Goals and pitfalls of gene-network inference methods:
A comparative study from virtual microarrays and network dynamics

J. Goñi[1,2], C. Rodríguez-Caso[3], R. V. Solé[3,4], P. Villoslada[2], A. Munteanu[3]

[1]Dept. of Physics and Applied Mathematics, University of Navarra, [2]Dept. of Neuroscience. Center for Applied Medical Research, University of Navarra, [3]Complex Systems Lab, GRIB - University Pompeu Fabra, [4]Santa Te Institute, New Mexico

Abstract

Correct identification of gene-gene interactions is a critical step in uncovering gene regulatory networks. Microarrays technology provides information about the state of thousands of genes. The meta-analysis of this data provides gene expression correlations required for genetic network inference. However, there is no unifying reliable framework of network inference, as the assumptions and results of these methods are quite diverse. Here we consider two important issues related to the capability of these methods to correctly infer the underlying gene activation/inhibition relations: 1) the network's *topology*; 2) genes' *intrinsic stochasticity* and the cell population heterogeneity of the samples, which combined with the experimental errors, result in noisy microarrays data of gene expression. We evaluate different approximations of gene network inference, such us statistical methods, information theory and Bayesian networks, using *in silico* gene expression data. This data has been obtained from deterministic and stochastic dynamical simulations of a set of pre-defined network topologies. By the presented framework, we provide a dynamical calibration study for these methods by applying them on 3 synthetic and one empirical meso-scale gene networks.

Introduction

We conducted a performance study of gene network inference by means of synthetic gene expression data. This *virtual microarrays* data (Fig. 1) was produced through a meso-scale approach (Bornholdt, S. Science 310,449, 2005; Kirschner, M. Cell 121, 503, 2005), more precisely by dynamic simulations of the gene network using a Boolean network approximation. We have used as model gene network the cell-cycle one from Li et al. (PNAS 101, 4781, 2004), a network characterised by **11** genes and its regulations. In addition, three networks of different topologies were also used, and in order to correctly compare among different topologies, the number of **11** genes was maintained (Fig. 2).

Virtual microarrays

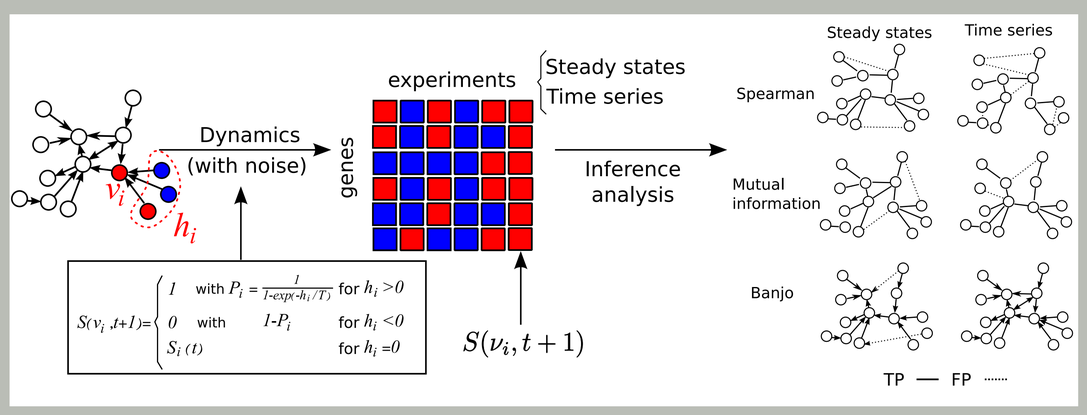

Figure: 1. Schematic view of the *virtual microarrays* conceptual approach. By simulating the network dynamics, two types of datasets have been produced: 1) *time-series*; 2) *attractors* and *steady-state series* data. The state $S_i(t+1)$ of gene i is controlled by the state of its j neighbours at previous time step through the sign of $h_i \equiv \sum a_{ij} S_j(t)$, with $a_{ij} \in \{-1, 1\}$ defining the inhibition/activation gene relations (Fig. 1), and the probability P_i depends on the intensity of noise T.

Network dynamics

The rules for obtaining the expression profiles of the gene networks (Fig. 1) are based on Li et al., where the gene states are $S = 0$ (inactive) and $S = 1$ (active). For the *time-series* approach, we recorded the time trajectories of the system's states from all possible initial conditions ($2^{11} = 2048$) until the corresponding steady-state. For the other approach (see Fig. 1), we restricted to these final steady-states, either just the system's attractors (once each) or all the system's steady states ($2^{11} = 2048$). The inference methods were applied on **20%** of the datasets from both approaches, considering that not all expression profiles are accessible in real genetic networks, and thus in real microarrays experiments.

Synthetic model-networks

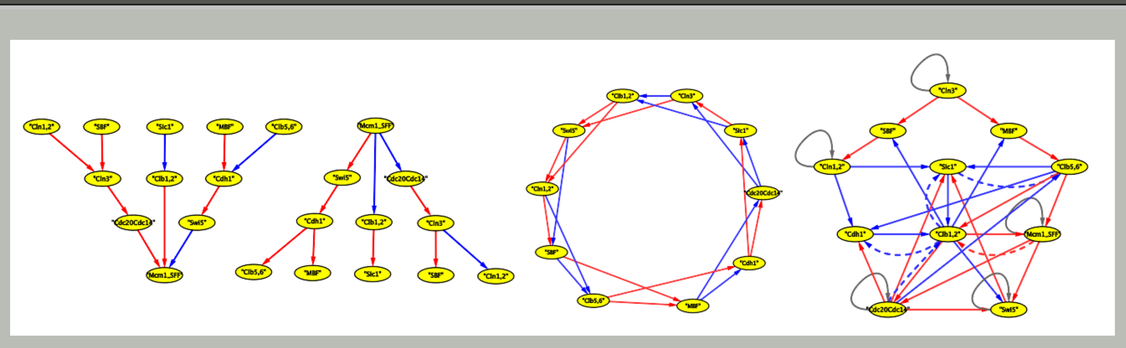

Figure: 2. Meso-scale network models analysed in the current study. From left to right, the convergent tree-like, divergent tree-like, regular circle and Li cell-cycle network of Li et al. The arrows represent: red = activation, blue = inhibition, grey = self-inhibition (see Li et al. for details).

Gene network inference and evaluation

The methods selected stand for the three major frameworks in the area of reverse engineering:

- **Statistical methods. Spearman correlation:** this coefficient provides signed correlations among variables, without inferring interaction directionality.
- **Information Theory. Mutual information (MI):** MI(x,y) measures x entropy reduction when y-value is known (usually measured in bits). It provides unsigned undirected associations.
- **Bayesian networks. Banjo:** Banjo is a gene network inference software developed by Yu et al. (Bioinformatics 20, 3594, 2004). It provides signed directed relationships.

Methods' performance was measured by the *accuracy* $\boldsymbol{Ac} = \frac{TP}{TP+FP}$ and the *sensitivity* $\boldsymbol{Se} = \frac{TP}{TP+FN}$ measures, with $\boldsymbol{TP}$, the number of true positives, and $\boldsymbol{FP}$, of false positives and $\boldsymbol{FN}$ of false negatives (Bansal et al. Mol. Syst. Biol. 3, 2007). In ideal conditions: $\boldsymbol{FN} = \boldsymbol{FP} = \boldsymbol{0}$, and thus $\boldsymbol{Ac}$ and $\boldsymbol{Se}$ reach maximum value, $\boldsymbol{Ac} = \boldsymbol{Se} = \boldsymbol{1}$ (green colour in Table 1).

		Spearman		MI		BANJO	
Dataset	T	Ac	Se	Ac	Se	Ac	Se
Convergent							
time	0	0.89 ± 0.09	0.76 ± 0.08	0.85 ± 0.08	0.84 ± 0.10	0.48 ± 0.15	0.69 ± 0.14
time	0.5	1.00 ± 0.00	0.54 ± 0.04	1.00 ± 0.00	0.56 ± 0.04	0.53 ± 0.06	0.45 ± 0.07
time	2	1.00 ± 0.00	0.79 ± 0.04	1.00 ± 0.00	0.78 ± 0.04	0.60 ± 0.12	0.70 ± 0.09
time	10	0.65 ± 0.07	0.92 ± 0.10	0.66 ± 0.10	0.92 ± 0.08	0.30 ± 0.08	1.00 ± 0.00
steady	0	1.00 / 1.00	0.68 / 0.71	0.98 / 1.00	0.70 / 0.66	0.57 / 0.40	0.90 / 1.00
steady	0.5	1.00 / 1.00	0.80 / 0.71	0.91 / 1.00	0.59 / 0.66	0.52 / 0.40	0.58 / 0.80
steady	2	0.86 / 0.90	0.93 / 0.90	0.79 / 0.90	0.74 / 0.81	0.54 / 0.40	0.73 / 0.80
steady	10	0.09 / 0.40	0.54 / 0.80	0.60 / 0.40	0.71 / 0.80	0.26 / 0.10	0.91 / 0.10
Divergent							
time	0	0.86 ± 0.06	0.23 ± 0.02	0.83 ± 0.05	0.25 ± 0.04	0.48 ± 0.13	0.50 ± 0.10
time	0.5	1.00 ± 0.00	0.18 ± 0.00	1.00 ± 0.00	0.18 ± 0.00	0.58 ± 0.14	0.62 ± 0.11
time	2	1.00 ± 0.00	0.38 ± 0.01	1.00 ± 0.00	0.40 ± 0.02	0.49 ± 0.14	0.68 ± 0.15
time	10	0.49 ± 0.10	0.66 ± 0.10	0.44 ± 0.08	0.69 ± 0.11	0.18 ± 0.12	0.80 ± 0.32
steady	0	1.00 / 0.80	0.38 / 0.53	1.00 / 0.80	0.39 / 0.61	0.50 / 0.30	0.93 / 1.00
steady	0.5	1.00 / 0.80	0.30 / 0.53	1.00 / 0.80	0.24 / 0.61	0.46 / 0.50	0.85 / 1.00
steady	2	0.71 / 0.90	0.60 / 0.64	1.00 / 0.80	0.21 / 0.66	0.56 / 0.40	0.79 / 0.80
steady	10	0.07 / 0.40	0.27 / 1.00	1.00 / 0.20	0.21 / 1.00	0.25 / 0.20	0.98 / 1.00
Circular							
time	0	0.79 ± 0.03	0.43 ± 0.02	0.79 ± 0.04	0.46 ± 0.02	0.30 ± 0.11	0.59 ± 0.18
time	0.5	0.86 ± 0.05	0.46 ± 0.02	0.83 ± 0.05	0.47 ± 0.03	0.18 ± 0.03	0.63 ± 0.21
time	2	0.69 ± 0.04	0.95 ± 0.05	0.66 ± 0.02	0.95 ± 0.04	0.29 ± 0.09	0.93 ± 0.06
time	10	0.16 ± 0.06	0.88 ± 0.12	0.17 ± 0.06	0.85 ± 0.14	0.04 ± 0.03	0.56 ± 0.49
steady	0	0.67 / 0.40	0.55 / 0.90	0.62 / 0.13	0.58 / 1.00	0.23 / 0.31	0.90 / 1.00
steady	0.5	0.62 / 0.09	0.77 / 1.00	0.15 / 0.00	0.84 / 0.00	0.28 / 0.13	0.46 / 0.60
steady	2	0.70 / 0.04	0.43 / 0.50	0.04 / 0.00	0.46 / 0.00	0.10 / 0.04	0.47 / 0.50
steady	10	0.50 / 0.00	0.45 / 0.00	0.02 / 0.00	0.25 / 0.00	0.00 / 0.00	0.00 / 0.00
Li							
time	0	0.92 ± 0.04	0.44 ± 0.01	0.93 ± 0.02	0.46 ± 0.01	0.37 ± 0.08	0.41 ± 0.07
time	0.5	0.97 ± 0.02	0.47 ± 0.01	0.95 ± 0.03	0.46 ± 0.01	0.29 ± 0.08	0.62 ± 0.13
time	2	1.00 ± 0.00	0.43 ± 0.00	1.00 ± 0.00	0.43 ± 0.00	0.35 ± 0.10	0.69 ± 0.10
time	10	0.96 ± 0.02	0.43 ± 0.00	0.96 ± 0.03	0.44 ± 0.01	0.46 ± 0.09	0.68 ± 0.04
steady	0	0.12 / 0.04	0.50 / 1.00	0.12 / 0.00	0.49 / 0.00	0.19 / 0.25	0.40 / 0.75
steady	0.5	0.75 / 0.00	0.52 / 0.00	0.91 / 0.00	0.48 / 0.00	0.20 / 0.00	0.40 / 0.00
steady	2	0.91 / 0.00	0.48 / 0.00	0.91 / 0.04	0.48 / 1.00	0.19 / 0.16	0.61 / 0.66
steady	10	0.91 / 0.00	0.48 / 0.00	0.91 / 0.00	0.48 / 0.00	0.21 / 0.12	0.48 / 0.42

Table: 1. Inference performance for the studied networks. From top to bottom: convergent-tree, divergent-tree, regular circular and Li cell-cycle network. *Dataset* = *time* (20% time-series) or *steady* (20% series/attractors); $\boldsymbol{T}$ = Temperature; *Ac* = Accuracy; *Se* = Sensitivity.
Greenish colours: $\boldsymbol{Ae}, \boldsymbol{Se} \approx \boldsymbol{1}$, Reddish colours: $\boldsymbol{Ae}, \boldsymbol{Se} \approx \boldsymbol{0}$.

Results and discussion

It can be seen that the methods' performance tightly depends on the underlying network topology. High values of both $\boldsymbol{Ac}$ and $\boldsymbol{Se}$ characterise tree-like topologies, and lower values are associated to regular networks and real high-connectivity networks. As a consequence, signalling networks are more accurately predicted by these methods. Additionally, there is an optimum level of stochasticity for the first two methods that leads to high values in both $\boldsymbol{Ac}$ and $\boldsymbol{Se}$. On the contrary, prediction based on the Banjo method appears to be independent on the noise level, and unexpectedly, shows $\boldsymbol{Se} > \boldsymbol{Ac}$. This emphasises the heterogeneity of performance depending on the inference method selected.