Learning Gaussian Graphical Models of Gene Networks with False Discovery Rate Control

Jose M. Peña

IFM, Linköping University, SE-58183 Linköping, Sweden
jmp@ifm.liu.se

Abstract. In many cases what matters is not whether a false discovery is made or not but the expected proportion of false discoveries among all the discoveries made, i.e. the so-called false discovery rate (FDR). We present an algorithm aiming at controlling the FDR of edges when learning Gaussian graphical models (GGMs). The algorithm is particularly suitable when dealing with more nodes than samples, e.g. when learning GGMs of gene networks from gene expression data. We illustrate this on the Rosetta compendium [8].

1 Introduction

Some models that have received increasing attention from the bioinformatics community as a means to gain insight into gene networks are Gaussian graphical models (GGMs) and variations thereof [4, 6, 9, 14, 25, 26, 29, 31, 32]. The GGM of a gene network represents the network as a Gaussian distribution over a set of random variables, each of them representing (the expression level of) a gene in the network. Learning the GGM reduces to learning the independence structure of the Gaussian distribution. This structure is represented as an undirected graph such that if two sets of nodes are separated by a third set of nodes in the graph, then (the expression level of) the corresponding sets of genes are independent given (the expression level of) the third set of genes in the gene network. Gene dependencies can also be read off a GGM [18]. A further advantage of GGMs is that there already exists a wealth of algorithms for learning GGMs from data. However, not all of them are applicable when the database contains fewer samples than nodes (i.e. $n < q$),[1] which is the case in most gene expression databases. Of the algorithms that are applicable when $n < q$, only the one proposed in [25] aims at controlling the false discovery rate (FDR), i.e. the expected proportion of falsely discovered edges among all the edges discovered. However, the correctness of this algorithm is neither proven nor fully supported by the experiments reported, e.g. see the results for sample size 50 in Figure 6 in [25].

In this paper, we present a modification of the incremental association Markov boundary (IAMB) algorithm [19, 28] aiming at controlling the FDR. Although

[1] We denote the number of nodes (i.e. genes) by q though it is customary to use p for this purpose. We reserve p to denote a probability distribution.

we have not yet succeeded in proving that the new algorithm controls the FDR, the experiments reported in this paper support this conjecture. Furthermore, the new algorithm is particulary suitable for those domains where $n < q$, which makes it attractive for learning GGMs of gene networks from gene expression data. We show that the new algorithm is indeed able to provide biologically insightful models by running it on the Rosetta compendium [8].

2 Preliminaries

The definitions and results in the following two paragraphs are taken from [11, 15, 27, 30]. We use the juxtaposition $\mathbf{XY}$ to denote $\mathbf{X} \cup \mathbf{Y}$, and X to denote the singleton $\{X\}$. Let $\mathbf{U}$ denote a set of q random variables. Unless otherwise stated, all the probability distributions and graphs in this paper are defined over $\mathbf{U}$. Let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{W}$ denote four mutually disjoint subsets of $\mathbf{U}$. We represent that $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$ in a probability distribution p by $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$, whereas we represent that $\mathbf{X}$ is dependent of $\mathbf{Y}$ given $\mathbf{Z}$ in p by $\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z}$. Any probability distribution satisfies the following four properties: Symmetry $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z} \Rightarrow \mathbf{Y} \perp \mathbf{X}|\mathbf{Z}$, decomposition $\mathbf{X} \perp \mathbf{YW}|\mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$, weak union $\mathbf{X} \perp \mathbf{YW}|\mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y}|\mathbf{ZW}$, and contraction $\mathbf{X} \perp \mathbf{Y}|\mathbf{ZW} \wedge \mathbf{X} \perp \mathbf{W}|\mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{YW}|\mathbf{Z}$. Any strictly positive probability distribution also satisfies intersection $\mathbf{X} \perp \mathbf{Y}|\mathbf{ZW} \wedge \mathbf{X} \perp \mathbf{W}|\mathbf{ZY} \Rightarrow \mathbf{X} \perp \mathbf{YW}|\mathbf{Z}$. Any Gaussian distribution also satisfies composition $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z} \wedge \mathbf{X} \perp \mathbf{W}|\mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{YW}|\mathbf{Z}$.

Let $sep(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ denote that $\mathbf{X}$ is separated from $\mathbf{Y}$ given $\mathbf{Z}$ in an undirected graph (UG) G, i.e. every path in G between $\mathbf{X}$ and $\mathbf{Y}$ contains some $Z \in \mathbf{Z}$. G is an undirected independence map of a probability distribution p when $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$ if $sep(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$. G is a minimal undirected independence (MUI) map of p when removing any edge from G makes it cease to be an independence map of p. MUI maps are also called Markov networks. Furthermore, p is faithful to G when $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$ iff $sep(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$. A Markov boundary of $X \in \mathbf{U}$ in p is any subset $MB(X)$ of $\mathbf{U} \setminus X$ such that (i) $X \perp \mathbf{U} \setminus MB(X) \setminus X|MB(X)$, and (ii) no proper subset of $MB(X)$ satisfies (i). If p satisfies the intersection property, then (i) $MB(X)$ is unique for each $X \in \mathbf{U}$, (ii) the MUI map G of p is unique, and (iii) two nodes X and Y are adjacent in G iff $X \in MB(Y)$ iff $Y \in MB(X)$ iff $X \not\perp Y|\mathbf{U} \setminus (XY)$. The MUI map of a Gaussian distribution p is usually called the Gaussian graphical model (GGM) of p. GGMs are also called covariance selection models. In a Gaussian distribution $Normal(\mu, \Sigma)$, $X \perp Y|\mathbf{Z}$ iff $\rho_{XY|\mathbf{Z}} = 0$, where $\rho_{XY|\mathbf{Z}} = \frac{-((\Sigma_{XY\mathbf{Z}})^{-1})_{XY}}{\sqrt{((\Sigma_{XY\mathbf{Z}})^{-1})_{XX}((\Sigma_{XY\mathbf{Z}})^{-1})_{YY}}}$ is the population partial correlation between X and Y given $\mathbf{Z}$.

Assume that a sample of size n from a Gaussian distribution $Normal(\mu, \Sigma)$ is available. Let $r_{XY|\mathbf{Z}}$ denote the sample partial correlation between X and Y given $\mathbf{Z}$, which is calculated as $\rho_{XY|\mathbf{Z}}$ but replacing $\Sigma_{XY\mathbf{Z}}$ by its maximum likelihood estimate based on the sample. Under the null hypothesis that $\rho_{XY|\mathbf{Z}} = 0$, the test statistic $\frac{1}{2} \log \frac{1+r_{XY|\mathbf{Z}}}{1-r_{XY|\mathbf{Z}}}$ has an asymptotic $Normal(0, \frac{1}{\sqrt{n-3-|\mathbf{Z}|}})$ distribution [2]. Moreover, this hypothesis test is consistent [10]. We call this test

Fisher's z-test. Under the null hypothesis that $\rho_{XY|\mathbf{Z}} = 0$, the test statistic $\frac{\sqrt{n-2-|\mathbf{Z}|}\cdot r_{XY|\mathbf{Z}}}{\sqrt{1-r^2_{XY|\mathbf{Z}}}}$ has an exact Student's t distribution with $n-2-|\mathbf{Z}|$ degrees of freedom [2]. We call this test Fisher's t-test. Note that Fisher's z-test and t-test are applicable only when $n > |XY\mathbf{Z}|$: These tests require $r_{XY|\mathbf{Z}}$ which in turn requires the maximum likelihood estimate of $\Sigma_{XY\mathbf{Z}}$, and this exists iff $n > |XY\mathbf{Z}|$ [11].

In many problems what matters is not whether a false discovery is made or not but the expected proportion of false discoveries among all the discoveries made. False discovery rate (FDR) control aims at controlling this proportion. Moreover, FDR control tends to have more power than familywise error rate control, which aims at controlling the probability of making some false discovery [3]. Consider testing m null hypotheses $H_0^1, \ldots, H_0^m$. The FDR is formally defined as the expected proportion of true null hypotheses among the null hypotheses rejected, i.e. $FDR = E[|F|/|D|]$ where $|D|$ is the number of null hypotheses rejected (i.e. discoveries) and $|F|$ is the number of true null hypotheses rejected (i.e. false discoveries). Let $p_1, \ldots, p_m$ denote p-values corresponding to $H_0^1, \ldots, H_0^m$. Moreover, let $p_{(i)}$ denote the i-th smallest p-value and $H_0^{(i)}$ its corresponding hypothesis. The following procedure controls the FDR at level α (i.e. $FDR \leq \alpha$) [3]: Reject $H_0^{(1)}, \ldots, H_0^{(j)}$ where j is the largest i for which $p_{(i)} \cdot \frac{m}{i} \cdot \sum_{k=1}^{m} \frac{1}{k} \leq \alpha$. We call this procedure BY.

3 Learning GGMs

In this section, we present three algorithms for learning GGMs from data. The third one is the main contribution of this paper, as it aims at learning GGMs with FDR control when $n < q$. Hereinafter, we assume that the GGM to learn is sparse, i.e. it contains only a small fraction of all the $q(q-1)/2$ possible edges. This assumption is widely accepted in bioinformatics for the GGM of a gene network.

3.1 EE Algorithm

One of the simplest algorithms for learning the GGM G of a Gaussian distribution p consists in making use of the fact that an edge $X - Y$ is in G iff $X \not\perp Y|\mathbf{U} \setminus (XY)$. We call this algorithm edge exclusion (EE), as the algorithm can be seen as starting from the complete graph and, then, excluding from it all the edges $X - Y$ for which $X \perp Y|\mathbf{U} \setminus (XY)$. Since EE performs a finite number of hypothesis tests, EE is consistent when the hypothesis tests are so. Recall from Section 2 that consistent hypothesis tests exist. Note that EE with Fisher's z-test or t-test is applicable only when $n > q$, since these tests are applicable only in this case (recall Section 2).

Since EE can be seen as performing simultaneous hypothesis tests, BY can be embedded in EE to control the FDR. Note that Fisher's z-test relies on the asymptotic probability distribution of the test statistic and, thus, may not return

Table 1. IAMB(X) and IAMBFDR(X).

IAMB(X)

```
1  MB = ∅
2  for i in 1..q − 1 do
3     p_i = pvalue(X ⊥ Y_i|MB \ Y_i)
4  for i in q − 1..1 do
5     if Y_(i) ∈ MB then
6        if p_(i) > α then
7           MB = MB \ Y_(i)
8           go to line 2
9  for i in 1..q − 1 do
10    if Y_(i) ∉ MB then
11       if p_(i) ≤ α then
12          MB = MB ∪ Y_(i)
13          go to line 2
14 return MB
```

IAMBFDR(X)

```
1  MB = ∅
2  for i in 1..q − 1 do
3     p_i = pvalue(X ⊥ Y_i|MB \ Y_i)
4  for i in q − 1..1 do
5     if Y_(i) ∈ MB then
6        if p_(i) · ((q−1)/i) · Σ_{k=1}^{q−1} 1/k > α then
7           MB = MB \ Y_(i)
8           go to line 2
9  for i in 1..q − 1 do
10    if Y_(i) ∉ MB then
11       if p_(i) · ((q−1)/i) · Σ_{k=1}^{q−1} 1/k ≤ α then
12          MB = MB ∪ Y_(i)
13          go to line 2
14 return MB
```

p-values but approximate p-values. This may cause that the FDR is controlled only approximately. Fisher's t-test, on the other hand, returns p-values and, thus, should be preferred in practice.

3.2 IAMB Algorithm

EE is based on the characterization of the GGM of a Gaussian distribution p as the UG G where an edge $X - Y$ is in G iff $X \not\perp Y|\mathbf{U} \setminus (XY)$. As a consequence, we have seen above that EE is applicable only when $n > q$. We now describe an algorithm that can be applied when $n < q$ under the sparsity assumption. The algorithm is based on the characterization in which an edge $X - Y$ is in G iff $Y \in MB(X)$. Therefore, we first introduce in Table 1 an algorithm for learning MBs that we call IAMB(X), because it is a modification of the incremental association Markov boundary algorithm studied in [19, 28]. IAMB(X) receives the target node X as input and returns an estimate of $MB(X)$ in MB as output. IAMB(X) first computes p-values for the null hypotheses $X \perp Y_i|MB \setminus Y_i$ with $Y_i \in \mathbf{U} \setminus X$. In the table, $p_{(i)}$ denotes the i-th smallest p-value and $Y_{(i)}$ the corresponding node. Then, IAMB(X) iterates two steps. The first step aims at removing false discoveries from MB by removing the node with the largest p-value if this is larger than α. The second step is run when the first step cannot remove any node from MB, and it aims at adding true discoveries to MB by adding the node with the smallest p-value if this is smaller than α. Note that after each node removal or addition, the p-values are recomputed. The original IAMB(X) executes step 2 while possible and only then executes step 1. This delay in removing nodes from MB may harm performance as the larger MB gets the less reliable the hypothesis tests tend to be. The modified version proposed here avoids this problem by keeping MB as small as possible at all times. We prove in [19] that the original IAMB(X) is consistent, i.e. its output converges in probability to a MB of X, if the hypothesis tests are consistent. The proof also applies to the modified version presented here. The proof relies on the fact that any Gaussian distribution satisfies the composition property. It is this property

what allows IAMB(X) to run forward, i.e. starting with $MB = \emptyset$. Recall from Section 2 that consistent hypothesis tests exist.

IAMB(X) immediately leads to an algorithm for learning the GGM G of p, which we just call IAMB: Run IAMB(X) for each $X \in \mathbf{U}$ and, then, link X and Y in G iff X is in the output of IAMB(Y) or Y is in the output of IAMB(X). Note that, in theory, X is in the output of IAMB(Y) iff Y is in the output of IAMB(X). However, in practice, this may not always be true, particulary when working in high-dimensional domains. That is why IAMB only requires one of the two statements to be true for linking X and Y in G. Obviously, IAMB is consistent under the same assumptions as IAMB(X), namely that the hypothesis tests are consistent.

The advantage of IAMB over EE is that it can be applied when $n < q$, because the largest dimension of the covariance matrix for which the maximum likelihood estimate is computed is not $q \times q$ but $s \times s$, where $s - 2$ is the size of the largest MB at line 3 of IAMB(X). We expect that $s \ll q$ under the sparsity assumption. It goes without saying that there are cases when $n < q$ where IAMB is not applicable either, namely those where $n < s$.

3.3 IAMBFDR Algorithm

Unfortunately, IAMB(X) cannot be seen as performing simultaneous hypothesis tests and, thus, BY cannot be embedded in IAMB(X) to control the FDR. In this section, we present a modification of IAMB(X) aiming at controlling the FDR. The modification is based on redefining $MB(X)$ as the set of nodes such that $Y \in MB(X)$ iff $X \not\perp Y | MB(X) \setminus Y$. We now prove that his redefinition is equivalent to the original definition given in Section 2. If $Y \in MB(X)$, then let us assume to the contrary $X \perp Y | MB(X) \setminus Y$. This together with $X \perp \mathbf{U} \setminus MB(X) \setminus X | MB(X)$ implies $X \perp (\mathbf{U} \setminus MB(X) \setminus X) Y | MB(X) \setminus Y$ by contraction, which contradicts the minimality property of $MB(X)$. On the other hand, if $Y \notin MB(X)$ then $X \perp \mathbf{U} \setminus MB(X) \setminus X | MB(X)$ implies $X \perp Y | MB(X) \setminus Y$ by decomposition.

Specifically, we modify IAMB(X) so that the nodes in the output MB are exactly those whose corresponding null hypotheses are rejected when running BY at level α with respect to the null hypotheses $X \perp Y | MB \setminus Y$. In other words, $Y \in MB$ iff $X \perp Y | MB \setminus Y$ according to BY at level α. To implement this modification, we modify the lines 6 and 11 of IAMB(X) as indicated in Table 1. Therefore, the two steps the modified IAMB(X), which we hereinafter call IAMBFDR(X), iterates are as follows. The first step removes from MB the node with the largest p-value if its corresponding null hypothesis is not rejected by BY at level α. The second step is run when the null hypotheses corresponding to all the nodes in MB are rejected by BY at level α, and it adds to MB the node with the smallest p-value among the nodes whose corresponding null hypotheses are rejected by BY at level α.

Finally, we can replace IAMB(X) by IAMBFDR(X) in IAMB and so obtain an algorithm for learning the GGM G of p. We call this algorithm IAMBFDR. It is easy to see that the proof of consistency of IAMB(X) also applies to

IAMBFDR(X) and, thus, IAMBFDR is consistent under the same assumptions as IAMB, namely that the hypothesis tests are consistent. Unfortunately, IAMBFDR does not control the FDR: If the true GGM is the empty graph, then the FDR gets arbitrarily close to 1 as q increases, as any edge discovered by IAMBFDR is a false discovery and the probability that IAMBFDR discovers some edge increases with q. However, if we redefine the FDR of IAMBFDR as the expected FDR of IAMBFDR(X) for $X \in \mathbf{U}$, then IAMBFDR does control the FDR if IAMBFDR(X) controls the FDR: If FDR_X denotes the FDR of IAMBFDR(X), then $E[FDR_X] = \sum_{X \in \mathbf{U}} \frac{1}{q} FDR_X \leq \frac{q}{q} \cdot \alpha$. Although we have not yet succeeded in proving that IAMBFDR(X) controls the FDR, the experiments reported in the next section support the conjecture that IAMBFDR controls the FDR in the latter sense.

4 Evaluation

In this section, we evaluate the performance of EE, IAMB and IAMBFDR on both simulated and gene expression data.

4.1 Simulated Data

We consider databases sampled from random GGMs. Specifically, we consider 100 databases with 50, 100, 500 and 1000 instances sampled from random GGMs with 300 nodes. To produce each of these 400 databases, we do not really sample a random GGM but a random Gaussian network (GN) [7]. The probability distribution so sampled is with probability one faithful to a GGM whose UG is the moral graph of the GN sampled [13]. So, this is a valid procedure for sampling random GGMs. Each GN sampled contains only 1 % of all the possible edges in order to model sparsity. The edges link uniformly drawn pairs of nodes. Each node follows a Gaussian distribution whose mean depends linearly on the value of its parents. For each node, the unconditional mean, the parental linear coefficients and the conditional standard deviation are uniformly drawn from [-3, 3], [-3, 3] and [1, 3], respectively. We do not claim that the databases sampled resemble gene expression databases, apart from some sample sizes and the sparsity of the models sampled. However, they make it possible to compute performance measures such as the power and FDR. This will provide us with some insight into the performance of the algorithms in the evaluation before we turn our attention to gene expression data in the next section.

Table 2 summarizes the results of our experiments with Fisher's t-test and $\alpha = 0.01, 0.05$. Each entry in the table is the average of 100 databases sampled from 100 GGMs randomly generated as indicated above. We do not report standard deviation values because they are very small. For EE, *power* is the fraction of edges in the GGM sampled that are in G, whereas FDR is the fraction of edges in G that are not in the GGM sampled. For IAMB(X) and IAMBFDR(X), $power_X$ is the fraction of nodes in $MB(X)$ that are in the output MB of IAMB(X) or IAMBFDR(X), $FDR_{X,1}$ is the fraction of nodes in

Table 2. Performance of the algorithms on simulated data.

		$\alpha = 0.01$						$\alpha = 0.05$					
n	algorithm	sec.	*power*	FDR	$\overline{power}$	$\overline{FDR_1}$	$\overline{FDR_2}$	sec.	*power*	FDR	$\overline{power}$	$\overline{FDR_1}$	$\overline{FDR_2}$
50	IAMB	4	0.49	–	0.45	0.53	0.19	4	–	–	–	–	–
	IAMBFDR	1	0.36	–	0.35	0.05	0.00	1	0.39	–	0.37	0.05	0.00
100	IAMB	4	0.59	–	0.52	0.46	0.19	42	0.65	–	0.57	0.82	0.37
	IAMBFDR	2	0.47	–	0.43	0.04	0.00	2	0.49	–	0.44	0.04	0.00
500	EE	0	0.46	0.00	0.52	–	–	0	0.49	0.00	0.55	–	–
	IAMB	9	0.78	–	0.68	0.37	0.22	24	0.83	–	0.73	0.70	0.44
	IAMBFDR	6	0.68	–	0.59	0.02	0.00	7	0.70	–	0.60	0.02	0.00
1000	EE	0	0.68	0.00	0.70	–	–	0	0.70	0.00	0.73	–	–
	IAMB	14	0.84	–	0.74	0.35	0.23	27	0.88	–	0.78	0.68	0.46
	IAMBFDR	10	0.76	–	0.66	0.02	0.00	11	0.77	–	0.67	0.02	0.00

MB that are not in $MB(X)$, and $FDR_{X,2}$ is the fraction of nodes Y in MB such that $X \perp Y | MB \setminus Y$. For IAMB and IAMBFDR, we report $\overline{power}$, $\overline{FDR_1}$ and $\overline{FDR_2}$ which denote the average of $power_X$, $FDR_{X,1}$ and $FDR_{X,2}$ over all $X \in \mathbf{U}$. As discussed in Section 3, EE controls FDR, whereas IAMBFDR aims at controlling $\overline{FDR_2}$. We also report EE's *power* for IAMB and IAMBFDR as well as $\overline{power}$ for EE, in order to assess the relative performance of the algorithms. Finally, we also report the runtimes of the algorithms in seconds (sec.). The runtimes correspond to C++ implementations of the algorithms run on a Pentium 2.0 GHz, 1 GB RAM and Windows XP.[2] We draw the following conclusions from the results in the table:

- As discussed above, EE is applicable only when $n > q$ which, as we will see in the next section, renders EE useless for learning GGMs of gene networks from most gene expression databases.
- In the cases where EE is applicable, EE controls FDR. This was expected as BY has been proven to control the FDR [3].
- IAMBFDR controls $\overline{FDR_2}$, though we currently lack a proof for this fact. IAMBFDR does not control $\overline{FDR_1}$, though it keeps it pretty low. The reason why IAMBFDR does not control $\overline{FDR_1}$ is in its iterative nature: If IAMBFDR fails to discover a node in $MB(X)$, then a node $Y \notin MB(X)$ may appear in the output MB of IAMB(X) or IAMBFDR(X). We think that this is a positive feature, as Y is informative about X because $X \not\perp Y | MB \setminus Y$ for Y to be included in MB. The average fraction of nodes in MB such that $Y \notin MB(X)$ but $X \not\perp Y | MB \setminus Y$ is $\overline{FDR_1} - \overline{FDR_2}$.
- IAMB controls neither $\overline{FDR_1}$ nor $\overline{FDR_2}$. As a matter of fact, the number of false discoveries made by IAMB(X) may get so large that the size of MB at line 3 exceeds $n - 3$, which implies that the hypothesis tests at that line cannot be run since the maximum likelihood estimates of the corresponding covariance matrices do not exist (recall Section 2). When this problem occurred, we aborted IAMB(X) and IAMB. With $\alpha = 0.05$, this problem occurred in the 100 databases with 50 samples, and in 26 databases with 100 samples. This problem also occurred when we applied IAMB to learn a

[2] These implementations are available at www.ifm.liu.se/∼jmp.

GGM of a gene network from gene expression data (see next section), which compromises the use of IAMB for such a task.
- IAMBFDR outperforms EE in terms of *power* whereas there is no clear winner in terms of $\overline{power}$. That IAMB outperforms the other two algorithms in terms of *power* and $\overline{power}$ is rather irrelevant, as it controls neither $\overline{FDR_1}$ nor $\overline{FDR_2}$. IAMBFDR is actually more powerful than what *power* and $\overline{power}$ indicate, as none of these measures takes into account the nodes $Y \in MB$ such that $Y \notin MB(X)$ but $X \not\perp Y | MB \setminus Y$ which, as discussed, above are informative about X.

In the light of the observations above, we conclude that IAMBFDR should be preferred to EE and IAMB: IAMBFDR offers FDR control while IAMB does not, moreover EE can only be run when $n > q$ in which case IAMBFDR is more powerful. Furthermore, the runtimes reported in Table 2 suggest that IAMBFDR scales to high-dimensional databases such as, for instance, gene expression databases. The next section confirms it. This is due to the fact that IAMBFDR exploits the composition property of Gaussian distributions to run forward, i.e. starting from the empty graph.

Finally, it is worth mentioning that we repeated all the experiments above with the unconditional means and the parental linear coefficients being uniformly drawn from [-1, 1], and the conditional standard deviations being equal to 1. The results obtained led us to the same conclusions as those above. As a sanity check, we also repeated all the experiments above with the sampled GGMs containing no edge. The results obtained confirmed that EE and IAMBFDR control the FDR even in such an extreme scenario whereas IAMB does not.

4.2 Rosetta Compendium

The Rosetta compendium [8] consists of 300 expression profiles of the yeast *Saccharomyces cerevisiae*, each containing expression levels for 6316 genes. Since for this database $n < q$, EE could not be run. Furthermore, the run of IAMB had to be aborted, since the problem discussed in the previous section occurred. Therefore, IAMBFDR was the only algorithm among those studied in this paper that could be run on the Rosetta compendium. Running IAMBFDR with Fisher's t-test and $\alpha = 0.01$ took 7.4 hours on a Pentium 2.4 GHz, 512 MB RAM and Windows 2000 (C++ implementation). The output contains 32641 edges, that is 0.16 % of all the possible edges.

In order to illustrate that the GGM learnt by IAMBFDR provides biological insight into the yeast gene network, we focus on the iron homeostasis pathway. Iron is an essential nutrient for virtually every organism, but it is also potentially toxic to cells. The iron homeostasis pathway regulates the uptake, storage, and utilization of iron so as to keep it at a non-toxic level. According to [12, 20, 21, 23, 24], yeast can use two different high-affinity mechanisms, reductive and non-reductive, to take up iron from the extracellular medium. The former mechanism is composed of the genes in the FRE family, responsible for iron reduction, and the iron transporters FTR1 and FET3, while the latter mechanisms consist of

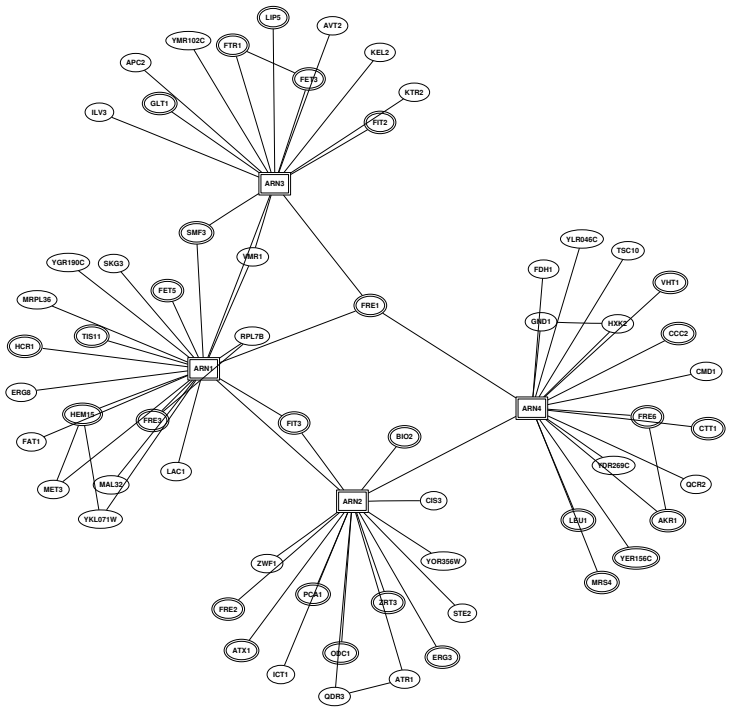

Fig. 1. Subgraph of GGM learnt by IAMBFDR that is induced by the genes that are adjacent to the four genes (square-shaped) involved in the non-reductive mechanism for iron uptake. Double-lined genes are related to iron homeostasis.

the iron transporters ARN1, ARN2, ARN3 and ARN4. The iron homeostasis pathway in yeast has been previously used in [16, 17] to evaluate different algorithms for learning gene network models from gene expression data. These two papers conclude that their algorithms provide biologically plausible models of the iron homeostasis pathway after finding that many genes from that pathway are connected to ARN1 through a path of length one or two. We here take a similar approach to validate the GGM learnt.

Figure 1 depicts the subgraph of the GGM learnt that is induced by the genes that are adjacent to the four genes in the non-reductive mechanism for iron uptake, i.e. ARN1, ARN2, ARN3 and ARN4. These four genes are square-shaped in the figure. In addition to these, the figure contains many other genes related to iron homeostasis. These genes are double-lined in the figure. We now elaborate on these genes. As discussed above, FRE1, FRE2, FRE3, FRE6, FTR1 and FET3 are involved in the reductive mechanism for iron uptake. According to the Gene Ontology search engine AmiGO [1], FET5, MRS4 and SMF3 are iron transporters, FIT2 and FIT3 facilitate iron transport, PCA1 is involved in iron homeostasis, and ATX1 and CCC2 are involved in copper transport and are required by FET3, which is part of the reductive mechanism. According to

[24], BIO2 is involved in biotin synthesis which is regulated by iron, GLT1 and ODC1 are involved in glutamate synthesis which is regulated by iron too, LIP5 is involved in lipoic acid synthesis and regulated by iron, and HEM15 is involved in heme synthesis and regulated by iron too. Also according to [24], TIS11 and the biotin transporter VHT1 are regulated by AFT1, the major iron-dependant transcription factor in yeast. Though AFT1 is not depicted in the subgraph in Figure 1, it is noteworthy that it is a neighbor of FET3 in the GGM learnt. The relation of the zinc transporter ZRT3 to iron homeostasis is documented in [23]. Finally, [5] provides statistical evidence that the following genes are related to iron homeostasis: LEU1, AKR1, HCR1, CTT1, ERG3 and YER156C. Besides, the paper confirms the relation of the first two genes through miniarray and quantitative PCR.

In summary, we have found evidence supporting the relation to iron homeostasis of 32 of the 64 genes in Figure 1. This means that, of the 60 genes that IAMBFDR linked to the four genes that we decided to study, 28 are related to iron homeostasis, which is a substantial fraction. Further evidence of the accuracy of the GGM learnt comes from the fact that these 60 genes are, according to the annotation tool g:Profiler [22], significantly enriched for several Gene Ontology terms that are related to iron homeostasis: GO:0055072 iron ion homeostasis (p-value $< 10^{-5}$), GO:0006825 copper ion transport (p-value $< 10^{-7}$), GO:0015891 siderophore transport (p-value $< 10^{-4}$), GO:0006826 iron ion transport (p-value $< 10^{-14}$), GO:0005506 iron ion binding (p-value $< 10^{-19}$), GO:0005507 copper ion binding (p-value $< 10^{-6}$), GO:0000293 ferric-chelate reductase activity (p-value $< 10^{-6}$), GO:0005375 copper ion transmembrane transporter activity (p-value $< 10^{-4}$), GO:0005381 iron ion transmembrane transporter activity (p-value $< 10^{-5}$), and GO:0043682 copper-transporting ATPase activity (p-value $= 10^{-4}$).

We think that the conclusions drawn in this section, together with those drawn in the previous section, prove that IAMBFDR is scalable and reliable for inferring GGMs of gene networks when $n < q$. Moreover, recall that neither EE nor IAMB could be run on the database used in this section.

5 Discussion

In this paper, we have proposed IAMBFDR, an algorithm for controlling the FDR when learning GGMs and $n < q$. We have shown that the algorithm works well in practice and scales to high-dimensional domains. In particulary, we have shown that IAMBFDR is able to provide biological insight in domains with thousands of genes but many fewer samples. Other works that propose algorithms for controlling the FDR when learning GGMs and $n < q$ are [25, 26]. However, the correctness of the algorithm proposed in the first paper is neither proven nor fully supported by the experiments reported (e.g. see the results for sample size 50 in Figure 6 in [25]), whereas the algorithm in the second paper does not really aim at controlling the FDR but the closely related local FDR. IAMBFDR resembles the algorithms proposed in [6, 14] in the sense that they all learn the

GGM of a gene network by learning the MB of each node. Specifically, [6] takes a Bayesian approach that combines elements from regression and graphical models whereas [14] uses the lasso method. However, the main difference between our algorithm and theirs is that the latter do not aim at controlling the FDR. For the algorithm proposed in [14], this can clearly be seen in the experimental results reported in Table 1 in that work and in Figure 3 in [26].

Acknowledgements

This work is funded by the Swedish Research Council (ref. VR-621-2005-4202). We thank the anonymous reviewers for their insightful comments.

References

1. http://amigo.geneontology.org/cgi-bin/amigo/go.cgi
2. Anderson,T.W. (1984) *An Introduction to Multivariate Statistical Analysis*. Wiley.
3. Benjamini,Y. and Yekutieli,D. (2001) The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics*, **29**, 1165-1188.
4. Castelo,R. and Roverato,A. (2006) A Robust Procedure for Gaussian Graphical Model Search from Microarray Data with p Larger than n. *Journal of Machine Learning Research*, **7**, 2621-2650.
5. De Freitas,J.M., Kim,J.H., Poynton,H., Su,T., Wintz,H., Fox,T., Holman,P., Loguinov,A., Keles,S., van der Laan,M., Vulpe,C. (2004) Exploratory and Confirmatory Gene Expression Profiling of *mac1Δ*. *Journal of Biological Chemistry*, **279**, 4450-4458.
6. Dobra,A., Hans,C., Jones,B., Nevins,J.R., Yao,G. and West,M. (2004) Sparse Graphical Models for Exploring Gene Expression Data. *Journal of Multivariate Analysis*, **90**, 196-212.
7. Geiger,D. and Heckerman,D. (1994) Learning Gaussian Networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 235-243.
8. Hughes,T.R. *et al.* (2000) Functional Discovery via a Compendium of Expression Profiles. *Cell*, **102**, 109-126.
9. Jones,B., Carvalho,C., Dobra,A., Hans,C., Carter,C. and West,M. (2005) Experiments in Stochastic Computation for High Dimensional Graphical Models. *Statistical Science*, **20**, 388-400.
10. Kalisch,M. and Bühlmann,P. (2007) Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, **8**, 613-636.
11. Lauritzen,S.L. (1996) *Graphical Models*. Oxford University Press.
12. Lesuisse,E., Blaiseau,P.L., Dancis,A. and Camadro,J.M. (2001) Siderophore Uptake and Use by the Yeast *Saccharomyces cerevisiae*. *Microbiology*, **147**, 289-298.
13. Meek,C. (1995) Strong Completeness and Faithfulness in Bayesian Networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 411-418.
14. Meinshausen,N. and Bühlmann,P. (2006) High-Dimensional Graphs and Variable Selection with the Lasso. *Annals of Statistics*, **34**, 1436-1462.
15. Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

16. Pe'er,D., Regev,A., Elidan,G. and Friedman,N. (2001) Inferring Subnetworks from Perturbed Expression Profiles. *Bioinformatics*, **17**, S215-S224.
17. Peña,J.M., Nilsson,R., Björkegren,J. and Tegnér,J. (2005) Growing Bayesian Network Models of Gene Networks from Seed Genes. *Bioinformatics*, **21**, ii224-ii229.
18. Peña,J.M., Nilsson,R., Björkegren,J. and Tegnér,J. (2006) Reading Dependencies from the Minimal Undirected Independence Map of a Graphoid that Satisfies Weak Transitivity. In *Proceedings of the Third European Workshop on Probabilistic Graphical Models*, 247-254.
19. Peña,J.M., Nilsson,R., Björkegren,J. and Tegnér,J. (2007) Towards Scalable and Data Efficient Learning of Markov Boundaries. *International Journal of Approximate Reasoning*, **45**, 211-232.
20. Philpott,C.C., Protchenko,O., Kim,Y.W., Boretsky,Y. and Shakoury-Elizeh,M. (2002) The Response to Iron Deprivation in *Saccharomyces cerevisiae*: Expression of Siderophore-Based Systems of Iron Uptake. *Biochemical Society Transactions*, **30**, 698-702.
21. Protchenko,O., Ferea,T., Rashford,J., Tiedeman,J., Brown,P.O., Botstein,D. and Philpott,C.C. (2001) Three Cell Wall Mannoproteins Facilitate the Uptake of Iron in *Saccharomyces cerevisiae*. *The Journal of Biological Chemistry*, **276**, 49244-49250.
22. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler - A Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments. *Nucleic Acids Research*, **35**, W193-W200.
23. Santos,R., Dancis,A., Eide,D., Camadro,J.M. and Lesuisse,E. (2003) Zinc Suppresses the Iron-Accumulation Phenotype of *Saccharomyces cerevisiae* Lacking the Yeast Frataxin Homologue (Yfh1). *Biochemical Journal*, **375**, 247-254.
24. Shakoury-Elizeh,M., Tiedeman,J., Rashford,J., Ferea,T., Demeter,J., Garcia,E., Rolfes,R., Brown,P.O., Botstein,D. and Philpott,C.C. (2004) Transcriptional Remodeling in Response to Iron Deprivation in *Saccharomyces cerevisiae*. *Molecular Biology of the Cell*, **15**, 1233-1243.
25. Schäfer,J. and Strimmer,K. (2005a) An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks. *Bioinformatics*, **21**, 754-764.
26. Schäfer,J. and Strimmer,K. (2005b) A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**.
27. Studený,M. (2005) *Probabilistic Conditional Independence Structures*. Springer.
28. Tsamardinos,I., Aliferis,C.F. and Statnikov,A. (2003) Algorithms for Large Scale Markov Blanket Discovery. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, 376-380.
29. Werhli,A.V., Grzegorczyk,M. and Husmeier,D. (2006) Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks. *Bioinformatics*, **22**, 2523-2531.
30. Whittaker,J. (1990) *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons.
31. Wille,A. and Bühlmann,P. (2006) Low-Order Conditional Independence Graphs for Inferring Genetic Networks. *Statistical Applications in Genetics and Molecular Biology*, **5**.
32. Wille,A., Zimmermann,P., Vranova,E., Fürholz,A., Laule,O., Bleuler,S., Hennig,L., Prelic,A., von Rohr,P., Thiele,L., Zitzler,E., Gruissem,W. and Bühlmann,P. (2004) Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in *Arabidopsis thaliana*. *Genome Biology*, **5**, 1-13.