Building Transcriptional Association Networks in Cytoscape with *RegNetC*

Isabel A. Nepomuceno-Chamorro,
Alfonso Marquez-Chamorro, and Jesus S. Aguilar-Ruiz

Abstract—The Regression Network plugin for Cytoscape (*RegNetC*) implements the RegNet algorithm for the inference of transcriptional association network from gene expression profiles. This algorithm is a model tree-based method to detect the relationship between each gene and the remaining genes simultaneously instead of analyzing individually each pair of genes as correlation-based methods do. Model trees are a very useful technique to estimate the gene expression value by regression models and favours localized similarities over more global similarity, which is one of the major drawbacks of correlation-based methods. Here, we present an integrated software suite, named *RegNetC*, as a Cytoscape plugin that can operate on its own as well. *RegNetC* facilitates, according to user-defined parameters, the resulted transcriptional gene association network in .sif format for visualization, analysis and interoperates with other Cytoscape plugins, which can be exported for publication figures. In addition to the network, the *RegNetC* plugin also provides the quantitative relationships between genes expression values of those genes involved in the inferred network, i.e., those defined by the regression models.

Index Terms—Systems biology, transcriptional association networks, gene expression profiles, linear regression, model tree

1 Introduction

STANDARD approaches to biomarker discovery from gene expression data are based on the identification of differentially expressed genes. However, these approaches are based on the assumption that genes work in isolation when it is known that genes don't act independently from a System Biology point of view. For this reason, inferring gene-gene associations involved in a biological function is a relevant task in the area of microarray data analysis. To infer these gene-transcriptional association networks from gene expression profiles, there are several models from the straightforward correlation methods to more sophisticated methods such as Bayesian networks [1], [2]. All of them are based on a statistical measure of dependency between pairs of genes. The idea behind these methods, named the guilt-by-association heuristics, indicates that co-expression means co-regulation, i.e., if two genes show similar expression behavior then they are supposed to follow the same regulatory regime.

Our previous work introduthe Regression Network (RegNet) algorithm [3]. While other methods analyze each pair of genes, our approach infers transcriptional association networks taken into account each gene as a target and the remaining genes as inputs to estimate the behaviour of that gene. Our approach is a model tree-based method that strongly favours localized similarities over more global similarity, which is one of the major drawbacks of correlation-based methods. In the work mentioned earlier we used an initial implementation to successfully model transcriptional associations networks from several microarray experiments, including synthetic data sets, the *Saccharomyces cerevisiae* cell cycle data set and *Escherichia coli* gene expression database. We have also used our software as a part of a prognosis model based on the discovery of clinically relevant transcriptional association networks to discover potential biomarker in cardiovascular disease [4]. Furthermore, we have used the initial implementation of RegNet to detect associations between genes in Alzheimer's disease. However, there are several aspects to be improved in the initial implementation: no graphical interface or visualization of results and no integration with other network tools. Here we present an improved and extended version of the RegNet algorithm as a Cytoscape plugin to incorporate the visualization tool in Cytoscape software [5] that can interoperate with other Cytoscape plugins [6].

- *I.A. Nepomuceno-Chamorro is with the Department of Lenguajes y Sistemas Informaticos, Universidad de Sevilla, Spain. E-mail: inepomuceno@us.es.*
- *A. Marquez and J.S. Aguilar-Ruiz are with School of Engineering, Pablo de Olavide University, Seville, Spain.*

2 *RegNetC*

2.1 Algorithm

The original RegNet approach was described in [3]. The method is divided into three steps. In the first step, each gene is analysed as a target by taking into account the remaining genes as inputs to a model tree algorithm that estimates the expression value of that gene by means of linear regression functions in the leaves of the tree. We used the implementation of the M5′ model tree algorithm provided by the Weka library. The M5′ algorithm builds several regression functions spread over separate areas of the search space, i.e., optimal partitions of gene expression samples. Each regression function represents localized similarities between specific groups of genes. Moreover, the algorithm builds regression models under all samples (global similarity) if the optimal subspace is defined by the complete set of gene expression samples. From the forest of trees built in the first step, the method extracts a set of hypothetical gene-gene associations in the second step. Only the model trees from the forest which have a relative error ε higher than a threshold value is taken into account. Finally, the third step involves building a graph model of transcriptional association network by assessing the significance of the set of hypothetical evidences using the Benjamini-Yekutieli (BY) procedure [7] to control the false discovery rate.

2.2 The Input/Output Data and Visual Capabilities

RegNetC infers gene association networks from gene expression profiles. Fig. 1 shows a screenshot of the user interface. The software supports input files in either csv or arff format for the expression data set. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes [8]. Both formats are described in the web page.

To run the tool the user should provide several parameters: the threshold value for the prunning phase, i.e., the ε allowed in the model trees; the α value for the BY statistical procedure. Finally, the minimum number of instances in each localized subspace is the default value set in the software provided by [9]. The transcriptional network inferred by the method is displayed in .sif file format. This format is used to visualize the network in the plugin using the powerful visualization functionality within the Cytoscape framework. The resulted network in this format allows the user to interoperate with other Cytoscape plugins. For instance, one can use the NOA plugin [10] to analyze the network-based enrichment using Gene Ontology annotations [11] to network links. In addition to the network, the *RegNetC* plugin also reports the quantitative relationships between genes expression values of those genes involved in the inferred network, i.e., those defined by the regression models built using the model tree technique.

3 An Example

To briefly illustrate the potential of our approach, we applied *RegNetC* to the microarray data sets of [12]. It consists of 33 Alzheimer's disease (AD) brain samples. We run our method on

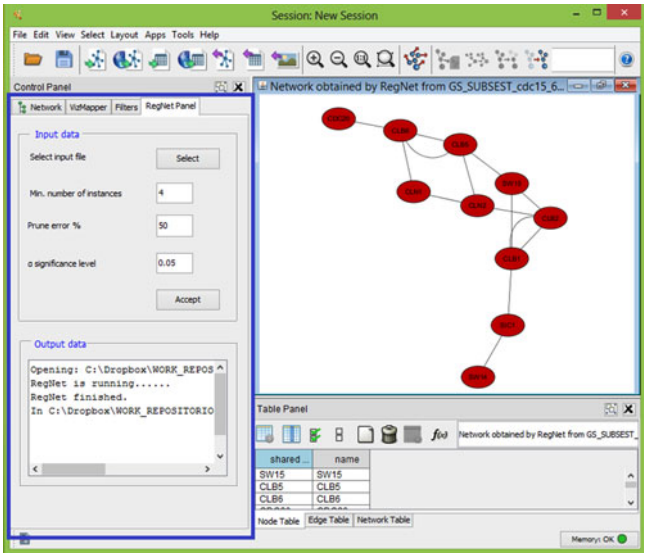

Fig. 1. The output panel of RegNetCscreenshot.

a subset of the data set which corresponds to 1,663 genes obtained from that data set using a pre-processing step described in [13]. From the network reported by *RegNetC* and using DAVID tool [14], we observed that genes from the network are related with various diseases: cardiovascular disease, Alzheimer's disease and type 2 diabetes. In an enrichment analysis using GO we detected several genes involved in immune, inflammatory or defense response. We can assert that we found more overlapping set of genes between our results and Miller's study (about AD and ageing) and the Kong's study (severe AD) than expected by chance (using the Hypergeometric test in two experiments: taking into account all and the 50 percent of the model tree.) We also found in the resulted network the inflammation-related gene named IFITM2 which is believed to be a culprit in AD pathogenesis [15]. Finally, it is known that the level of metal ion metabolism is closely associated with AD and we found metal protein-related genes named CHGB, MTF1 and MT1M in the network. In short, the approach is able to discover hidden associations due to the capability of analyzing local similarities by means of regression trees, and these can be checked by integrating RegNetC as a Cytoscape plugin.

4 Conclusions

The *RegNetC* is an integrated software suite implemented in Java for the inference of regression networks from gene expression profiles presented under Cytoscape as a plugin. This software tool improves some aspects of the original version: graphical interface or visualization of results and integration with other network tools. An advantage of being a Cytoscape plugin is that it allows the user to interoperate with other Cytoscape plugins to analyze the resulted network. As a future work we are planning to integrate our software in the community NetworkInference.org. Further tutorials, information and examples are available at http://www.lsi.us.es/~isanepo/toolRegNet/.

Acknowledgments

This paper has been partially supported by the Spanish government under project TIN2007-68084-C00 and Andalusian Government under project P11-TIC-7528. Isabel A. Nepomuceno-Chamorro and Alfonso Marquez-Chamorro are the corresponding authors for this paper.

References

[1] W. Lee and W. Tzou, "Computational methods for discovering gene networks from expression data," *Briefings Bioinform.*, vol. 10, no. 4, pp. 408–423, 2009.

[2] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo, "How to infer gene networks from expression profiles," *Molecular Syst. Biol.*, vol. 3, no. 1, pp. 1–10, 2007.

[3] I. Nepomuceno-Chamorro, J. Aguilar-Ruiz and J. Riquelme, "Inferring gene regression networks with model trees," *BMC Bioinformatics*, vol. 11, no. 1, p. 517, 2010.

[4] I. Nepomuceno-Chamorro, F. Azuaje, Y. Devaux, P. V. Nazarov, A. Muller, J. S. Aguilar-Ruiz, and D. R. Wagner, "Prognostic transcriptional association networks: A new supervised approach based on regression trees," *Bioinformatics*, vol. 27, no. 2, pp. 252–258, Jan. 2011.

[5] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, 2003.

[6] R. Saito, M. Smoot, K. Ono, J. Ruscheinski, P. Wang, S. Lotia, A. Pico, G. Bader, and T. Ideker, "A travel guide to cytoscape plugins," *Nature Method*, vol. 9, no. 11, pp. 1069–1076, 2012.

[7] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, no. 4, pp. 1165–1188, 2001.

[8] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[10] C. Zhang, J. Wang, K. Hanspers, D. Xu, L. Chen, and A. R. Pico, "NOA: A cytoscape plugin for network ontology analysis," *Bioinformatics*, vol. 29, no. 16, pp. 2066–2067, 2013.

[11] T. G. O. Consortium, "Gene ontology: Tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–34, 2000.

[12] T. Dunckley, T. Beach, K. Ramsey, A. Grover, D. Mastroeni, D. Walker, B. LaFleur, K. Coon, K. Brown, R. Caselli, W. Kukull, R. Higdon, D. McKeel, J. Morris, C. Hulette, D. Schmechel, E. Reiman, J. Rogers, and D. Stephanai, "Gene expression correlates of neurofibrillary tangles in alzheimers disease," *Neurobiol. Aging*, vol. 27, no. 10, pp. 1359–1371, Oct. 2006.

[13] M. Ray, J. Ruan, and W. Zhang, "Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases," *Genome Biol.*, vol. 9, no. 10, p. R148, Oct. 2008.

[14] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using David bioinformatics resources," *Nat. Protocols*, vol. 4, no. 1, pp. 1754–2189, 2009.

[15] W. Kong, X. Mou, Q. Liu, Z. Chen, C. R. Vanderburg, J. T. Rogers, and X. Huang, "Independent component analysis of Alzheimer's DNA microarray gene expression data," *Molecular Neurodegeneration*, vol. 4, article 5, 2009.