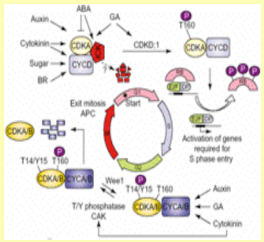
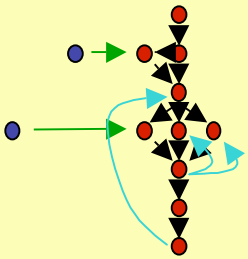
A modified graphical gaussian model approach for genetic regulatory networks

Anja Wille, SfS
Reverse Engineering Project, ETHZ

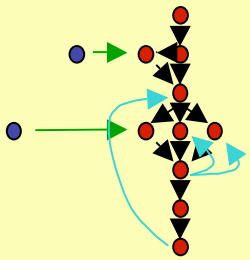
Introduction

biological scenarios





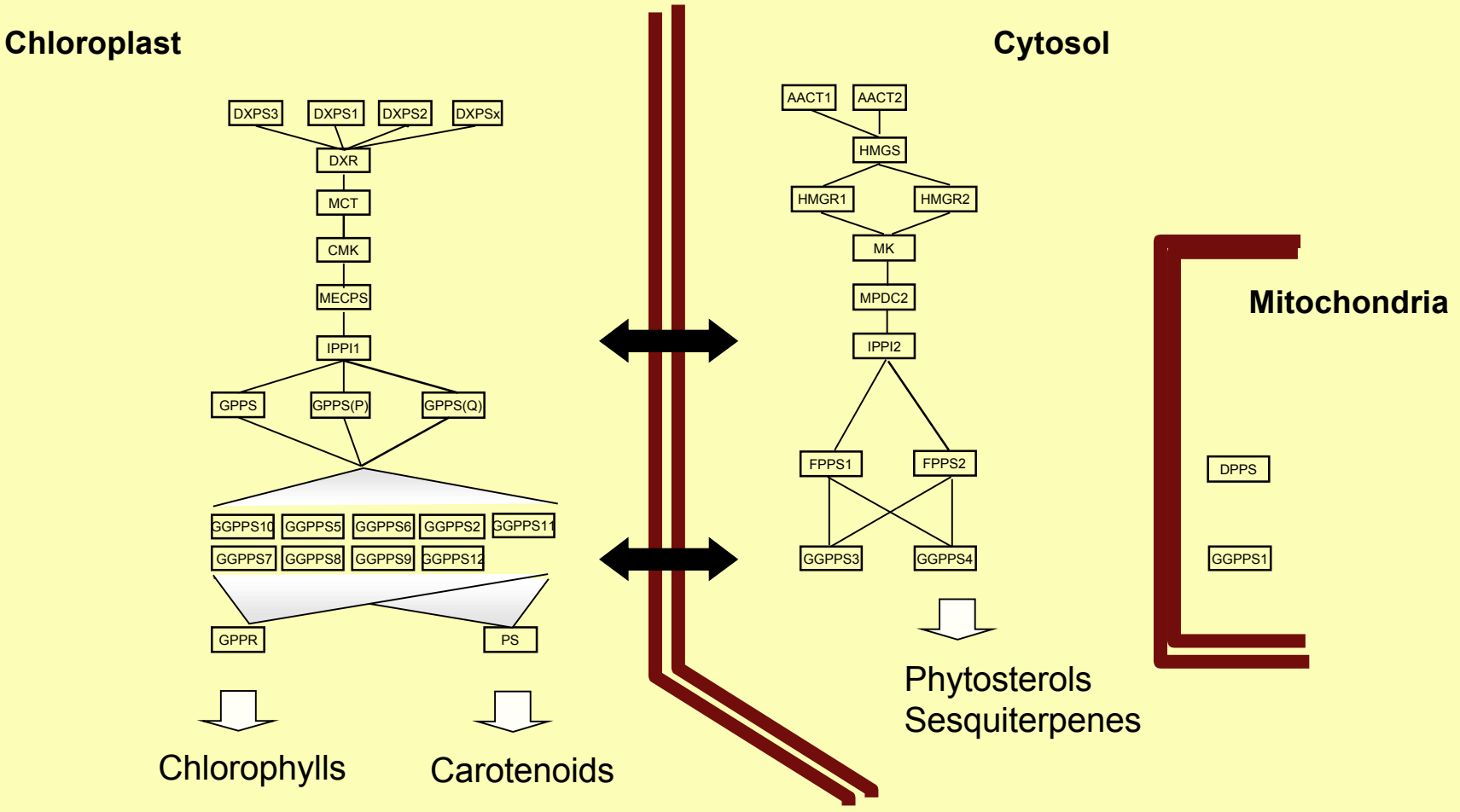
modified GGMs



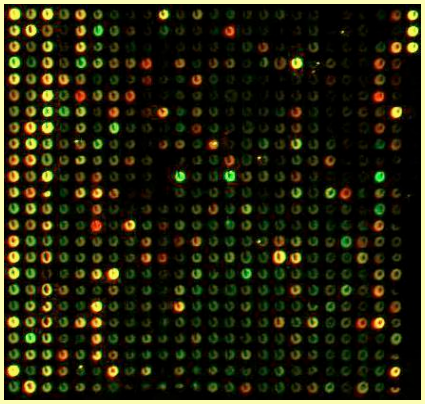
graphical gaussian models
(GGMs)

Biological scenarios

Isoprenoid pathways



Genetic regulation

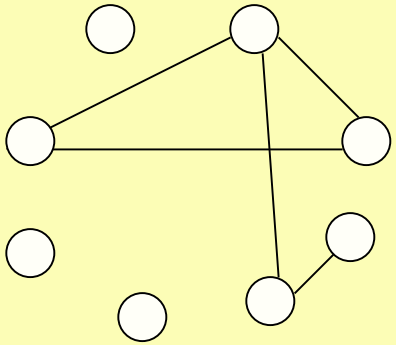


118 observations

39 genes

… signals …

graphical model to estimate conditional dependence between genes



Graphical Gaussian Models

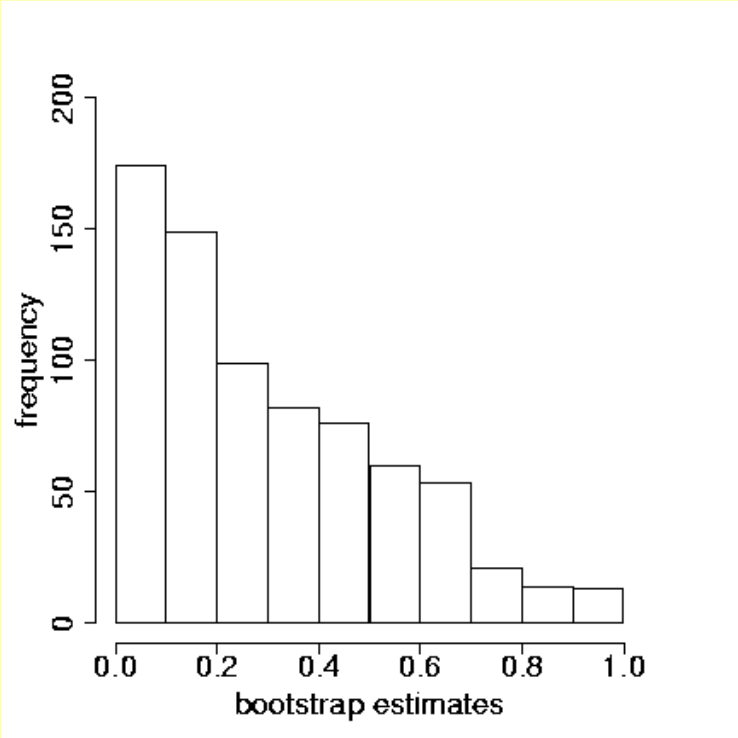
- undirected
- random variables follow a multivariate normal distribution
- log likelihood:

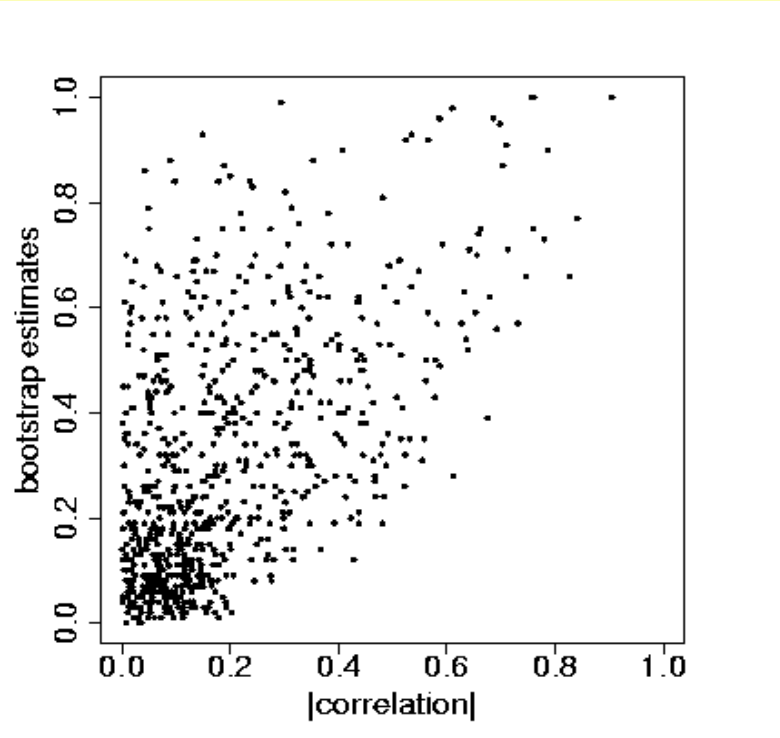
$$l(\mu, \Omega) = -\frac{N}{2}(q \ln(2\pi) + \ln|\Omega| + tr(\Omega S) + (\bar{y} - \mu)' \Omega (\bar{y} - \mu))$$

for a sample of N observations $y^1,\ldots,y^N$ with
sample mean $\bar{y}$, sample covariance matrix S
and precision matrix $\Omega = \Sigma^{-1}$

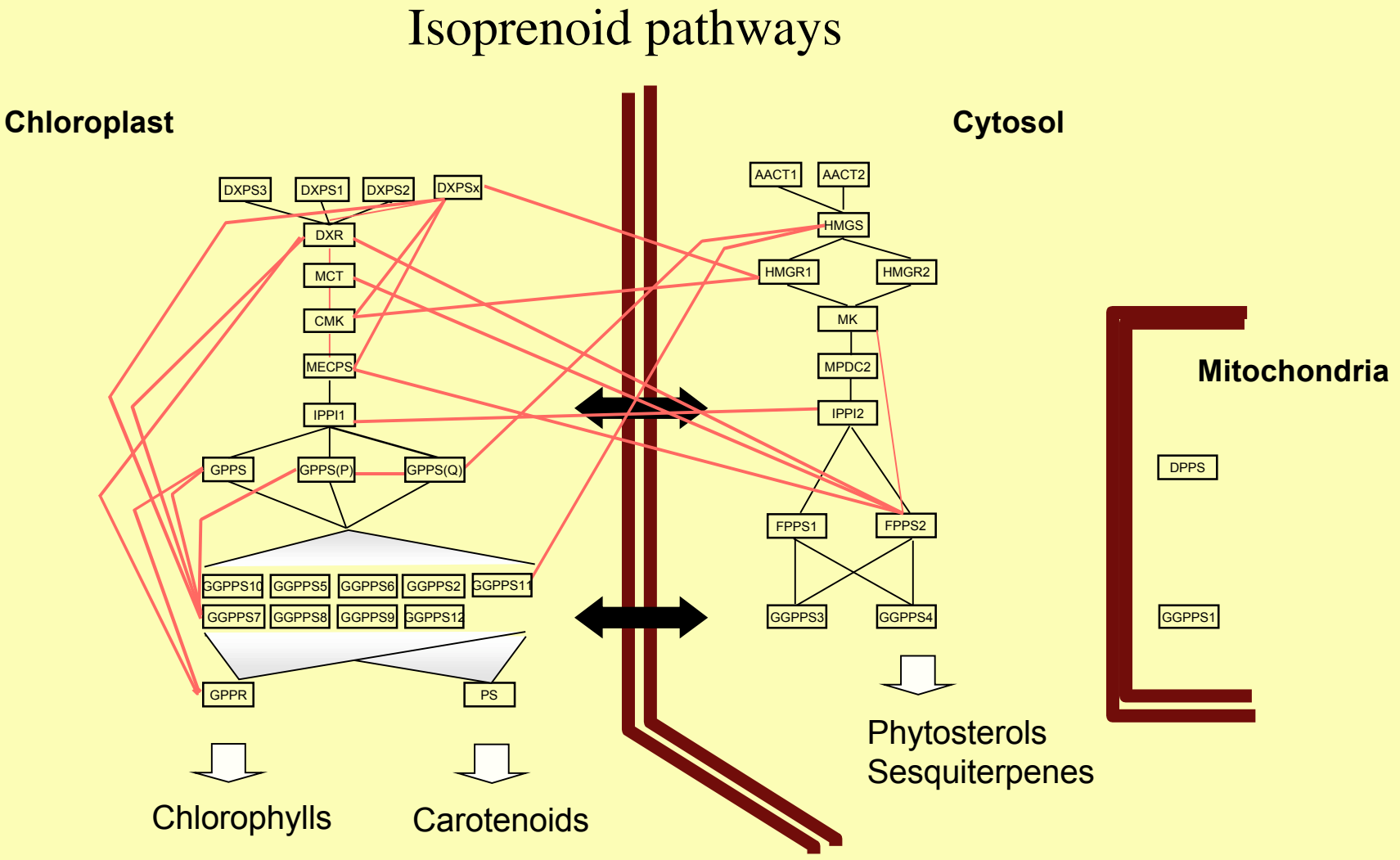
- partial correlation coefficient $\rho_{ijlrest}$ for gene pair ij

Bootstrapping and pairwise correlation





Application to isoprenoid pathways



Problems

- matrix inversion rank-sensitive
- genes ↑ ⇒ exponentially increasing number of models

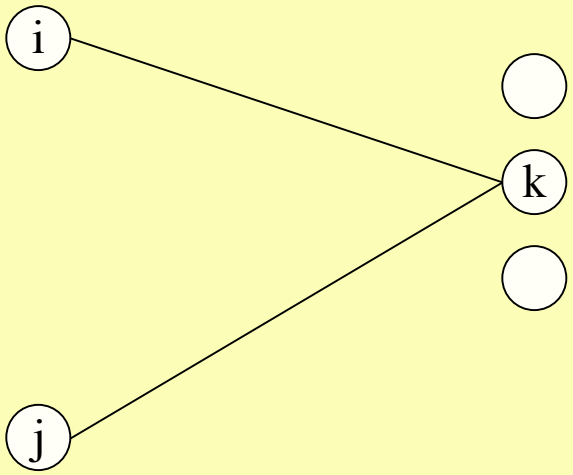
  ⇒ difficult to interpret
- how to interpret high partial correlation accompanied with low pairwise correlation
- efficient procedure for attaching new genes needed

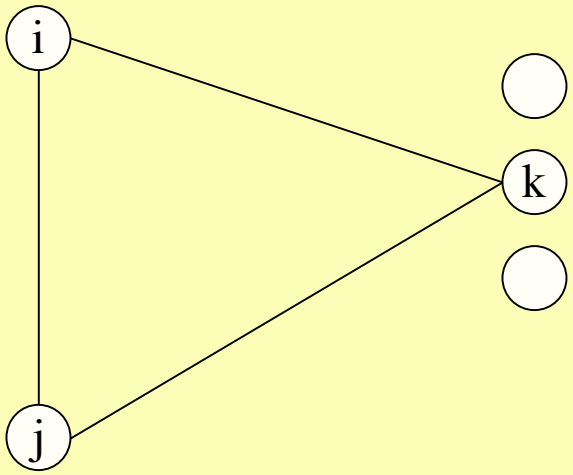
Outline for modified GGM approach

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

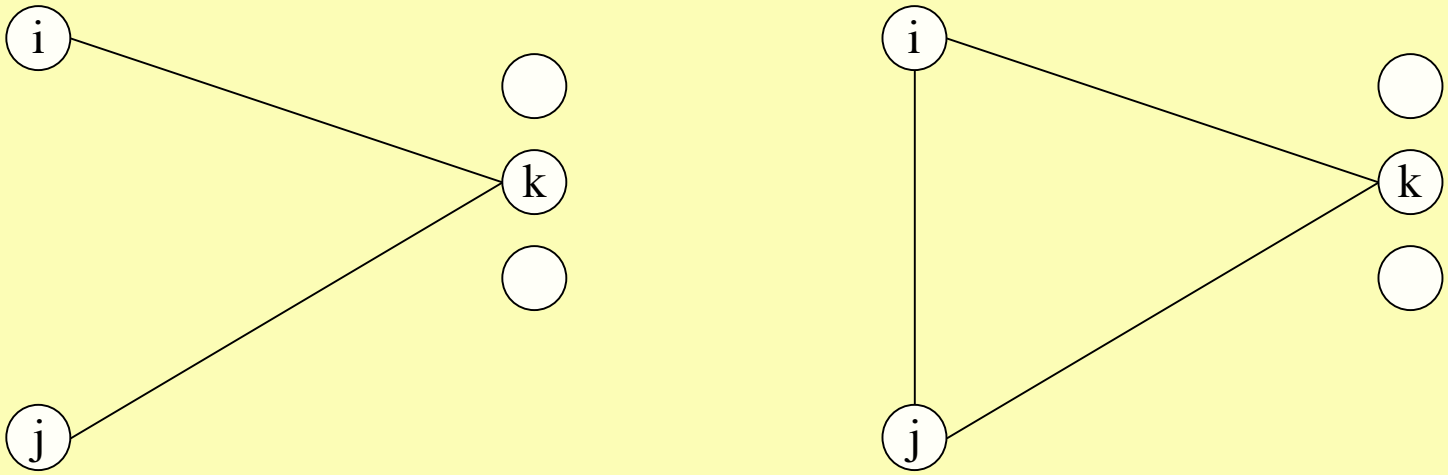




Focus on genepair ij

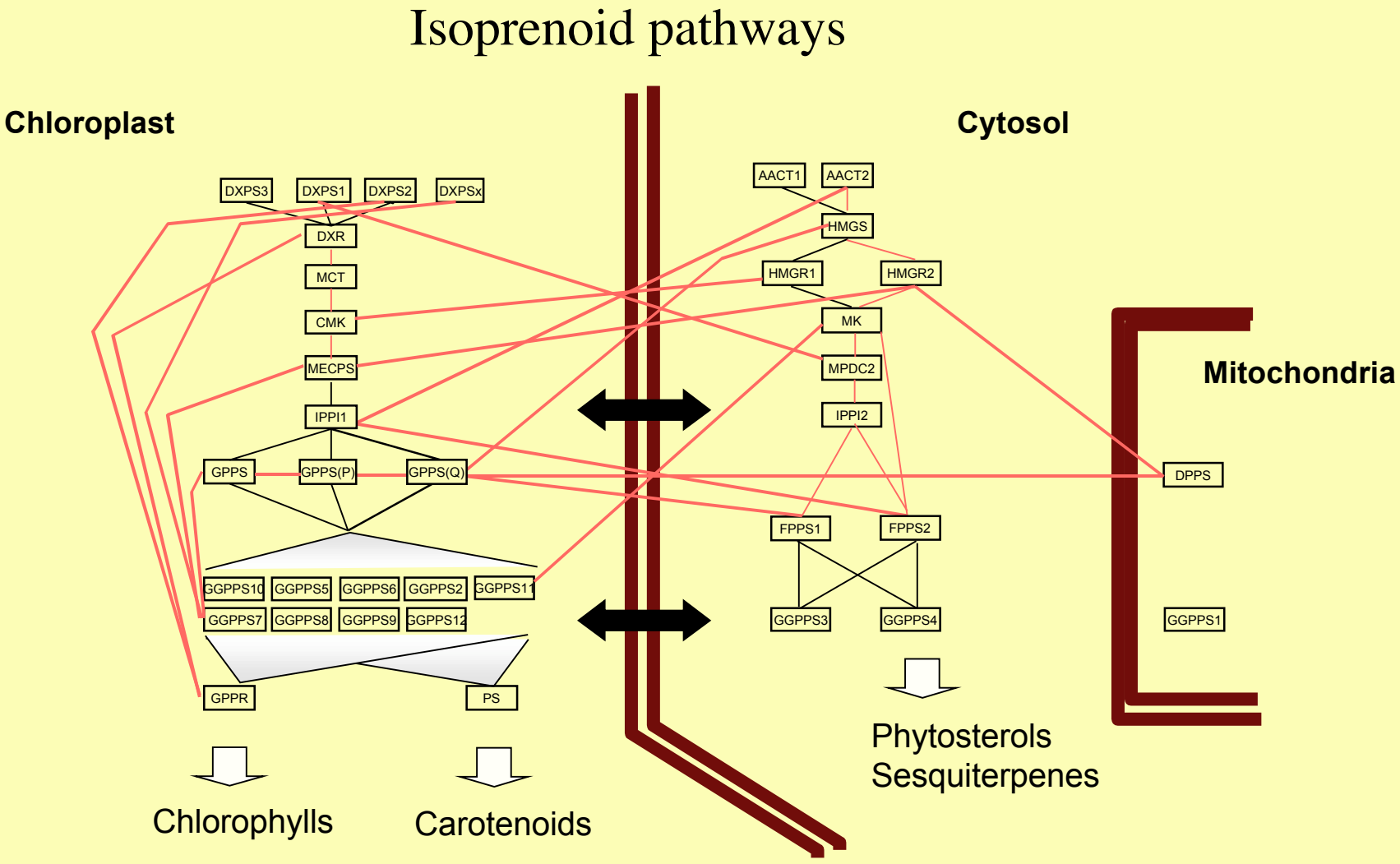
- $p_{ij|k}$ is p-value from deviance test $\delta_{ij|k} \neq 0$ versus $\delta_{ij|k} = 0$

Frequentist approach (cont'd)

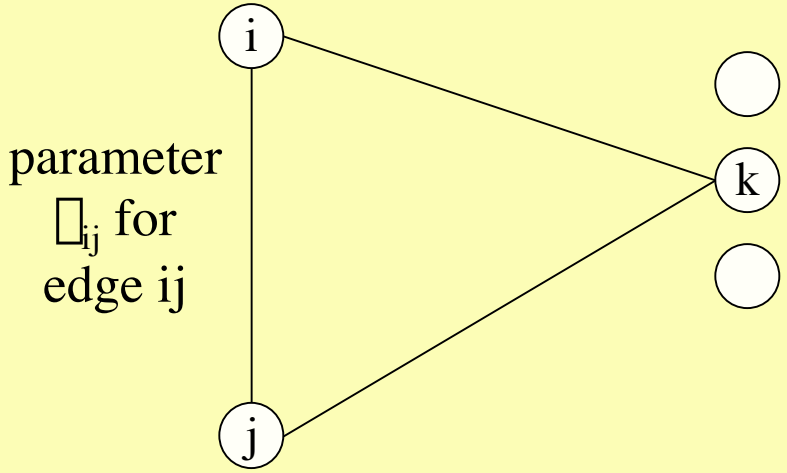


1) Form $p_{ij,max}$ = max ($p_{ij|k}$ for all genes $k \neq i,j$)
2) Adjust $p_{ij,max}$ according to Bonferroni-Holm or FDR
3) If the adjusted value $p_{ij,max} < 0.05$, draw edge between i and j

Application to isoprenoid pathway



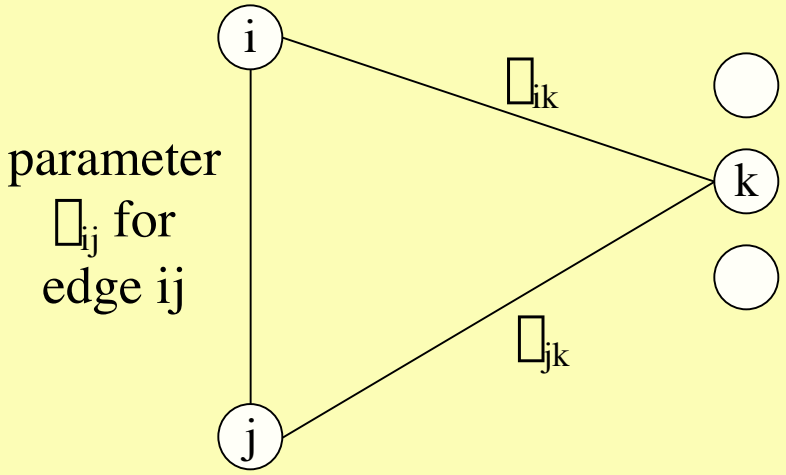
Likelihood approach with parameters θ



Estimate $\theta = \{\theta_{ij}$ for all i,j$\}$ in a maximum likelihood approach

$$L(\theta) = \sum_{g \in G} L(\theta \mid g)P(g)$$

EM algorithm



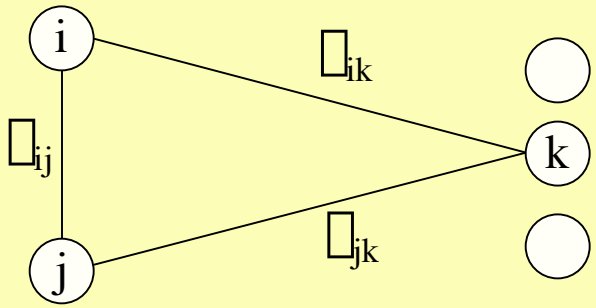
$$Q(\theta \mid \theta^t) = \sum_{g \in G} L(\theta \mid g) P(g \mid \theta^t, y) \quad \text{to be maximized}$$

$$\downarrow$$

$$\theta_{ij}^{t+1} = \prod_{k \neq i,j} \frac{\sum_{g \mid g_{ij}=1} \theta_{ik}^{t\, g_{ik}} (1-\theta_{ik}^t)^{1-g_{ik}} \theta_{jk}^{t\, g_{jk}} (1-\theta_{jk}^t)^{1-g_{jk}} \cdot L(\theta_g, \theta_g)}{\sum_{g} \theta_{ik}^{t\, g_{ik}} (1-\theta_{ik}^t)^{1-g_{ik}} \theta_{jk}^{t\, g_{jk}} (1-\theta_{jk}^t)^{1-g_{jk}} \cdot L(\theta_g, \theta_g)}$$

Simplification

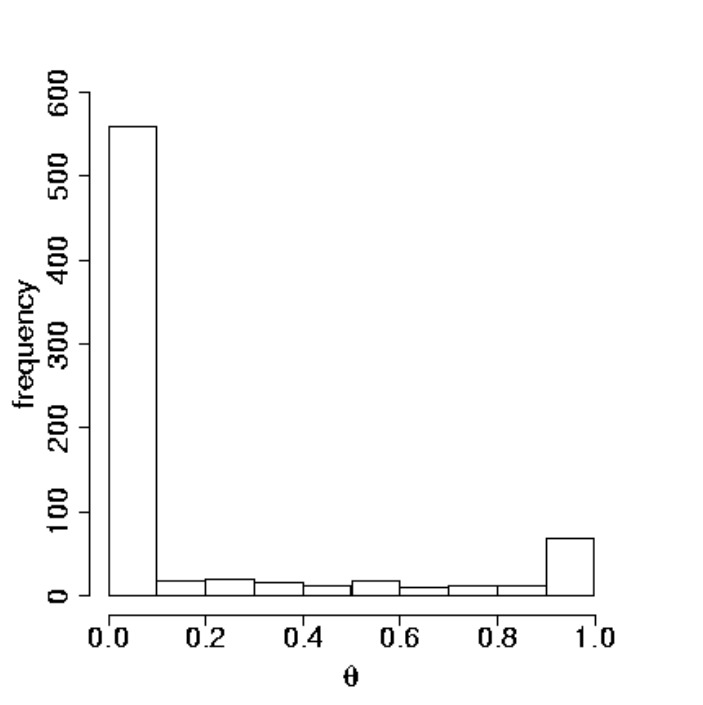
$$\theta_{ij}^{t+1} = \sum_{k \neq i,j} \frac{\sum_{g \mid g_{ij}=1} \theta_{ik}^{t\,g_{ik}} (1-\theta_{ik}^{t})^{1-g_{ik}} \theta_{jk}^{t\,g_{jk}} (1-\theta_{jk}^{t})^{1-g_{jk}} \cdot L(\beta_g, \sigma_g)}{\sum_{g} \theta_{ik}^{t\,g_{ik}} (1-\theta_{ik}^{t})^{1-g_{ik}} \theta_{jk}^{t\,g_{jk}} (1-\theta_{jk}^{t})^{1-g_{jk}} \cdot L(\beta_g, \sigma_g)}$$

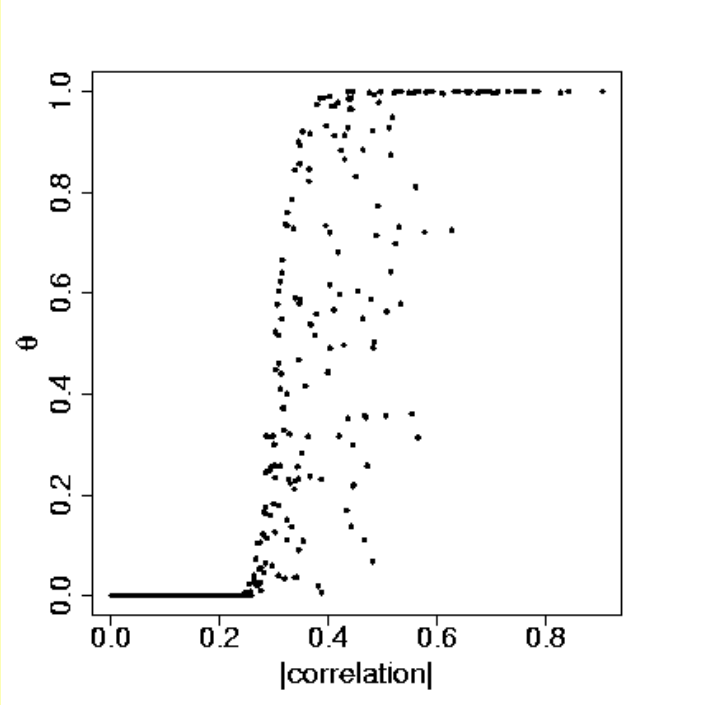


not all GGMs with i,j,k considered

$$\theta_{ij}^{t+1} = \sum_{k \neq i,j} \theta_{ik}^{t}\theta_{jk}^{t} \cdot P(\beta_{ij|k} > 0 \mid y) + (1 - \theta_{ik}^{t}\theta_{jk}^{t}) \cdot P(\beta_{ij} > 0 \mid y)$$

Distribution of θ_{ij}

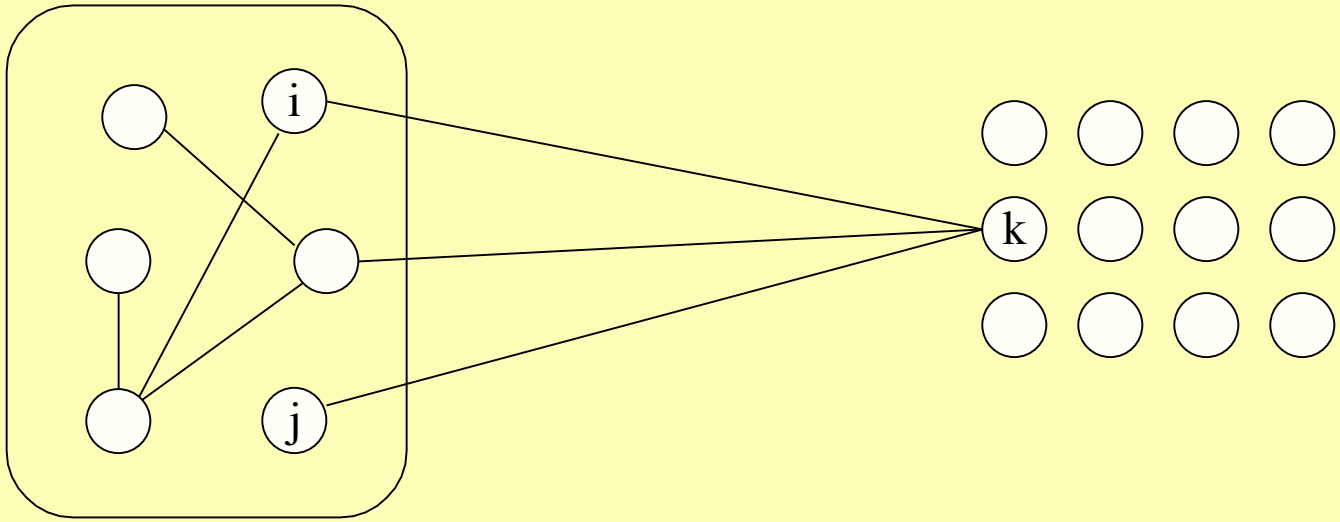




Conclusions

- 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 □-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



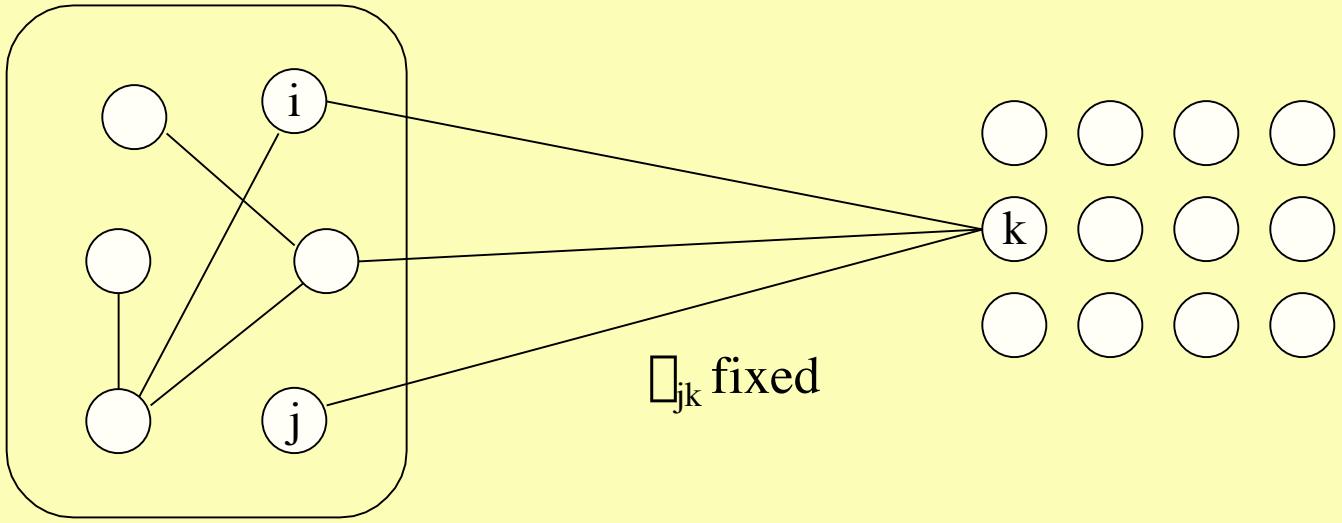
genetic network

- small number <100
- model edges

attach additional genes

- possibly 1000s
- which one “explain” edges?

Attaching additional genes

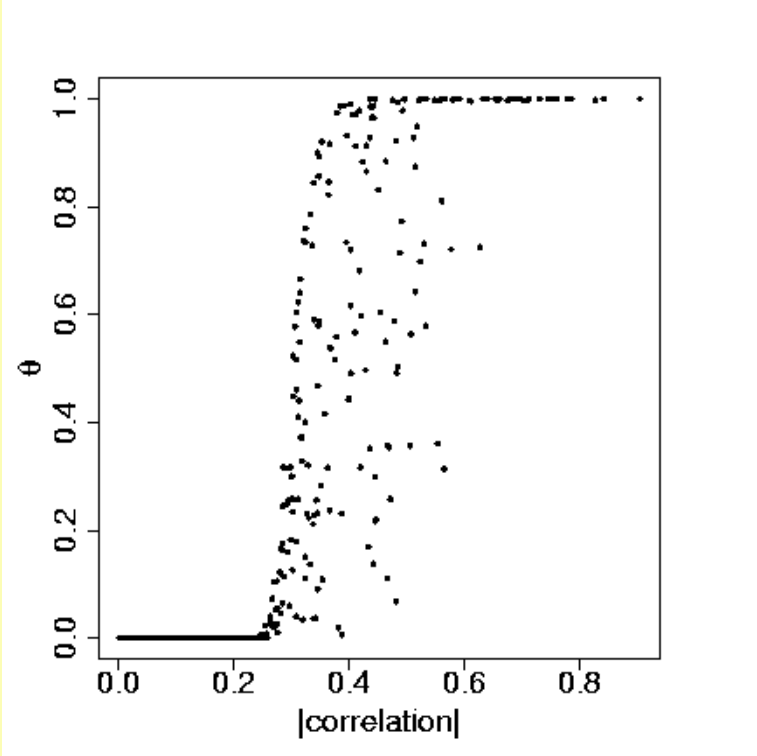


For additional genes k

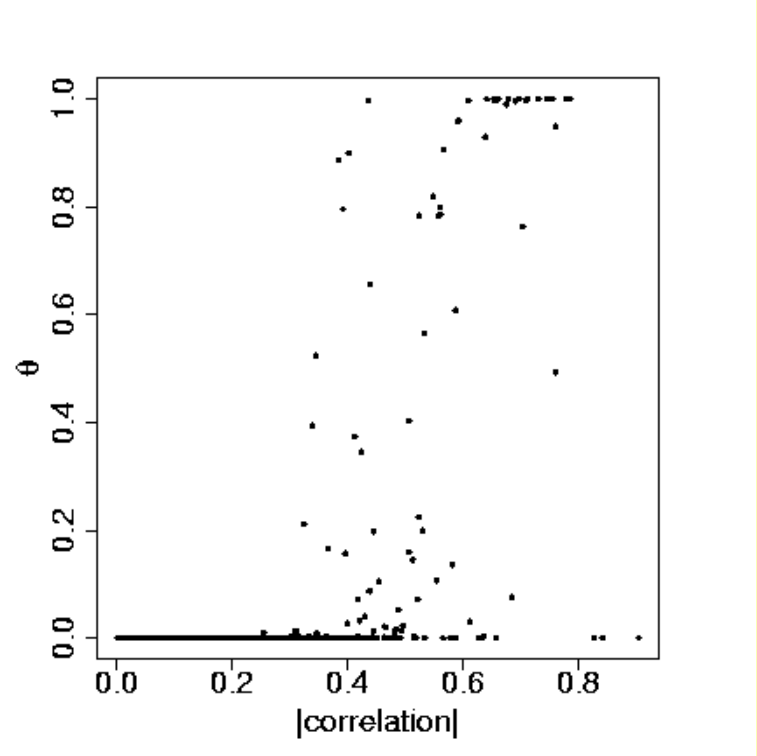
- include in computation of θ_{ij} but keep θ_{ik} and θ_{jk} fixed
- θ_{ij} decreases
- count how often k decreases θ_{ij}, validate

Attaching genes from other pathways

without additional genes



with additional genes

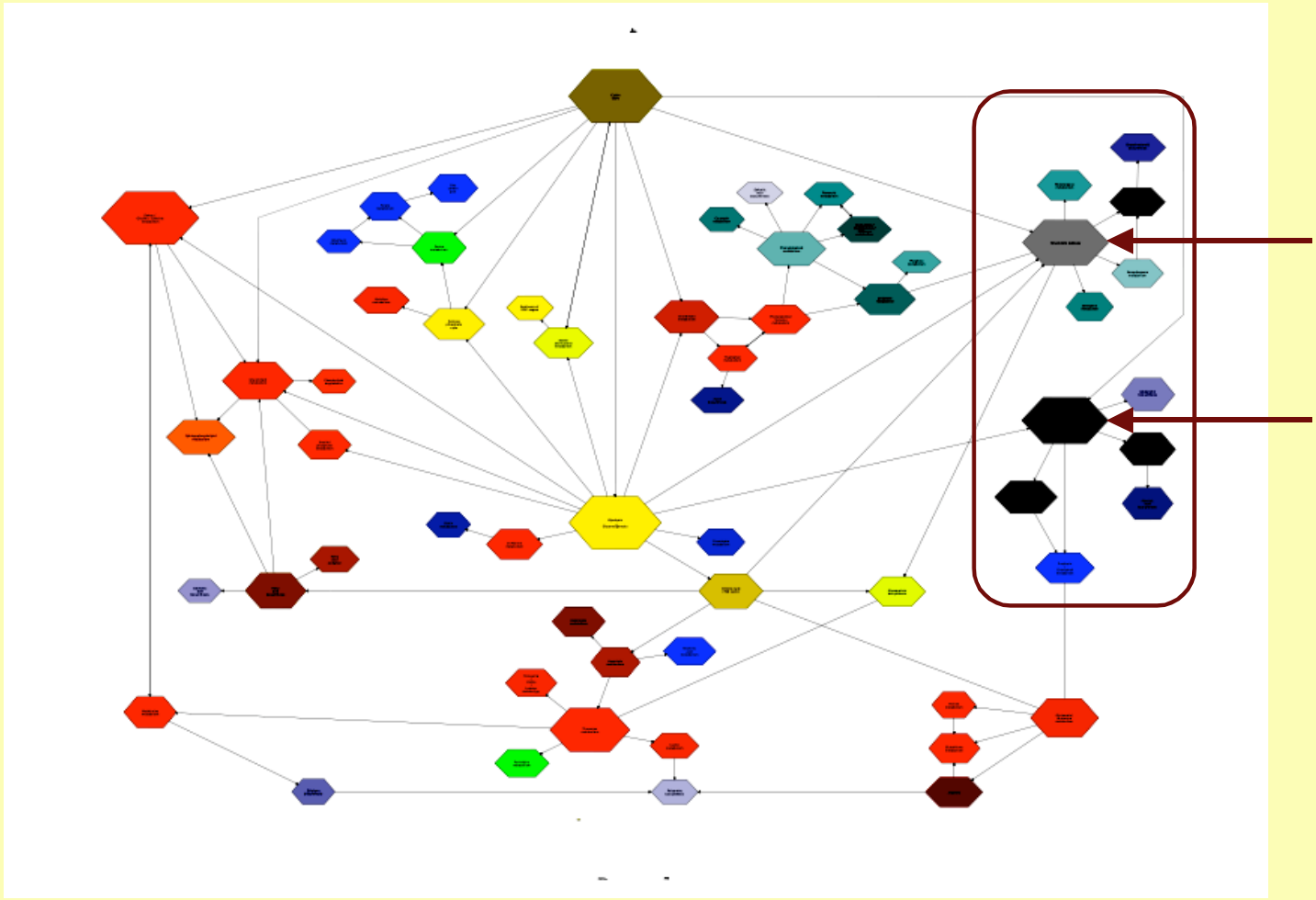


Attaching genes from other pathways

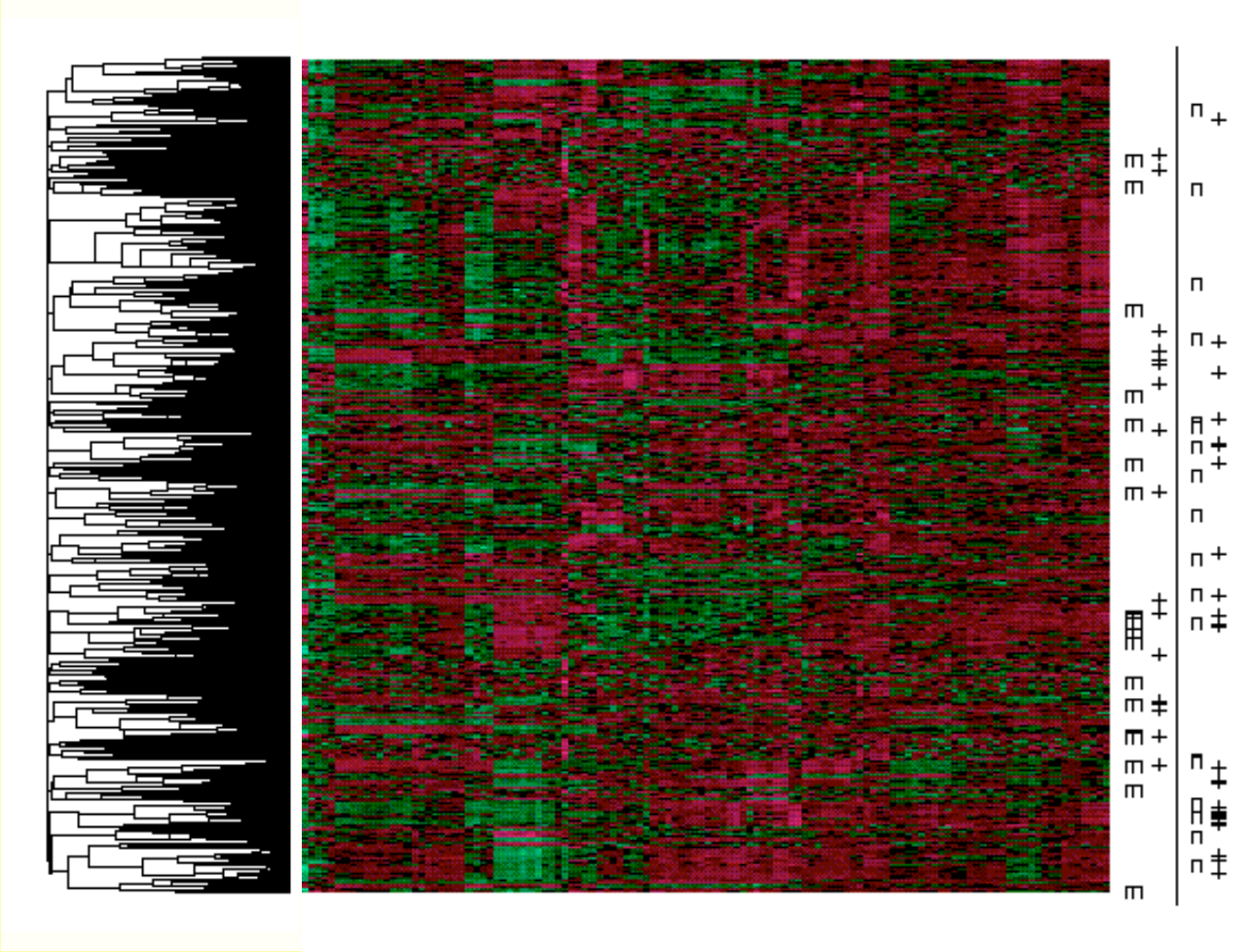
both pathways	chloroplast	cytoplasm
carotenoid*	carotenoid*	tocopherol*
tocopherol*	onecarbonpool	phytosterol*
calvin cycle	chlorophyll*	
chlorophyll*	calvin cycle	
abscisic acid*		
onecarbonpool		
phytosterol*		

downstream pathways marked by *

Attaching genes from other pathways



Hierarchical clustering

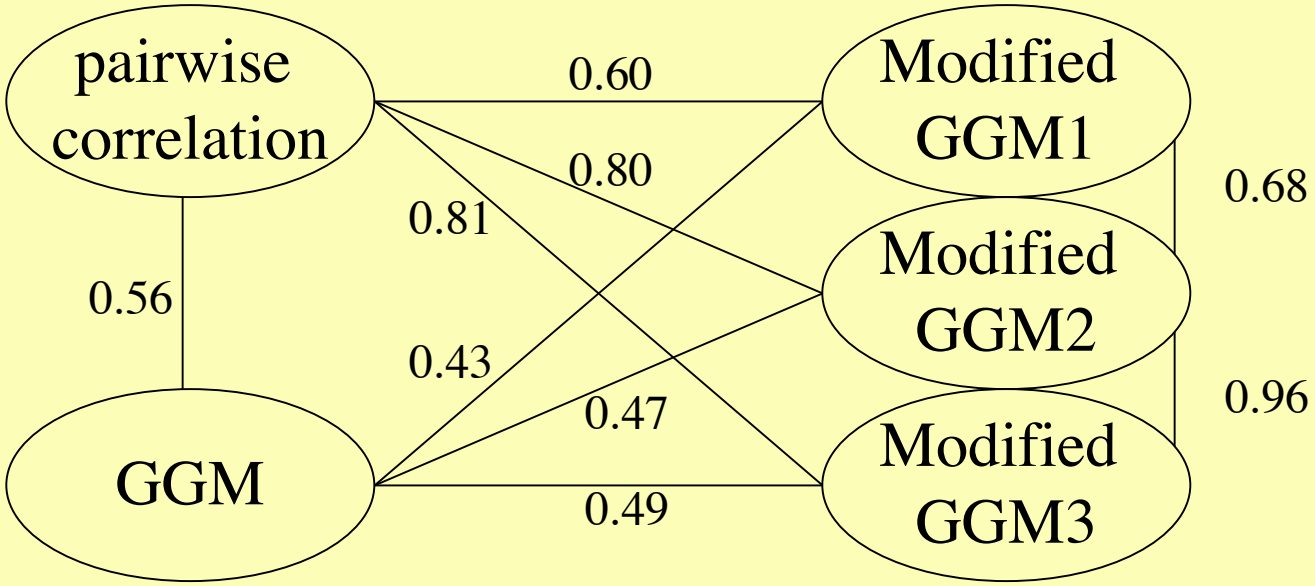


Conclusions

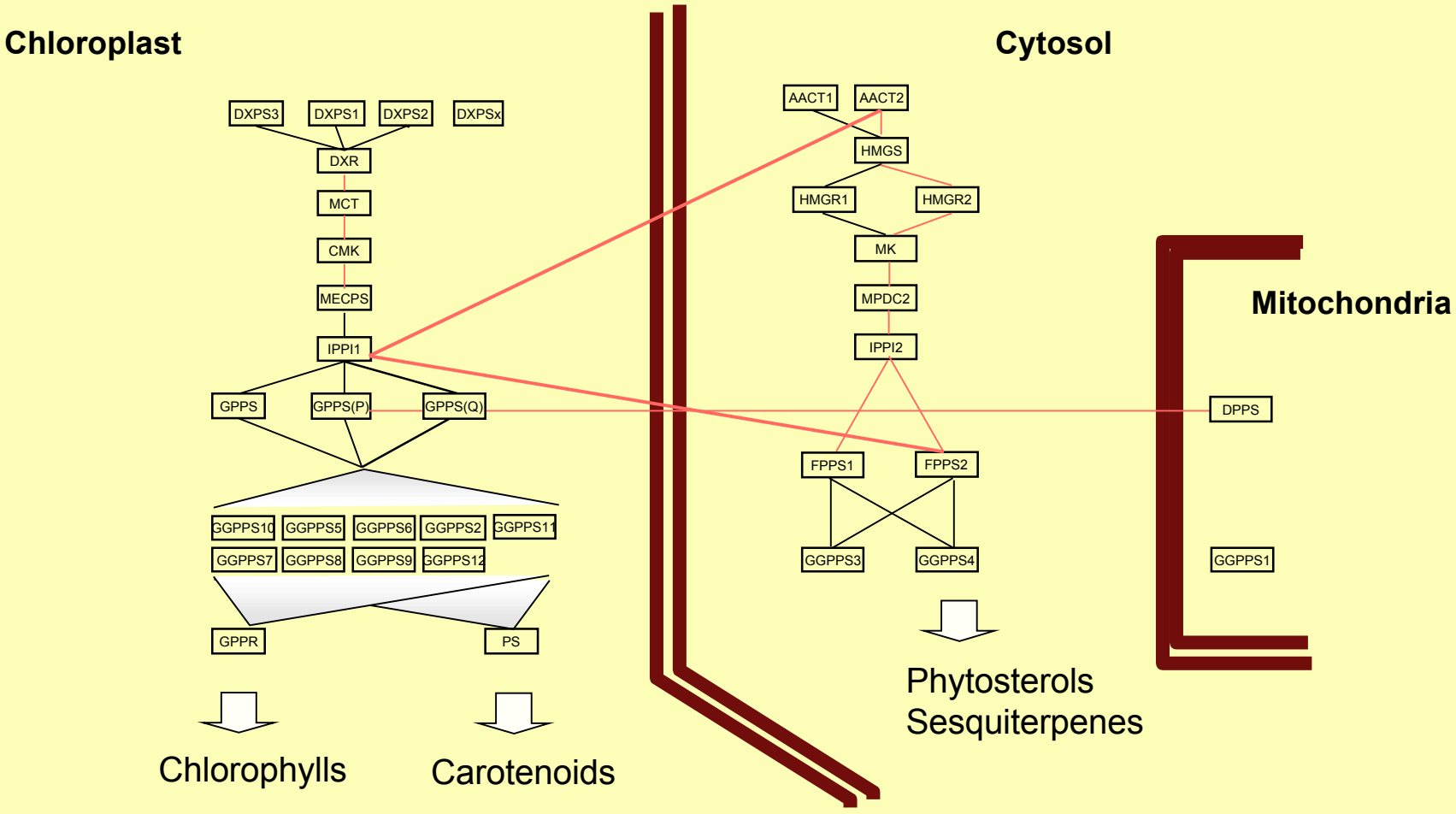
Modified GGMs

- 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



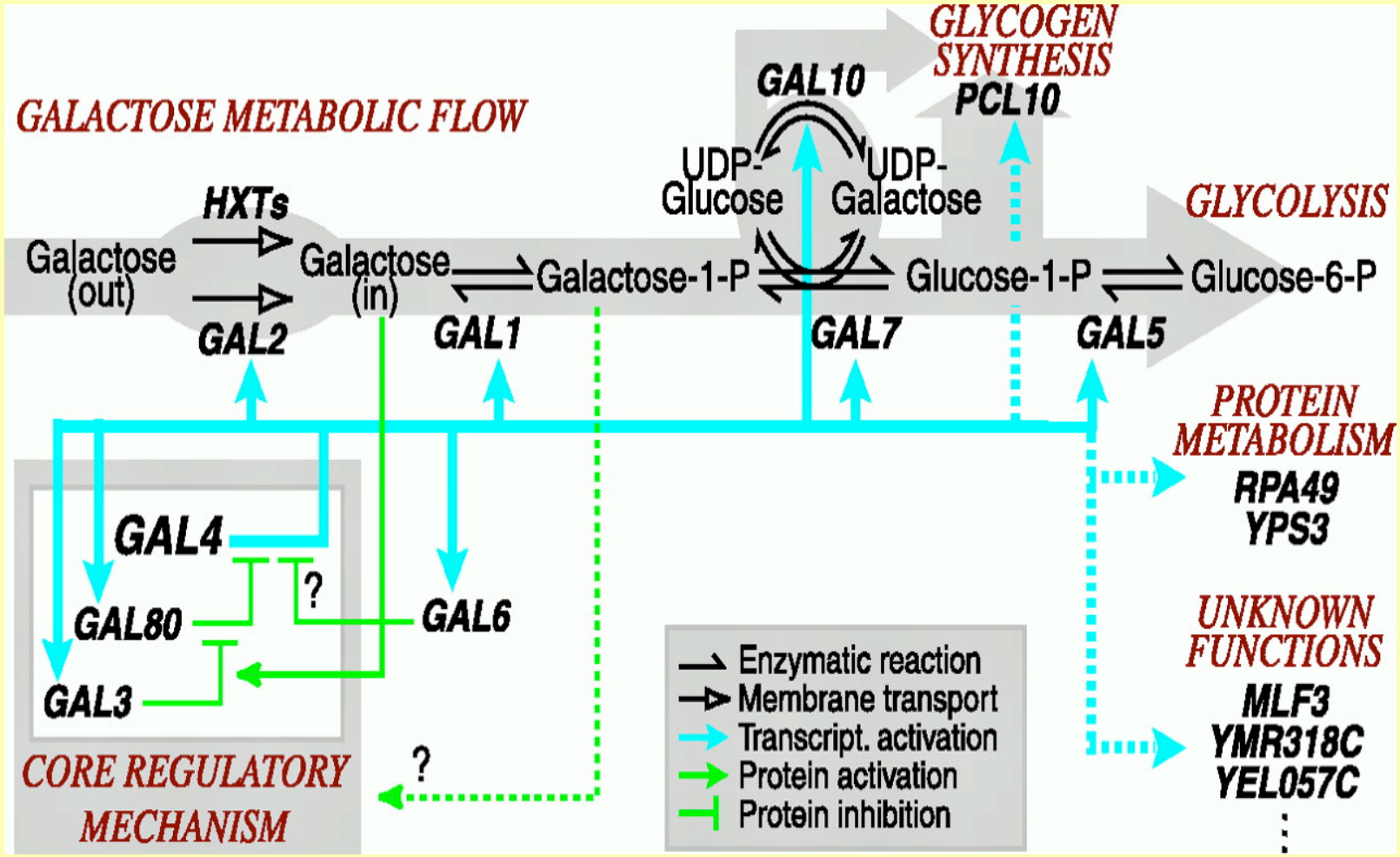
Consistent results



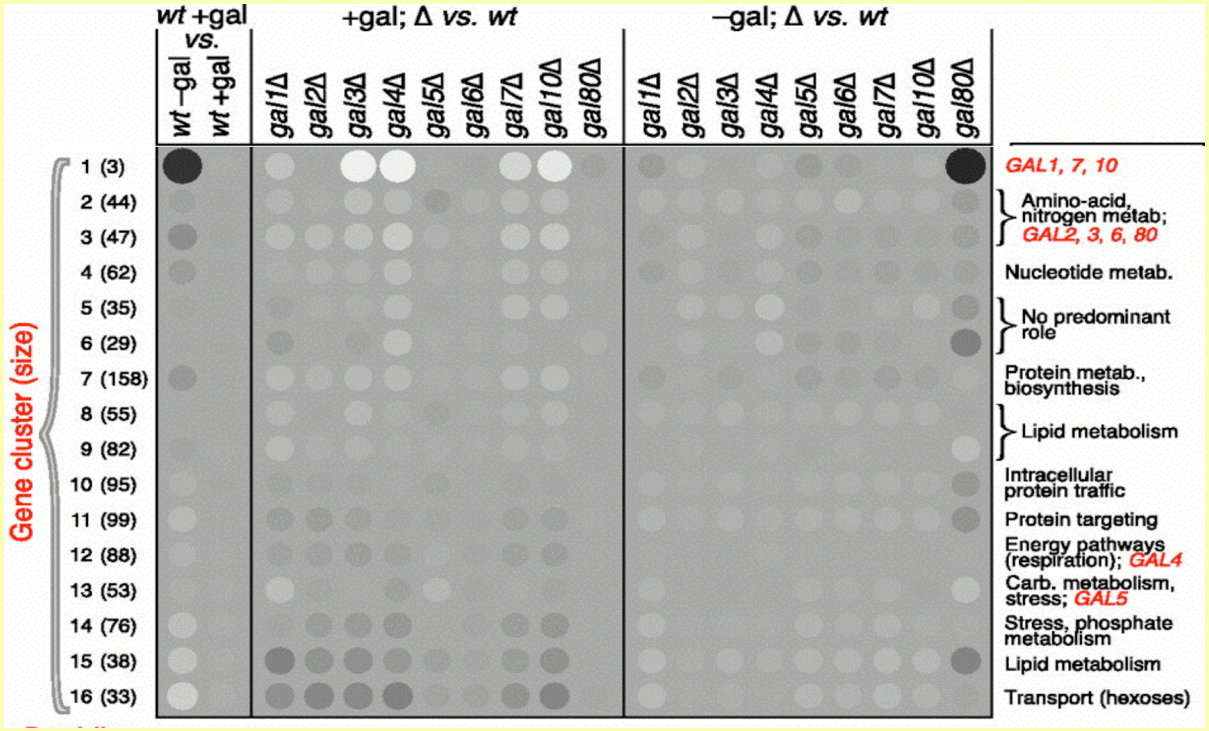
Acknowledgments

- Peter Buehlmann
- Stefan Bleuler, Amela Prelic, Eckhard Zitzler
- Philip Zimmermann, Lars Hennig, Willi Gruissem

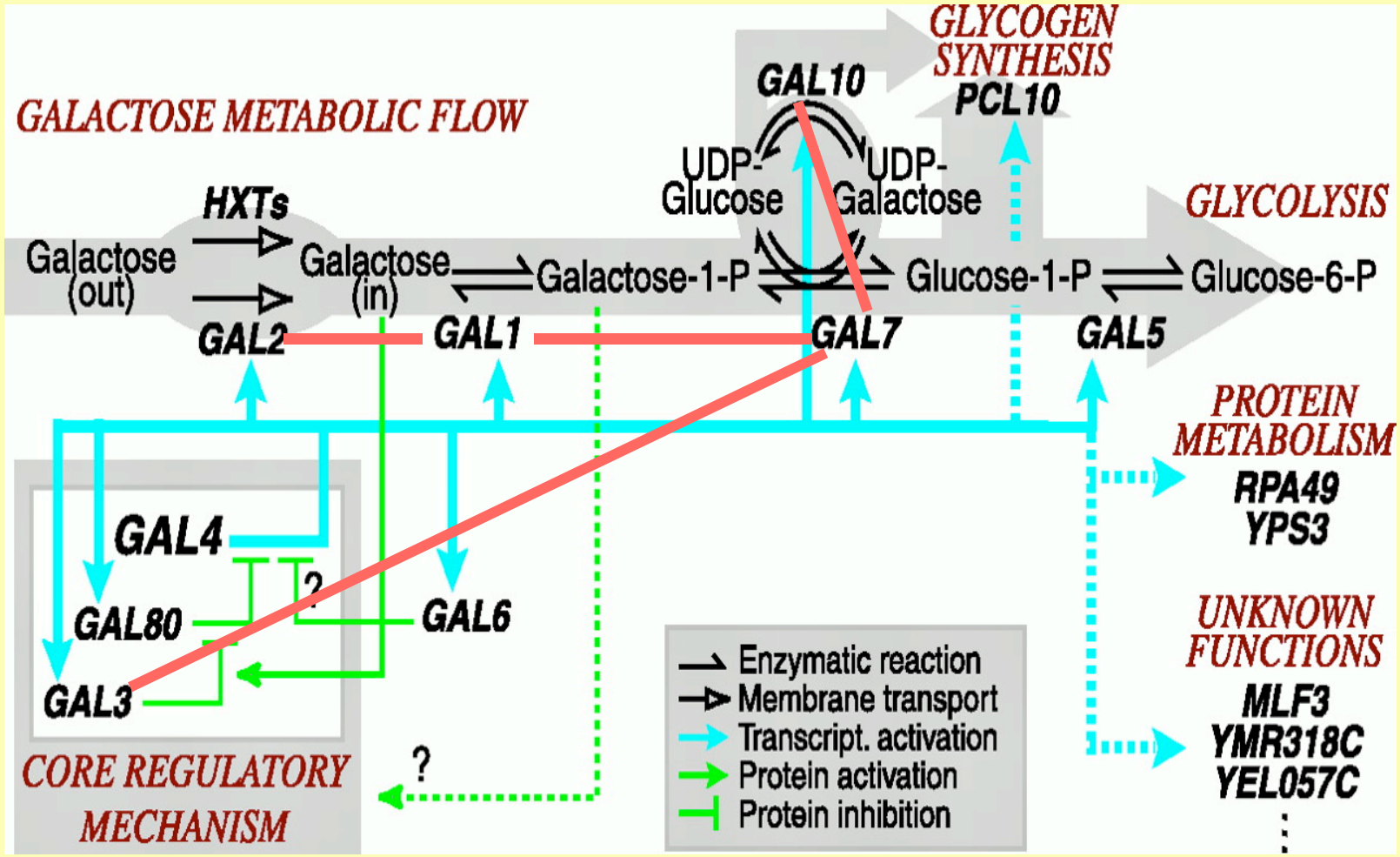
Galactose pathway in yeast



Galactose pathway in yeast



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

$$l(\theta,\Sigma) = -\frac{N}{2}(q\ln(2\pi) + \ln|\Sigma| + tr(\Sigma S) + (\bar{y}-\theta)'\Sigma(\bar{y}-\theta))$$

$$L(\theta) = \sum_{g \in G} L(\theta \mid g)P(g)$$

$$Q(\theta \mid \theta^t) = \sum_{g \in G} L(\theta \mid g)P(g \mid \theta^t, y)$$

$$\theta_{ij}^{t+1} = \sum_{k \neq i,j} \theta_{ik}^t \theta_{jk}^t \cdot P(\theta_{ij|k} > 0 \mid y) + (1-\theta_{ik}^t\theta_{jk}^t) \cdot P(\theta_{ij} > 0 \mid y)$$

$$\theta_{ij}^{t+1} = \sum_{k \neq i,j} \frac{\sum_{g \mid g_{ij}=1} \theta_{ik}^{t\,g_{ik}}(1-\theta_{ik}^t)^{1-g_{ik}}\theta_{jk}^{t\,g_{jk}}(1-\theta_{jk}^t)^{1-g_{jk}} \cdot L(\theta_g, \Sigma_g)}{\sum_g \theta_{ik}^{t\,g_{ik}}(1-\theta_{ik}^t)^{1-g_{ik}}\theta_{jk}^{t\,g_{jk}}(1-\theta_{jk}^t)^{1-g_{jk}} \cdot L(\theta_g, \Sigma_g)}$$