Sparse graphical gaussian modeling for genetic regulatory network inference

Anja Wille[*1,5], Philip Zimmermann[*4], Eva Vranová[4], Andreas Fürholz[4], Oliver Laule[4], Stefan Bleuler[2], Lars Hennig[4], Amela Prelić[2], Peter von Rohr[3], Lothar Thiele[2], Eckart Zitzler[2], Wilhelm Gruissem[4] and Peter Bühlmann[5]

Reverse Engineering Group, Swiss Federal Institute of Technology (ETH), Zurich
[1]Colab, Swiss Federal Institute of Technology (ETH), Zurich
[2]Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology (ETH), Zurich
[3]Institute of Computational Science, Swiss Federal Institute of Technology (ETH), Zurich
[4]Institute for Plant Sciences and Functional Genomics Center Zurich, Swiss Federal Institute of Technology (ETH), Zurich
[5]Seminar for Statistics, Swiss Federal Institute of Technology (ETH), Zurich

Email: Anja Wille*- awille@inf.ethz.ch; Philip Zimmermann*- philip.zimmermann@ipw.biol.ethz.ch;

*Corresponding author

Abstract

Background: DNA microarray technologies have made it possible to simultaneously measure expression levels of thousands of genes. In recent years, graphical models have become increasingly popular for generating and evaluating hypotheses on genetic control mechanisms and network topologies based on these expression profiles. When applied to a large number of genes, however, several limitations of graphical models come to the fore.

Methods: To overcome these limitations, we propose a novel graphical gaussian modeling (GGM) approach in which modeling is carried out in small subnetworks with three genes only. These subnetworks are then combined for inference on the complete network. We present two versions of our method, a frequentist approach and a likelihood approach based on latent random graphs, and find that both versions outperform the standard graphical gaussian modeling in a simulation study.

Results: As a main application, we employ our method to infer a gene network for isoprenoid biosynthesis in *Arabidopsis thaliana*. We detect modules of closely connected genes and candidate genes for a possible crosstalk between both isoprenoid pathways. The discovered structures are further validated with the help of domain knowledge by identifying genes of downstream pathways that fit well in this network.

Background

Graphical models [1, 2] form a probabilistic tool to analyze and visualize conditional relationships between random variables. Random variables are represented by vertices of a graph and conditional dependencies between them are encoded by edges. The structure of the conditional relationships among random variables can be exhaustively explored with the help of the so-called Markov properties [1, 2].

When applied to gene expression data in order to infer genetic regulatory networks, the vertices stand for genes whose conditional dependence patterns are examined based on their expression profiles. Graphical modeling of genetic networks can be carried out with directed and undirected edges, with discretized and continuous data. Over the last few years, graphical models, in particular Bayesian networks, have become increasingly popular in reverse engineering of genetic regulatory networks [3–6].

Graphical models are powerful for a small number of genes. As the number of genes increases, however, reliable estimates of conditional dependencies require many more observations than are usually available in gene expression profiling. Furthermore, since the number of models grows superexponentially with the number of genes only a small subset of models can be tested [6]. Most importantly, a large number of genes often entails a large number of spurious edges in the model [7]. The interpretation of the graph within a conditional independence framework is then rendered difficult [8]. Even a search for local dependence structures and subnetworks with high statistical support [3] provides no guarantee against the detection of numerous spurious features.

Some of the aforementioned problems may be circumvented by restricting the number of possible models or edges [6, 9] or by exploiting prior knowledge on the network structure. So far, however, this prior knowledge is difficult to obtain.

As an alternative approach for modeling genetic networks with many genes, we propose not to condition on all genes at a time. Instead, we apply graphical modeling only to small subnetworks with three genes to explore the dependence between two of the genes conditional on the third one. These subnetworks are then combined for inference on the complete network. This modified graphical modeling approach makes it possible to include many genes in the network while studying dependence patterns in a more complex and exhaustive way than with only pairwise correlation-based relationships.

For an independent validation of our method, we compare our modified GGM approach with the conventional graphical modeling in a simulation study. We show at the end of the Result section that our approach outperforms the standard method in simulation settings with many genes and few observations. For a further evaluation with real data, we apply our approach to the galactose utilization data from [10] to detect galactose regulated genes in *Saccharomyces cerevisiae*.

The main aim of this methodological work, however, was to elucidate the regulatory network of the two isoprenoid biosynthesis pathways in *Arabidopsis thaliana* (reviewed in [11]). The bigger part of this paper is therefore devoted to the inference and biological interpretation of a genetic regulatory network of these two pathways. To motivate our novel modeling strategy, we first describe the problems that we encountered with standard GGMs before presenting the results of our modified GGM approach.

Results

Isoprenoids serve numerous biochemical functions in plants, e.g. as components of membranes (sterols), as photosynthetic pigments (carotenoids and chlorophylls) or as hormones (gibberellins). Isoprenoids are synthesized through condensation of the 5 carbon intermediates isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). In higher plants, two distinct pathways for the formation of IPP and DMAPP exist, one in the cytosol and the other in the chloroplast. The cytosolic pathway, often described as mevalonate or MVA pathway, starts from acetyl-Coa to form IPP via several steps including the intermediate mevalonate (MVA). In contrast, the plastidial (non-mevalonate or MEP) pathway involves condensation of pyruvate and glyceraldehyde-3-phosphate via several intermediates to form IPP and DMAPP. Whereas the MVA pathway is responsible for the synthesis of sterols, sesquiterpenes, and the side chain of ubiqinone, the MEP pathway is employed for the synthesis of isoprenes, carotenoids, and the side chains of chlorophyll and plastoquinone. Although both pathways operate independently under normal conditions, interaction between them has been repeatedly reported [12,13]. Reduced flux through the MVA pathway after treatment with lovastatin can be partially compensated for by the MEP pathway. However, inhibition of the MEP pathway in seedlings leads to reduced levels in carotenoids and chlorophylls indicating a predominantly unidirectional transport of isoprenoid intermediates from the chloroplast to the cytosol [12,14], although some reports indicate that an import of isoprenoid intermediates into the chloroplast also takes place [15–17].

Application of standard GGM to isoprenoid pathways in *Arabidopsis thaliana*

In order to gain better insight into the crosstalk between both pathways on the transcriptional level, gene expression patterns were monitored under various experimental conditions using 118 GeneChip® (Affymetrix) microarrays. For the construction of the genetic regulatory network, we focused on 40 genes, 16 of which were assigned to the cytosolic pathway, 19 to the plastidal pathway and 5 encode proteins located in the mitochondrion. These 40 genes comprise not only genes of known function but also genes whose encoded proteins exhibited high homology to proteins of known functions. For reference, we adopt the notation from [18] (see Table 1).

The genetic interaction network among these genes was first constructed employing graphical gaussian modeling with backward selection under the Bayesian Information Criterion (BIC) [19]. This was carried out with the MIM 3.1 program [20] (see the Method section for further details). The obtained network had 178 (out of 780) edges - too many to single out biologically relevant structures. Therefore, bootstrap resampling was applied to determine the statistical confidence of the edges in the model (Figure 1b). For the bootstrap edge probabilities, only a cutoff level as high as 0.8 led to a reasonably low number of selected edges (31 edges, Figure 2). However, a comparison between bootstrap edge probabilities and the pairwise correlation coefficients suggested that for such a high cutoff level, many true edges may be missed. For example, the gene AACT2 appears to be completely independent from all genes in the model although it is strongly correlated to MK, MPDC1 and FPPS2.

The just described phenomenon was already observed in a simulation study of [21] and may be related to the surprisingly frequent appearance of edges with a low absolute pairwise correlation coefficient but a high bootstrap estimate (Figure 1c). Although there is no concise explanation for this pattern, one conjecture would be that the simultaneous conditioning on many variables introduces many spurious edges with little absolute pairwise correlation but high absolute partial correlation into the model. Our modification for GGMs is to improve upon this drawback.

Application of our modified GGM approaches

As described in more detail in the method section, our approach aims at modeling dependencies between two genes by taking the effect of other genes separately into account. In the hope to identify immediate co-regulation between genes, an edge is drawn between two genes i and j when their pairwise correlation is not the effect of a third gene. Each edge has therefore a clear interpretation.

We have developed two versions of our method, a frequentist approach in which each edge is tested for presence or absence, and a likelihood approach with parameters θ_{ij} which describe the probability for an edge between i and j in a latent random graph. One main benefit of the second version over full graphical models is that one can easily test on a large scale how well additional genes can be incorporated into the network. This allows to select additional candidate genes for the network in a fast and efficient way.

We have applied and tested our modified GGM approaches by constructing a regulatory network of the 40 genes in the isoprenoid pathways in *Arabidopsis thaliana* and by attaching 795 additional genes from 56 other metabolic pathways to it.

Figure 3 shows the network model obtained from the frequentist modified GGM approach. Since we find a module with strongly interconnected genes in each of the two pathways, we split up the graph into two subgraphs each displaying the subnetwork of one module and its neighbors. Our finding provides a further example that within a pathway, potentially many consecutive or closely positioned genes are jointly regulated [22].

In the MEP pathway, the genes DXR, MCT, CMK, and MECPS are nearly fully connected (upper panel of Figure 3). From this group of genes, there are a few edges to genes in the MVA pathway. Among these genes, AACT1 and HMGR1 form candidates for a crosstalk between the MEP and the MVA pathway because these genes have no further connection to the MVA pathway. Their correlation to DXR, MCT, CMK, MECPS is always negative.

Similarly, the genes AACT2, HMGS, HMGR2, MK, MPDC1, FPPS1 and FPPS2 share many edges in the MVA pathway (lower panel of Figure 3). The subgroup AACT2, MK, MPDC1, FPPS2 is completely interconnected. From these genes, we find edges to IPPI1 and GGPPS12 in the MEP pathway. Whereas IPPI1 is positively correlated to AACT2, MK, MPDC1 and FPPS2, GGPPS12 displays negative correlation to the four genes.

In contrast to the conventional graphical model, we could now identify the connection between AACT2 and MK, MPDC1 and FPPS2. In general, we found a better agreement between the absolute pairwise correlation and the selected edges (frequentist approach) or the probability parameters θ (latent random graph approach). Figures 4a and b show the selected edges and θ-values as a function of the absolute pairwise correlation.

Attaching additional pathway genes to the network

Following the construction of the isoprenoid genetic network, 795 additional genes from 56 metabolic pathways were incorporated. Among these were also genes from downstream pathways of the two isoprenoid biosynthesis pathways such as phytosterol biosynthesis, mono- and diterpene metabolism, porphyrin/chlorophyll metabolism, carotenoid biosynthesis, plastoquinone biosynthesis etc. Using the second version of our method, i.e. the latent random graph approach, we compared θ-values for all gene pairs in the network with and without attaching these additional genes (Figure 4b and c). As was expected, the parameters θ for the edge probabilities decreased if additional genes were included in the isoprenoid network (see Method section). If for a gene pair i, j, θ_{ij} dropped for more than 0.3 it was assumed that the dependence between i and j could be "explained" by some of the additional genes.

To find these genes out of all additionally tested candidates k, GGMs with genes i, j and k were formed. A gene k was considered to explain the dependency between i and j when an edge between i and j was not supported in the GGM, i.e. when the null hypothesis $\rho_{ij|k} = 0$ was accepted in the corresponding likelihood ratio test. k was then taken to "attach well" to the gene pair i, j.

Thus, for each gene pair i, j whose parameter θ_{ij} dropped for more than 0.3, we obtained a list of well-attaching genes. Genes appearing significantly frequently in these lists of well-attaching genes were assumed to connect well to the complete genetic network. We tested for significance by randomization: For each gene pair i, j, a randomized list of well-attaching genes was formed with the same size as the original gene list. In order to explore which pathways attach significantly well to the MVA and MEP pathway, the portion of genes from each of the 56 pathways was summed over all gene pairs i, j. These sums were then compared for the originally attached genes and the sums of randomly attached genes in 100 data sets.

Table 2 shows the pathways whose genes were found to attach significantly frequently to the MVA pathway, the MEP pathway or both pathways. Interestingly, from all 56 considered metabolic pathways, we find predominantly genes from downstream pathways to fit well into the isoprenoid network. These results suggest a close regulatory connection between isoprenoid biosynthesis genes and groups of downstream genes. On the one side, we find strong connections between the MEP pathway and the plastoquinone, the carotenoid and chlorophyll pathways (experimentally supported by [11, 12, 23]). On the other side, the plastoquinone and phytosterol biosynthesis pathways appear to be closely related to the genetic network of the MVA pathway.

On a metabolic level, our results are substantiated by earlier labeling experiments using [1^{13}C] glucose which revealed that sterols were formed via the MVA pathway, while plastidic isoprenoids (β-carotene, lutein, phytol and plastoquinone-9) were synthesized using intermediates from the MEP pathway [23]. Moreover, incorporation of [1-^{13}C]- and [2,3,4,5-$^{13}C_4$]1-deoxy-D-xylulose into β-carotene, lutein and phytol indicated that the carotenoid and chlorophyll biosynthesis pathways proceed from intermediates obtained via the MEP pathway [24].

In contrast, a close connection between the MVA and the MEP pathways could not be detected. This suggests that a crosstalk on the transcriptional level may be restricted to single genes in both pathways.

In a further analysis step, we examined to which gene pairs the 4 identified pathways (plastoquinone, carotenoid, chlorophyll, and phytosterols) attached. Genes from the plastoquinone pathway were predominantly linked to the genes DXR, MCT, CMK, GGPPS11, GGPPS12, AACT1, HMGR1, and FPPS1 supporting the hypothesis that the genes AACT1 are HMGR1 are involved in a communication between the MEP and the MVA pathway.

Genes from the carotenoid pathway attached to DXPS2, HDS, HDR, GGPPS11, DPPS2 and PPDS2 whereas the chlorophyll biosynthesis appears to be related to DXPS2, DXPS3, DXR, CMK, MCT, HDS, HDR, GGPPS11 and GGPPS12. Genes from the phytosterol pathway attach to the genes FPPS1, HMGS, DPPS2, PPDS1 and PPDS2.

Incorporating 795 additional genes into the isoprenoid genetic network would not have been feasible with standard GGMs since the graphical model would have to be newly fitted for each additional gene. Also, hierarchical clustering would not have been an appropriate tool for detecting the similarities in the correlation patterns between the two isoprenoid metabolisms and their downstream pathways. Figure 5 shows the hierarchical clustering of the 40 isoprenoid genes and 795 additional pathway genes based on the distance measure $1 - |\sigma_{ij}|$ where σ_{ij} denotes the pairwise correlation between genes i and j.

In the left respectively right column of the axis of the heatmap, the positions of the mevalonate pathway genes (labeled with "m") respectively the positions of the non-mevalonate pathway genes (labeled with "n") are displayed. The symbols "+" represent the positions of the genes from the downstream pathways identified in Table 2. From Figure 5 it can be easily seen that there is no clear assignment between genes of the isoprenoid biosynthesis and genes of the downstream pathways in the hierarchical clustering.

Simulation study

For an independent comparison between the modified and the conventional GGM approaches, we simulated gene expression data with 40 genes and 100 observations. This simulation framework corresponds to the data for the isoprenoid biosynthesis and is thought to be only exemplary at this point. An extensive simulation study is currently under way and will be presented elsewhere.

Following recent findings on the topology of metabolic and protein networks [25, 26], we simulated scale-free networks in which the fraction of nodes with k edges decays as a power law $\propto k^{-\gamma}$. For metabolic and protein networks, γ is usually estimated to range between 2 and 3, which would result in very sparse networks with fewer edges than nodes in our simulation settings. In order to also allow for denser networks, we generated 100 graphs each for $\gamma = 0.5$, 1.5, and 2.5. With 40 nodes, these graphs then comprised 88.3, 49.7 and 30.5 edges on average. For each edge, the conditional dependence of the corresponding gene pairs was modeled with a latent random variable in a structural equation model as described by [27]. Further details are of technical nature and are omitted here. The use of latent random variables enabled us to model partial correlation coefficients according to the previously defined network structure while ensuring positive definiteness of the complete partial correlation matrix. This matrix was then transformed into a covariance matrix Σ from which synthetic gene expression data with 100 observations were sampled according to a multivariate normal distribution $N(0, \Sigma)$.

The performance of the graphical modeling approaches was monitored using the rate of true and false positives in receiver operator characteristics (ROC) curves (see [7] for a short introduction). Since for the standard graphical model, bootstrapping would have been to time consuming, we ranked all edges according to their sequential removal in the backward selection process. Figure 6a shows the ROC curves for the graphical modeling with backward selection and the modified graphical modeling approaches (frequentist and latent random graph approach). We also included the ROC curve for network inference with pairwise correlation coefficients. It can be seen that the modified GGM approaches outperform the conventional graphical modeling. Both, the frequentist and the latent random graph method show a similar performance. Also, it should be noted that a simple measure such as the pairwise correlation can be quite powerful in detecting conditional dependencies between genes.

ROC curves depict the true positive rate as a function of the false negative rate. However, in our setting where the false positive edges by far outnumber the true positive ones, the proportion of true positives among the selected edges is also of interest (Figure 6b). Note that this proportion is the complementary false discovery rate 1-FDR [28]. Figure 6b provides further evidence that the modified GGM approaches have a better performance than the standard graphical gaussian modeling.

Application to galactose utilization in *Saccharomyces cerevisiae*

For further evaluation, we applied our approach to the galactose utilization data set from [10] to detect galactose regulated genes in *Saccharomyces cerevisiae.* [10] used self-organizing maps to cluster 997 genes with significant expression changes in 20 systematic perturbation experiments of the galactose pathway. From the 9 galactose genes under investigation, two subgroups with 3 respectively 4 genes were found in 2 of the 16 clusters. 9 of the 87 genes in these 2 clusters carried GAL4p binding sites and are thus candidates genes for regulation by the transcription factor GAL4p. Among these candidate genes, GCY1 and PCL10 are known to be targets of GAL4p [29] and YMR318C has been implicated in another binding site study [30].

After incorporating all yeast genes into our network of the 9 galactose genes, 13 genes were found to attach significantly well. Among these, GCY1 and PCL10 were also detected. Furthermore, 3 out of the remaining 11 candidate genes (MLF3, YEL057C, YPL066W) had GAL4p binding sites. These genes were also identified by [10]. This result shows once more that with our approach, we are not only able to model the dependence between genes but also find genes whose expression profiles fit well to the original genes in the model. In contrast to [10], we did not have to rely on gene clusters with a high occurrence of galactose genes to find these genes.

Conclusions

Analysis of gene expression patterns, for example cluster analysis, often focuses on coexpression and pairwise correlation between genes. Graphical models are based on a more sophisticated measure of conditional dependence among genes. However, with this measure, modeling is restricted to a small number of genes. With a larger set of genes, it is rather difficult to interpret the model and to generate hypotheses on the regulation of genetic networks.

In our approaches, in the search for significant co-regulation between two genes, all other genes in the model are also taken into account. However, the effect of these genes is examined separately, one gene at a time. Due to this simplification, modeling can include a larger number of genes. Also, each edge has a clear interpretation representing a pair of significantly correlated genes whose dependence cannot be explained by a third gene in the model. Our frequentist method carries resemblance to the first two steps in the SGS and PC algorithm [27]. By restricting the modeling to subnetworks with three genes, we avoid the statistically unreliable and computationally costly search for conditional independence in large subsets as in the SGS algorithm. Also, we avoid to remove edges in a stepwise fashion as in the PC algorithm. Therefore, we do not run the risk of mistakenly removing an edge at an early stage leading to improved stability in the modeling process.

For the isoprenoid biosynthesis pathways in *Arabidopsis thaliana*, we constructed a genetic network and identified candidate genes for a crosstalk between both pathways. Interestingly, both positive and negative correlations were found between the identified candidates genes and the corresponding pathways. AACT1 and HMGR1, key genes of the MVA pathway, were found to be negatively correlated to the module of connected genes in the MEP pathway. This suggests that in the experimental conditions tested, AACT1 and HMGR1 may respond differently (than the MEP pathway genes) to environmental conditions, or that they possess a different organ-specific expression profile. In either case, expression within both groups seems to be mutually exclusive. On the other hand, a positive correlation was identified between IPPI1 and members of the MVA pathway, suggesting that this enzyme controls the steady-state levels of IPP and DMAPP in the plastid when a high transfer of intermediates between plastid and cytosol take place.

Although we have considered only metabolic genes in this analysis, the method can be extended to identify genes encoding other types of proteins belonging to the same transcription module. In fact, transcription factors and other regulator proteins, as well as structural proteins such as transporters have been shown to be often found in the same expression module [22]. Our results suggest that the expression of genes belonging to the chlorophyll and carotenoid biosynthesis pathways is controlled by a module possibly including genes from the MEP pathway.

Vice versa, the expression of genes in the phytosterol pathway appears to be influenced by genes from the MVA pathway. For the downstream regulation of plastoquinone biosynthesis, however, genes from both pathways seem to be involved. This finding stands in line with the dual localization of enzymes from the plastoquinone in either the plastid or the cytosol. The regulation of this pathway may therefore be dependent on processes happening on the metabolic and regulatory level in both compartments.

We have shown in a simulation study that for gene expression data with many genes and few observations, the modified GGM approaches have a better performance in recovering conditional dependence structures than conventional graphical gaussian modeling. However, a final evaluation of our inferred network for the isoprenoid biosynthesis pathways in *Arabidopsis thaliana* can only be conducted based on additional knowledge and biological experiments. At this stage, the use of domain knowledge has provided some means for network validation. Since genes from the respective downstream pathways were significantly more often attached to the isoprenoid network than candidate genes from other pathways, we are quite confident that our method can grasp modularity in the dependence structure within groups of genes and also between groups of genes. Such modularity would have been difficult to detect by standard graphical modeling or clustering.

Methods

Graphical gaussian models

Let q be the number of genes in the network, n be the number of observations for each gene. The vector of log scaled gene expression values, $Y = (Y_1, \ldots, Y_q)$ is assumed to follow a multivariate normal distribution $N(\mu, \Sigma)$ with mean $\mu = (\mu_1, \ldots, \mu_q)$ and covariance matrix Σ. The partial correlation coefficients $\rho_{ij|rest}$ which measure the correlation between genes i and j conditional on all other genes in the model are calculated as

$$\rho_{ij|rest} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}},$$

where ω_{ij}, $i, j = 1, \ldots, q$ are the elements of the precision matrix $\Omega = \Sigma^{-1}$.

Using likelihood methods, each partial correlation coefficients $\rho_{ij|rest}$ can be estimated and tested against the null hypothesis $\rho_{ij|rest} = 0$ [1]. An edge between genes i and j is drawn if the null hypothesis is rejected. Since the estimation of the partial correlation coefficients involves matrix inversion, estimators are very sensitive to the rank of the matrix. If the model comprises many genes, estimates are only reliable for a large number of observations.

Commonly, the modeling of the graph is carried out in a stepwise backward manner starting from the full model from which edges are removed consecutively. The process stops when no further improvement can be achieved by removal of an additional edge. The final model is usually evaluated by bootstrapping to exclude spurious edges in the model.

Modified GGM approaches

Let i, j be a pair of genes. The sample Pearson's correlation coefficient σ_{ij} is the commonly used measure for coexpression. For examining possible effects of other genes k on σ_{ij}, we consider GGMs for all triples of genes i, j, k with $k \neq i, j$. For each k, the partial correlation coefficient $\rho_{ij|k}$ is computed and compared to σ_{ij}. If the expression level of k is independent of i and j, the partial correlation coefficient would not differ from σ_{ij}. If on the other hand, the correlation between i and j is caused by k since k co-regulates both genes, one would expect $\rho_{ij|k}$ to be close to 0. Here, we use the terminology, that k "explains" the correlation between i and j.

In order to combine the different $\rho_{ij|k}$ values in a biologically and statistically meaningful way, we define an edge between i and j if $\rho_{ij|k} \neq 0$ for all remaining genes k. In particular, if there is at least one k with $\rho_{ij|k} = 0$, no edge between i and j is drawn since the correlation between i and j may be the effect of k. Our approach can be implemented as a frequentist approach in which each edge is tested for presence or absence or alternatively, as a likelihood approach with parameters θ_{ij} which describe the probability for an edge between i and j in a latent random graph.

Frequentist approach

For the gene pair i, j and all remaining genes k, p-values $p_{ij|k}$ are obtained from the likelihood ratio test of the null hypothesis $\rho_{ij|k} = 0$. In order to combine the different p-values $p_{ij|k}$, we simply test whether a third gene k exists that "explains" the correlation between i and j. For this purpose, we apply the following procedure:

1) For each pair i, j form the maximum p-value

$$p_{ij,\max} = \max\{p_{ij|k}, k \neq i, j\}.$$

2) Adjust each $p_{ij,\max}$ according to standard multiple testing procedures such as FDR [28].

3) If the adjusted $p_{ij,\max}$-value is smaller than 0.05 draw an edge between the genes i and j, otherwise omit it.

The correction for multiple testing in step 2 is carried out with respect to the possible number of edges $\frac{q(q-1)}{2}$ in the model. Implicitly, multiple testing over all genes k is also involved in step 1. However, since the maximum over all $p_{ij|k}$ is considered, a multiple testing correction is not necessary.

Latent random graph approach

The frequentist approach has the disadvantage that a connection between two genes i and j is either considered to be present or absent. Also, it is not taken into account whether an edge between i and k respectively j and k is truly present when we test for $\rho_{ij|k} = 0$. In our second method, we introduce a parameter θ_{ij} as the probability for an edge between two genes i and j in a latent random graph model. Let θ be the parameter vector of θ_{ij} for all $1 \leq i < j \leq q$ and $y = (y^1, \ldots, y^n)$ be a sample of n observations. For estimating θ, we maximize the log-likelihood $L(\theta) = \log P_\theta(y)$ via the EM-algorithm [31].

Let θ^t be a current estimate of θ. Further, let g be the unobserved graph encoded as adjacency matrix with $g_{ij} \in \{0, 1\}$ depending on whether there is an edge between gene i and j or not. In the E-step of the EM-algorithm, the conditional expectation of the complete data log-likelihood is determined with respect to the conditional distribution $p(g|y, \theta^t)$,

$$E_\theta(\log P_\theta(g, y)|y, \theta^t) = \sum_g \log P_\theta(g, y) p(g|y, \theta^t). \tag{1}$$

By assuming independence between edges, equation (1) becomes

$$E_\theta(\log P_\theta(g, y)|y, \theta^t) = \sum_g \log P_\theta(g, y) \prod_{i<j} p(g_{ij}|y, \theta^t), \tag{2}$$

and further, after replacing

$$\log P_\theta(g, y) = \sum_{i<j} g_{ij} \log \theta_{ij} + (1 - g_{ij}) \log(1 - \theta_{ij}),$$

and summing out equation (2) we find

$$\begin{aligned} E_\theta(\log P_\theta(g)|y, \theta^t) = & \\ \sum_{i<j} \Big(& P(g_{ij} = 1|y, \theta^t) \log \theta_{ij} + \\ & P(g_{ij} = 0|y, \theta^t) \log(1 - \theta_{ij})\Big). \end{aligned} \tag{3}$$

$P(g_{ij} = 1|y, \theta^t)$ and $P(g_{ij} = 0|y, \theta^t)$ at the right side of equation (3) are approximated by the statistical evidence of edge i, j in GGMs with genes i, j and k. Since we only want to estimate the effect of k on the correlation between i and j, we distinguish only the two cases whether k is a common neighbor of i and j, e.g. $g_{ik} = 1$ and $g_{jk} = 1$, or not. When k is a common neighbor, we test $\rho_{ij|k} \neq 0$ versus $\rho_{ij|k} = 0$. When k is not a common neighbor of i and j, we test $\sigma_{ij} \neq 0$ versus $\sigma_{ij} = 0$ for the pairwise correlation coefficients instead. Thus, we obtain

$$\begin{aligned} P(g_{ij} = 1|y, \theta^t) \approx \prod_{k \neq i,j} \Big(& \theta^t_{ik} \theta^t_{jk} \cdot \hat{P}(\rho_{ij|k} \neq 0|y) + \\ & (1 - \theta^t_{ik} \theta^t_{jk}) \cdot \hat{P}(\sigma_{ij} \neq 0|y)\Big), \end{aligned} \tag{4}$$

where $\hat{P}(\rho_{ij|k} \neq 0|y)$ and $\hat{P}(\sigma_{ij} \neq 0|y)$ are p-values of the corresponding likelihood ratio tests. After replacing (4) in equation (3), the M-step of the EM-algorithm, that is the maximization of $E_\theta(\log P_\theta(g)|y, \theta^t)$ with respect to θ, leads to an iterative updating scheme $\theta^t \to \theta^{t+1}$ with

$$\begin{aligned} \theta^{t+1}_{ij} = \prod_{k \neq i,j} \Big(& \theta^t_{ik} \theta^t_{jk} \cdot P(\rho_{ij|k} \neq 0|y) + \\ & (1 - \theta^t_{ik} \theta^t_{jk}) \cdot P(\sigma_{ij} \neq 0|y)\Big). \end{aligned} \tag{5}$$

In summary, we determine the probability parameters θ as follows

1) For gene pairs i, j, compute $P(\rho_{ij|k} \neq 0)$ and $P(\sigma_{ij} \neq 0)$ for all genes $k \neq i, j$.

2) Starting with θ^0, apply iteratively equation (5) until the error $|\theta^{t+1} - \theta^t|$ drops below a pre-specified value, for example 10^{-6}.

Our latent random graph approach also enables us to fit a large number of additional genes into a constructed genetic network. In this case, for a gene pair i, j in step 1 of the analysis, the partial correlation coefficients $\rho_{ij|k}$ are not only computed and tested for genes k in the model but also for the additional candidate genes. However, the iteration in step 2 is not extended to these candidate genes. In other words, θ_{ij} is only iteratively updated in equation (5) if both genes i, j are in the original model. For candidate genes k, θ_{ik} and θ_{jk} are kept fixed at a pre-specified value, e.g. 1, and are not re-estimated in the EM-iteration process.

This outline introduces a second level into the modeling process. At the first level, the network between the original genes is constructed. At the second level, we test how additional candidate genes influence the parameters θ. If these candidates have an effect on the correlation between i and j, θ_{ij} will decrease. Thus, by comparing the original network with the network inferred from allowing for additional genes in step 1, we can determine which candidate genes lower the θ-values and, accordingly, fit well into the network.

References

1. Edwards D: *Introduction to Graphical Modelling*. Springer Verlag; 2nd edition 2000.
2. Lauritzen S: *Graphical Models*. Oxford University Press 1996.
3. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7**(3-4):601–620.
4. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: **Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.** In *Pac Symp Biocomput*, PSB01 2001:422–433.
5. Toh H, Horimoto K: **Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling.** *Bioinformatics* 2002, **18**(2):287–297.
6. Wang J, Myklebost O, Hovig E: **MGraph: graphical models for microarray data analysis.** *Bioinformatics* 2003, **19**(17):2210–2211.
7. Husmeier D: **Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks.** *Bioinformatics* 2003, **19**(17):2271–2282.
8. Waddell PJ, Kishino H: **Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data.** *Genome Informatics* 2000, **11**:129–140.
9. Friedman N, Nachman I, Pe'er D: **Learning Bayesian network structure from massive datasets: The "Sparse Candidate" algorithm**. In *Proc Fifteenth Conf on Uncertainty in Artificial Intelligence*, UAI 1999:206–215.
10. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, **292**(5518):929–934.
11. Rodriguez-Concepcion M, Boronat A: **Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics.** *Plant Physiol* 2002, **130**(3):1079–1089.
12. Laule O, Fürholz A, Chang HS, Zhu T, Wang X, Heifetz PB, Gruissem W, Lange M: **Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in Arabidopsis thaliana.** *Proc Natl Acad Sci U S A* 2003, **100**(11):6866–6871.
13. Rodriguez-Concepcion M, Fores O, Martinez-Garcia JF, Gonzalez V, Phillips M, Ferrer A, Boronat A: **Distinct light-mediated pathways regulate the biosynthesis and exchange of isoprenoid precursors during Arabidopsis seedling development.** *Plant Cell* 2004, **16**:144–156.
14. Bick JA, Lange BM: **Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane.** *Arch Biochem Biophys* 2003, **415**(2):146–154.
15. Kasahara H, Hanada A, Kuzuyama T, Takagi M, Kamiya Y, Yamaguchi S: **Contribution of the mevalonate and methylerythritol phosphate pathways to the biosynthesis of gibberellins in Arabidopsis.** *J Biol Chem* 2002, **277**(47):45188–45194.
16. Nagata N, Suzuki M, Yoshida S, Muranaka T: **Mevalonic acid partially restores chloroplast and etioplast development in Arabidopsis lacking the non-mevalonate pathway.** *Planta* 2002, **216**(2):345–350.
17. Hemmerlin A, Hoeffler JF, Meyer O, Tritsch D, Kagan IA, Grosdemange-Billiard C, Rohmer M, Bach TJ: **Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells.** *J Biol Chem* 2003, **278**(29):26666–26676.
18. Lange B, Ghassemian M: **Genome organization in Arabidopsis thaliana: a survey for genes involved in isoprenoid and chlorophyll metabolism.** *Plant Mol Biol* 2003, **51**(6):925–948.
19. Schwarz G: **Estimating the dimension of a model**. *Ann of Statistics* 1978, **6**:461–464.
20. **MIM 3.1 Student Version**[http://www.hypergraph.dk].
21. Friedman N, Goldszmidt M, Wyner A: **Data Analysis with Bayesian Networks: A Bootstrap Approach**. In *Proc Fifteenth Conf on Uncertainty in Artificial Intelligence*, UAI 1999:196–205.
22. Ihmels J, Levy R, Barkai N: **Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae.** *Nat Biotechnol* 2004, **22**:86–92.
23. Lichtenthaler HK, Schwender J, Disch A, Rohmer M: **Biosynthesis of isoprenoids in higher plant chloroplasts proceeds via a mevalonate-independent pathway.** *FEBS Lett* 1997, **400**(3):271–274.
24. Arigoni D, Sagner S, Latzel C, Eisenreich W, Bacher A, Zenk MH: **Terpenoid biosynthesis from 1-deoxy-D-xylulose in higher plants by intramolecular skeletal rearrangement.** *Proc Natl Acad Sci U S A* 1997, **94**(20):10600–10605.
25. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296**(5569):910–913.
26. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**(6804):651–654.
27. Spirtes P, Glymour C, Scheines R: *Causation, Prediction, and Search*. MIT Press, 2nd edition 2000.

28. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.**

29. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290**(5500):2306–2309.

30. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16**(10):939–945.

31. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm**. *Journal of the Royal Statistical Society Series B* 1977, **39**:1–38.

32. **TargetP prediction of subcellular location**[http://www.cbs.dtu.dk/services/TargetP/].

Figures

Figure 1

Bootstrapped GGM of the isoprenoid pathway. a) comparison between absolute pairwise correlation coefficients and presence of edges (0 and 1 denote absent and present edges respectively), b) histogram of the bootstrap edge probabilities, c) comparison between absolute pairwise correlation coefficients and bootstrap edge probabilities for all 780 possible edges.

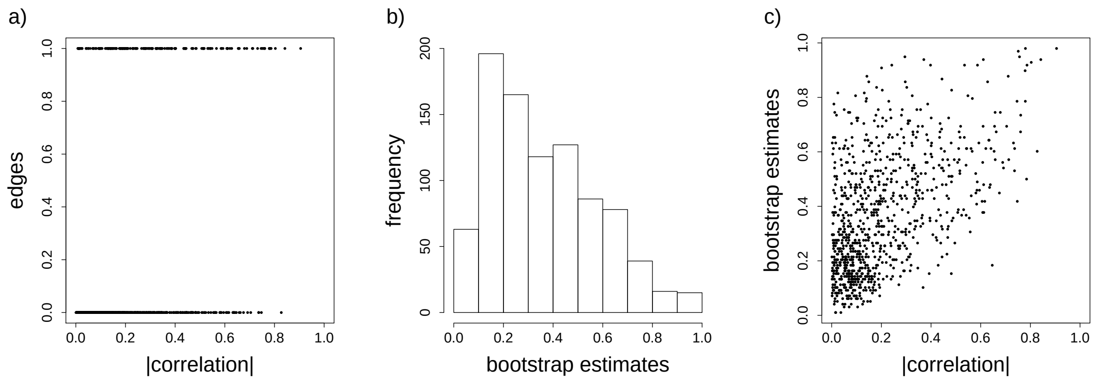


Figure 2

Bootstrapped GGM of the isoprenoid pathway (cutoff 0.8). Dotted directed edges mark the metabolic network and are not part of the GGM.

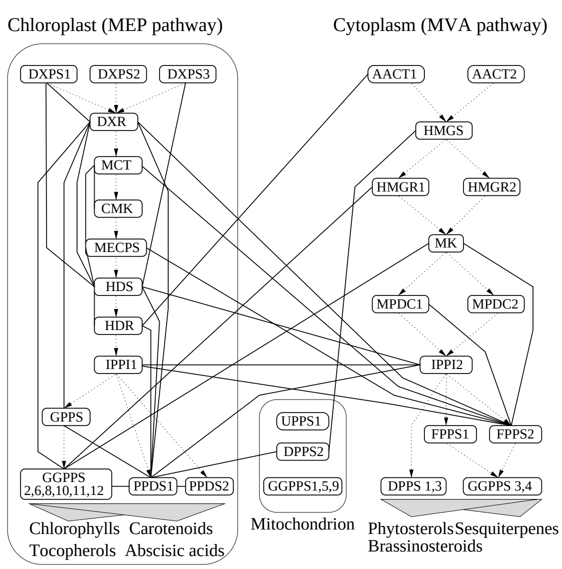


Figure 3
Dependencies between genes of the isoprenoid pathways according to the frequentist modified GGM method. Upper panel: subgraph of the gene module in the MEP pathway, lower panel: subgraph of the gene module in the MVA pathway.

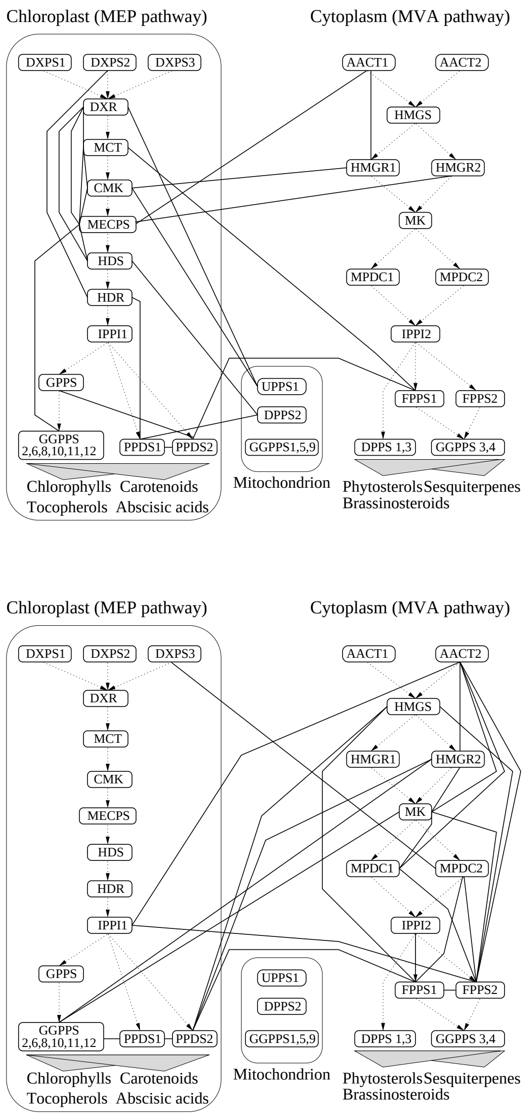


Figure 4

Comparison of the absolute pairwise correlation coefficients and the modifide GGM approaches. a) selected edges in the frequentist modified GGM approach (0 and 1 denote absent and present edges respectively), b) θ-values in the latent random graph approach, c) θ-values after attaching 795 genes from other pathways.

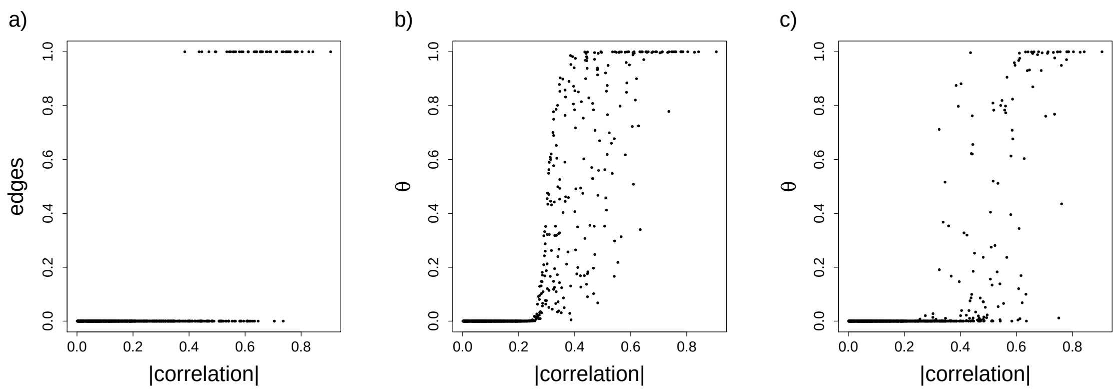


Figure 5

Hierarchical clustering of 40 genes involved in the isoprenoid pathway and 795 genes from other pathways. Positions of the genes from the MEV pathway (m) and the plastoquinone and phytosterol pathways (+) are indicated in the left column of the axis on the right side. The positions of the genes from the MEP pathway (n) and the plastoquinone, carotenoid and chlorophyll pathways (+) are indicated in right column of the axis.

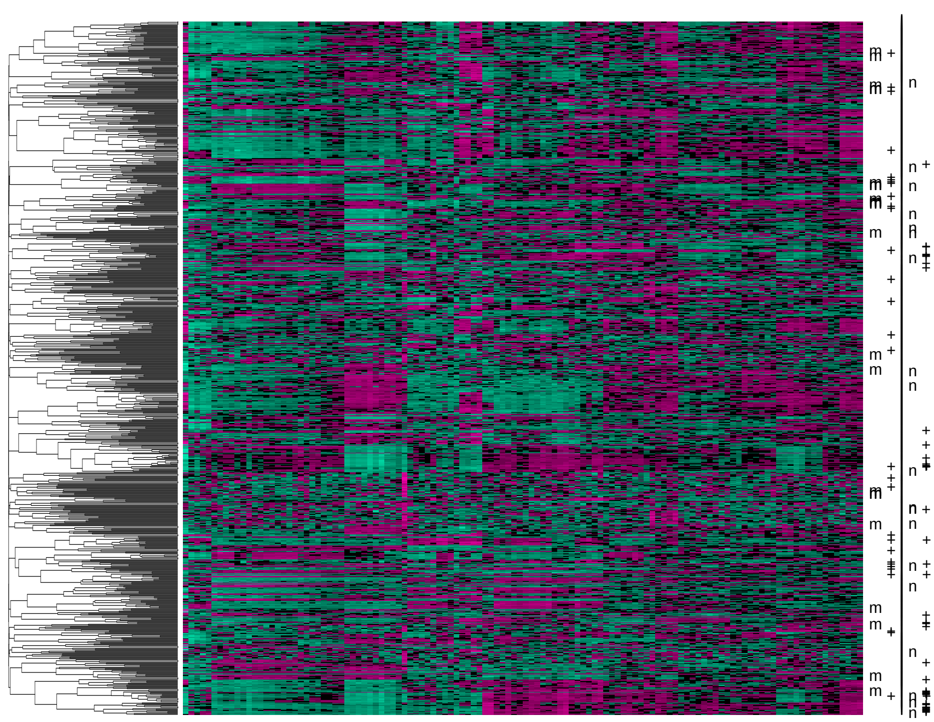
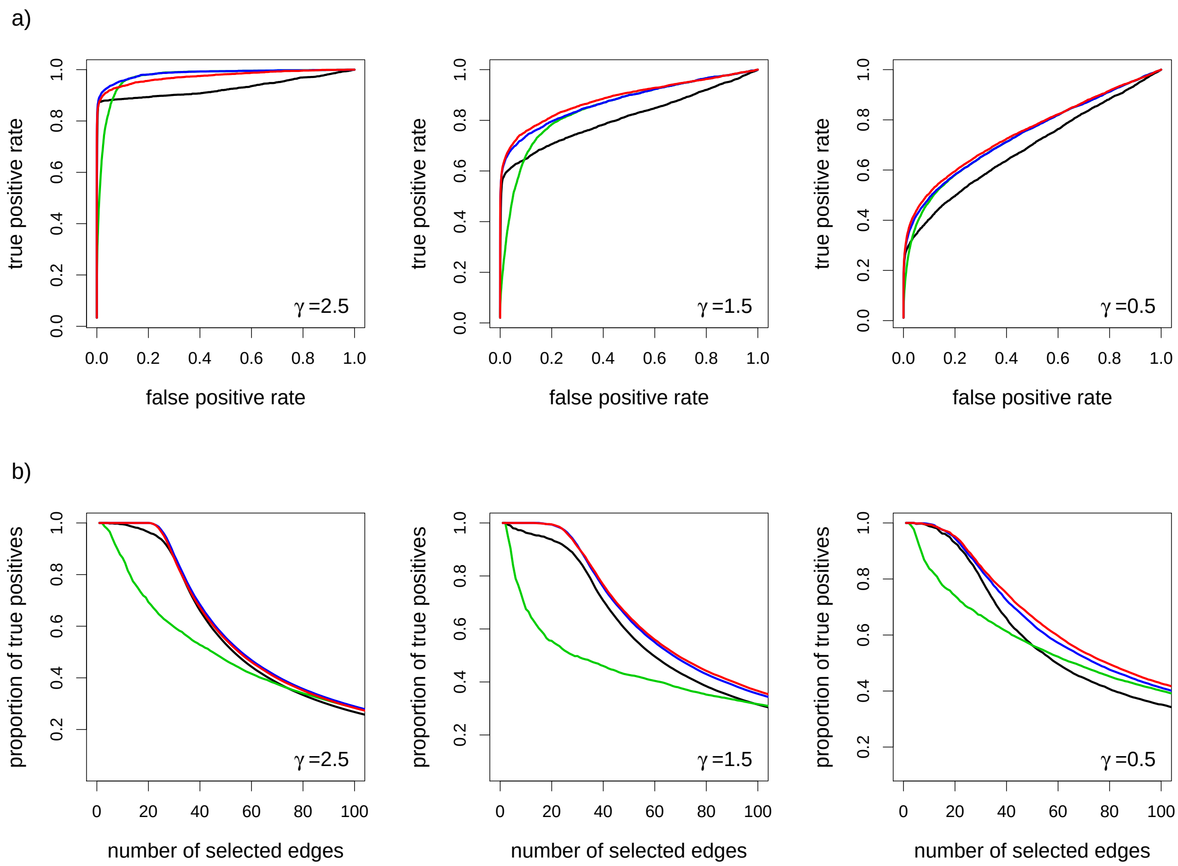


Figure 6

Performance of different GGM approaches. a) ROC curves and b) proportion of true positive edges as a function of the number of selected edges for the different graphical modeling strategies. Black line: the standard GGM, red line: frequentist modified GGM approach, blue line: latent random graph modified GGM approach, green line: pairwise correlation. Sparse networks with fewer edges as nodes ($\gamma = 2.5$) are represented in the left column, networks with approximately as many edges as nodes ($\gamma = 1.5$) are represented in the middle column, and networks with approximately twice as many edges as nodes ($\gamma = 0.5$) are in the right column.



Tables

Table 1

Genes coding for enzymes in the two isoprenoid pathways. Subcellular locations are pooled from experimental data, the TargetP data base [32], and [18]: C - cytoplasm, ER - endoplasmic reticulum, M - mitochondrion, P - chloroplast. Experimentally verified subcellular locations are marked with an asterisk(*).

Name	AGI number	Subcellular location
AACT1	At5g47720	C
AACT2	At5g48230	C
CMK	At2g26930	P
DPPS1	At2g23410	C/ER
DPPS2	At5g58770	M
DPPS3	At5g58780	ER
DXPS1	At3g21500	P
DXPS2	At4g15560	P*
DXPS3	At5g11380	P
DXR	At5g62790	P*
FPPS1	At4g17190	C
FPPS2	At5g47770	C/M*
GGPPS1	At1g49530	M*
GGPPS2	At2g18620	P
GGPPS3	At2g18640	C/ER*
GGPPS4	At2g23800	C/ER*
GGPPS5	At3g14510	M
GGPPS6	At3g14530	P
GGPPS7	At3g14550	P*
GGPPS8	At3g20160	C/ER
GGPPS9	At3g29430	M
GGPPS10	At3g32040	P
GGPPS11	At4g36810	P*
GGPPS12	At4g38460	P
GPPS	At2g34630	P*
HDR	At4g34350	P
HDS	At5g60600	P*
HMGR1	At1g76490	C/ER*
HMGR2	At2g17370	C/ER*
HMGS	At4g11820	C
IPPI1	At3g02780	P
IPPI2	At5g16440	C
MCT	At2g02500	P*
MECPS	At1g63970	P
MK	At5g27450	C
MPDC1	At2g38700	C
MPDC2	At3g54250	C
PPDS1	At1g17050	P
PPDS2	At1g78510	P
UPPS1	At2g17570	M

Table 2

Pathways whose genes attach significantly well to the isoprenoid pathways. Downstream pathways are marked with an asterisk (*). The calvin cycle is also metabolically linked to the isoprenoid pathways.

both isoprenoid pathways	MEP pathway	MVA pathway
plastoquinone*	plastoquinone*	plastoquinone*
carotenoid*	carotenoid*	phytosterol*
calvin cycle	porphyrin/chlorophyll*	
histidine	one carbon pool	
one carbon pool	calvin cycle	
tocopherol*		
porphyrin/chlorophyll*		