Integration of function hierarchy data for gene function prediction

Miguel Romero, Jorge Finke, and Camilo Rocha

Department of Electronics and Computer Science
Pontificia Universidad Javeriana de Cali, Colombia
{miguel.romero,jfinke,camilo.rocha}@javerianacali.edu.co

Developments like high-throughput sequencing technologies have enable the identification of a large number of genes (and gene products), but many of these genes have no known function (Ranganathan et al., 2019). Identifying functions associated to genes is a key step to better understand how genes relate to the genome as a whole. Most efforts address this task as a binary classification problem based on gene expression data or protein interaction data.

Gene co-expression networks (GCN) integrate large transcriptional data sets and have been used to infer biological information (e.g., biological processes or pathways) based on highly correlated expression patterns between genes (Oti et al., 2008; van Dam et al., 2017; Vandepoele et al., 2009). Co-expressed genes, i.e., genes with similar expression profiles, tend to share the same function or be related to the same regulatory pathway (Emamjomeh et al., 2017; Serin et al., 2016; Zhou et al., 2002).

Traditional approaches to annotating gene functions consider each function (biological processes) independently, ignoring the relations and dependencies between them. However, function assignment to genes or gene products must obey the *true-path rule*: if a function is associated to a gene (or gene product), all its ancestors must also be associated to that gene (or gene product) (Jiang et al., 2008). Therefore, to predict whether a gene (or gene product) is associated to a particular function, its ancestors in the function hierarchy must be considered. Ignoring such a hierarchy often leads to inconsistencies in the prediction results.

This work introduces a novel model for functional annotation of genes based on the information from gene co-expression networks and the function hierarchical structure defined by Gene Ontology (GO, http://geneontology.org/). The proposed approach considers the function hierarchy for the classification task, instead of each term independently. In other words, the approach guarantees that the *true-path* rule is satisfied. The model is an extension of the approach presented in (Romero et al., 2020), which uses machine learning techniques to classify whether a gene is associated to a function. Although the work in (Romero et al., 2020) fails to account for function hierarchy, the authors show that for some biological processes structural properties of the gene co-expression network may improve the performance of the classification. The proposed model take into account structural properties and features learned from the gene co-expression network, alongside with the known gene functional information.

The performance of the proposed model is compared to the probabilistic Hierarchical Binomial-Neighborhood (HBN) model presented in (Jiang et al., 2008). Both models are applied to rice (*Oryza sativa Japonica*).

Bibliography

Emamjomeh, A., Saboori Robat, E., Zahiri, J., Solouki, M., & Khosravi, P. (2017). Gene co-expression network reconstruction: A review on computational methods for inferring functional information from plant-based expression data. *Plant Biotechnology Reports*, *11*(2), 71–86.

Jiang, X., Nariai, N., Steffen, M., Kasif, S., & Kolaczyk, E. D. (2008). Integration of relational and hierarchical network information for protein function prediction. *BMC Bioinformatics*, *9*(1), 350.

Oti, M., van Reeuwijk, J., Huynen, M. A., & Brunner, H. G. (2008). Conserved co-expression for candidate disease gene prioritization. *BMC Bioinformatics*, *9*(1), 208.

Ranganathan, S., Gribskov, M. R., Nakai, K., & Schönbach, C. (2019). *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier. OCLC: 1052465484.

Romero, M., Finke, J., Quimbaya, M., & Rocha, C. (2020). In-silico Gene Annotation Prediction Using the Co-expression Network Structure. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.) *Complex Networks and Their Applications VIII*, vol. 882, (pp. 802–812). Cham: Springer International Publishing.

Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M., & Ligterink, W. (2016). Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in Plant Science*, *7*.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, (p. bbw139).

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., & Van de Peer, Y. (2009). Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. *Plant Physiology*, *150*(2), 535–546.

Zhou, X., Kao, M.-C. J., & Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*, *99*(20), 12783–12788.