Discovering Candidates for Gene Network Expansion by Variable Subsetting and Ranking Aggregation

Luca Erculiani* Francesca Galante* Caterina Gallo* Francesco Asnicar* Luca Masera* Paolo Morettin* Nadir Sella* Thomas Tolio* Giulia Malacarne† Kristof Engelen† Andrea Argentini‡ Valter Cavecchia§ Claudio Moser† Enrico Blanzieri*

We present a method that produces a list of genes that are candidates for Network Expansion by Subsetting and Ranking Aggregation (NESRA) and its application to gene regulatory networks. Our group has recently developed gene@home [3], a BOINC project [1] that permits to search for candidate genes for the expansion of a gene regulatory network using gene expression data. The project adopts intensive variable-subsetting strategies enabled by the computational power provided by the volunteers who join the project by means of the BOINC client, and exploits the PC algorithm for discovering putative causal relationships within each subset of variables. The PC algorithm, whose name derives from the initials of its authors [7] and PC* [2] are algorithms that discover causal relationships among variables. In particular, PC is based on the systematic testing for conditional independence of variables given subsets of other variables, comprehensively presented and evaluated by Kalish and colleagues [4] who proposed it also for gene network reconstruction [5]. NESRA is an algorithm which runs as a postprocessor of the gene@home project that has: 1) a procedure that systematically subsets the variables, runs the PC and ranks the genes; the subsetting is iterated several times and a ranked list of candidates is produced by counting the number of times a relationship is found; 2) several ranking steps are executed with different values of the dimension of the subsets and with different number of iterations producing several ranked lists; 3) the ranked lists are aggregated by using a state-of-the-art ranking aggregator. Here we show that a single ranking step is enough to outperform PC and PC*, but with some dependency on the parameters. Moreover, we show that the output ranking aggregation method is better that the average performance of the single ranking steps. Evaluations are done by means of the gene@home project on *Arabidopsis thaliana* including a comparison against ARACNE [6] (Table 1).

Method	k=5	k=10	k=20	k=55
NESRA	0.90	0.80	0.60	0.42
ARACNE	0.2	0.3	0.35	0.45

Table 1: *A. thaliana*, Expansion of the Flower Organ Specification Gene Regulatory Network. NESRA and ARACNE (default parameters) precision for different values k of the length of the gene list.

References

[1] D. P. Anderson. BOINC: A system for public-resource computing and storage. In *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, GRID '04, pages 4–10, Washington, DC, USA, 2004. IEEE Computer Society.

[2] D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *arXiv preprint arXiv:1211.3295*, 2012.

[3] gene@home. `http://gene.disi.unitn.it/test/`, April 2015.

[4] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, 8:613–636, May 2007.

[5] M. H. Maathuis et al. Predicting causal effects in large-scale systems from observational data. *Nat. Methods*, 7(4):247–248, Apr 2010.

[6] A. A. Margolin et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

[7] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

*University of Trento, Italy.

†CRI, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy.

‡Ghent University and VIB, Ghent, Belgium.

§CNR-IMEM, Trento, Italy.