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 (Borovkova, 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**

![Virtual microarrays](image)

**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 $S_i$ of gene $i$ is controlled by the state of its $j$ neighbours at previous time step through the sign of $\theta_i = \sum_j \theta_{ij} S_j$, with $\theta_{ij} \in (-1,1)$ 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 (211 = 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 (211 = 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**

![Synthetic model-networks](image)

**Figure:** 2. Meso-scale network models analysed in the current study. From left to right, the convergent tree-like, divergent tree-like, regular cube 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).

**Results and discussion**

It can be seen that the methods’ performance tightly depends on the underlying network topology. High values of both Ac and 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 Ac and Se. On the contrary, prediction based on the Banjo method appears to be independent on the noise level, and unexpectedly, shows Se > Ac. This emphasises the heterogeneity of performance depending on the inference method selected.

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

![Table: Inference performance for the studied networks.](image)

**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); $T$ = Temperature; $Ac$ = Accuracy; $Se$ = Sensitivity. Greenish colours: $Ac$, $Se$ = 1; Reddish colours: $Ac$, $Se$ = 0.