Chapter 5
Gene Regulatory Network Inference from Systems Genetics Data Using Tree-Based Methods

Vân Anh Huynh-Thu, Louis Wehenkel and Pierre Geurts

Abstract One of the pressing open problems of computational systems biology is the elucidation of the topology of gene regulatory networks (GRNs). In an attempt to solve this problem, the idea of systems genetics is to exploit the natural variations that exist between the DNA sequences of related individuals and that can represent the randomized and multifactorial perturbations necessary to recover GRNs. In this chapter, we present new methods, called GENIE3-SG-joint and GENIE3-SG-sep, for the inference of GRNs from systems genetics data. Experiments on the artificial data of the StatSeq benchmark and of the DREAM5 *Systems Genetics* challenge show that exploiting jointly expression and genetic data is very helpful for recovering GRNs, and one of our methods outperforms by a large extent the official best performing method of the DREAM5 challenge.

5.1 Introduction

Networks are commonly used in biological research to represent information. In this chapter, we focus on GRNs. These networks represent regulatory interactions among genes that happen at the level of transcription, through transcription factors. They often offer a simplified view of gene regulation, and are usually represented by graphs where each node corresponds to a gene and an edge directed from one gene to another gene indicates that the first gene codes for a transcription factor that regulates the rate of transcription of the second gene.

V. A. Huynh-Thu · L. Wehenkel · P. Geurts (✉)
Department of EE and CS and GIGA-R , University of Liège, Liège, Belgium
e-mail: P.Geurts@ulg.ac.be

V. A. Huynh-Thu
e-mail: vahuynh@ulg.ac.be

L. Wehenkel
e-mail: L.Wehenkel@ulg.ac.be

A. de la Fuente (ed.), *Gene Network Inference*,
DOI: 10.1007/978-3-642-45161-4_5,

Edges in regulatory networks can be directed or undirected. An undirected edge connecting two genes indicates that there exists a transcriptional regulatory interaction between these two genes, while a directed edge means furthermore that the source gene regulates the expression of the target gene. Edges can also be signed. When a gene is connected to another gene, a positive (resp. negative) sign indicates that the former is an activator (resp. repressor) of the latter. In this chapter, we focus on directed unsigned networks. The targeted networks are thus graphs with p nodes, where an edge directed from one gene i to another gene j indicates that gene i (directly) regulates, either positively or negatively, the expression of gene j $(i, j = 1, \ldots, p)$.

The problem of the inference of regulatory networks has been studied for many years in the literature and many algorithms already exist. The authors De Smet and Marchal (2010) proposed a categorization of these methods. First, they distinguish supervised from unsupervised methods. Supervised methods exploit prior partial knowledge of the network to guide the network inference, while unsupervised methods do not assume any prior knowledge. There are also direct methods, which consider only individual interactions, and module-based methods, which search for sets of genes that are regulated by the same transcription factors. Finally, non-integrative methods only use expression data for the inference, while integrative methods also use other kinds of information besides expression data, e.g. counts of sequence motifs that serve as binding sites for transcription factors.

Among integrative methods, one can also find methods exploiting systems genetics data. The goal of systems genetics is to exploit the natural variations that exist between the DNA sequences of related individuals in a segregating population and that can represent the randomized and multifactorial perturbations necessary to recover (GRNs) (Jansen and Nap 2001; Jansen 2003). In such a study, two strains that are widely separated in terms of genetic background are crossed and their children are self-crossed during several generations in order to produce a recombinant inbred line (RIL) segregating population. The genomes of the individuals of this population comprise random segments of the genomes of the two original parents and genetic differences can therefore be detected between them, representing multifactorial genetic perturbations. Each individual is then analyzed by microarray expression profiling as well as by genetic marker analysis.

Multiple methods have been developed to infer GRNs from systems genetics data. Several methods infer causal regulatory relationships among pairs of genes, including procedures that rely on statistical tests to identify causal links (Chen et al., 2007) and approaches based on the fitting of causal models (Kulp and Jagalur 2006; Schadt et al. 2005). Other methods are based on the analysis of the correlation between expression profiles of genes located in a particular genomic region and expression profiles of genes that are potentially affected by the markers located in this region (Bing and Hoeschele, 2005). Methods that study the regulatory relationships at a systems-level include approaches based on Bayesian networks (Zhu et al. 2007; Li et al. 2005; Vignes et al. 2011), structural equation models (Li et al. 2006; Liu et al. 2008), and the orientation of the edges of an undirected network using genetic markers as causal anchors (Aten et al. 2008; Chaibub Neto et al. 2008). Random Forests have also been

successfully used for expression quantitative trait loci (eQTL) mapping (Michaelson et al., 2010).

In this chapter, we propose new methods, based on ensembles of regression trees, for the inference of regulatory networks from systems genetics data. According to the categories of De Smet and Marchal (2010), these methods are direct (we do not search for modules) and unsupervised (we do not assume any prior knowledge of the network).

The chapter is structured as follows. Section 5.2 describes our network inference methods. Section 5.3 shows the results obtained with these methods on the StatSeq compendium as well as on the datasets of the DREAM5 *Systems Genetics* challenge. Finally, Sect. 5.4 concludes the chapter and discusses some ideas for further developments.

5.2 Methods

We assume that we have at our disposal a dataset containing the steady-state expression levels of p genes measured in N individuals, as well as the genotype value of one genetic marker for each of these genes in the same N individuals:

$$LS = \{(\mathbf{e}_1, \mathbf{m}_1), (\mathbf{e}_2, \mathbf{m}_2), \ldots, (\mathbf{e}_N, \mathbf{m}_N)\}, \tag{5.1}$$

where $\mathbf{e_k} \in \mathbb{R}^{\mathbf{p}}$ and $\mathbf{m_k} \in \{\mathbf{0}, \mathbf{1}\}^{\mathbf{p}}, \mathbf{k} = \mathbf{1}, \ldots, \mathbf{N}$ are, respectively, the vectors of expression levels and genotype values of the p genes in the kth individual:

$$\begin{cases} \mathbf{e}_k = (e_k^1, e_k^2, \ldots, e_k^p)^\top, \\ \mathbf{m}_k = (m_k^1, m_k^2, \ldots, m_k^p)^\top. \end{cases} \tag{5.2}$$

Note that we suppose that the individuals come from a RIL population and are hence homozygous. Each genetic marker can thus have two possible genotype values only.

From this dataset, our goal is to infer a gene regulatory network, i.e., to make a prediction of the underlying regulatory links between genes. Many network inference algorithms work first by providing a ranking of the potential regulatory links from the most to the less significant. A practical network prediction is then obtained by setting a threshold on this ranking. In this chapter, we focus only on the first task and the question of the choice of an optimal confidence threshold, although important, is left open.

A network inference algorithm is thus defined in this chapter as a procedure that assigns weights $w_{i,j} \geq 0 (i, j = 1, \ldots, p)$ to putative regulatory links from any gene i to any gene j, with the aim of yielding larger values for weights that correspond to actual regulatory interactions.

To infer GRNs from systems genetics data, we propose two extensions of a method called GENIE3 (Huynh-Thu et al., 2010) that exploits tree-based ensemble methods for the inference of networks from expression data. As in the GENIE3 procedure,

our two extensions decompose the problem of recovering a network of p genes into p feature selection subproblems, where each of these subproblems consists in identifying the regulators of one of the genes of the network. This idea has also been exploited in other methods, such as MRNET (Meyer et al., 2007), the Graphical Lasso (Meinshausen and Bühlmann, 2006), or the meta-analysis developed by Vignes et al. (2011).

5.2.1 Network Inference as a Feature Selection Problem

To infer GRNs from systems genetics data, the two procedures that we propose make the assumption that the expression of each gene j in a given individual is a function of the expression and genotype values of the other genes of the network in the same individual (plus some random noise). The first procedure, called **GENIE3-SG-joint**, learns a single predictive model from both expression and genetic data, while the second procedure, called **GENIE3-SG-sep**, learns two separate predictive models, one based on the genetic markers and the other based on the expression data. Both methods then compute, for each gene $i \neq j$, two scores $w^e_{i,j}$ and $w^m_{i,j}$, measuring, respectively, the importances of the expression and of the marker of gene i when predicting the expression of gene j. Depending on the method, the computation of $w^e_{i,j}$ and $w^m_{i,j}$ is different. These two scores are then aggregated to obtain a single weight $w_{i,j}$ for the regulatory link directed from gene i to gene j.

We first describe the procedures for training the predictive models and computing the importance scores. We then discuss aggregation techniques, which are common to both approaches, to obtain the final weights.

5.2.1.1 GENIE3-SG-Joint

The GENIE3-SG-joint procedure assumes that a unique model f_j explains the expression of a gene j in a given individual, knowing the expression levels and the genotype values of the different genes of the network:

$$e^j_k = f_j(\mathbf{e}^{-j}_k, \mathbf{m}_k) + \varepsilon_k, \forall k, \tag{5.3}$$

where ε_k is a random noise and $\mathbf{e^{-j}_k}$ is the vector containing the expression levels of all the genes except gene j in the kth individual:

$$\mathbf{e}^{-j}_k = (e^1_k, \ldots, e^{j-1}_k, e^{j+1}_k, \ldots, e^p_k)^\top. \tag{5.4}$$

We further make the assumption that the function f_j only exploits the expression levels in $\mathbf{e^{-j}_k}$ and/or the genotype values in $\mathbf{m_k}$ of the genes that are direct regulators of gene j, i.e., genes that are directly connected to gene j in the targeted network.

Notice that $\mathbf{m_k}$ contains the genotype value of gene j. Indeed, it often happens that a genetic marker contributes to the expression of the gene in which it is located (*cis*-acting polymorphism). Including the marker $\mathbf{m^j}$ of gene j in the input variables thus avoids to wrongly attribute to another regulator the part of the expression of gene j that is actually explained by $\mathbf{m^j}$.

Recovering the regulatory links pointing to gene j thus amounts to finding those genes whose expression and/or genetic marker are predictive of the expression of the target gene. In machine learning terminology, this can be considered as a feature selection problem (in regression) for which many solutions exist (Guyon and Elisseeff 2003; Saeys et al. 2007). We assume here the use of a feature ranking technique that, instead of directly returning a feature subset, yields a ranking of the features from the most relevant to the least relevant for predicting the output.

The GENIE3-SG-joint procedure is illustrated in Fig. 5.1 and works as follows:

- For $j = 1$ to p:
  - Generate the learning sample of input–output pairs for gene j:

$$LS^j = \{((\mathbf{e}_k^{-j}, \mathbf{m}_k), e_k^j), k = 1, \ldots, N\}. \tag{5.5}$$

  - Use a feature ranking technique on LS^j to compute confidence levels $w_{i,j}^e (i \neq j)$ and $w_{i,j}^m$, $i = 1, \ldots, p$, respectively, for the expression and the genetic marker of input gene i.
  - Aggregate $w_{i,j}^e$ and $w_{i,j}^m$ to get a weight $w_{i,j}$ for each gene $i \neq j$ (see Sect. 5.2.1.3).
- Use $w_{i,j}$ as weight for the regulatory link $i \rightarrow j$ and get a ranking of all links.

5.2.1.2 GENIE3-SG-Sep

In the second proposed procedure, GENIE3-SG-sep, we assume that two different models f_j^e and f_j^m can both explain the expression of a gene j in a given individual, either from the expression levels of the other genes, or from the genotype values:

$$\begin{cases} e_k^j = f_j^e(\mathbf{e}_k^{-j}) + \varepsilon_k, \forall k, \\ e_k^j = f_j^m(\mathbf{m}_k) + \varepsilon'_k, \forall k. \end{cases} \tag{5.6}$$

The functions f_j^e and f_j^m are therefore, respectively, learned from two different learning samples. The method is illustrated in Fig. 5.2 and works as follows:

- For $j = 1$ to p:
  - Generate two learning samples of input–output pairs for gene j:

$$\begin{aligned} LS_e^j &= \{(\mathbf{e}_k^{-j}, e_k^j), k = 1, \ldots, N\}, \\ LS_m^j &= \{(\mathbf{m}_k, e_k^j), k = 1, \ldots, N\}. \end{aligned} \tag{5.7}$$

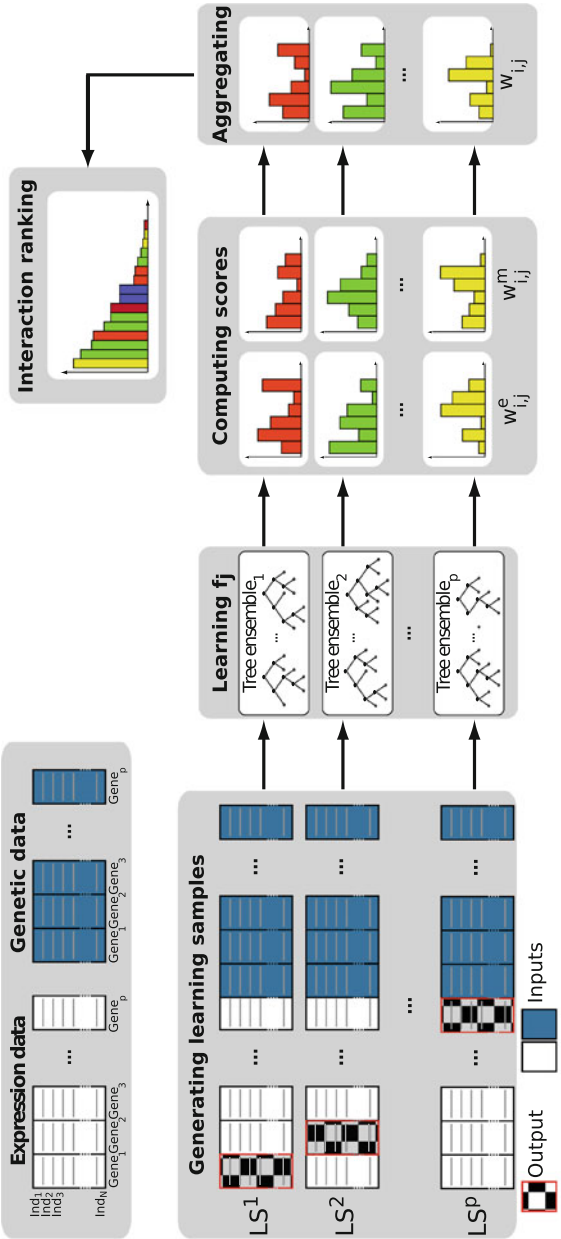

Fig. 5.1 GENIE3-SG-joint procedure. For each gene $j = 1, \ldots, p$, a learning sample LS^j is generated with expression levels of gene j as output values and expression levels and genotypes values of all the other genes as input values. A function f_j is learned from LS^j and confidence levels $w^e_{i,j}$ and $w^m_{i,j}$ are computed for the expression and genotype value of each input gene i respectively. These levels are then aggregated for each input gene and a ranking of all regulatory links is obtained

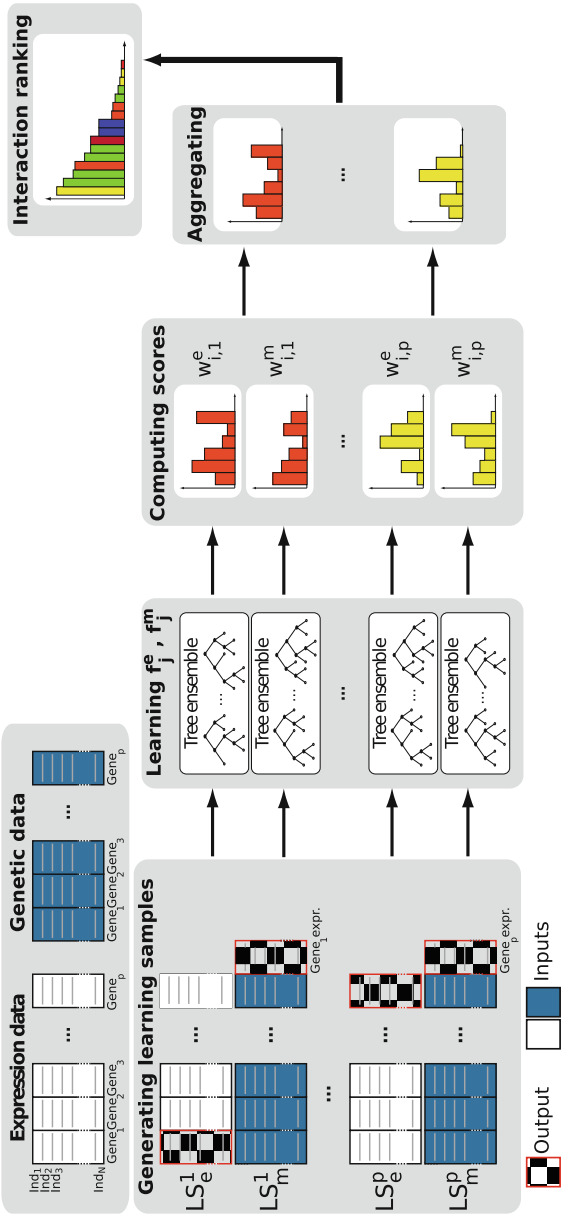

Fig. 5.2 GENIE3-SG-sep procedure. For each gene $j = 1, \ldots, p$, two learning samples LS_e^j and LS_m^j are generated. In both learning samples, the output values are the expression levels of gene j. In LS_e^j the input values are the expression levels of all the other genes, while in LS_m^j the input values are the genotypes values. Functions f_j^e and f_j^m are, respectively, learned from LS_e^j and LS_m^j, and confidence levels $w_{i,j}^e$ and $w_{i,j}^m$ are computed for the expression and genotype value of each input gene i respectively. These levels are then aggregated for each input gene and a ranking of all regulatory links is obtained

  - Use a feature ranking technique on LS_e^j to compute the confidence level $w_{i,j}^e$ of the expression of input gene i, $\forall i \neq j$.
  - Use a feature ranking technique on LS_m^j to compute the confidence level $w_{i,j}^m$ of the genetic marker of input gene i, $\forall i$.
  - Aggregate $w_{i,j}^e$ and $w_{i,j}^m$ to get a weight $w_{i,j}$ for each gene $i \neq j$ (see Sect. 5.2.1.3).
- Use $w_{i,j}$ as weight for the regulatory link $i \rightarrow j$ and get a ranking of all links.

5.2.1.3 Weight Aggregation

In both procedures GENIE3-SG-joint and GENIE3-SG-sep, we obtain for each input gene i, two separate importance scores $w_{i,j}^e$ and $w_{i,j}^m$, corresponding, respectively, to the expression and the marker of gene i. We propose two procedures to aggregate these two scores and hence obtain a ranking of regulatory interactions. In the first procedure, the final weight of the edge directed from gene i to gene j is given by the *sum* of the importance scores:

$$w_{i,j} = w_{i,j}^e + w_{i,j}^m. \tag{5.8}$$

The edge will thus have a high weight if *either* the marker *or* the expression of gene i is predictive of the expression of gene j. In the second aggregation procedure, we consider the *product* of the importance scores:

$$w_{i,j} = w_{i,j}^e \times w_{i,j}^m. \tag{5.9}$$

The edge directed from gene i to gene j will thus have a high weight if the marker *and* the expression of gene i are *both* predictive of the expression of gene j.

5.2.2 Feature Ranking with Tree-Based Methods

As in the original GENIE3 method (Huynh-Thu et al., 2010), GENIE3-SG-joint and GENIE3-SG-sep exploit the embedded feature ranking mechanism of tree-based ensemble methods to compute the weights $w_{i,j}^e$ and $w_{i,j}^m$. These methods are described below.

5.2.2.1 Tree-Based Ensemble Methods

Among supervised learning methods, which allow to learn a predictive model or function f_j from observed data, one can find methods based on regression trees (Breiman et al., 1984). The basic idea of regression trees is to recursively split the learning sample with binary tests each based on one input variable, trying to reduce

as much as possible the variance of the output variable in the resulting subsets of samples. Candidate splits for numerical variables typically compare the input variable values with a threshold that is determined during the tree growing.

Single trees are usually much improved by ensemble methods, which average the predictions of several trees, such as Bagging (Breiman, 1996) or Random Forests (Breiman, 2001). In a Bagging ensemble, each tree is built from a bootstrap sample of the original learning sample. The Random Forests method adds an extra level of randomization compared to the Bagging; at each test node, K attributes are selected at random among all candidate attributes before determining the best split.

5.2.2.2 Variable Importance Measure

One of the most interesting characteristics of tree-based methods is the possibility to compute from a tree a variable importance measure that allows us to rank the input features according to their relevance for predicting the output. In our experiments, we consider a measure that computes, at each test node $\mathcal{N}$, the total reduction of the variance of the output variable due to the split, defined by Breiman et al. (1984):

$$I(\mathcal{N}) = \#S.\mathrm{Var}(S) - \#S_t.\mathrm{Var}(S_t) - \#S_f.\mathrm{Var}(S_f), \tag{5.10}$$

where S denotes the set of samples that reach node $\mathcal{N}$, S_t (resp. S_f) denotes its subset for which the test is true (resp. false), Var(.) is the variance of the output variable in a subset, and # denotes the cardinality of a set of samples. For a single tree, the overall importance w of one variable is then computed by summing the I values of all tree nodes where this variable is used to split. Those attributes that are not selected at all obtain a zero value of their importance, and those that are selected close to the root node of the tree typically obtain high scores. Attribute importance measures can be easily extended to ensembles, simply by averaging importance scores over all trees in the ensemble. The resulting importance measure is then even more reliable because of the variance reduction effect resulting from this averaging (Hastie et al., 2009).

5.2.2.3 Regulatory Link Ranking

In the GENIE3-SG-joint and GENIE3-SG-sep procedures, the different tree-based models that are generated yield importance scores $w^e_{i,j}$ and $w^m_{i,j}$ for each pair of genes (i, j), computed as sums of variance reductions in the form (5.10). The sum of the importance scores of all input features for a tree is usually very close to the initial total variance of the output. In the case of the GENIE3-SG-joint procedure, we thus have for each target gene j:

$$\sum_{i \neq j}^{p} w^e_{i,j} + \sum_{i=1}^{p} w^m_{i,j} \approx N.\mathrm{Var}_j(LS'^j), \tag{5.11}$$

where LS'^j is the learning sample from which the tree was built (i.e., a bootstrap sample of LS^j for the Random Forests and Bagging methods) and $\mathrm{Var}_j(LS'^j)$ is the variance of the target gene j estimated in the corresponding learning sample.

Similarly, for the GENIE3-SG-sep procedure, we have:

$$\begin{cases} \sum_{i \neq j} w^e_{i,j} \approx N.\mathrm{Var}_j(LS'^j_e), \\ \sum_i w^m_{i,j} \approx N.\mathrm{Var}_j(LS'^j_m), \end{cases} \tag{5.12}$$

where LS'^j_e and LS'^j_m are the learning samples generated from the expression and genotype data respectively.

As a consequence, if we trivially use the scores $w^e_{i,j}$ and $w^m_{i,j}$ to order the regulatory links, this is likely to introduce a positive bias for regulatory links directed towards the most highly variable genes. To avoid this bias, we first normalize the expression of the target gene j so that it has a unit variance in the training set (LS^j for GENIE3-SG-joint, LS^j_e and LS^j_m for GENIE3-SG-sep), before applying the tree-based ensemble method:

$$\mathbf{e}^j \leftarrow \frac{\mathbf{e}^j}{\sigma^j}, \ \forall j, \tag{5.13}$$

where $\mathbf{e}^j \in \mathbb{R}^N$ is the vector of expression levels of gene j in all N experiments and σ^j denotes its standard deviation. This normalization indeed implies that the different importance scores inferred from different models predicting the different gene expressions are comparable.

5.2.2.4 Computational Complexity

The computational complexity of the Random Forests and Bagging algorithms is $O(TKN \log N)$, where T is the number of trees, N is the dataset size, and K is the number of randomly selected variables at each node of a tree (in the case of Bagging, K is equal to the number of input variables). The complexities of GENIE3-SG-joint and GENIE3-SG-sep are thus of the order of $O(pTKN \log N)$ since these methods require to build, respectively, one and two ensemble(s) of trees for each of the p genes. The complexities are thus log linear with respect to the number of measurements and, at worst, quadratic with respect to the number of genes (when $K = 2p - 1$ for GENIE3-SG-joint and $K = p$ for GENIE3-SG-sep).

To give an idea of the computing times, with our MATLAB®[1] implementations of the methods, GENIE3-SG-sep and GENIE3-SG-joint take, respectively, about 1 and 3 h to infer a network of 1,000 genes from 300 individuals, when K is fixed to the square root of the number of input variables and 1,000 trees are grown in each ensemble. In the worst-case scenario (5,000 genes, 900 individuals, and K equal to the number of input variables), GENIE3-SG-sep and GENIE3-SG-joint would,

[1] http://www.mathworks.com/.

respectively, take 4 months and more than a year to infer the network on a single computer. To reduce computing times, the two algorithms can be trivially parallelized on a computing grid (with one separate computing process for each gene and/or tree).

5.3 Results

After a presentation of the performance metrics, this section presents the results that we obtained when we applied the proposed procedures to two series of synthetic datasets: the StatSeq datasets (Sect. 5.3.2) and the datasets of the DREAM5 *Systems Genetics* challenge (Sect. 5.3.3).

5.3.1 Performance Metrics

Each of our algorithms provides a ranking of the regulatory links from the most confident to the less confident. To evaluate such a ranking independently of the choice of a specific threshold, we used the precision–recall (PR) curve and the area under this curve (AUPR). The PR curve plots, for different thresholds on the weights of the links, the proportion of true positives among all predictions (precision) versus the percentage of true positives among those to be retrieved (recall). A perfect ranking, i.e., a ranking where all the positives are located at the top of the list, yields an AUPR equal to one, while a random ranking results in an AUPR close to the proportion of positives (i.e., close to zero since the proportion of true links among all possible links in a network is usually very low).

5.3.2 Experiments on the StatSeq Datasets

5.3.2.1 Description of the Data

The StatSeq compendium[2] comprises 72 datasets generated from nine different networks. These networks can be divided into three groups of networks of 100, 1,000, and 5,000 genes respectively. For each individual in a dataset, the gene expression levels are provided as well as the genotype value of one genetic marker for each gene. For each of the nine networks, datasets have been generated under eight different setting configurations, by combining different population sizes (300 or 900 individuals), distances between the genetic markers (large or small), and heritability (large or small), as shown in Table 5.1. All networks and datasets were generated using SysGenSIM[3] 1.0.2 (Pinna et al., 2011). The reader can refer to Chap. 1 for details about the StatSeq compendium.

[2] http://sysgensim.sourceforge.net/datasets.html.

[3] http://sysgensim.sourceforge.net/.

Table 5.1 Setting configurations for data simulation

Configuration	Marker distance*	Heritability	Population size
1	$\sim\mathcal{N}(5, 1)$	High	300
2	$\sim\mathcal{N}(5, 1)$	High	900
3	$\sim\mathcal{N}(5, 1)$	Low	300
4	$\sim\mathcal{N}(5, 1)$	Low	900
5	$\sim\mathcal{N}(1, 0.1)$	High	300
6	$\sim\mathcal{N}(1, 0.1)$	High	900
7	$\sim\mathcal{N}(1, 0.1)$	Low	300
8	$\sim\mathcal{N}(1, 0.1)$	Low	900

*Means and standard deviations are expressed in centimorgans

5.3.2.2 Comparison of Tree-Based Methods

We applied the GENIE3-SG-joint and GENIE3-SG-sep methods on the StatSeq datasets, using the Random Forests algorithm with the main parameter K fixed to the square root of the number of input variables, as well as the Bagging procedure (equivalent to Random Forests with K fixed to the number of input variables). Ensembles of 1,000 trees were grown in each case, except when we used Bagging to infer networks of 5,000 genes. In that case, only 100 trees were grown in order to reduce the computational burden.[4]

Figure 5.3 shows the AUPR scores obtained with the different combinations. Bagging typically yields better performances than Random Forests, whatever the combination. Lower scores are obtained only with GENIE3-SG-joint (using the product of the weights $w^e_{i,j}$ and $w^m_{i,j}$) for some networks. Therefore, all results shown in the remainder of this chapter will be those obtained with the Bagging procedure.

5.3.2.3 Performance of the GENIE3 Methods

GENIE3-SG-joint versus GENIE3-SG-sep

Figure 5.4 shows the performances of the different GENIE3 procedures. Given an aggregation procedure (either sum or product of the importance scores), better performances are obtained when two separate models are, respectively, learned from the two types of data (GENIE3-SG-sep), instead of one single model (GENIE3-SG-joint). The worse performance of GENIE3-SG-joint can be potentially explained by the fact that when the inputs comprise continuous and discrete variables (with a low number of categories), the Bagging method has a positive bias for the continuous variables when selecting a variable at a test node (Strobl et al., 2007). Indeed, since a continuous variable provides more possible cut-points than a variable with a low number of categories, it has more chance to provide the highest variance reduction

[4] Note that, for the smaller networks, we do not observe significant differences in performance when reducing the number of trees from 1,000 to 100.

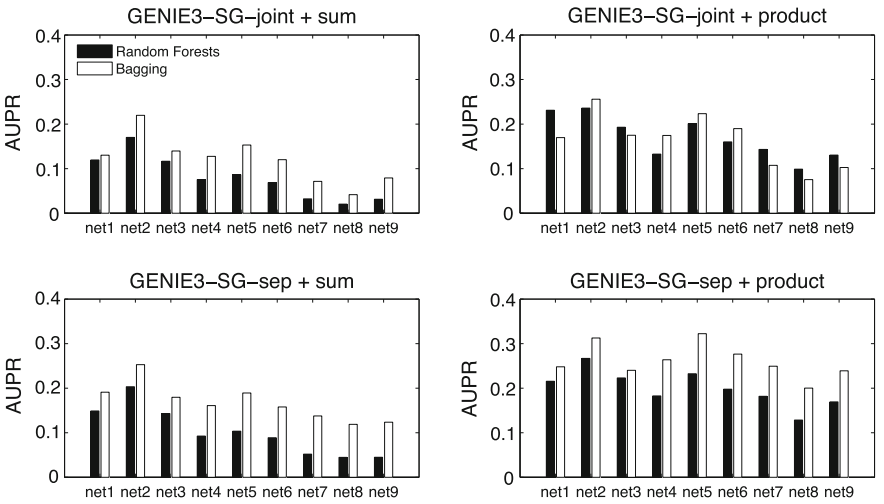

Fig. 5.3 Comparison of tree-based methods. The Bagging method typically yields better performances than Random Forests. The AUPR values of each method were averaged over the eight datasets corresponding to each of the nine networks

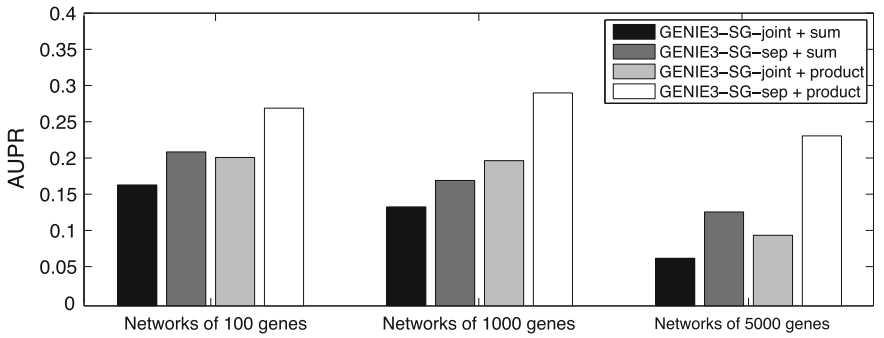

Fig. 5.4 Performances of inference methods. GENIE3-SG-sep yields better performances than GENIE3-SG-joint and higher AUPR scores are obtained by taking the product of the weights $w^e_{i,j}$ and $w^m_{i,j}$ rather than their sum. The AUPR scores of each method were averaged over the 24 datasets related to each network size

on the local node, and hence to be selected for the test, even if it is actually less or equally informative globally. Therefore, in GENIE3-SG-joint, which learns a joint model from the gene expression values (continuous variables) and from the genotype values (discrete variables), the importance $w^m_{i,j}$ of the marker of each gene i is systematically lower than the importance $w^e_{i,j}$ of its expression, as shown in Fig. 5.5. Moreover, the ranking of interactions obtained from the importances $w^m_{i,j}$ is significantly less accurate with GENIE3-SG-joint compared to GENIE3-SG-sep, while the rankings obtained from $w^e_{i,j}$ are equally good (Fig. 5.6).

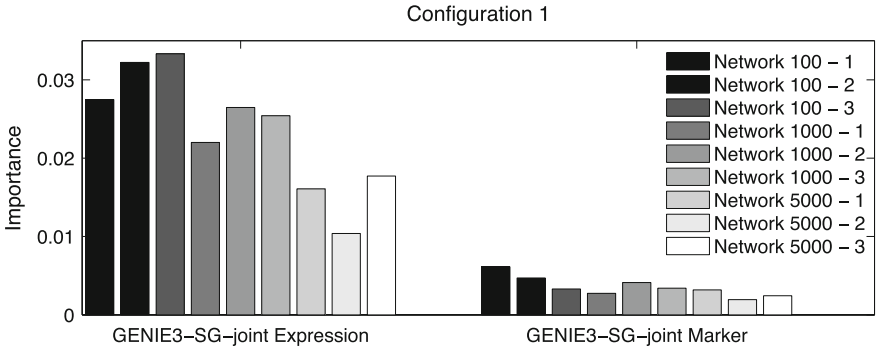

Fig. 5.5 Importance of expressions and markers. This figure shows, for each network, the average weight $w^e_{i,j}$ obtained from the expression profiles, as well as the average weight $w^m_{i,j}$ obtained from the markers, both computed over the edges $i \rightarrow j$ that are part of the gold standard network. The weights $w^e_{i,j}$ and $w^m_{i,j}$ are those obtained on the datasets simulated with the setting configuration 1 (large marker distance, high heritability, and small population size). Despite the high heritability, the tree-based importance values of the genetic markers, as computed in GENIE3-SG-joint, are typically much lower than those obtained from the expression data

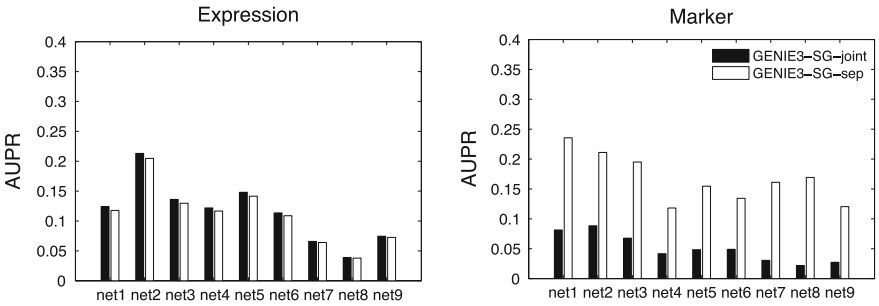

Fig. 5.6 AUPR scores of expressions and markers. This figure shows the AUPR scores obtained when the weight of each edge $i \rightarrow j$ is the importance $w^e_{i,j}$ obtained from the expression data (*left*) and the importance $w^m_{i,j}$ obtained from the genetic markers (*right*). The ranking of interactions obtained from the markers is significantly less accurate with GENIE3-SG-joint, compared to GENIE3-SG-sep. The AUPR values of each method were averaged over the eight datasets corresponding to each of the nine networks

Aggregation procedures

For both procedures GENIE3-SG-joint and GENIE3-SG-sep, higher scores are obtained when the importance scores $w^e_{i,j}$ and $w^m_{i,j}$ are aggregated by taking their product rather than their sum (Fig. 5.4), i.e., when we consider that both the genetic marker and the expression of a regulating gene are important for the prediction of the expression of a target gene. This conservative aggregation procedure allows to give a lower weight to a lot of false edges, since many of them can still have a high value of $w^e_{i,j}$ or a high value of $w^m_{i,j}$ without having a regulatory effect (Fig. 5.7). By contrast, high values for both $w^e_{i,j}$ and $w^m_{i,j}$ are obtained only for true edges.

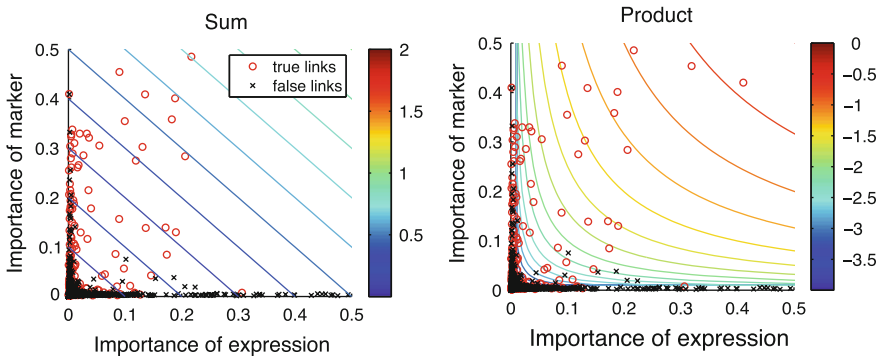

Fig. 5.7 Scatter plot of importances of expressions and markers returned by GENIE3-SG-sep. The red circles correspond to the true edges of the gold standard network, while the black crosses correspond to the false edges (i.e., edges that are not part of the gold standard). In addition, each plot shows the contour lines of the sum and product of $w^e_{i,j}$ and $w^m_{i,j}$ (*left* and *right* figures respectively). The values of the product are shown in a logarithmic scale. These results are those obtained on the first network of 100 genes (Network 100-1), under configuration 1 (high heritability, large marker distance, and small population size). Taking the product of the importance scores allows to give a lower weight to a lot of false edges

5.3.2.4 Influence of Population Size, Heritability, and Marker Distance

Figure 5.8 shows the AUPR scores obtained by the different GENIE3 procedures on the networks of 1,000 genes, for each setting configuration of the simulation runs. As expected, the performance of each method improves when the number of individuals for which data are available increases. The scores also indicate that genetic markers are much more informative than expression data for the inference of the networks when the median heritability as well as the distance between the markers are both high (configurations 1 and 2). In these configurations, only exploiting genetic data ("GENIE3 on markers") results in significantly more accurate predictions than learning from expression data alone ("GENIE3 on expression"). This result is not surprising since a higher heritability means that a higher proportion of the variance of the expression data is actually explained by the genetic markers. Moreover, a higher distance between the markers implies an increased rate of chromosomal crossovers between the different markers, and hence more meaningful multifactorial perturbations (i.e., genetic variations) between the individuals, helping to recover the networks in a more accurate way. By contrast, when the heritability and the marker distance are both small (configurations 7 and 8), expression data are more informative than the markers. In the remaining configurations (configurations 3–6), the performance obtained from gene expression is not very different from the one obtained from genetic markers. Nevertheless, it seems that expression and genetic data contain different and complementary information about the underlying networks, since in all configurations the predictions can be highly improved when both types of data

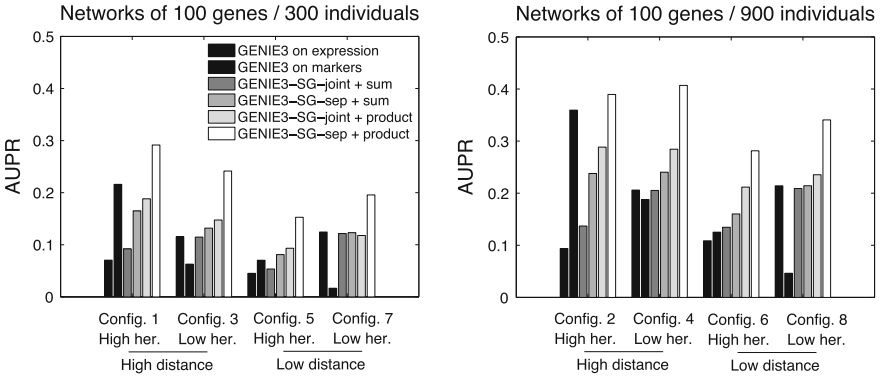

Fig. 5.8 Performances of inference methods for each setting configuration. Predictions can be highly improved when expression and genetic data are integrated, the highest AUPR scores being obtained by GENIE3-SG-sep, using the product of the weights $w^e_{i,j}$ and $w^m_{i,j}$. GENIE3 on expression: weight of edge $i \rightarrow j$ is the importance $w^e_{i,j}$ of the expression of gene i as computed in GENIE3-SG-sep. GENIE3 on markers: weight of edge $i \rightarrow j$ is the importance $w^m_{i,j}$ of the genetic marker of gene i as computed in GENIE3-SG-sep. Her.: heritability. The AUPR values of each method were averaged, for each configuration, over the three datasets related to the networks of 1,000 genes

are integrated, the best results being achieved by far by GENIE3-SG-sep using the product of the weights $w^e_{i,j}$ and $w^m_{i,j}$.

The PR curves related to the first network of 1,000 genes (Network 1000-1) are plotted in Fig. 5.9, for the configuration 1 (high heritability, large marker distance, and small population size). As an example, the 500 first regulatory links obtained with GENIE3-SG-sep (product) yield a precision of 77 % and a recall of 12 %. Increasing the number of considered edges to 1,000 allows us to recover more true edges (recall equal to 18 %), with however, a decrease in precision (57 %). Limiting the network to the first 200 links allows to keep a precision higher than 90 % (recall equal to 6 %). On the other hand, more than 800,000 links have to be considered to obtain a recall higher than 90 % (precision equal to 0.4 %), which is of course of no practical interest.

5.3.2.5 Direction of the Edges

One interesting feature of the GENIE3 methods is their potential ability to predict directed networks. To assess the ability of each method to predict link directions, we computed the error rate on the direction of the edges, i.e., the proportion of edges $i \rightarrow j$ in the gold standard network such that there is no edge $j \rightarrow i$ and for which the method wrongly predicts $w_{i,j} < w_{j,i}$. The error rates are shown in Fig. 5.10. Compared to exploiting expression data alone, using information about genetic markers greatly helps for the prediction of the direction of the edges. However, there is no significant difference between the different methods exploiting the markers. As an example, the GENIE3-SG-sep (product) method yields an average error rate of 27 %.

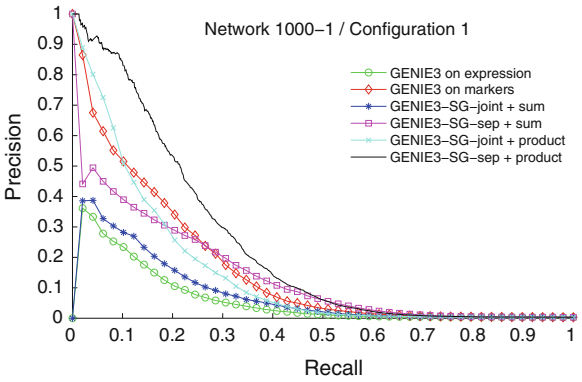

Fig. 5.9 Precision–recall curves. These PR curves were obtained for the first network of 1,000 genes (Network 1000-1), under configuration 1 (high heritability, large marker distance, and small population size). As an example, the 500 first regulatory links obtained with GENIE3-SG-sep (product) yield a precision of 77 % and a recall of 12 %. GENIE3 on expression: weight of edge $i \rightarrow j$ is the importance $w^e_{i,j}$ of the expression of gene i as computed in GENIE3-SG-sep. GENIE3 on markers: weight of edge $i \rightarrow j$ is the importance $w^m_{i,j}$ of the genetic marker of gene i as computed in GENIE3-SG-sep

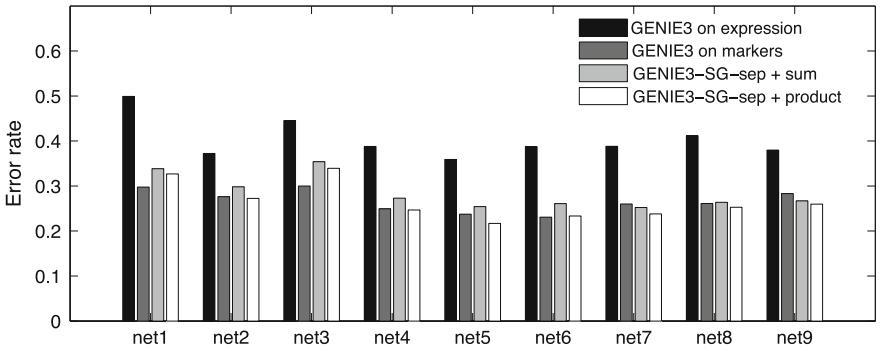

Fig. 5.10 Error rates on edge directionality. Using information about genetic markers greatly helps for the prediction of the direction of the edges. GENIE3 on expression: weight of edge $i \rightarrow j$ is the importance $w^e_{i,j}$ of the expression of gene i as computed in GENIE3-SG-sep. GENIE3 on markers: weight of edge $i \rightarrow j$ is the importance $w^m_{i,j}$ of the genetic marker of gene i as computed in GENIE3-SG-sep. The error rates of each method were averaged over the eight setting configurations of Table 5.1 corresponding to each of the nine networks

5.3.2.6 Interactions Types

We adopted the same evaluation protocol as in Vignes et al. (2011), and analyzed the performance of the GENIE3-SG-sep (product) method, as a function of the type of interactions. We labeled a gene "*cis*" if the corresponding genetic marker is detected in its promoter region, and "*trans*" if the marker is in its coding region. The gene

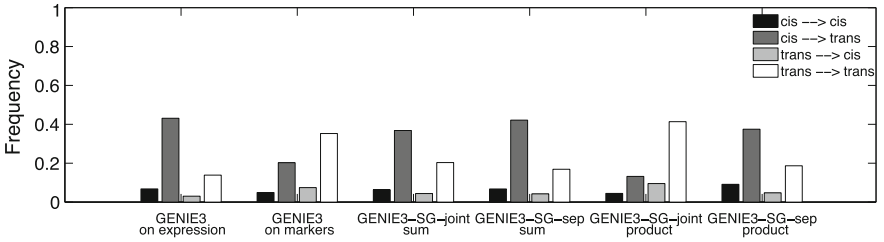

Fig. 5.11 Frequency of the different interaction types. For each method, we selected the first 500 regulatory links of the predicted ranking and for each type of interaction, we computed its frequency among the true interactions of the gold standard that are retrieved. Whatever the method, the top of the ranking typically contains links directed toward *trans* genes. The predictions were obtained from the datasets simulated under configuration 2 (high heritability, large marker distance, and large population size), and the interactions frequencies were averaged over the nine networks

classification was obtained by performing an analysis of variance, as described in Vignes et al. (2011). Genes with a corrected p-value lower than 0.001 were identified as *cis*, and those with an uncorrected p-value higher than 0.1 were identified as *trans*. Using the datasets simulated under configuration 2 (high heritability, large marker distance, and large population size), 23 % of the genes were predicted as *cis* on average, which is close to the actual proportion of 25 % announced in the description of the data, and 63 % of the genes were predicted as *trans*. The classification of genes was used to define four types of interactions: *cis* → *cis*, *cis* → *trans*, *trans* → *cis*, and *trans* → *trans*. For example, an interaction of type *cis* → *trans* is a regulatory link directed from a *cis* gene to *trans* gene. Figure 5.11 shows that the top-ranked predicted interactions typically contain links that are directed toward *trans* genes, whatever the method used. As explained in Vignes et al. (2011), these interactions are predicted more reliably since the target gene does not undergo a *cis*-effect and hence the variation of its expression is only due to the regulating gene (plus the noise).

Cis → *trans* interactions are more frequently predicted than *trans* → *trans* interactions by all methods except GENIE3 on markers and GENIE3-SG-joint with the product. This difference can be explained. In *trans* → *trans* interactions, the impact of the marker of the regulating gene on the expression of the target gene is indeed more direct than in *cis* → *trans* interactions, where the marker only affects the expression of the target gene through the expression of the regulating gene. This leads to higher scores for *trans* → *trans* interactions when GENIE3 is applied on markers only. In GENIE3-SG-joint, the scores of markers and expressions are more balanced for *trans* → *trans* interactions than for *cis* → *trans*, as in *trans* → *trans* interactions both the marker and the expression of the regulating gene are directly and independently affecting the expression of the target gene. This eventually leads to higher scores for *trans* → *trans* interactions when taking the product of marker and expression scores.

5.3.3 The DREAM5 Systems Genetics Challenge

5.3.3.1 Description of the Challenge

The Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative organizes an annual reverse engineering competition that comprises several challenges[5] (Marbach 2012; Prill et al. 2010; Stolovitzky et al. 2009, 2007). We report here our results on the DREAM5 *Systems Genetics* challenge.[6] This challenge concerned the inference of *in silico* regulatory networks from systems genetics data. It was divided into three sub-challenges. The goal of each sub-challenge was to infer five networks from populations of 100, 300, and 999 individuals respectively. Each of the 15 networks contained 1,000 genes and were of increasing connectivity within each sub-challenge. For each individual, expression levels of all the genes were provided, as well as the genotype value of one genetic marker for each gene. All data of the challenge were generated using a preliminary version of SysGenSIM. However, the information about the configuration used to run the simulations was not provided to the challenge participants.

5.3.3.2 Performance of the GENIE3 Methods

Figure 5.12 shows the performances of the different methods. We observe results similar to those obtained on the StatSeq datasets: the performance increases with the number of individuals, better performances are obtained with GENIE3-SG-sep compared to GENIE3-SG-joint, and the product of importance scores $w^e_{i,j}$ and $w^m_{i,j}$ also yields higher AUPR scores than the sum.

5.3.3.3 Comparison with the DREAM5 Best Performer

Figure 5.13 compares, in terms of AUPR scores, GENIE3-SG-sep to the procedure that was used by the official best performing team of the DREAM5 *Systems Genetics* challenge. This procedure is a meta-analysis of different methods, respectively, based on Dantzig regression (Candès and Tao, 2007), LASSO regression (Tibshirani, 1996), and static Bayesian network learning (Friedman et al., 2000). The procedure is described in detail in Vignes et al. (2011). The AUPR scores indicate that our procedure significantly outperforms the meta-analysis for each of the networks.

[5] http://www.the-dream-project.org/.

[6] http://wiki.c2b2.columbia.edu/dream/index.php/D5c3.

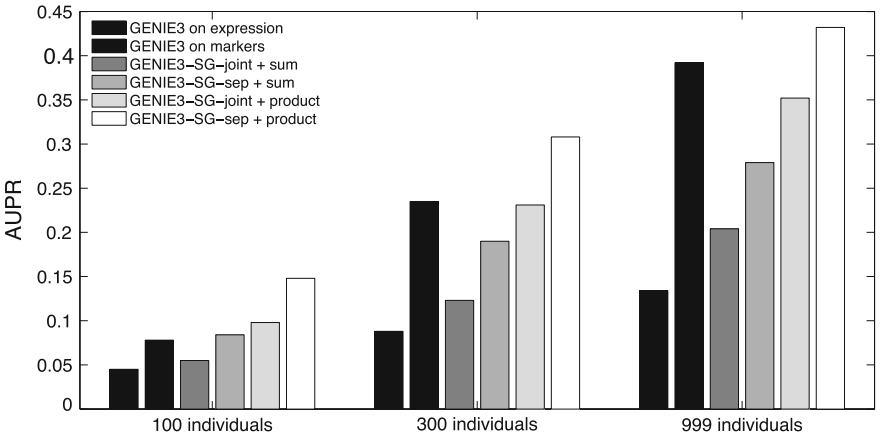

Fig. 5.12 AUPR scores for the DREAM5 Systems Genetics challenge. As for the StatSeq networks, the best predictions are obtained with GENIE3-SG-sep, using the product of the weights $w^e_{i,j}$ and $w^m_{i,j}$. GENIE3 on expression: weight of edge $i \rightarrow j$ is the importance $w^e_{i,j}$ of the expression of gene i as computed in GENIE3-SG-sep. GENIE3 on markers: weight of edge $i \rightarrow j$ is the importance $w^m_{i,j}$ of the genetic marker of gene i as computed in GENIE3-SG-sep. The AUPR values of each method were averaged over the five networks of each sub-challenge

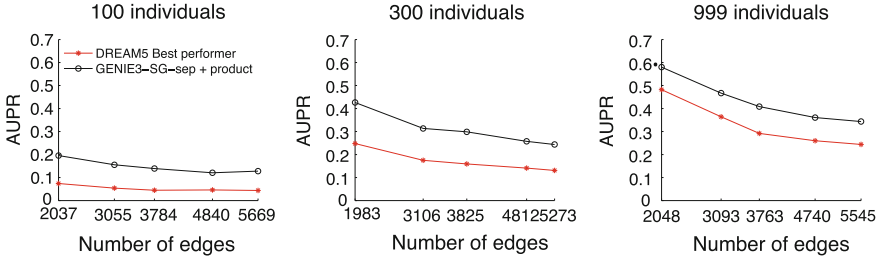

Fig. 5.13 Comparison with the best performer and influence of network density. The GENIE3-gen-sep (product) method outperforms the procedure of the official best performer of the challenge. The performance of both methods, however, decreases when the number of edges in the network increases

5.3.3.4 Influence of Network Density

Besides the study of the effect of the dataset size (number of individuals) on the predictions returned by inference methods, the DREAM5 *Systems Genetics* challenge was also designed to study the effect of the connectivity of a network on the predictions. Figure 5.13 shows the effect of the network density on the predictions. Clearly, in each sub-challenge, the ability of the methods to recover a network tends to decrease as the number of edges in the network increases and regulatory interactions become more complex.

5.4 Discussion

In this chapter, we proposed two procedures, GENIE3-SG-joint and GENIE3-SG-sep, that infer (GRNs) from systems genetics data. Both procedures decompose the problem of inferring a regulatory network of p genes into p different feature selection problems, the goal of each being to retrieve the regulators of one of the genes of the network. Each feature selection problem is then solved by applying a tree-based ensemble method in order to obtain a model predicting the expression of one gene j. In the GENIE3-SG-joint procedure, a single predictive model is learned from expression and genetic data, while in GENIE3-SG-sep, two separate predictive models are learned, one based on the genetic markers and the other based on the expression data. Both methods then compute, for each gene $i \neq j$, two scores $w^e_{i,j}$ and $w^m_{i,j}$, measuring, respectively, the importances of the expression and of the marker of gene i when predicting the expression of gene j. These two scores are then aggregated, by computing either their sum or their product, to obtain a single weight $w_{i,j}$ for the regulatory link directed from gene i to gene j.

The artificial datasets of the StatSeq benchmark were simulated using different setting configurations, i.e., by combining different values of the number of individuals, distance between the markers, and heritability, allowing us to check under which configurations our different methods perform best. Results showed that depending on the marker distance and heritability, genetic markers bring more or less information about the regulatory networks than expression data, and combining the two types of data can be highly helpful for their recovering. GENIE3-SG-sep, using the product of the weight $w^e_{i,j}$ and $w^m_{i,j}$, yields the best performances, whatever the configuration. This method also yields the best performances when recovering the networks of the DREAM5 *Systems Genetics* challenge, and actually outperforms the official best performing algorithm of the challenge.

The StatSeq datasets and the DREAM5 challenge allowed us to make a first evaluation of the performances of our different procedures on systems genetics data. However, these benchmarks are solely based on networks and data that are artificial. As future works, we thus would like to apply our methods on real datasets. Datasets related to various organisms are publicly available, such as the *S. cerevisiae* dataset of Brem and Kruglyak (2005). However, in our different procedures, we assume that each gene whose expression is measured in N individuals is also analyzed for one single genetic marker in each of these N individuals. Unfortunately, this situation is usually not encountered in real datasets. We will thus have to modify our methods in order to deal with missing data, and also to establish a procedure to aggregate the importance scores of different genetic markers related to the same gene.

Finally, although we exploited tree-based ensemble methods, the frameworks of GENIE3-SG-joint and GENIE3-SG-sep are general, and other feature ranking techniques could have been used as well. In the future, we thus plan to apply and compare different ranking techniques, and check which of them permit the best exploitation of expression and genetic data.

Acknowledgments The authors would like to thank Jimmy Vandel for useful discussions on the topic during a visit in February 2011. This work was partially funded by the Interuniversity Attraction Poles Programme (IAP P6/25 BIOMAGNET and IAP P7 DYSCO), initiated by the Belgian State, Science Policy Office, and by the European Network of Excellence PASCAL2. Pierre Geurts is a research associate of FNRS, Belgium. The authors thank the GIGA Bioinformatics platform and the SEGI (University of Liège) for providing computing resources.

References

Aten JE, Fuller TF, Lusis AJ, Horvath S (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Syst Biol 2:34

Bing N, Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics 170:533–542

Breiman L (1996) Bagging predictors. Mach Learn 24:123–124

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman JH, Olsen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International, California

Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci USA 102:1572–1577

Candès E, Tao T (2007) The dantzig selector: Statistical estimation when p is much larger than n. Ann Stat 35:2313–2351

Neto Chaibub E, Ferrara CT, Attie AD, Yandeli BS (2008) Inferring causal phenotype networks from segregating populations. Genetics 179:1089–1100

Chen LS, Emmert-Streib F, Storey JD (2007) Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biol 8:R219

De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Microbiol 8:717–729

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comp Biol 7:601–620

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. JMLR 3:1157–1182

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Prediction, inference and data mining. Springer Verlag, Second Edition

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS ONE 5:e12776

Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4:145–151

Jansen RC, Nap J-P (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388–391

Kulp DC, Jagalur M (2006) Causal inference of regulator-target pairs by gene mapping of expression phenotypes. BMC Genomics 7:125

Li. H, Lu L, Manly KF, Chesler EJ, Bao L, Wang J, Zhou M, Williams RW, Cu i Y (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Gen 14:1119–1125

Li R, Tsaih S-W, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA (2006) Structural model analysis of multiple quantitative traits. PLoS Genet 2:e114

Liu B, de la Fuente A, Hoeschele I (2008) Gene network inference via structural equation modeling in genetical genomics experiments. Genetics 178:1763–1776

Marbach D, Costello J (2012) C., Küffner, R., Vega, N., Prill, R. J., Camacho, D. M., Allison, K. R., the DREAM5 Consortium, Kellis, M., Collins, J. J., Stolovitzky, G.: Wisdom of crowds for robust gene network inference. Nat Methods 9:796–804

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the Lasso. Ann Stat 34:1436–1462

Meyer PE, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol 2007:79879

Michaelson JJ, Alberts R, Schughart K, Beyer A (2010) Data-driven assessment of eQTL mapping methods. BMC Genomics 11:502

Pinna A, Soranzo N, Hoeschele I, de la Fuente A (2011) Simulating systems genetics data with SysGenSIM. Bioinformatics 27:2459–2462

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G (2010) Towards a rigorous assessment of systems Biology models: the DREAM3 challenges. PLoS ONE 5:e9202

Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507–2517

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717

Stolovitzky G, Monroe D, Califano A (2007) Dialogue on Reverse-Engineering assessment and methods: the DREAM of high-throughput pathway inference. Ann NY Acad Sci 1115:11–22

Stolovitzky G, Prill RJ, Califano A (2009) Lessons from the DREAM2 challenges. Ann NY Acad Sci 1158:159–195

Strobl C, Boulesteix A-L, Zeileis A, Horthorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform 8:25

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B 58:267–288

Vignes M, Vandel J, Allouche D, Ramadan-Alban N, Cierco-Ayrolles C, Schiex T, Mangin B, de Givry S (2011) Gene regulatory network reconstruction using bayesian networks, the Dantzig selector, the Lasso and their meta-analysis. PLoS ONE 6:e29165

Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, Sachs JR, Schadt EE (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. PLoS Comput Biol 3:e69