# A simulator of gene regulatory network evolution for validation of biomarker identification methods

Di Camillo B[(1)], Martini M[(1)], Toffolo G[(1)]

[(1)] Department of Information Engineering, University of Padova, Padova

## Motivation

Class prediction and feature selection are two strictly paired tasks in biomarker identification from high throughput expression data derived from different phenotypes. Although good performance, i.e. the percentage of subjects classified in the correct group, was reached in several studies, usually, the lists of biomarkers only partially overlap when the experimental protocols are replicated. The reasons for these differences are imputable to different causes: -Datasets present a low number of subjects (order of tens) with respect to the number of variables (order of tens of thousands) under analysis; -Groups of non-healthy subjects are highly non homogeneous; -Some pathologies, e.g. cancer, are correlated with the malfunctioning of one or more physiologic pathways, rather than with few genes. The above issues have been addressed in different works [1, 2]; however, it is not easy to validate classification methods using real data, due to the lack of both numerous samples and sufficient knowledge of the regulatory network underlying expression data. Our goal is to build a realistic simulator, that, by exploiting a regulatory network model that contains some impaired regulation in the pathways of the non-healthy subjects, generates data that: 1) have statistical distribution similar to real data, 2) when undergo a biomarker discovery algorithm, gives similar results to the one obtained with real data.

## Methods

A procedure similar to that described in [3] is used to generate a population of healthy subjects. Subjects are modeled as regulatory networks of N=200 genes, exploiting a simulator described in [4]. Topology is randomly generated, constrained to have scale-free distribution of the degree and clustering coefficient independent of the number of nodes in the network. Once the topology is defined, differential equations are used to simulate the dynamic of the system that, by explicitly representing interactions among the regulators of each gene, can be characterized by a finite number of possible steady states. The regulatory network is characterized by: - A connectivity matrix W with weights wij representing the affinity of the promoter region of gene i for transcription factor j. Since wij can in principle be mapped to specific nucleotide sequences in the enhancer regions of gene i, W also represents the genotype of the subject. - A finite number of steady states, which characterize the phenotype of each individual. We consider an initial population of M=100 identical individuals; subsequent generations are produced by random pairing of individuals. Offspring are created by randomly selecting rows of the connectivity matrix from each parent with equal probability and mutating the nonzero wij with probability equal to 5% of their number. Subjects with at least one mutated wij survive and can generate offspring only if their phenotype does not change with respect to the original population. At each generation, we simulate 100 individuals, independently on the number of subjects survived in the previous generation. Evolution proceeds for 100 generations, a time sufficient to have a final population of 100 subjects with the same phenotype but different genotype. This population is divided in two groups of 50 individuals, and for one group, we simulate a pathological condition by knocking out two genes chosen among the ones with highest out-degree. Biomarkers are defined as the KO genes and the genes they directly regulate.

## Results

Using opportune simulator parameter setting we were able to generate two groups of subjects with distributions similar to those observed in a real Affymetrix dataset (HGu133plus) of 29 leukemicand 13 control subjects (p= 0.90 and 0.93 respectively, using Wilcoxon-test). Moreover, the non-healthy population shows a within group variability higher than the healthy population, as observed in real data, due to the fact that KO produces different effects on subjects with different connectivity weights wij. Moreover, the results obtained by applying Support Vector Machine (SVM) to simulated data are similar to those reported for real data. With few subjects per group (<10), SVM shows good performance in classifying the subjects in the correct group (precision=0.92, sensitivity=1 estimated using leave 1 out cross validation);

however, fails to identify biomarkers (precision=0.45, sensitivity=0.52). Robust classification approaches, as the one proposed in [1], improve SVM performance in producing more stable and accurate lists of putative biomarkers, however, they are not always able to identify them (precision=0.70, sensitivity=0.83);. In conclusion, the simulator provides an useful test bed to validate biomarker discovery methods and to develop new methods that accounts for the regulatory network underlying gene expression.

## References

[1] Furlanello et al, IEEE/ACM Trans. on Comp. Biol. and Bioinf., 2005
[2] Segal et al. Nature Genetics, 2004
[3] Siegal and Bergman, PNAS 2002
[4] Di Camillo et al., PNYAS 2009

**Contact :** dicamill@dei.unipd.it