A modified graphical gaussian model approach for genetic regulatory networks

Anja Wille, SfS
Reverse Engineering Project, ETHZ
Introduction

modified GGMs

biological scenarios

graphical gaussian models (GGMs)
Genetic regulation

39 genes → 118 observations

... signals ...

graphical model to estimate conditional dependence between genes
Graphical Gaussian Models

• undirected

• random variables follow a multivariate normal distribution

• log likelihood:

\[
l(N, S) = \frac{N}{2}(q \ln(2\pi) + \ln |S| + \text{tr}(S) + (\bar{y} S \bar{y})' \bar{y} (\bar{y} S \bar{y}))
\]

for a sample of N observations \(y^1, \ldots, y^N\) with sample mean \(\bar{y}\), sample covariance matrix \(S\) and precision matrix \(\tilde{\Psi} = \Psi^{-1}\)

• partial correlation coefficient \(\tilde{\Psi}_{ij|\text{rest}}\) for gene pair \(ij\)
Bootstrapping and pairwise correlation
Application to isoprenoid pathways

Isoprenoid pathways

Chloroplast

Mitochondria

Cytosol

Phytosterols

Sesquiterpenes

Chlorophylls

Carotenoids

DXPS1

DXPS3

DXPS

DXR

MCT

CMK

MECP

IPPI

GPPS

GGPPS

AACT

HMGR

MK

MPDC

FPPS

GGPPS

DPPS
Problems

• matrix inversion rank-sensitive
• genes $\uparrow$ exponentially increasing number of models
  □ difficult to interpret
• how to interpret high partial correlation accompanied with low pairwise correlation
• efficient procedure for attaching new genes needed
Outline for modified GGM approach

For each pair of genes i,j:

• take pairwise correlation into account

• fit GGMs with gene triples i,j,k for all remaining genes k to study the partial correlation

• combine GGMs and pairwise correlation for inference on edge ij

• 3 methods that differ with respect to statistical framework and computational costs
Frequentist approach

Focus on genepair $ij$
• $p_{ijkl}$ is $p$-value from deviance test $\square_{ijkl} \neq 0$ versus $\square_{ijkl} = 0$
Frequentist approach (cont’d)

1) Form \( p_{ij,\text{max}} = \max (p_{ijkl} \text{ for all genes } k \neq i,j) \)
2) Adjust \( p_{ij,\text{max}} \) according to Bonferroni-Holm or FDR
3) If the adjusted value \( p_{ij,\text{max}} < 0.05 \), draw edge between \( i \) and \( j \)
Application to isoprenoid pathway

Isoprenoid pathways

Chloroplast

- DXP53
- DXP51
- DXP52
- DXP54
- DXR
- MCT
- CMK
- MECPS
- IPP11
- GPPS
- GPPS(P)
- GPPS(O)
- GGPPS11
- GGPPS6
- GGPPS2
- GGPPS1
- GPPR
- PS

Cytosol

- AACT1
- AACT2
- HMGS
- HMGR1
- HMGR2
- MK
- MPDC2
- IPP12
- FPPS1
- FPPS2
- GGPPS3
- GGPPS4

Mitochondria

- DPPS
- GGPPS1

Phytosterols

Sesquiterpenes

Chlorophylls

Carotenoids
Likelihood approach with parameters $\square$

Estimate $\square = \{\square_{ij} \text{ for all } i,j\}$ in a maximum likelihood approach

$$L(\square) = \prod_{g \in G} L(\square | g) P(g)$$
EM algorithm

\[
Q(g|G) = \prod_{g \in G} L(g|G)P(g|G',y) \text{ to be maximized}
\]

\[
\begin{align*}
Q^{t+1}_{ij} &= \frac{\prod_{k \neq i,j} g_{ik} (1 - g_{ik})^{1 - g_{ik}} g_{jk} (1 - g_{jk})^{1 - g_{jk}} \cdot L(g, g)}{\prod_{k \neq i,j} g_{ik} (1 - g_{ik})^{1 - g_{ik}} g_{jk} (1 - g_{jk})^{1 - g_{jk}} \cdot L(g, g)} \\
\end{align*}
\]
Simplification

\[
\mathbb{Q}_{ij}^{t+1} = \mathbb{Q}_{g|ij} = 1 \cdot \frac{\mathbb{Q}_{ik} \mathbb{Q}_{jk} \cdot P(\bigcap_{ij} > 0 \mid y) + (1 - \mathbb{Q}_{ik} \mathbb{Q}_{jk}) \cdot P(\bigcap_{ij} > 0 \mid y)}{\mathbb{Q}_{ik} \mathbb{Q}_{jk} \cdot P(\bigcap_{ij} > 0 \mid y) + (1 - \mathbb{Q}_{ik} \mathbb{Q}_{jk}) \cdot P(\bigcap_{ij} > 0 \mid y)}
\]

not all GGMs with \(i,j,k\) considered
Distribution of $\square_{ij}$
Conclusions

• GGM of gene triples used to look whether correlation between two genes can be “explained” by a third one

• frequentist approach simple, can be applied to many genes

• approach with \( q \)-parameters requires iteration, tested for up to 70 genes

• a large set of additional genes can be attached to constructed network
Modeling at two levels

genetic network
• small number <100
• model edges

attach additional genes
• possibly 1000s
• which one “explain” edges?
Attaching additional genes

For additional genes $k$

- include in computation of $q_{ij}$ but keep $q_{ik}$ and $q_{jk}$ fixed
- $q_{ij}$ decreases
- count how often $k$ decreases $q_{ij}$, validate
Attaching genes from other pathways

without additional genes

with additional genes
### Attaching genes from other pathways

<table>
<thead>
<tr>
<th>both pathways</th>
<th>chloroplast</th>
<th>cytoplasm</th>
</tr>
</thead>
<tbody>
<tr>
<td>carotenoid*</td>
<td>carotenoid*</td>
<td>tocopherol*</td>
</tr>
<tr>
<td>tocopherol*</td>
<td>onecarbonpool</td>
<td>phytosterol*</td>
</tr>
<tr>
<td>calvin cycle</td>
<td>chlorophyll*</td>
<td></td>
</tr>
<tr>
<td>chlorophyll*</td>
<td>calvin cycle</td>
<td></td>
</tr>
<tr>
<td>abscisic acid*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>onecarbonpool</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phytosterol*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

downstream pathways marked by *
Attaching genes from other pathways
Hierarchical clustering
Conclusions

Modified GGMs

• model dependence between genes
• combine GGMs and pairwise correlation for inference on edge $ij$
• different statistical design, computational cost
• additional genes can be fitted into the model
• similarities in expression patterns between groups of genes can be identified (also verified in yeast data)
Comparison of different methods

- pairwise correlation
- GGM

- Modified GGM1: 0.68
- Modified GGM2: 0.96
- Modified GGM3: 0.68

- GGM: 0.56
- 0.60
- 0.80
- 0.81
- 0.43
- 0.47
- 0.49
Consistent results

Chloroplast
- DXP83
- DXPS1
- DXPS2
- DXPS4
- DXR
- MCT
- CMK
- MECP5
- IPP11
- GPPS
- GPPS(P)
- GPPS(O)
- GGPPS1
- GGPPS2
- GGPPS3
- GGPPS4
- GGPPS5
- GGPPS6
- GGPPS7
- GGPPS8
- GGPPS9
- GGPPS10
- GGPPS11
- GGPPS12
- GPPR
- PS
- Chlorophylls
- Carotenoids

Cytosol
- AACT1
- AACT2
- HMGR1
- HMGR2
- MK
- MPDC2
- IPP12
- FPPS1
- FPPS2
- GPPS3
- GPPS4
- DPPS
- Phytosterols
- Sesquiterpenes

Mitochondria
Acknowledgments

• Peter Buehlmann
• Stefan Bleuler, Amela Prelic, Eckhard Zitzler
• Philip Zimmermann, Lars Hennig, Willi Gruissem
Galactose pathway in yeast

GALACTOSE METABOLIC FLOW

Galactose (out) → HXTs → Galactose (in) → GAL2 → Galactose → Galactose-1-P → GAL1 → UDP-Galactose → UDP-Glucose → Glucose-1-P → GAL7 → Glucose-1-P → Glucose-6-P → PCL10 → GLYCOCEN SYNTHESIS

GLYCOLYSIS

PROTEIN METABOLISM

CORE REGULATORY MECHANISM

GAL4 → GAL80 → GAL3

Enzymatic reaction
Membrane transport
Transcript. activation
Protein activation
Protein inhibition

GAL10

RPA49
YPS3
MLF3
YMR318C
YEL057C
Galactose pathway in yeast

<table>
<thead>
<tr>
<th>Gene Cluster (size)</th>
<th>wt +gal vs. wt -gal</th>
<th>+gal; Δ vs. wt</th>
<th>-gal; Δ vs. wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 (44)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 (47)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 (62)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 (35)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 (29)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 (158)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 (55)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 (82)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 (95)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 (99)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 (88)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 (53)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 (76)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 (38)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 (33)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **GAL1, 7, 10**
  - Amino-acid, nitrogen metab.:
    - **GAL2, 3, 6, 80**
  - Nucleotide metab.
  - No predominant role
  - Protein metab., biosynthesis
  - Lipid metabolism
  - Intracellular protein traffic
  - Protein targeting
  - Energy pathways (respiration);
  - **GAL4**
  - Carb. metabolism, stress; **GAL5**
  - Stress, phosphate metabolism
  - Lipid metabolism
  - Transport (hexoses)
Network for galactose pathway
Network for galactose pathway

Genes that attached to network:

- GCY1
- FAR1
- YDR010C
- YEL057C
- YPL066W
- YOR121C
- MLF3
- YLL058W
- YJL212C
- PCL10
- LYS1
- YBR139W
- YCR059C
\[
\ell(\mathcal{D}, \mathcal{D}) = \frac{N}{2} \left( q \ln(2 \mathcal{D}) + \ln |\mathcal{D}| + \text{tr}(\mathcal{D} S) + (\bar{y} \mathcal{D} \mathcal{D})' \mathcal{D} (\bar{y} \mathcal{D} \mathcal{D}) \right)
\]

\[
L(\mathcal{D}) = \prod_{g \in \mathcal{G}} L(\mathcal{D} | g) P(g)
\]

\[
Q(\mathcal{D} | \mathcal{D}^t) = \prod_{g \in \mathcal{G}} L(\mathcal{D} | g) P(g | \mathcal{D}^t, y)
\]

\[
\mathcal{D}_{ij}^{t+1} = \prod_{k \neq i, j} \mathcal{D}_{ik}^{t} \mathcal{D}_{jk}^{t} \cdot P(\mathcal{D}_{ij}^{t} > 0 | y) + (1 - \mathcal{D}_{ik}^{t} \mathcal{D}_{jk}^{t}) \cdot P(\mathcal{D}_{ij}^{t} > 0 | y)
\]

\[
\mathcal{D}_{ij}^{t+1} = \prod_{k \neq i, j} \mathcal{D}_{ik}^{t} \mathcal{D}_{jk}^{t} \cdot \frac{g_{ik} (1 - g_{ik})^{1-g_{ik}} \mathcal{D}_{ij}^{t} g_{jk} (1 - g_{jk})^{1-g_{jk}} \cdot L(\mathcal{D}_{g}, \mathcal{D}_{g})}{g_{ik} (1 - g_{ik})^{1-g_{ik}} \mathcal{D}_{ij}^{t} g_{jk} (1 - g_{jk})^{1-g_{jk}} \cdot L(\mathcal{D}_{g}, \mathcal{D}_{g})}
\]