NES^{2}RA: A TOOL FOR GRAPEVINE TRANSCRIPTOMIC DATA MINING

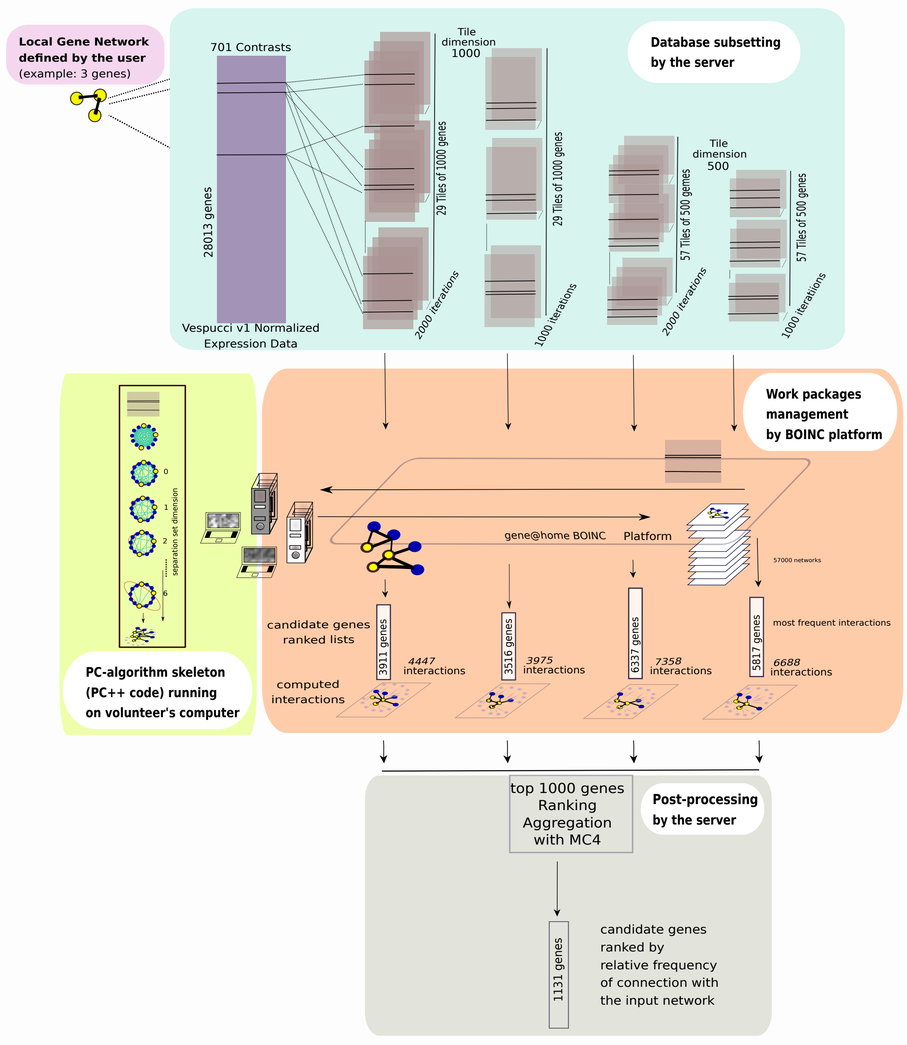
Stefania Pilati*, Giulia Malacarne, Samuel Valentini, Francesco Asnicar, Luca Masera, Marco Moretto, Paolo Sonego, Valter Cavecchia, Enrico Blanzieri and Claudio Moser.

FONDAZIONE E. MACH, Research and Innovation Center, via E. Mach, 1 - 38010 San Michele all'Adige (TN)- Italy
CNR, Institute of Materials for Electronics and Magnetism, via alla Cascata, 56/C - 38123 Trento - Italy
UNIVERSITY OF TRENTO, Dept of Information Engineering and Computer Science, via Sommarive, 9 - 38123 Trento -Italy
*stefania.pilati@fmach.it

Introduction

The development of “omics” technologies to study gene expression has revolutionized our perspective from the single gene to the *gene network* level. However, the complexity of the system biology approach requires appropriate mathematical, computational and statistical tools to analyze data and extract information. Grapevine transcriptomic data obtained with both microarrays and RNAseq technologies have been collected into the Vitis Expression Studies Platform Using COLOMBOS Compendia Instances (VESPUCCI, Moretto et al., 2016). Here we present the application of the algorithm of Network Expansion by Subsetting and Ranking Aggregation (NES^{2}RA, Asnicar et al., 2016) to the VESPUCCI database in order to expand four Local Gene Netowrks (LGNs) related to the grapevine response to climate changes. NES^{2}RA is based on the systematic and iterative application of the PC algorithm - aimed at identifying causal relationships from observational data - on subsets of the input data. To overcome the computational power requirement of NES^{2}RA algorithm, it has been run as part of the gene@home project, a distributed computation project which relies on thousands of volunteers’ computers managed by the TN-Grid, an infrastructure based on BOINC system (Asnicar et al., 2015).

Blocks scheme of the architecture of NES^{2}RA



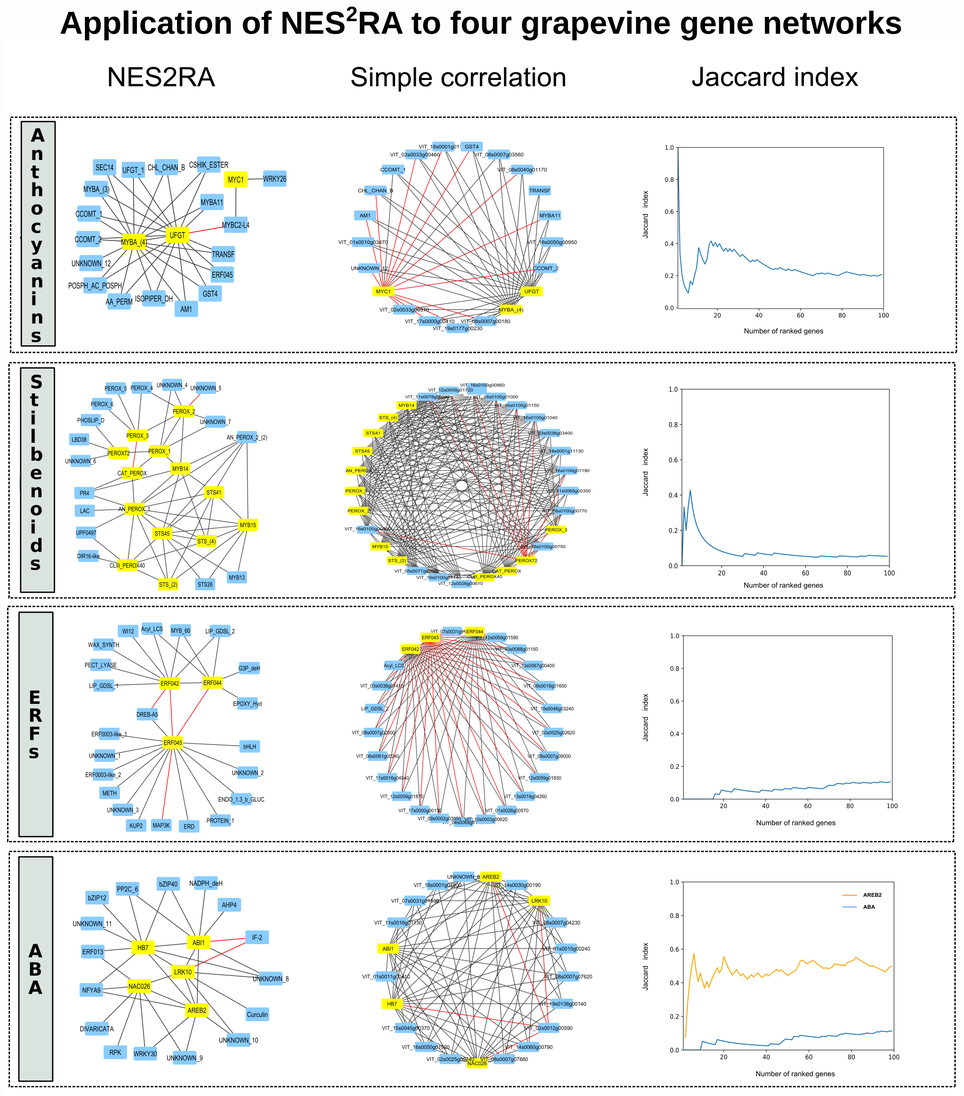
° Every time one LGN is provided, NES^{2}RA algorithm randomly divides the VESPUCCI dataset into tiles of equal number of genes (subsetting), to be then processed by the PC-algorithm. The random subsetting of all the genes in the genome is repeated for a given number of iterations.

° Tiles are divided into work packages by the BOINC platform, which are distributed to thousands of volunteers’ computers.

° PC-algorithm (Spirtes and Glymour, 1991) is based on a systematic test for conditional independence to retain significant relations between pairs of genes. It starts from a fully connected network and removes interactions between genes, if it finds a set of genes that supports that interaction (i.e., separation set).

° The networks found by the PC-algorithm are then post-processed off-line to determine and aggregate the final expansion gene lists. List aggregation is done by the ranking aggregation method Markov Chain 4 (MC4) (Dwork et al., 2001), since it has been shown to yield the most precise results (Asnicar et al., 2016), considering the first 1,000 genes of each expansion output list.

Application of NES^{2}RA to four grapevine gene networks



Networks visualization and comparison between NES^{2}RA and simple correlation analyses. Four LGNs related to the grapevine response to climate changes have been expanded using NES^{2}RA or simple correlation, as described in Malacarne et al., 2018. Networks have been visualized in Cytoscape using the interaction files produced by NES^{2}RA in one case (left column) and the Z-test for correlation (P-value < 0.05) in the other case. While in simple correlation networks the genes are almost fully connected, the number of interactions retained by NES^{2}RA is considerably reduced, allowing to focus on the most likely gene interactions.
Jaccard similarity index curves (right column) were calculated to compare the expansion gene lists obtained. In the case of LGNs formed by just one gene, results are quite similar (orange line in the bottom plot, about 60%), conversely for the other LGN expansions quite different trends and lower similarity were observed, suggesting that when a network rather than a single gene is expanded, the two approaches identify different sets of genes.

Conclusion

The present study proposes NES^{2}RA algorithm as a suitable tool to mine grapevine trasncriptomic data in order to highlight biologically relevant relationships among genes. The networks obtained integrating the information about the genes and their interactions found by NES^{2}RA provide an *in silico* hint to identify new genes of partially known metabolic and/or regulatory networks, as shown here. Beside, we are testing new applications, such as for example, the use of NES^{2}RA to discriminate among the isoforms of a gene family by exploiting the gene interaction information, in addition to the more common sequence structure analysis. Finally, we are developing a strategy to make NES^{2}RA available as a web tool that the biologist can interrogate in real time. One possibility under investigation consists in separating the gene expansion step, by pre-computing it, and the post-processing ranking step, computed on demand.