BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Biological Networks
Title	The direction of causation: inferring directed relationships by post-processing causal gene networks discovered from omics data.
All Authors	Pelagalli C(1), Blanzieri E(1,2), Cavecchia V(2), Pilati S(3), Tebaldi T(4)

Affiliation

(1) DISI, University of Trento, Italy
(2) CNR-IMEM, Trento, Italy
(3) Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy
(4) CIBIO, University of Trento, Italy

Motivation

Research on gene expression data is essential to identify functional interactions between genes in a species, by reconstructing gene regulatory networks. The gene@home project implemented the OneGenE method [1] to discover novel candidate genes that could be associated with an initial gene regulatory network, by performing a so-called task of gene network expansion. However, the expansion lists produced by OneGeneE do not determine the causal direction of the putative relationships between genes. We consider here to post-process the expansion lists by means of the MPC-stable algorithm (a version of PC algorithm [2]) that explores and tests the relationships and, possibly, orients them causally. The application of MPC-stable to gene transcriptomic data aims at giving causal direction to the top associations found among genes by OneGenE. This procedure builds, as a working hypothesis, a model for a directed gene network in silico: it could shed light on gene functional dependencies and be a useful tool for the biologist, by testing, and possibly discovering, relationships of cause and effect between gene transcripts.

Methods

The OneGenE method expands a candidate gene of an organism, by applying a systematic and iterative version of the PC algorithm's skeleton function, given an input matrix, which contains a chosen subset of genes, including the candidate gene, and their observational expression data. Since this method is applied to both Homo sapiens and Vitis vinifera genes, input matrices are retrieved from two different resources: the FANTOM5 dataset for human genes and the VESPUCCI compendium for grapevine. After building the complete undirected graph, connecting every pair of genes with each other, the function computes the graph skeleton: undirected edges are cut according to the conditional independence test, performed on all couples, by means of their expression data. If gene X and gene Y pass the test, the edge between them is cut and the subset S of genes, conditioned on in the independence test, is included in their separation set. If X and Y fail the test, the edge is preserved. Note that since $S \subseteq adj(G, X)\backslash\{Y\}$, the S cardinality increases from zero to $|adj(G, X)\backslash\{Y\}|$. Each iteration of this function is RAM and time consuming: furthermore, separation sets need to be stored in order to proceed with the following orientation steps of the PC algorithm. The expansion list follows a decreasing order: the first ranking gene has the highest relative frequency and the strongest correlation with the candidate gene. At high relative frequency the correlation between a high-ranking gene and the candidate gene can be interpreted as corresponding to a direct causal relationship. The MPC-stable algorithm, which is a modified version of the original PC algorithm, is applied to a subset of the candidate gene expansion list, by adding to the candidate gene the first thirty genes in the ranking and a variable number of the subsequent genes on the list. In successive runs of the algorithm, this number is increased until it coincides with all the genes of the candidate gene expansion list. The MPC-stable applies the above-mentioned skeleton function and proceeds to orient as many of the remaining edges as possible in two major steps: v-structure orientation and orientation rules. If an edge between two genes X and Y is oriented in the graph, as $X \rightarrow Y$, X is interpreted as a direct cause of Y. The output should be a completed partially directed acyclic graph (CPDAG), where nodes represent genes and directed edges reflect causal relationships. The candidate gene and the first thirty genes of its expansion list are tracked in the multiple iterations of the MPC-stable algorithm, to check if they are present as nodes in the output graphs and in which causal relationship they stand with each other, by checking edges and their direction. Note that the MPC-stable is applied to an increasing input of genes, so edges between the candidate gene and the first thirty genes may vary in number and orientation. The MPC-stable code used is a package available in the R environment.

Results

The MPC-stable algorithm has been used to investigate the gene network of a V. vinifera transcription factor and a human lysophosphatidylcholine (fatty acids) transporter. Biological validation of the obtained graphs and analysis results has been performed by integrating additional experimental data.
Since edge directions and node connections are the main focus of this work, results, which may be interesting for further consideration by specialists in bioinformatics, are shown in the form of Cytoscape graphs, in accordance with the algorithm output.

Info

References:
[1] Blanzieri, E., Tebaldi, T., Cavecchia, V., Asnicar, F., Masera, L., Tome, G., Nigro, E.,

Colasurdo, E., Ciciani, M., Mazzoni, C., Pilati, S.: A Computing System for Discovering Causal Relationships Among Human Genes to Improve Drug Repositioning. IEEE Transactions on Emerging Topics in Computing 9(4), 1667–1682 (Oct 2021), https://ieeexplore.ieee.org/document/9224179/
[2] Spirtes, P., Glymour, C.N., Scheines, R.: Causation, prediction, and search. Adaptive computation and machine learning, MIT Press, Cambridge, Mass, 2nd ed edn. (2000)

filename	-
Figure	
-	
Availability	-

Corresponding Author

Name, Surname	Enrico, Blanzieri
Email	enrico.blanzieri@unitn.it
Submitted on	29.04.2022