Gene network reconstruction from microarray data

F. Jaffezic (1) and G. Tosser-Klopp (2)
1 INRA, Jouy-en-Josas, France
2 INRA, Toulouse, France

Gene network reconstruction from microarray data

F. Jaffrézic[1] and G. Tosser-Klopp[2]

[1]INRA, Jouy-en-Josas, France - [2] INRA, Toulouse, France

Eadgene workshop, November 2008

Outline

Outline

Introduction

Gene network reconstruction from microarray data.

Two main approaches :

1) Bayesian networks.

2) Graphical Gaussian models.

Introduction

Bayesian networks :

- Directed acyclic graph.

$\Rightarrow$ No feedback loop available

- Very computationally intensive.
- No R package available ?

Introduction

Graphical Gaussian models :

- Undirected graphs.
- Very computationally efficient.
- R package GeneNet (K. Strimmer)
Schäfer and Strimmer (2005).

Introduction

Wehrli et al. (*Bioinformatics*, 2006) : Comparison study between Bayesian Networks and Graphical Gaussian models.

Both methods were found to have quite similar performances for gene network reconstruction from microarray data.

For simplicity of implementation ⇒ Graphical Gaussian models.

Outline

Graphical Gaussian models

Let $\mathbf{X}$ be the observed data matrix with N rows (samples) and G columns (genes).

$\mathbf{X}$ is supposed to be drawn from a multivariate normal distribution $\mathcal{N}_G(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_G)'$ and positive-definite covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{(1 \leq i,j, \leq G)}$.

Graphical Gaussian models

Covariance parameters σ_{ij} can also be written as :

$$\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$$

where σ_i^2 and σ_j^2 are the variance terms for genes i and j, respectively and ρ_{ij} is the Pearson correlation coefficient between genes i and j.

Graphical Gaussian models

Let $\mathbf{P}$ be the Pearson correlation matrix $\mathbf{P} = (\rho_{ij})_{(1 \leq i,j, \leq G)}$.

A high correlation coefficient ρ_{ij} between two genes may indicate either :

1) a direct interaction between genes i and j.
2) an indirect interaction between these two genes.
3) a regulation of the two genes by a common gene.

Graphical Gaussian models

For network reconstruction, we are only interested in direct interactions, represented by the partial correlation matrix $\boldsymbol{\Pi} = (\pi_{ij})$.

Coefficient π_{ij} represents the correlation between two genes i and j conditioned on all the other genes.

Graphical Gaussian models

It can be shown that partial correlation matrix $\mathbf{\Pi}$ is related to the inverse of the covariance matrix $\mathbf{\Sigma}$ as follows :

$$\pi_{ij} = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}}$$

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

Graphical Gaussian models

Construction of a GGM network.

1) Estimate the empirical covariance matrix $\hat{\mathbf{S}} = (\hat{\sigma}_{ij}) = \frac{1}{N-1}(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})$.

2) Estimate the partial correlation matrix using the previous equations.

3) Statistical tests to determine the partial correlation coefficients significantly different from 0, which correspond to the significant edges of the graph.

Graphical Gaussian models

Procedure only applicable when sample size N larger than the number of variables G, otherwise the sample covariance matrix is not positive-definite and cannot be inverted, which prevents a direct computation of the partial correlation matrix.

In microarray analyses, $N << G$ and small sample size N.

Graphical Gaussian models

Shrunk estimate of the covariance matrix using a James-Stein estimator (Schäfer and Strimmmer, 2005).

Objectives :

1) Construct a positive-definite matrix (all eigenvalues > 0).

2) Well-conditioned covariance matrix (ratio of its maximum and minimum singular value is not too large).

$\Rightarrow$ The matrix has full rank and can easily be inverted.

Graphical Gaussian models

Shrinkage estimation of the covariance matrix.

Let λ be a shrinkage coefficient ($\lambda \in [0, 1]$)

$\mathbf{\Sigma}^* = \lambda\mathbf{T} + (1 - \lambda)\hat{\mathbf{S}}$ where $\hat{\mathbf{S}}$ is the estimated empirical covariance matrix.

Shrinkage parameter λ : chosen to minimize the mean-squared error (MSE) and determined analytically (Ledoit and Wolf, 2003).

Graphical Gaussian models

Choice for matrix $\mathbf{T}$: Several possibilities.

Schäfer and Strimmer (2005) 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, the shrinkage parameter λ can be estimated analytically as :

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

where s_{ij} are the empirical covariance parameters.

Graphical Gaussian models

Test of the significant edges.

Network structure is usually sparse.

An edge-specific local FDR was defined using the estimated partial correlation coefficients (Schäfer and Strimmer, 2005).

An edge is considered significant if its local FDR is smaller than 0.2 (Efron, 2005).

Outline

Chicken infection data set

GeneNet package.

1. List of DE genes between MM8 and MA8 :

85 differentially expressed genes at a 5% BH threshold.

No missing values allowed in GeneNet ⇒ 58 genes.

Application

Network reconstruction using the expression values for each condition independently.

No significant edges at a 20% local FDR threshold for either MM8 or MA8 expression values.

Application

2. List of DE genes between MM8 and MM24 :

116 differentially expressed genes at a 1% BH threshold.

No missing values ⇒ 85 genes.

Application

Network reconstruction using the expression values for each condition independently.

For MM8 :

2356 significant edges among the 85 genes at a 20% local FDR threshold.

1964 significant edges at a 5% local FDR threshold.

Gene network reconstruction from microarray data

F. Jaffezic (1) and G. Tosser-Klopp (2)
1 INRA, Jouy-en-Josas, France
2 INRA, Toulouse, France

Application

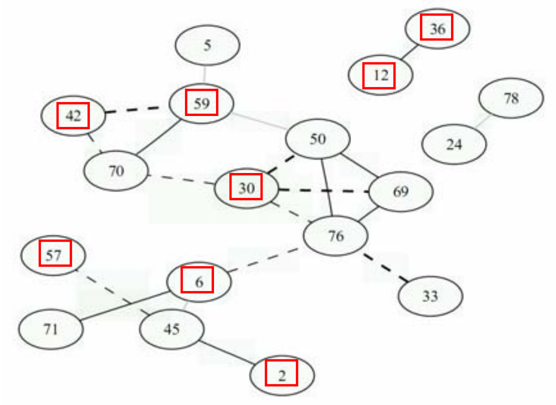
For MM24 :

1760 significant edges among the 85 genes at a 20% local FDR threshold.

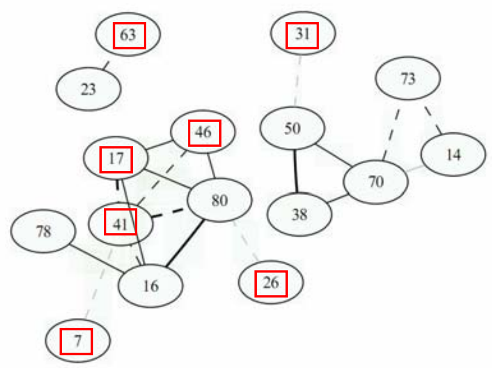
1156 significant edges at a 5% local FDR threshold.

Graph of the 20 most significant edges for both conditions.

Gene network for condition MM8



Gene network for condition MM24



Application

Very little overlapping between the two networks.

Only genes 50, 70 and 78 remain in both graphs.

No link in common between the two graphs.

⇒ Evolution of the gene connections over time.

Biological interpretation

In collaboration with Gwenola Tosser-Klopp (INRA, Toulouse).

For condition MM8 :

Among the 18 genes present in the graph only 8 were annotated.

For condition MM24 :

Among the 16 genes present in the graph only 7 were annotated.

Biological interpretation

For both graphs, the links found here between annotated genes were not confirmed with either Ingenuity or Pathway Studio.

⇒ Further biological validation required (using knock-out experiments, etc.).

Discussion

Gene network reconstruction only based on expression data from this experiment.

Integrate prior biological information :

Gene relationships already found in the literature.

Combine expression data from several studies to have more power and accuracy in the edge detection.

Gene network reconstruction from microarray data

Introduction

Graphical Gaussian models

Application

Discussion

Model only linear dependency between genes.

Non linear relationships taken into account when using entropy instead of partial correlations ?

Other methods and R packages for gene network reconstruction in time course studies :
VAR1 models, state-space models with hidden states.