COST STSM Reference Number: COST-STSM-FP1202-15054

Period: 2013-10-18 to 2013-10-31

COST Action: FP1202

Scientific Report

Unravelling gene regulatory networks in a non-model tree

Applicant: Pedro M. Barros, PhD, GPlantS Lab., ITQB-UNL, Oeiras, Portugal

Host: Dr. Kai Sohn, Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB, Stuttgart, Germany

1. Scope of the STSM

Identifying genes associated with adaptive traits can help to understand how species have adapted to their environment and to predict how they will respond to future climatic changes. However, our knowledge regarding the genetic mechanisms involved in adaptation to biotic and abiotic stimuli is still limited in non-model species, particularly in trees.

The goal of this STSM was to learn new tools to explore the data previously generated by the Portuguese CorOakEST consortium on the sequencing of the cork oak (*Quercus suber*) transcriptome (Pereira-Leal *et al.*, submitted). In this project, RNA collected from different tissues and developmental stages, as well as from plants exposed to abiotic stresses and to biotic interactions was analyzed by 454 pyrosequencing and the data generated was used to establish the CorkOak ESTs Database (www.corkoakdb.org) (Pereira-Leal *et al.*, submitted). In our lab we are currently focused on the identification of regulatory genes putatively involved in the cross-talk between biotic and abiotic stress. For this purpose, data generated from the non-normalized libraries obtained for the different stress conditions tested (root infection by *Phytophthora cinnamomi*, and drought, cold or heat stresses) is being analysed using both bioinformatic and molecular biology tools. During this STSM in Fraunhofer IGB we tested several bioinformatic routines for gene network inference based on RNA-seq data, including statistical methods for construction of co-expression networks and graph-clustering algorithms.

2. Main achievements

The proposed STSM took place in the Fraunhofer Institute for Interfacial Engineering and Biotechnology during 2 weeks, under the supervision of Dr. Kai Sohn (head of the lab) and the bioinformaticians Dr. Phillip Stevens and Dr. Stefan Lorenz. The group of Dr. Kai Sohn has already conducted work in collaboration with our laboratory on the transcriptomic analysis of drought stress response in *Jatropha curcas* (purging nut), a soft-woody oil-seed bearing plant of the Euphorbiaceae family that has emerged as a potential source of biodiesel. This work has generated a set of RNA-seq libraries from root and leaf tissue under well-watered, mild drought and severe drought conditions.

The construction of a gene expression network starts with the calculation of a correlation matrix for a selected group of genes (e.g. genes differentially expressed for a specific condition), which in network inference are called nodes. The correlation matrix is based on the pairwise comparison of the gene expression profiles among a complete set of RNA-seq libraries, which results in a fully connected network among all selected genes. A high value of the similarity measure between two expression profiles usually indicates coexpression at least in some conditions, which could be related to direct or indirect regulation of one gene by the other, or co-regulation of both by a third gene (Veiga *et al.*, 2010). The complete set of gene links is then filtered for their strength using specific threshold values (Veiga *et al.*, 2010). The resulting relevance network may be later characterized by the identification of specific clusters of genes sharing high coexpression levels, and assessing the abundance of gene ontology (GO) terms within those modules.

As a starting point to test and select the best approach to construct gene networks, we decided to first use the data obtained from *J. curcas*, since the data already available for this species (12 RNA-seq libraries obtained for leaf tissues and 10 for root tissue) is more extensive than the one available for cork oak until now (6 RNA-seq libraries for leaf tissue and 7 for root tissue). This allowed us to test different approaches for network construction, for example, checking the effect of using only libraries for one tissue, or using all libraries together.

Most of the analysis was conducted using Cytoscape v3.2.0, a software environment gathering a complete set of tools for gene network prediction. The WGCNA package for R environment (Langfelder & Horvath, 2008) was also tested, but most of the trial analyses were conducted using Cytoscape since it allowed the use of different correlation measures (Pearson's, Spearman's or Kendall's correlation) and the direct visualization of the networks. Several networks have been constructed for specific groups of genes (e.g. differentially expressed genes in roots during mild drought stress), using two different sets of libraries (e.g. only libraries from root tissue or the complete set of libraries). We also tested different correlation coefficients to build the gene correlation matrices, namely Pearson's *r* and Kendall's *tau*, selecting specific correlation thresholds, so that only nodes with a correlation above the threshold would be connected in the network.

During these analyses we have paid special attention to the node degree distribution, which is basically the distribution of correlations between all gene pairs. In biological networks, node degree distribution approximates a power-law distribution (Barabasi & Albert, 1999). However, in some of the tested cases, particularly when a lower number of libraries were used, this was not the case (Figure 1a). When using the whole set of libraries independently from the tissue, the node degree distribution closely fitted a power-law distribution (Figure 1b). Thus, the construction of biological networks with a reduced set of data should be made with special caution and crucial steps such as the thresholding should be fine-tuned. After the construction of the genetic networks, we used a Cytoscape plug-in named ClusterONE (Nepusz & Paccanaro, 2012) to identify clusters of highly correlated genes. To associate each cluster specific function or biological processes, a functional enrichment analysis was performed using Blast2GO (Conesa & Götz, 2008).

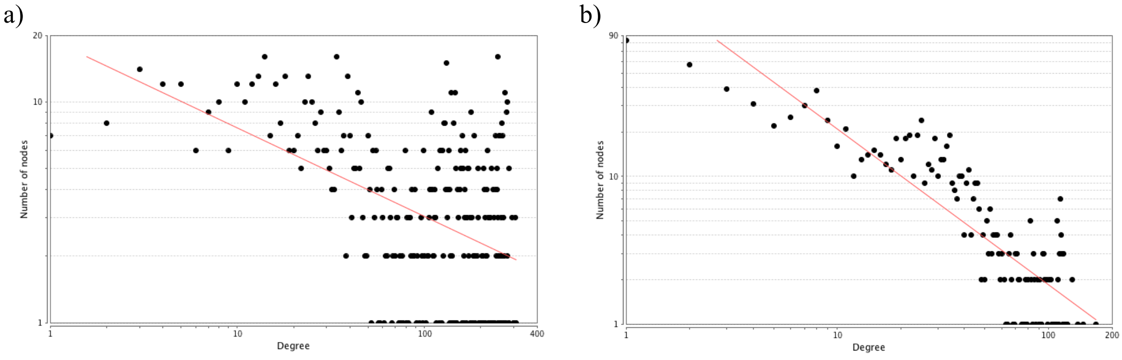

Figure 1: Node degree distribution of two gene networks obtained for the same set of genes, but using different a reduced set (a) or a large set (b) of RNA-seq libraries to calculate pairwise correlations. In b), this distribution closely fits the power-law distribution. The increase in the number of libraries consequently allowed a better estimation of biological variability among different conditions, modifying the shape of the node degree distribution.

2.1 Preliminary gene expression network of stress response in cork oak

Still during the STSM, we build a preliminary network for cork oak, using a set of 1000 genes, which in different stress conditions show differential expression when comparing to the correspondent control libraries (Figure 2). For this, we have used a Pearson correlation coefficient and a soft thresholding option provided by the WGCNA package. All the genes were annotated using Blast2GO for the prediction of protein-coding genes and corresponding GO terms, based in homology searches in plant databases. This preliminary network is divided in several clusters and some of them are enriched with genes associated to particular molecular function or biological processes. Three examples are highlighted in Figure 2. Cluster 1 is composed of 107 genes, being enriched with genes associated with response to external stimulus (GO:0009605) and secondary metabolism (GO:0019748). Cluster 2 (89 genes) shows a significant enrichment in genes related to sequence-specific DNA binding transcription factor activity (GO:0003700 and GO:0003677) and also catalytic activity (GO:0003824). Several genes from cluster 2 show homology to transcription factor families with a predicted role in stress response (e.g. WRKY, Zink finger). A deeper analysis of the genes associated in this cluster may give hints of putative targets of these transcription factors. Finally, Cluster 3 includes 84 correlated genes, among which a large proportion associates to the plasma membrane (GO:0005886) or cell wall (GO:0005618). Some of these genes are also associated with hydrolase or transferase activity (GO:0016787 and GO:0016740, respectively). Nevertheless, we still need to further explore this network and characterize the remaining clusters, particularly the large cluster in the centre, which contain the genes with highest number of connections in the network.

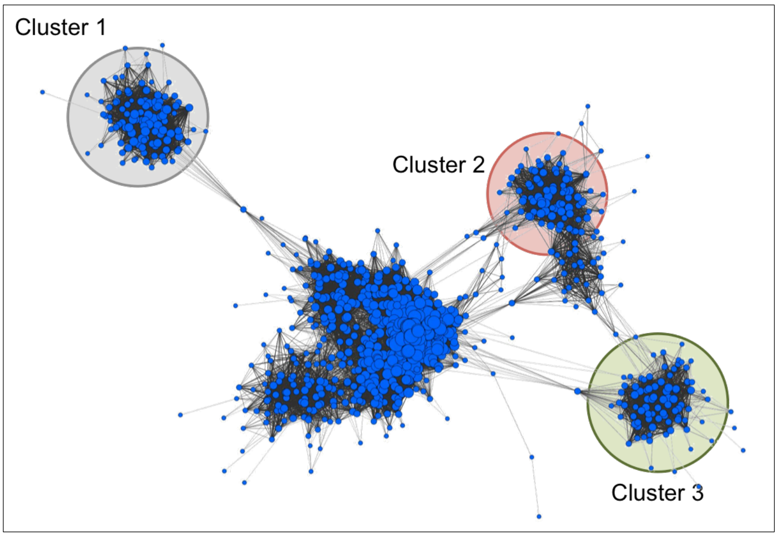

Figure 2: Coexpression network of stress responsive genes showing differential regulation in cork oak tissues under during biotic (infection with *Phytophthora cinnamomi*) and abiotic (drought, cold and/or heat) stress assays. Each blue circle represents a gene (or node). Gene pairs are connected by a gray line (edge) if the level of expression correlation is above the chosen threshold. The number of edges connecting to surrounding nodes defines the size of each node. Three clusters identified by ClusterOne are also highlighted.

3. Future perspectives

The STSM described in the report provided advanced hands-on learning of some of the tools necessary for gene regulatory network inference based on gene expression data. It allowed the identification of critical bottlenecks in this type of analysis, such as the choice of the suitable correlation measure to compare gene expression patterns, the selection of the correlation threshold for network inference and node degree distribution.

Using the experience gathered during this STSM, a gene expression network of stress response in cork oak was built, allowing the identification of specific clusters showing high correlation coefficient. One of these clusters was enriched with genes associated with DNA binding, and further analysis will be conducted to investigate if those genes can be regulators of other genes within the cluster. Still, new networks should be constructed testing the different correlation measures and thresholds. Since this analysis deals with predictions of gene interactions we need to bear in mind that although, a high value of the similarity measure between two expression profiles may indicate coexpression at least in some conditions, it can also appear by chance. Therefore we should explore these data as much as possible and experiment with tweaking the network construction and module detection parameters. Only in this way we may obtain better (biologically more relevant) results of the analysis.

The construction of a preliminary gene network for stress response in cork oak will allow the identification of candidate regulatory genes involved in stress adaptation and survival. These candidate genes will be targeted for functional characterization, by validating gene expression profiles and screening for molecular regulators or targets. In addition these candidate genes could be later used to assess the genetic variability occurring in natural population growing in different areas, allowing the identification of specific signatures of natural selection.

References:

Barabasi AL, & Albert R (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.

Conesa A, & Götz S (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics, 2008.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559.

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, 9(5), 471–472.

Pereira-Leal J.B., *et al.* (submitted) A comprehensive assessment of the transcriptome of the Cork Oak (*Quercus suber*) through EST sequencing.

Veiga, D.F.T., Dutta, B., Balázsi, G. (2010). Network inference and network response identification: moving genome-scale data to the next level of biological discovery. Mol BioSyst 6(3), 469–480.