

# Extended Bayesian scores for reconstructing gene regulatory networks

Jimmy Vandel, Brigitte Mangin, Matthieu Vignes, and Simon de Givry

INRA - BIA, Toulouse, France,  
jimmy.vandel@toulouse.inra.fr

**Abstract.** The discovery of regulatory networks is an important aspect in the post genomic research. We compared various scoring criteria in the context of structure learning in Bayesian networks. We show that adding a uniform node in-degree prior helps improve precision without deteriorate sensitivity for typical regulatory networks encountered in prokaryote and eukaryote organisms. Experiments are performed on simulated highly non-linear genetical genomics data. Taking into account specific biological information about DNA markers can further enhance the reconstruction process. We performed a large comparison with existing approaches of gene regulatory network inference. Our preliminary results show that the best approach exploits an extended normalized maximum likelihood score in the framework of discrete Bayesian networks.

**Keywords:** structure learning, discrete Bayesian networks, extended Bayesian information criterion, normalized maximum likelihood, gene regulation, genetical genomics.

## 1 Introduction

The post-genomic era shifted the main biological focus from *single-gene* to *genome-wide* approaches. High throughput data available from new technologies allowed to get inside main features of gene expression and its regulation and, as a result, to discover the existence of nonrandom and well-defined structures that determine a network of complex (highly non linear) interactions called *Gene Regulatory Network* (GRN). The ultimate goal is to find direct causal influences between gene-activities. Knowing the complete GRN of the cell or organism allows to better understand and control complex biological phenomenon as demonstrated e.g., obesity and bone traits in mouse [20], and flowering time in *Arabidopsis thaliana* [16].

In practice, DNA arrays simultaneously measure the level of mRNAs product in a (set of) living cell(s). We assume steady-state data. These mRNA concentrations define the *gene-activities* or gene expression levels.

In addition to gene-activities data, we want to exploit DNA marker genotype data. The so-called *genetical genomics* approach entails an analysis of high throughput mRNA expression data in a pedigree of genetically distinct individuals [15] [7](Chap. 4). Therefore, our observed data for structure inference will be both continuous (i.e., gene-activities) and discrete (i.e., genotypes).

Among many frameworks used to infer GRN we choose graphical probabilistic model and more specific *static* Bayesian Networks (BN) [6]. Learning BN structure is still an open problem and several approaches try to solve it. One of them consists in exploring the space of BN structures and evaluates each structure with a specific scoring criterion in order to select which structure maximizes the score. It is this score-based approach we use in the following study, we compare several scores in an extended form to favour specific structures.

In Section 2 we present existing approaches to learn GRN. Next we explore in Section 3 existing scoring criteria, including a combination of two very recent ones, resulting in a new criterion which is more robust than classical one to small sample size [26, 27] and which allows to add *a priori* information on the selected model [3]. Then, in Section 4, we present our simulated data in the field of genetical genomics. Finally, we report our experimental results comparing network inference methods in Section 5 and conclude.

## 2 State-of-the-art

Existing approaches for network inference can be classified in three categories [1]. The first category measure (partial) correlations for all pairs of gene-activities and keep in the final graph gene pairs which correlate above a user-defined significance threshold. Due to symmetry measure this methods cannot determine the direction of edges, resulting in undirected GRN, also called *Co-Expression Networks* (CEN). CEN inference software ParCorA [9] uses Pearson and Spearman (rank) correlations. ARACNE [19] and CLR [10] are based on Mutual Information instead.

The second type of methods relies on linear models. The general framework of *Gaussian Graphical Model* (GGM) assumes the data are drawn from a multivariate normal distribution. Among the existing approaches based on GGM are GeneNet [23], Simone [5] (it assumes a modular network with constant connectivity inside each module), and GGMSelect [12] which first builds a family of candidate graphs from the data and then selects one graph among this family according to a dedicated criterion. The difficulty of learning a global structure has conducted to a simpler gene-by-gene regression, by regressing with the Lasso each gene-activity against the others [22], in the so-called *Structural Equation Modeling* (SEM) framework (similar to deterministic ordinary differential equations in the case of steady-state data without noise). This approach was tested in the context of genetical genomics by [17] where structure inference is based on a penalized likelihood ratio and an adaptation of *Occam's window model selection* [18].

The third class of methods consists in probabilistic discrete graphical models. According to the steady-state data, we focus our study on *static* Bayesian networks, composed of a *directed acyclic graph* (DAG) with a conditional probability (possibly modeling a highly non-linear gene interaction) associated to each node in the graph. On the contrary to linear models, Bayesian networks can represent any complex gene interaction, except cyclic gene interactions. This framework was first applied to gene expression data in [11]. It was further tested on genetical genomics data by [28] with the genetic information used as prior information. Our goal is to test the Bayesian net-

work approach on genetical genomics data by modeling the genetic information as extra variables in the DAG.

As all of these methods cannot (or only partially) distinguish the direction of interactions, we study and evaluate our methods for undirected GRNs in this paper.

### 3 Extended Bayesian scores

A *Bayesian network* [6] denoted by  $B = (\mathcal{G}, \mathbf{P}_{\mathcal{G}})$  is composed of a directed acyclic graph  $\mathcal{G} = (\mathbf{X}, \mathbf{E})$  with nodes representing  $p$  random discrete variables  $\mathbf{X} = \{X_1, \dots, X_p\}$ , linked by a set of directed edges  $\mathbf{E}$ , and a set of conditional probability distributions  $\mathbf{P}_{\mathcal{G}} = \{P_1, \dots, P_p\}$  defined by the topology of the graph:  $P_i = \mathbb{P}(X_i | Pa(X_i))$  where  $Pa(X_i) = \{X_j \in \mathbf{X} | \overrightarrow{(X_j, X_i)} \in \mathbf{E}\}$  is the set of parent nodes of  $X_i$  in  $\mathcal{G}$ . Let  $a = \max_{i=1}^p |Pa(X_i)|$  be the maximum number of parents for a node. A Bayesian network  $B$  represents a joint probability distribution on  $\mathbf{X}$  such that:

$$\mathbb{P}(\mathbf{X}) = \prod_{i=1}^p \mathbb{P}(X_i | Pa(X_i)) \quad (1)$$

The conditional probability distributions  $\mathbf{P}_{\mathcal{G}}$  are determined by a set of parameters,  $\theta$ , via the equation:

$$\mathbb{P}(X_i = k | Pa(X_i) = j) = \theta_{ijk}$$

where  $k$  is a value of  $X_i$ , and  $j$  is a value configuration of the parent set  $Pa(X_i)$ . The number of independent parameters in  $\theta$  is called the *dimension* of  $B$  and it is noted  $Dim(B) = \sum_{i=1}^p Dim(P_i)$  with  $Dim(P_i) = (r_i - 1)q_i$  where  $r_i$  is the domain size of variable  $X_i$  and  $q_i = \prod_{X_j \in Pa(X_i)} r_j$ , is the product of the parental domains of  $X_i$ .

Learning the structure of a Bayesian network consists in finding a DAG  $\mathcal{G}$  maximizing  $\mathbb{P}(\mathcal{G} | \mathbf{D})$  where  $\mathbf{D}$  represents the observed data. We have:

$$\mathbb{P}(\mathcal{G} | \mathbf{D}) = \frac{\mathbb{P}(\mathbf{D} | \mathcal{G}) \mathbb{P}(\mathcal{G})}{\mathbb{P}(\mathbf{D})} \propto \mathbb{P}(\mathbf{D} | \mathcal{G}) \mathbb{P}(\mathcal{G}) \quad (2)$$

The first term  $\mathbb{P}(\mathbf{D} | \mathcal{G})$  of Equation 2, called the *marginal likelihood*, can be rewritten by integrating parameters out:

$$\mathbb{P}(\mathbf{D} | \mathcal{G}) = \int_{\theta} \mathbb{P}(\mathbf{D} | \mathcal{G}, \theta) \mathbb{P}(\theta | \mathcal{G}) d\theta \quad (3)$$

Assuming a uniform prior on the parameters, this integral can be simplified as shown in [4] using Laplace approximation (for large sample size  $n$ ) resulting in the *Bayesian Information Criterion* (BIC) [24] which combines a maximum likelihood term with a penalty term on the dimension of the model, as expressed below:

$$BIC(\mathcal{G}) = \log(\mathbb{P}(\mathbf{D} | \mathcal{G}, \theta^{ML})) - \frac{1}{2} \log(n) Dim(B_{\mathcal{G}}) \approx \log(\mathbb{P}(\mathbf{D} | \mathcal{G}))$$

where  $B_{\mathcal{G}} = (\mathcal{G}, \theta^{ML})$  is the Bayesian network defined by the proposed structure  $\mathcal{G}$  with parameters  $\theta^{ML}$  estimated by following the maximum likelihood principle.

#### IV

BIC is derived through asymptotics and its behavior is suboptimal for small sample sizes. On the contrary, the popular Bayesian Dirichlet criterion (BDeu<sup>1</sup>) assumes that the parameter vectors  $\theta_{ij}$  are independent of each other and distributed according to Dirichlet distributions with hyper-parameters  $\alpha_{ijk} = \frac{\alpha}{r_i q_i}$ , resulting in the following score:

$$BDeu(\mathcal{G}) = \mathbb{P}(\mathbf{D}|\mathcal{G}) = \prod_{i=1}^p \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(n_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})}$$

with  $n_{ijk}$ , the number of occurrences of the configuration ( $X_i = k, Pa(X_i) = j$ ) in the  $n$  samples, and  $n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$ .

Unfortunately, the posterior probability distribution of the network structures is very sensitive to the choice of the so-called *equivalent sample size* parameter  $\alpha$  [25].

A recent scoring criterion based on the *normalized maximum likelihood* (NML) distribution has been proposed in [26, 27] to avoid this sensitivity problem,

$$NML(\mathcal{G}) = \mathbb{P}(\mathbf{D}|\mathcal{G}) = \frac{\mathbb{P}(\mathbf{D}|\mathcal{G}, \theta^{ML})}{\sum_{\mathbf{D}'} \mathbb{P}(\mathbf{D}'|\mathcal{G}, \theta^{ML})}$$

where the normalization is over all data sets  $\mathbf{D}'$  of the same sample size  $n$ . Although the number of data sets is exponential, this score can be locally computed for each variable  $X_i \in \mathbf{X}$  in linear time as described in [26, 27], assuming a  $r_i$ -ary multinomial variable. Multiplying this local score over all the variables in  $\mathbf{X}$  defines a global score called the *factorized normalized maximum likelihood* (fNML). In this sense fNML can be seen as an alternative way to define the marginal likelihood of Equation 3. Asymptotically fNML behaves like BIC [27].

Instead of choosing a uniform prior on the space of DAGs in Equation 2, we propose to choose a uniform prior on the number of parents (i.e. node in-degree) for every node. The motivation behind this is that the number of possible DAGs grows exponentially with the number of parents (up to  $p/2$  parents), resulting in the selection of dense networks which do not correspond to the current knowledge of existing real gene networks which are sparse graphs [2]. As done in [3], we define a sparsity prior local to each node in  $\mathcal{G}$ :

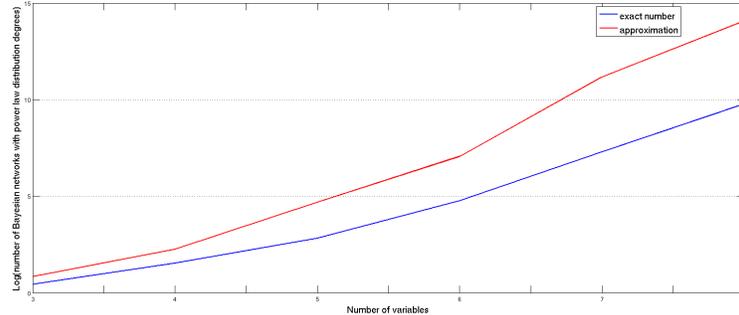
$$\mathbb{P}(\mathcal{G}_i) \propto \tau(X_i)^{-\gamma}$$

with  $\gamma \in [0, 1]$ ,  $\mathcal{G}_i$  is the restricted graph to  $X_i$  composed of variables  $\mathbf{X} = \{X_i \cup Pa(X_i)\}$  linked by  $\mathbf{E} = \{(\overrightarrow{X_j}, \overrightarrow{X_i}) | X_j \in Pa(X_i)\}$  and  $\tau(X_i) = C_{a_i}^{p-1}$  is the number of different parent sets of size  $a_i = |Pa(X_i)|$  among  $p-1$  potential parents. The value  $\gamma = 0$  corresponds to a uniform prior on the space of DAGs. The value  $\gamma = 1$  corresponds to a uniform node in-degree prior. In order to get a global prior, we assume restricted graph independence,  $\mathbb{P}(\mathcal{G}) \approx \prod_{i=1}^p \mathbb{P}(\mathcal{G}_i)$ .

Clearly, our approximation overestimates the number of DAGs having a fixed sequence of in-degrees  $a_i, \forall i \in \{1, \dots, p\}$ . E.g., there exists no DAG with only two nodes having one parent each (it would be cyclic), whereas it has a prior probability proportional to one in our approach. Figure 1 compares the true number of DAGs with

<sup>1</sup> Symbol *eu* stands for *Equivalent Uniform*, i.e. this criterion gives the same score for Markov equivalent Bayesian networks and assumes a uniform prior on the  $\theta$  parameters.

a varying number of nodes and a fixed sequence of in-degrees following a power law distribution (such that there are a total of  $p$  edges in average) with our approximation. Using  $\gamma < 1$  (e.g.,  $\gamma = 0.7$ ) would probably be more correct. It remains an open question if the true number can be expressed analytically.



**Fig. 1.** Comparison on true (blue) and approximate (red) logarithm number of DAGs with a fixed sequence of node in-degrees following a power law distribution, varying the number of nodes in the x axis.

In the sequel, we note  $BIC_\gamma$ ,  $BDeu_\gamma$ , and  $fNML_\gamma$  the previous scoring criteria extended by the  $\gamma$ -based sparsity prior defined above.

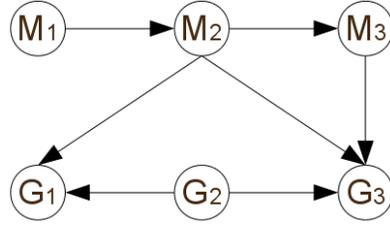
## 4 Simulated genetical genomic data

### 4.1 Bayesian network model description and specific biological information

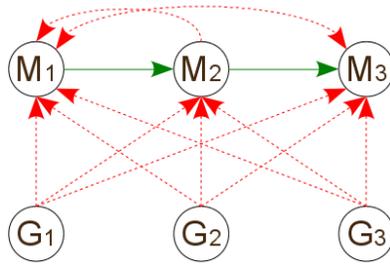
The set of discrete random variables  $\mathbf{X}$  is composed of one variable per gene-activity, denoted  $G_i$ , and one variable for each genetic marker, denoted  $M_i$ ,  $\forall i \in \{1, \dots, p\}$  with  $p$  the number of genes. We associate for each gene a single genetic marker as we see in figure 2.

Although we have  $2p$  variables in our model, we can reduce the search space thanks to biological knowledges. First ones due to genetic linkage between markers on the same chromosome and biological facts are presented in figure 3. Seconds depend on specific biological information about the genetic marker positions inside their associated gene since they are in our case used to represent *effective mutations* on the DNA, i.e., mutations which have an impact on gene expressions as we present in figure 4.

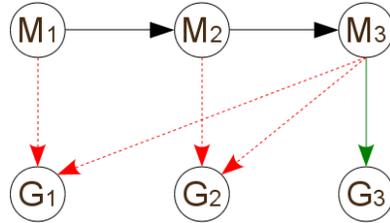
These restrictions on edges are simply applicable on BN but we also apply them for the SEM Lasso approach [22]. In order to ban some edges, we remove corresponding regressors (marker or gene variable) from the potential regressor list and in order to force edges, we make a first linear regression with the known regulators then we apply the standard method on the estimated residuals of this regression (and we also remove the known regressors from the list).



**Fig. 2.** Example of 3 genes network with regulation from  $G_2$  to  $G_1$  and  $G_3$ . We assume genetic linkage between marker ( $M_i \rightarrow M_{i+1}$ ) and consider two distinct marker positions. First we consider a marker located in the gene promoter region (e.g.  $M_3$ ), the mutation observed for this marker will only modify the expression level of the gene so we have  $M_3 \rightarrow G_3$ . The other situation is a mutation occurring in a coding region (*exon*) (e.g.  $M_1$  &  $M_2$ ) which modifies the regulation activity of the gene but not its proper expression level. In this case for each  $G_i \rightarrow G_j$  with  $i \in \{1, 2\}$  we have  $M_i \rightarrow G_j$ .



**Fig. 3.** If we assume that we know marker positions, we can depict them as a first-order Markov chain following their position on the chromosome by fixing (in green)  $M_i \rightarrow M_{i+1}$  and banning (in red) others edges between markers  $M_i \rightarrow M_j \forall j \neq i+1$ . We also assume there are no regulations from any gene-activity towards any genetic marker (which would have no biological meaning)  $G_i \rightarrow M_j \forall i, j$ .



**Fig. 4.** For each marker  $M_i$  in a promoter region ( $M_3$  here) we can fix in our model  $M_i \rightarrow G_i$  and forbidding any other edges  $M_i \rightarrow G_j, \forall j \neq i$ . In the case of marker  $M_i$  is in a promoter region ( $M_1$  and  $M_2$ ) we only ban  $M_i \rightarrow G_i$ .

## 4.2 Data simulator

The lack of data combining gene expressions and genotypes of segregating individuals leads us to generate artificial datasets with a complete control on the parameters (level of noise, interaction intensity, network topology,...). Gene networks are commonly assumed to be scale free [2], for this reason we used artificial networks described in [21] with such a structure. This kind of network is characterized by a power law on node connectivity, which means that a small number of nodes are highly connected to others (called hubs) while the vast majority have only few connections.

To simulate genes behavior we used non linear ordinary differential equations for each gene as described in [17]. The following equation defines the mRNA concentration

of gene  $i$  by simulating both gene and marker interactions in a realistic way. We have:

$$\frac{dG_i}{dt} = V_i \prod_{G_j \in \text{Inh}(G_i)} Z_j \left( \frac{1}{1 + G_j} \right) \prod_{G_k \in \text{Act}(G_i)} Z_k \left( 1 + \frac{G_k}{G_k + 1} \right) - G_i + \theta_i G_i$$

with  $V_i$  defines gene  $i$  activity rate, affected if the mutation is in promoter region ( $V_i = 0.75$  or  $1$  depending on mutation presence or absence),  $\text{Inh}(G_i)$  (resp.  $\text{Act}(G_i)$ ) are inhibitors (resp. activators) concentrations,  $Z_j$  (resp.  $Z_k$ ), inhibition (resp. activation) activity rate of regulators  $j$  (resp.  $k$ ) affected if the associated mutations are in exons ( $Z_{j/k} = 0.75$  or  $1$  depending on mutation presence or absence) and  $\theta_i$ , a gaussian noise with mean  $0$  and variance  $0.1$ .

We used the COPASI software [14] to define the steady state measurement of each gene expression level. We performed the steady state measurements for each individual of a RIL population (RIL generator from [8]) to get samples with large genetic diversity and so varying gene-activities between individuals.

In order to test our approach we used 50-node networks and generated for each of them a population of 500 individuals. We obtained with COPASI continuous gene-activities and discrete genotypes (boolean value indicating if there is a mutation or not) at every genetic marker. Because of discrete BN, we have to discretize gene-activities.

We have chosen an adaptive method depending on the expression level distribution of each gene. As we could observed, our datasets have complex distributions, so we have to distinguish 2 cases. If we detect a bimodal distribution, we use an adapted *k-means* algorithm to get a three-class discretization which ensures a minimum class size (5% sample size) and a maximum size for extremum classes (30%). In the case of multimodal distribution, we use the more general case of gaussian mixture model search to find a maximum of four classes.

## 5 Experimental results

In order to validate different aspects of this study we decompose our results in three distinct parts. First, we test the impact of extended criteria in the score-based structure learning algorithm *Greedy Search* (GS) [6]. Next, we consider taking into account specific biological knowledge as described in Section 4.1, and finally we perform a vast comparison with existing methods for inferring gene regulatory networks from expression and genotype data.

Before evaluating the network we make a post processing operation as we are only interested in gene regulation, we reduce the  $2p$  nodes of our learnt BN to a simple CEN network, by projecting all  $M_i \rightarrow G_j$  as a  $G_i \rightarrow G_j \forall j \neq i$ , and then keeping only  $G_i$  nodes, removing also edge orientations.

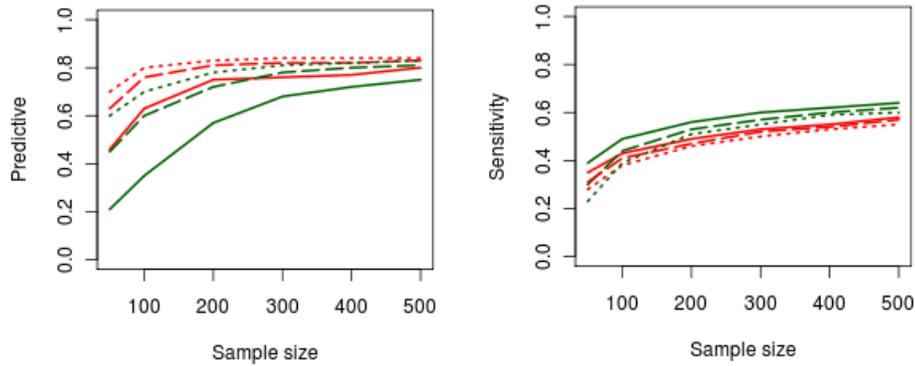
To evaluate each graph we use sensitivity (or recall)  $\frac{TP}{(TP+FN)}$  and positive predictive value (or accuracy)  $\frac{TP}{(TP+FP)}$  metrics [1] where TP (true positive) is the number of correct edges in the learnt network, FN (false negative) for missing edges absent from the learnt network and present in the true network, and FP (false positive) is the number of edges present in the learnt network and absent from the true network. In this study we do not consider edge orientations.

All results presented in this paper are the mean over 50 artificial networks.

### 5.1 Evaluation of extended Bayesian criteria

In figure 5 we observe a similar behavior for BIC and fNML scores. For a small sample size, increasing  $\gamma$  allows to improve significantly prediction of the learnt network with a moderate loss in sensitivity. This is coherent with the fact that higher  $\gamma$  promote low connectivity of the network especially in the case of small sample size. Instead of learning poor confident edges which mostly represent spurious edges, extended score will not select them increasing in this way precision. However sensitivity decreases, showing a loss of a few true regulations. Even if the trade-off between precision/sensitivity is clearly better for  $\gamma = 0.5$  compared to  $\gamma = 0$ , it decreases between  $\gamma = 0.5$  and  $\gamma = 1$  which leads to imagine an optimal  $\gamma$  value. Finally with more samples, all criteria converge due to asymptotically equivalence of these scores.

For clarity of graphics, we do not present the results for BDeu score which gave performance similar to BIC except in the case of 50 sample size where BDeu performed worse.



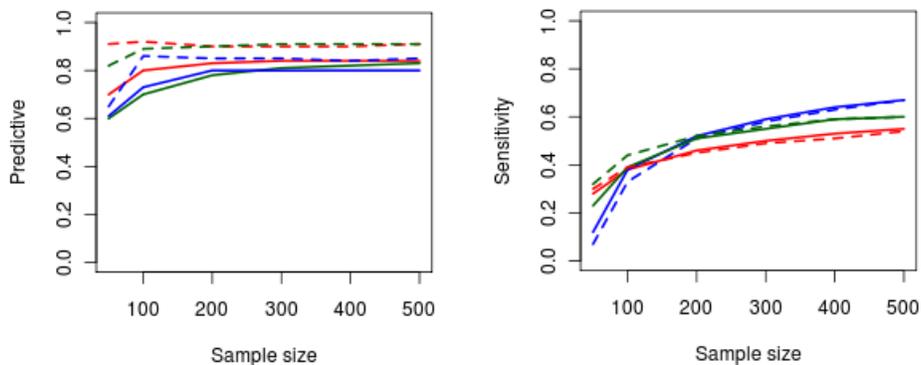
**Fig. 5. Comparison of two extended Bayesian criteria.** The figures show for extended scores fNML (green) and BIC (red), the evolution of the predictive value (to the left) and the sensitivity (to the right) of learnt networks with varying sample sizes. For each extended score, we have three different  $\gamma$  values:  $\gamma = 0$  (solid line),  $\gamma = 0.5$  (dash line) and  $\gamma = 1$  (point line).

### 5.2 Evaluation of adding specific biological information

By fixing/forbidding edges as presented in Section 4.1, we can restrict the space of DAGs to explore. In figure 6 we show the consequence of this approach. It is not surprising to see an improvement of the predictive metric for all methods since we give correct information about the true network even if this information do not give explicitly any regulation between genes. This phenomenon is more important for small sample size except for the SEM Lasso approach with 50 individuals caused by the poor performance of this method in this configuration. However we have also a residual loss of true interactions for BIC and Lasso methods like when we apply extended score. For BIC

criterion this loss is very limited (or null half of the time) and more visible for Lasso but without a clear explanation of this behavior. Only the  $fNML$  score do not respect this evolution, both metrics are improved with edge information which makes this score interesting for further investigation about integrating additional expert information.

In the same manner  $BDeu$  and  $BIC$  got similar results.



**Fig. 6. Impact of adding specific biological information.** The figures show for three approaches  $BIC_1$  (red),  $fNML_1$  (green), and SEM Lasso (blue), the evolution of the predictive value (to the left) and the sensitivity (to the right) of learnt networks with varying sample sizes. For each criterion, two curves represent the results without the extra information about edges (solid line) and with this information (dash line).

### 5.3 Global evaluation with existing network inference methods

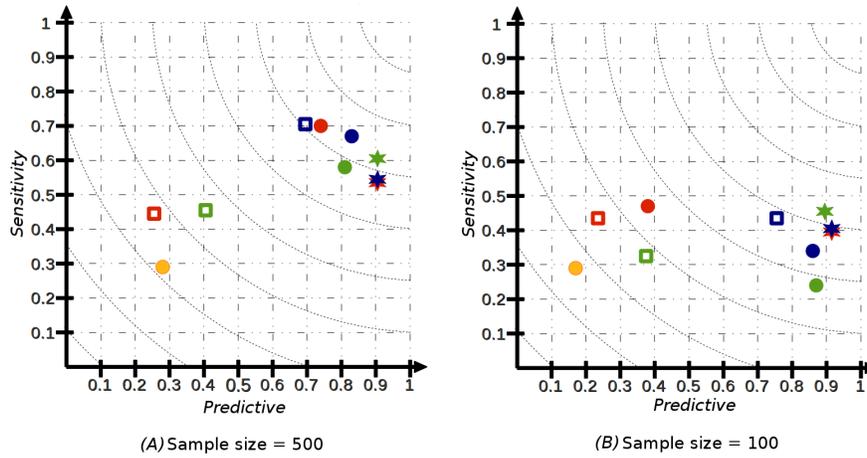
In this part we try to make a large comparison with different approaches proposed in the literature to infer gene networks. Methods used are summarized in figure 7 and are freely available on the web or on author's demand. We also give for each method the parameter settings fixed by ourself whereas other parameters are left to their default value. Based on previous comparison we used  $BIC_1$ ,  $BDeu_1$ ,  $fNML_1$  and SEM Lasso with including specific biological information (i.e., restriction on edges as presented in Section 4.1).

We see in figure 8 a great performance disparity of the methods, we distinguish in each case two groups the poor and the (relative) good results. There is also no outperforming method which means that all approaches have the same difficulties with a poor sensitivity in small sample size case. Hubs are the most problematic structure in network for all methods, they are link to a majority of edges and are known to be difficult to learn since some regulation dominate their behavior and hide others.

ARACNE and CLR have poor results in both situations whereas ParCorA perform well, computation of partial correlation give to it a serious advantage and seems to prove that Spearman correlation (same result with Pearson correlation) is more competitive in our simulation.

Software	Description	Parameters	Ref
Banjo(v2.2)	Bayesian network inference	$BDeu_{\alpha} = 1$	[13]
ARACNE	Mutual Information correlation	$threshold = 0.15$	[19]
CLR(v1.2)	Mutual Information correlation	$threshold = 4$	[10]
ParCorA	Spearman correlation	$pcutoff = 0.001; lborder$	[9]
Simone(v1.0)	Gaussian graphical model	$modularnumber = 2$	[5]
GGMselect(v0.1)	Gaussian graphical model	"C01" family	[12]
GeneNet(v1.2.4)	Gaussian graphical model	$cutoff = 0.95$	[23]
SEM Lasso	Lasso regression	$\alpha_{meinshausen} = 0.1$	[22]

Fig. 7. Features of tested methods



**Fig. 8. Predictive value (horizontal axe) and sensitivity (vertical axe) for sample size=500 (to the left) and 100 (to the right).** Methods are classified in 3 categories: **1-Co-expression network** ARACNE (red square), CLR (green square), ParCorA (blue square); **2-Linear model** Simone (yellow circle), GeneNet (red circle), GGMSelect (green circle), SEM Lasso (blue circle); **3-Bayesian network** Banjo with  $BIC_1$  (red star),  $BDeu_1$  (blue star), and  $fNML_1$  (green star).

Simone have the worst performance of the comparative study, maybe due to the modular network expected (we tested with several modular number 1 to 6 without significant improvements) which not correspond to the real network. GeneNet appear to be very sensitive to sample size with a huge performance decrease when we reduce the sample size and finally does not seem to be able to manage real dataset where number of sample is small compare to gene number.

SEM Lasso and GGMselect have a same behavior in both configurations with best results for SEM Lasso due to edges assumptions, these methods are robust concerning their predictive value but less concern their sensitivity with more than 50% decrease in the 100 individuals case.

BN with each of the 3 scores is the most robust method,  $BIC_1$  and  $BDeu_1$  performed similarly whereas  $fNML_1$  outperformed the both. This last score is the better choice

for 100 individuals case (the more realistic of our test) and is not too far from the best for 500 individuals. We have to note the general good performance of ParCorA since without any a priori, it performed very well even if it is less suitable to take in account expert knowledge as BN can.

## 6 Conclusion

In conclusion, we have studied different scoring criteria in the framework of discrete Bayesian networks, including the combination of two recent ones [3, 27],  $fNML_1$ , which obtained the best results on our simulated genetical genomics dataset for medium sample size ( $n = 100$  samples with  $p = 50$  genes and 50 markers) compared to a large set of existing gene regulatory network inference methods. Adding a sparsity prior greatly improves the accuracy of our methods. It should be noted that except for our approach and the SEM Lasso, other approaches used genotype information on their own, without the additional knowledge of promoter/coding region mutations which may be difficult to acquire in practice.

Still, it remains to validate our approach on real data and to study causality in GRNs.

## References

1. M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3(78), 2007.
2. A. Barabasi and Z. Oltvai. Network biology: Understanding the cell's functional organization. *NATURE REVIEWS GENETICS*, 5(2):101–115, FEB 2004.
3. J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–711, 2008.
4. D. Chickering and D. Heckermann. Learning bayesian networks is np-complete. *In learning from data: AI and Statistics*, 1996.
5. J. Chiquet, A. Smith, G. Grasseau, C. Matias, and C. Ambroise. Simone: Statistical inference for modular networks. *Bioinformatics*, 25(3):417–418, 2009.
6. A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.
7. S. Das, D. Caragea, W. H. Hsu, and S. M. Welch, editors. *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*. IGI Global, Hershey, New York, 2010.
8. S. de Givry, M. Bouchez, P. Chabrier, D. Milan, and T. Schiex. Carthagene: multipopulation integrated genetic and radiated hybrid mapping. *Bioinformatics*, 21(8):1703–1704, 2005.
9. A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.
10. J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5, 2007.
11. N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyse expression data. *Journal of computational biology*, 7(3/4):601–620, 2000.
12. C. Giraud, S. Huet, and N. Verzelen. Graph selection with ggmselect. Technical report, Ecole Polytechnique, 2009.

13. A. Hartemink. Reverse engineering gene regulatory networks. *Nature Biotechnology*, 23:554–555, 2005.
14. S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. Copasi—a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.
15. R. Jansen and J. Nap. Genetical genomics : the added value from segregation. *Trends in genetics*, 17(7):388–391, July 2001.
16. J. J. B. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, L. B. Snoek, A. J. M. Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen. Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences*, 104(5):1708–1713, 2007.
17. B. Liu, A. de la Fuente, and I. Hoeschele. Gene network inference via structural equation modeling in genetical genomics experiments. *GENETICS*, 178(3):1763–1776, 2008.
18. D. Madigan and A. Raftery. Model selection and accounting for model uncertainty in graphical model using occam’s window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.
19. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1), 2006.
20. M. Mehrabian, H. Allayee, J. Stockton, P. Y. Lum, T. A. Drake, L. W. Castellani, M. Suh, C. Armour, S. Edwards, J. Lamb, A. J. Lusis, and E. E. Schadt. Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat Genet*, 37(11):1224–1233, 2005.
21. P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19:ii122–ii129, 2003.
22. M. Nicolai and B. Peter. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.
23. J. Schäfer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.
24. G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 1978.
25. T. Silander, P. Kontkanen, and P. Myllymäki. On sensitivity of the map bayesian network structure to the equivalent sample size parameter. In *Proc. of UAI-07*, pages 360–367, Vancouver, Canada, 2007.
26. T. Silander, T. Roos, P. Kontkanen, and P. Myllymäki. Factorized normalized maximum likelihood criterion for learning bayesian network structures. In *4th European Workshop on Probabilistic Graphical Models*, Hirtshals, Denmark, 2008.
27. T. Silander, T. Roos, and P. Myllymäki. Learning locally minimax optimal bayesian networks. *International Journal of Approximate Reasoning*, 51(5):544 – 557, 2010.
28. J. Zhu, M. C. Wiener, C. Zhang, A. Fridman, E. Minch, P. Y. Lum, J. R. Sachs, and E. E. Schadt. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS COMPUTATIONAL BIOLOGY*, 3(4):692–703, 2007.