F., Nelson, S.F., Mischel, P.S. (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc. Natl. Acad. Sci. U. S. A., 103: 17402-17407.

[14] Ghazalpour, A., Doss, S., Zhang, B., Plaisier, C., Wang, S., Schadt, E., Thomas, A., Drake, T., Lusis, A., Horvath, S. (2006) Integrating genetics and network analysis to characterize genes related to mouse weight. PloS Genetics., 2: e130.

[15] Fuller, T., Ghazalpour, A., Aten, J., Drake, T., Lusis, A., Horvath, S. (2007) Weighted gene co-expression network analysis strategies applied to mouse weight. Mammalian Genome., 6: 463–472.

[16] Oldham, M., Horvath, S., Geschwind, D. (2006) Conservation and evolution of gene co-expression networks in human and chimpanzee brains. Proc. Natl. Acad. Sci. U.S.A., 103: 17973–17978.

[17] Miller, J.A., Oldham, M.C., Geschwind, D.H. (2008) A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. J. Neurosci., 28: 1410–1420.

[18] Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., Geschwind, D.H. (2008) Functional organization of the transcriptome in human brain. Nature Neuroscience., 11: 1271–1282.

[19] Keller, M.P., Choi, Y., Wang, P., Davis, B.D., Rabaglia, M.E., Oler, A.T., Stapleton, D.S., Argmann, C., Schueler, K.L., Edwards, S., Steinberg, H.A., Neto, E.C., Kleinhanz, R., Turner, S., Hellerstein, M.K., Schadt, E.E., Yandell, B.S., Kendzierski, C., Attie, A.D. (2008) A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res., 18: 706–716.

[20] Presson, A., Sobel, E., Papp, J., Suarez, C., Whistler, T., Rajeevan, M., Vernon, S., Horvath, S. (2008) Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. BMC Syst. Biol., 2: 95.

[21] Weston, D., Gunter, L., Rogers, A., Wullschleger, S. (2008) Connecting genes, coexpression modules, and molecular signatures to environmental stress phenotypes in plants. BMC Syst. Biol., 2: 12.

[22] https://pubmed.ncbi.nlm.nih.gov/?term="WGCNA".

[23] Yan, S., Wu, G. (2016) Network analysis of fine particulate matter PM2.5 emissions in China. Sci. Rep., 6: 33227.

[24] Yan, S., Wu, G. (2017) SO2 Emissions in China – their network and hierarchical structures. Sci. Rep., 7: 46216.

[25] http://igraph.org/

[26] Tutorials for the WGCNA package (2021) https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/index.html

[27] The R Core Team Version 3.6.3 (2020-02-29) R: A Language and Environment for Statistical Computing.

[28] Package ‘igraph’ October 2020, igraph.pdf, https://cran.r-project.org/web/packages/igraph/igraph.pdf, cluster_edge_betweenness Description