Semantic Similarity Integration for Gene Network Inference

Roger Verzola Peres de Lima, Fábio Fernandes da Rocha Vicente

Federal University of Technology - Paraná

Abstract

Genes are fundamental elements in the dynamic of biological systems. Finding out how genes interact with the dynamic of biological systems may foster not only a better understanding of living beings, but also the possible genetic manipulations of said living beings with a specific aim. Therefore, the inference of gene regulatory networks is of great importance. However, an issue in that field is the small amount of samples available when compared to the amount of variables, severely limiting the inference power of purely statistical methods. In this work a method that contours that difficulty by uniting both quantitative data and qualitative data is proposed. Our method combines two types of data: gene expression and gene ontology. The criterion function calculates the mean conditional entropy over the normalized gene expression data and the GFD-Net over the genes' ontology annotations. GFD-Net is a method that gives a numerical score to the functional dissimilarity of a gene network based on gene ontology. The proposed algorithm to use is the Sequential Forward Feature Selection (SFFS) due to its easy implementation and deterministic nature. Therefore, the proposed method is made of two parts: an algorithm that selects a suboptimal set of genes (called predictors) that may interact with the target gene; and a criterion function that said algorithm will use to determine which subset of genes the predictor set will be. Running said algorithm over every target gene enables us to form a gene regulatory network by creating an edge between each predictor set and their respective targets. The method aims to use two distinct forms of evaluation unifying, therefore, both semantic and quantitative measures.

Funding: CAPES, CNPq, Fundação Araucária