BMC Proceedings

Research

Open Access

Gene network reconstruction from microarray data

Florence Jaffrezic*[1] and Gwenola Tosser-Klopp[2]

Address: [1]INRA AgroParisTech, Animal Genetics and Integrative Biology, Populations Statistics Genomes, 78350 Jouy-en-Josas, France and [2]Laboratoire de Génétique Cellulaire, INRA, UMR444, F-31326 Castanet-Tolosan, France

Email: Florence Jaffrezic* - florence.jaffrezic@jouy.inra.fr; Gwenola Tosser-Klopp - gwenola.tosser@toulouse.inra.fr

* Corresponding author

from EADGENE and SABRE Post-analyses Workshop
Lelystad, The Netherlands. 12–14 November 2008

Published: 16 July 2009

BMC Proceedings 2009, **3**(Suppl 4):S12 doi:10.1186/1753-6561-3-S4-S12

This article is available from: http://www.biomedcentral.com/1753-6561/3/S4/S12

Abstract

Background: Often, software available for biological pathways reconstruction rely on literature search to find links between genes. The aim of this study is to reconstruct gene networks from microarray data, using Graphical Gaussian models.

Results: The *GeneNet* R package was applied to the Eadgene chicken infection data set. No significant edges were found for the list of differentially expressed genes between conditions MM8 and MA8. On the other hand, a large number of significant edges were found among 85 differentially expressed genes between conditions MM8 and MM24.

Conclusion: Many edges were inferred from the microarray data. Most of them could, however, not be validated using other pathway reconstruction software. This was partly due to the fact that a quite large proportion of the differentially expressed genes were not annotated. Further biological validation is therefore needed for these networks, using for example in vitro invalidation of genes.

Introduction

Two main approaches have been proposed in the literature for gene network reconstruction from microarray data, namely Bayesian networks and Graphical Gaussian models. Bayesian networks are directed acyclic graphs, i.e. no feedback loop is possible. They are usually very computationally intensive and, as far as we are aware of, no R package is available for large-scale gene network reconstruction using Bayesian networks. On the other hand, Graphical Gaussian models are undirected graphs and are very computationally efficient. An R package is available for gene network reconstruction from microarray data using Graphical Gaussian models, namely *GeneNet* [1].

Werhli et al. [2] presented a comparison study between Bayesian networks and Graphical Gaussian models for gene network reconstruction. They concluded that both methods provided quite similar results for network reconstruction based on observed microarray data. We therefore chose, in this study, to base inference on Graphical Gaussian models.

Graphical Gaussian models

Let $\boldsymbol{X}$ be the observed data matrix with N rows, corresponding to the number of samples, and G columns, corresponding to the number of genes. $\boldsymbol{X}$ is supposed to follow a multivariate normal distribution $\mathcal{N}_G(\mu, \Sigma)$, with mean vector $\mu = (\mu_1, ..., \mu_G)'$ and positive-definite covariance matrix $\Sigma = (\sigma_{ij})_{(1 \le i,j \le G)}$.

Covariance parameters σ_{ij} can also be written as: $\sigma_{ij} = \rho_{ij}\text{-}\sigma_i\sigma_j$, where σ_i^2 and σ_j^2 are the variance terms for genes i and j, respectively. Parameter ρ_{ij} corresponds to the Pearson correlation coefficient between genes i and j.

Let $\boldsymbol{P}$ be the Pearson correlation matrix: $\boldsymbol{P} = (\rho_{ij})_{(1 \le i,j \le G)}$. A high correlation coefficient between two genes may indicate either [3]: *i*) a direct interaction between genes *i* and *j*; ii) an indirect interaction between these two genes; iii) a regulation of the two genes by a common gene. For network reconstruction we are only interested in direct interactions, represented by the partial correlation matrix $\Pi = (\pi_{ij})_{(1 \le i,j \le G)}$. Coefficient π_{ij} represents the correlation between two genes i and j conditionally on all the other genes. It can be shown [3] that partial correlation matrix Π is related to the inverse of the covariance matrix Σ as follows:

$$\pi_{ij} = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}} \qquad (1)$$

with $\Sigma^{-1} = (\omega_{ij})$, for $1 \le i, j \le G$.

Several steps are required for the construction of a Graphical Gaussian model network. First, the empirical covariance matrix has to be estimated:

$$\hat{S} = \frac{1}{N-1}(X - \bar{X})'(X - \bar{X})$$

Second, the partial correlation matrix has to be calculated using the previous equations. Finally, statistical tests can be performed to determine the partial correlation coefficients that are different from 0, and which correspond to the significant edges of the graph.

The procedure described above is, however, only applicable when sample size N is larger than the number of variables G. In fact, the sample covariance matrix is otherwise not positive-definite and cannot be inverted, which prevents a direct computation of the partial correlation matrix. In microarray experiments, however, we are very often in situations where sample size N is much smaller than the total number of genes G.

Schäfer and Strimmer [1] therefore proposed to use a shrunk estimate of the covariance matrix using a James-Stein estimator. The aim of this approach is to construct a well conditioned positive-definite matrix so that the matrix has full rank and can easily be inverted.

Let λ be a shrinkage coefficient ($\lambda \in [0, 1]$). The shrunk covariance matrix Σ^* is obtained as:

$$\boldsymbol{\Sigma}^* = \lambda \boldsymbol{T} + (1 - \lambda)\boldsymbol{S} \qquad (2)$$

where $\hat{S}$ is the estimated empirical covariance matrix. Shrinkage parameter λ is chosen to minimize the mean-squared error (MSE) and can be determined analytically [1].

There are several possibilities for the choice of matrix $\boldsymbol{T}$. Schäfer and Strimmer [1] recommend for gene network reconstruction to shrink the correlation terms towards zero and to leave the diagonal terms as estimated by the empirical variances. In this case, shrinkage parameter λ can be estimated analytically as:

$$\hat{\lambda} = \frac{\sum_{i \neq j} \widehat{\mathrm{Var}}(s_{ij})}{\sum_{i \neq j} s_{ij}^2} \qquad (3)$$

where s_{ij} are the empirical covariance parameters.

An edge-specific local FDR procedure was then defined, based on the estimated partial correlation coefficients. As recommended by Efron [4], an edge is considered significant if its local FDR value is smaller than 20%.

Application

The *GeneNet* R package [1] was applied to the Eadgene chicken infection data set [5] and the R code used to produce these analyses is available from the first author. We considered here the lists of differentially expressed genes obtained for two sets of conditions. In condition MA, chickens were infected at two weeks of age with a parasite called Eimeria maxima and two weeks later with the parasite called Eimeria acervulina, and in condition MM, chickens were infected first with E. maxima and afterwards with the same parasite E. maxima. Two time points were sampled post infection: 8 hours and 24 hours. At a 5% Benjamini-Hochberg (BH) threshold, 85 genes were found differentially expressed between groups MM and MA at 8 hours post infection, whereas 800 genes were found differentially expressed at a 5% BH threshold for condition MM between the two time points 8 and 24 hours. Due to the quite small number of biological replicates per condition (5 animals), network inference can only be performed on a few dozens of genes. For condi-

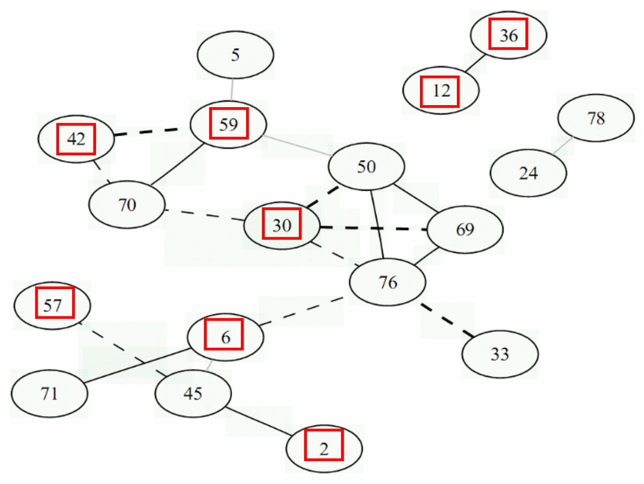

Figure 1
Gene network for condition MM8. Gene network for condition MM8 obtained with the *GeneNet* R package, for the 20 most significant edges. Solid lines represent positive relationships, dotted lines are negative relationships. The line intensities represent the strength of the relationships. Bold lines are stronger and light grey lines are weaker. Red squares represent annotated genes.

tions MM8 and MM24, we therefore considered a more stringent BH threshold, with 116 differentially expressed genes at a 1% Benjamini-Hochberg threshold.

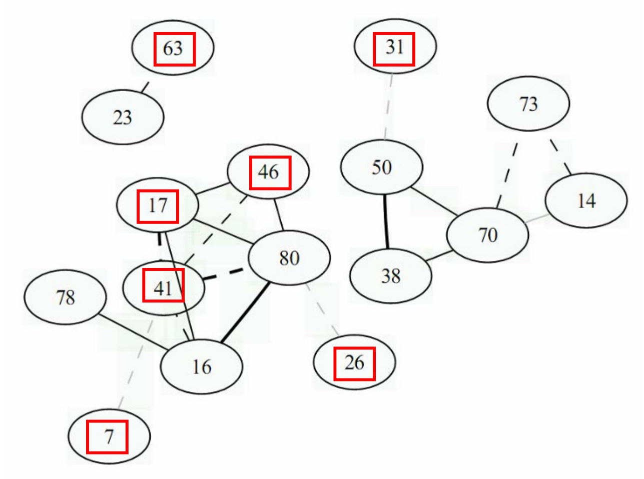

Figure 2
Gene network for condition MM24. Gene network for condition MM24 obtained with the *GeneNet* R package, for the 20 most significant edges. Legend for this graph is the same as for Figure 1.

As no missing values are allowed in *GeneNet*, 58 genes were used for network reconstruction among the list of differentially expressed genes between conditions MM8 and MA8. Network inference was performed using the expression values for each condition independently. For this first analysis, no significant edges were found at the recommended 20% local FDR [4] threshold, for either condition MM8 or MA8.

For conditions MM8 and MM24, as no missing values are allowed, network reconstruction was based on 85 genes among the 116 found differentially expressed at a 1% Benjamini-Hochberg threshold.

For expression values observed in condition MM8, 2356 edges were found significant at the 20% local FDR threshold among the 85 genes, and even 1964 edges were found significant at the more stringent 5% local FDR threshold. Similarly, a very large number of edges were found significant between these 85 genes for condition MM24. In fact, 1760 edges were found significant at the 20% local FDR threshold, and 1156 at the 5% local FDR threshold.

Figures 1 and 2 show the graphs of the 20 most significant edges for both conditions, and Tables 1 and 2 provide the correspondence between the gene numbers given in the figures and their RIGG names and Human orthologs. It can be seen that there is very little overlapping between the two networks. In fact, only three genes were found in common in both graphs, and no link was conserved between both graphs. Furthermore, for condition MM8, among the 18 genes present in the graph, only 8 were annotated and for condition MM24, only 7 genes were annotated, which made it very difficult to validate the links found here using literature based pathway reconstruction software. Among the annotated genes present in these graphs, no links were found between them using either Ingenuity or Pathway Studio. Further biological validation is therefore required for this experiment using, for example, in vitro invalidation of genes, in order to confirm the links inferred here based on the gene expression measurements.

Table 1: Names of the annotated genes for condition MM8.

Gene number	RIGG name	Human ortholog HGNC
2	RIGG00270	C9orf80
6	RIGG01146	C6orf106
12	RIGG02317	SLC30A6
30	RIGG07776	SH2B2
36	RIGG09586	EFR3B
42	RIGG11710	SGK2
57	RIGG15302	GOLGB1
59	RIGG15630	PPP1R12A

Table 2: Names of the annotated genes for condition MM24.

Gene number	RIGG name	Human ortholog HGNC
7	RIGG01243	BAZ1B
17	RIGG03141	AP000775.4
26	RIGG07311	PRKCQ
31	RIGG08254	ABHD6
41	RIGG11399	C15orf27
46	RIGG12572	CCNY
63	RIGG15936	DTX4

Discussion

Gene network reconstruction was based here on the expression data from this experiment only. It would be interesting to integrate some prior biological knowledge such as gene relationships already found in the literature, or to combine expression values from several studies, to have more power and accuracy for the edge detection.

Biological validation of the edges inferred here with Graphical Gaussian models was very difficult, mainly due to the lack of annotation for the lists of differentially expressed genes. An important effort therefore has to be made in the near future to obtain a more complete annotation of the chicken genome and other livestock species.

In the approach presented here, and based on partial correlations, it is only possible to model linear dependencies between genes. In order to take into account non linear relationships it may be possible, as suggested by Hausser and Strimmer [6] to use entropy instead of partial correlations to infer the edges between genes.

As only two time points were available in this study, static networks were considered here using the expression values at each time point separately. Several methods have recently been proposed for gene network reconstruction in time course studies, mainly based on VAR1 models [7]. If additional time points were added in the future to this experiment, it would be interesting to use these methods to study the gene relationships over time.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

F. Jaffrezic was in charge of the statistical analysis of the data and G. Tosser-Klopp of the biological interpretation of the results.

Acknowledgements

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 4, 2009: EADGENE and SABRE Post-analyses Workshop. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S4.

References

1. Schäfer J, Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Stat Appl Genet Mol Biol* 2005, **4:**32.
2. Werhli AV, Grzegorczyk M, Husmeier D: **Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian models.** *Bioinformatics* 2006, **22:**2523-2531.
3. Whittaker J: *Graphical Models in Applied Multivariate Statistics New York: Wiley*; 1990.
4. Efron B: **Local false discovery rates.** *Technical report Department of Statistics, Stanford University*; 2005.
5. Swinkels W, Cornelissen J, Rebel A: **Immune reactions after a homologous or heterologous challenge of broilers primed with Eimeria maxima.** 2009 in press.
6. Hausser J, Strimmer K: **Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks.** 2009 in press.
7. Opgen-Rhein R, Strimmer K: **Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process.** *BMC Bioinformatics* 2007, **8(Suppl 2):**S3.