A comparative study of Gaussian Graphical Model approaches for genomic data

Roberto Anglani

Institute of Intelligent Systems for Automation, CNR-ISSIA, Bari, Italy

in collaboration with PF Stifanelli, TM Creanza, VC Liuzzi, S Mukherjee, N Ancona

1st International Workshop on Pattern Recognition in Proteomics, Structural Biology and Bioinformatics. PR PS BB 2011 Ravenna, Italy, 13. Sept 2011

Motivation

A living cell is a **complex** system

Genes and gene products **interact** in complicated patterns

controlled by **biochemical** interactions and **regulatory** activities

SYSTEM BIOLOGY TASKS

Uncovering the interaction pictures

Modelling functional interactions between genes, proteins and transcriptional factors in a Gene Regulatory Network **(GRN)**

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

Motivation

Complexity needs **mathematical modelling**

High-throughput technologies provide **huge amounts of data**

Theoretical and computational approaches are necessary to model gene regulatory networks

FOCUS

Stochastic tools: **Graphical models**

Study and visualize the **conditional independence** structure between random variables (e.g. microarray data)

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

Scope

Preliminary investigation on isoprenoid pathways in A. thaliana

1 **Compare** different theoretical approaches for the study of the **conditional dependencies**

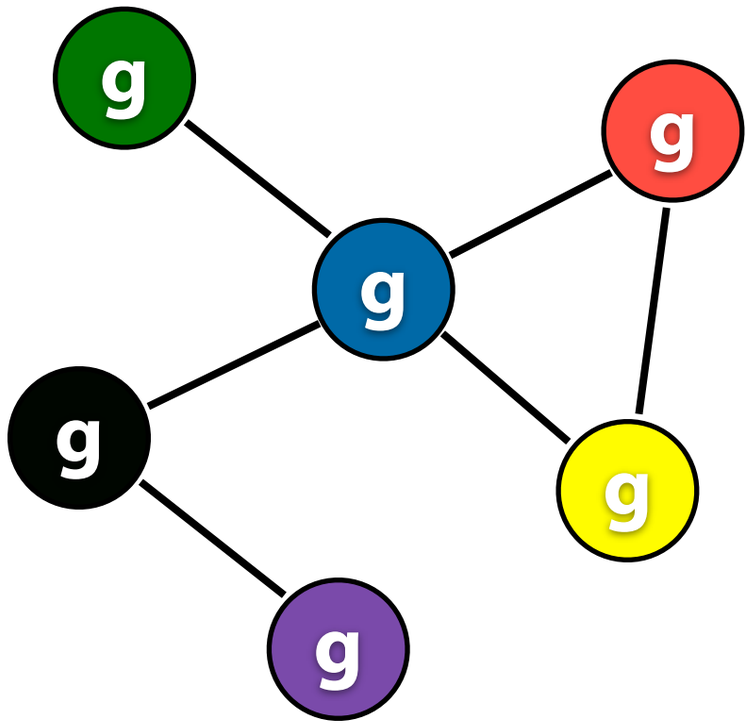
2 **Infer** a gene network for the isoprenoid biosinthesis **pathways** in *A. thaliana*

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

1 **Compare** different theoretical approaches for the study of the **conditional dependencies**

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

1.0 Graphical models



GRAPH $G = (V,E)$

VERTICES genes

EDGES conditional dependencies

ADVANTAGE **powerful** tool for **small # of genes** (wrt # observations)

SHORTCOMING high-throughput data **# genes p >> # samples n**

PROBLEM for any statistical **inference** for the **reliability** of inferred GRNs

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

1.1 GGMs with pairwise Markov property

UNDIRECTED GRAPHS

In this study we consider only
undirected Gaussian graphs
with **pairwise Markov property**

p-VARIATE NORMAL DISTRIBUTION

$$\mathbf{X} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$$

$$(i,j) \notin E \Leftrightarrow X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$$

$$\Leftrightarrow$$

ABSENCE OF EDGE

$$\rho_{ij \cdot V \setminus \{i,j\}} = 0$$

1.2 Facing n<<p problem

Partial correlation matrix is then
crucial for study of the edge structure

HOW TO SOLVE n << p PROBLEM?

NEGLECT MULTI-GENE EFFECTS

Reducing # of genes or gene lists

Toh & Horimoto (2002)

Evaluating only **limited-order** correlation

Wille & Bulhman (2004), Castelo & Roverato (2006), Gilbert & Dudoit (2009)

OK MULTI-GENE EFFECTS

Regularized estimates of precision matrix

Yuan & Lin (2007), Friedman & Tibshirani (2008), Witten & Tibshirani (2009)

Pseudoinv. estimates of precision matrix

Schaffer & Strimmer (2005)

1.3 Moore-Penrose Pseudoinverse

DATASET w/ *n* SAMPLES *p* VARIABLES

$n < p$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

ESTIMATE OF COVARIANCE $= \mathbf{S}$

ESTIMATE OF INV. COVAR. $= \hat{\mathbf{\Theta}}$

PINV

Moore-Penrose pseudoinverse

The precision matrix **Θ** is obtained as pseudoinverse of **S**, by using the **Singular Value Decomposition**

$$\rho_{ij \cdot V \setminus \{i,j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \qquad i \neq j$$

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

1.4 L_2 penalization

L2C

Cov-regularized method

The precision matrix **Θ** is obtained from maximization of a log-likelihood function with a **L_2 penalization**

Witten & Tibshirani (2009)

$$L(\mathbf{\Theta}) = \log\det\mathbf{\Theta} - \mathrm{Tr}(\mathbf{S\Theta}) - \lambda\|\mathbf{\Theta}\|_F^2 \qquad (\lambda > 0)$$

EIGENVALUE PROBLEM

$$\mathbf{\Theta}^{-1} - 2\lambda\mathbf{\Theta} = \mathbf{S} \quad \Rightarrow \quad \theta_i^{\pm} = -\frac{s_i}{4\lambda} \pm \frac{\sqrt{s_i^2 + 8\lambda}}{4\lambda}$$

$$\hat{\mathbf{\Theta}} = \sum_i \theta_i^{+}\mathbf{u}_i\mathbf{u}_i^{\top}$$

$$\|\mathbf{\Theta}\|_F^2 = \mathrm{tr}(\mathbf{\Theta}^{\top}\mathbf{\Theta})$$

CHOICE OF THE PARAMETER λ

λ that maximizes penalized log-likelihood: we carry out 20 random splits of the dataset in a training and a validation sets and then we evaluate the log-likelihood over the validation set

Friedman & Tibshirani (2008)

1.5 Regularized Least Squares

RCM

Residual corr. method

Given **RLS** estimates of the variables $\mathbf{X}_i$ and $\mathbf{X}_j$, we evaluate Pearson correlation between the **residuals**

REGRESSION MODEL

$$\mathbf{X}_i = \langle \boldsymbol{\beta}_{(i)}, \mathbf{X}_{\setminus i \setminus j} \rangle + b_i \qquad \mathbf{X}_j = \langle \boldsymbol{\beta}_{(j)}, \mathbf{X}_{\setminus i \setminus j} \rangle + b_j$$

REGULARIZED LEAST SQUARES

$$\min_{\beta \in \mathbb{R}^{p-2}} \frac{1}{n} \|\mathbf{X}_i - \boldsymbol{\beta}_{(i)} \mathbf{X}_{\setminus i \setminus j}\|_2^2 + \lambda \|\boldsymbol{\beta}_{(i)}\|_2^2$$

RESIDUAL VECTORS

$$\mathbf{r}_i = \tilde{\mathbf{X}}_i - \mathbf{X}_i \qquad \mathbf{r}_j = \tilde{\mathbf{X}}_j - \mathbf{X}_j$$

PARTIAL CORR MATRIX

$$\rho_{ij \cdot V \setminus \{i,j\}} = \frac{\mathrm{cov}(\mathbf{r}_i, \mathbf{r}_j)}{\sqrt{\mathrm{var}(\mathbf{r}_i)\mathrm{var}(\mathbf{r}_j)}} = r_{\mathbf{r}_i \mathbf{r}_j}$$

CHOICE OF THE PARAMETER λ

minimization of the Leave-One-Out cross validation errors

1.6 A comparative study

GENERATED DATASETS

p	50	200	400
n	20	200	500

from multivariate Gaussian distribution $N(\mathbf{0}, \mathbf{\Sigma_{gs}})$, $\mathbf{\Sigma_{gs}} = \mathbf{\Theta_{gs}^{-1}}$

STRUCTURE AND SPARSITY OF Θ_{gs}^{-1}

RANDOM	HUBS	*p(p-1)/2*
CLIQUES		*2p*

RANDOM

off-diagonal terms are set randomly to a fixed value $\theta_{ik} = \theta$

HUBS

we partition the columns into disjoint groups G_k
index *k* indicates the *k*-th column chosen as **central** in each group.
off-diagonal terms $\theta_{ik} = \theta$ if $i \in G_k$, otherwise $\theta_{ik} = 0$

CLIQUES

fully connected hubs

ACCURACY AND TIMING

For each pattern, for each inferring method, we evaluate timing and AUC performances
(Accuracy of classification of edges and non-edges)

Friedman & Tibshirani (2010)

1.7 Results of comparative study

p = 400

		ℓ_{2C}			PINV			RCM		
	n	AUC	AUC std	T (s)	AUC	AUC std	T (s)	AUC	AUC std	T (s)
r	500	0.998	0.0001	38.86	0.987	0.0006	0.161	0.999	0.0001	8343
h	500	1.000	0.0000	83.74	0.999	0.0000	0.164	1.000	0.0000	6468
c	500	0.995	0.0002	84.95	0.963	0.0014	0.164	0.996	0.0002	6449
r	200	0.976	0.0003	38.44	0.581	0.0161	0.111	0.984	0.0006	3566
h	200	1.000	0.0000	81.13	0.806	0.0150	0.115	0.999	0.0001	3555
c	200	0.936	0.0008	82.02	0.587	0.0049	0.121	0.923	0.0009	3747
r	20	0.808	0.0011	39.03	0.929	0.0018	0.093	0.924	0.0017	105
h	20	0.999	0.0001	82.03	1.000	0.0000	0.091	0.999	0.0000	106
c	20	0.668	0.0014	82.13	0.659	0.0014	0.091	0.659	0.0014	108

p = 200

		ℓ_{2C}			PINV			RCM		
	n	AUC	AUC std	T (s)	AUC	AUC std	T (s)	AUC	AUC std	T (s)
r	500	0.999	0.0001	5.807	0.999	0.0001	0.0377	0.999	0.0001	807
h	500	1.000	0.0000	10.655	1.000	0.0000	0.0376	1.000	0.0000	450
c	500	0.996	0.0002	10.821	0.999	0.0001	0.0439	0.999	0.0000	436
r	200	0.986	0.0003	5.592	0.703	0.0067	0.0310	0.990	0.0007	861
h	200	1.000	0.0000	10.425	0.748	0.0124	0.0309	0.999	0.0003	856
c	200	0.944	0.0010	10.529	0.612	0.0064	0.0336	0.950	0.0008	1028
r	20	0.784	0.0016	6.150	0.880	0.0048	0.0187	0.871	0.0046	24.5
h	20	0.999	0.0001	10.574	0.999	0.0002	0.0182	0.999	0.0001	27.9
c	20	0.669	0.0016	10.545	0.649	0.0017	0.0189	0.654	0.0017	25.3

Schaffer & Strimmer (2005)

2 **Infer** a gene network

for the isoprenoid biosinthesis **pathways** in *A. thaliana*

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

2.1 Isoprenoid pathways in *A. Thaliana*

ISOPRENOIDS

group of plant natural products.

FUNCTIONS

membrane components, hormones and plant defence compounds, etc.

MVA AND MPE PATHWAYS

They are synthesized through two different routes that take place in **two distinct cellular compartments**.

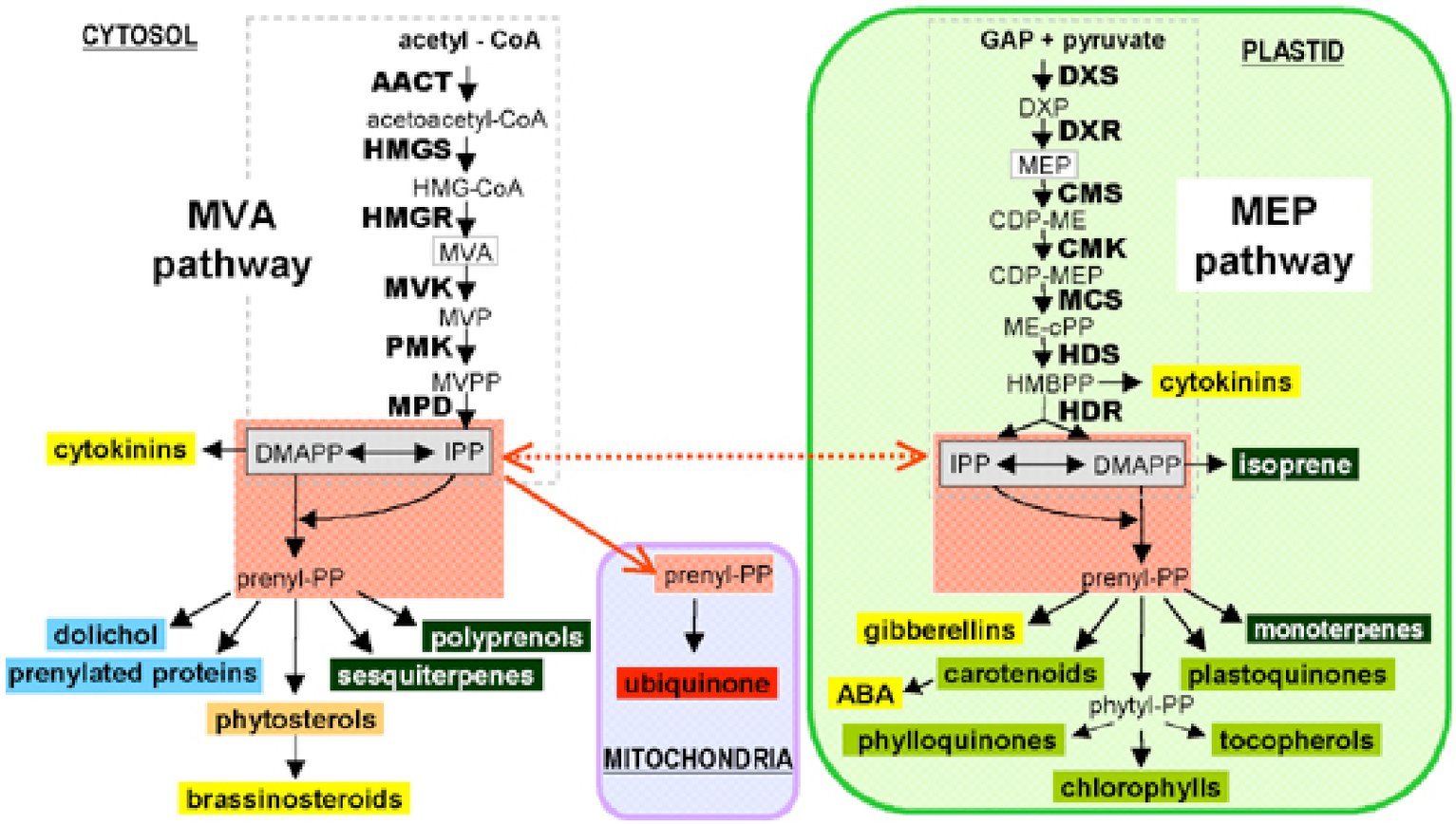


image from Universitat de Barcelona website http://www.bq.ub.es/~mrodrigu/RESEARCH.htm

Laule *et al.*, *PNAS* (2003)

Evidence of interactions at metabolic level
Gene expression levels **do not respond** to the single inhibition of the two pathways

Wille & Bulhman, *Genome Biology* (2004)

Beyond **one-gene** approach, a GRN has been inferred (795 gene expr. levels from other 56 pathways). It has been shown the possible presence of various connections between genes in the two pathways, i.e. **possible crosstalk at trascriptional level**

2.2 Inferring the crosstalk

DATASET — Wille (2004)

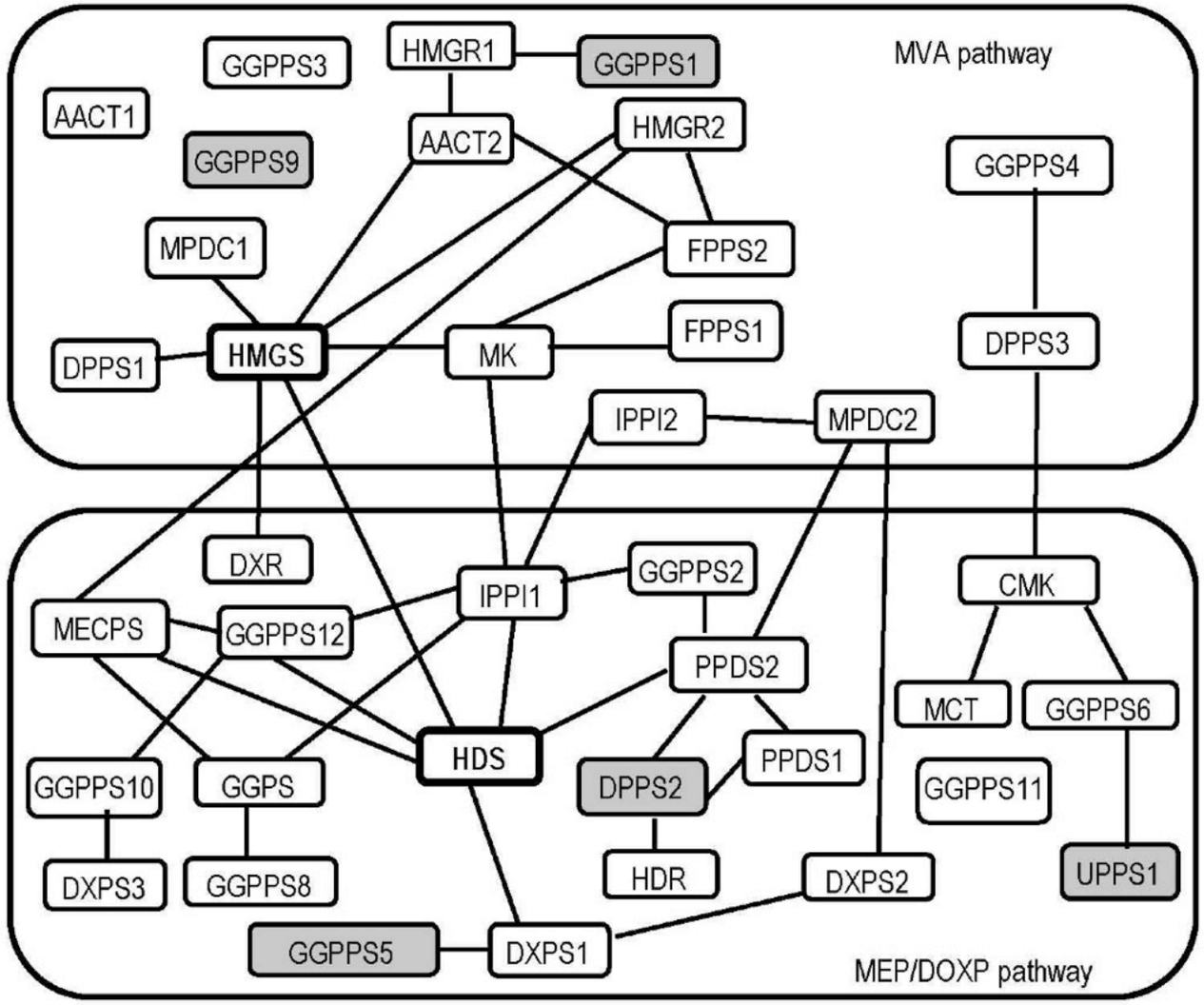
39 genes E.L. — **MVA & MEP**

+ — **118 observations**

795 genes E.L. — **56 PATHWAYS**

METHOD — L_{2C}

GRAPH SELECTION — 95% **bootstrap c.i.** of the statistics.



1. For each pathway: a module with **strongly interconnected** and positively correlated genes
2. Two strong candidate **hub** genes for the cross-talk between the pathways: **HMGS** and **HDS**
3. The negative correlation between HMGS and HDS means that **they respond differently** to the several tested experimental conditions: **possible evidence of a cross-talk**

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

Conclusions

1. We have provided a preliminary comparative study of three methods to obtain **estimates of partial correlation matrix**, in the regime ***n << p***
2. On the basis of the best AUC and timing performances, we have applied a **covariance-regularized method** (with L_2 penalty) to infer a gene network for isoprenoid biosynthesis pathways in *Arabidopsis thaliana*
3. We have found the evidence of **cross-talk** between the two pathways MVA and MEP, as expected in literature

Outlook

I. **Improving** inferring methods (e.g. novel algorithms for a more accurate edge selections) and applications to cancer or human disease
II. Investigation based *a priori* on **real network** properties (scale-free and small-world topologies, etc.)

R Anglani, PR PS BB 2011 - 13. Sept 2011 - A comparative study of GGM approaches for genomic data

Thanks for your attention

No problem is too small or too trivial if we can really do something about it. (R. P. Feynman)

Bootstrap confidence interval

Generate 100 resamples of the observed dataset (of equal size of the observed data set), obtained by random sampling with replacement from the original dataset.

We build a distribution for each element of the rho matrix, and we consider a non-edge if the zero value is contained in the 95% confidence interval.

Area Under the ROC curve

AUC is equal to the probability that a classifier will rank randomly chosen positive istance higher than a randomly chosen negative one.

A singular value decomposition of a $m \times q$ matrix M, is $M = U\Lambda V^*$, where U is a $m \times m$ unitary matrix, Λ is $m \times q$ diagonal matrix with nonnegative real numbers on the diagonal and V^* is a $q \times q$ unitary matrix (transpose conjugate of V). Then, the pseudoinverse of M is $M^+ = V\Lambda^+U^*$, where Λ^+ is obtained by replacing each diagonal element with its reciprocal and then transposing the matrix.

20 random splits

Split in X9 and X1 then evaluate S9. Then for a fixed window of lambda and evaluate Theta9 from the penalized log-likelihood.

For 20 splits we evaluate the log-likelihood the log-likelihood without penalty using S1, according to a fixed window of lambda values.

For each value of lambda we evaluate the average of the log-likelihood value over the 20 splits.

Then we choose among the window of lambdas, the one that maximizes the log-likelihood.

Then we use this lambda to evaluate the final precision matrix over the original dataset.