GeNICE: A Novel Framework for Gene Network Inference by Clustering, Exhaustive Search, and Multivariate Analysis

Ricardo de Souza Jacomini[1], David Correa Martins Jr[2], Felipe Leno da Silva[1], Anna Helena Reali Costa[1]

1 ESCOLA POLITÉCNICA DA USP
2 UFABC

Abstract

Gene network (GN) inference from temporal gene expression data is a crucial and challenging problem in systems biology. Expression data sets usually consist of dozens of temporal samples, while networks consist of thousands of genes, thus rendering many inference methods unfeasible in practice. To improve the scalability of GN inference methods, we propose a novel framework called GeNICE, based on probabilistic GNs; the main novelty is the introduction of a clustering procedure to group genes with related expression profiles and to provide an approximate solution with reduced computational complexity. We use the defined clusters to perform an exhaustive search to retrieve the best predictor gene subsets for each target gene, according to multivariate criterion functions. GeNICE greatly reduces the search space because predictor candidates are restricted to one gene per cluster. Finally, a multivariate analysis is performed for each defined predictor subset to retrieve minimal subsets and to simplify the network. In our experiments with in silico generated data sets, GeNICE achieved substantial computational time reduction when compared to solutions without the clustering step, while preserving the gene expression prediction accuracy even when the number of clusters is small (about 50) relative to the number of genes (order of thousands). For a Plasmodium falciparum microarray data set, the prediction accuracy achieved by GeNICE was roughly 97%, while the respective topologies involving glycolytic and apicoplast seed genes had a very large intramodularity, very small interconnection between modules, and some module hub genes, reflecting small-world and scale-free topological properties, as expected.

Funding: CNPq, FAPESP