rags2ridges
Network-Based Modeling and Analysis of High-Dimensional Omics Data

Carel F.W. Peeters
Dept. of Epidemiology & Biostatistics
VU University medical center, Amsterdam
cf.peeters@vumc.nl

Bioinformatics Seminar CCA
Amsterdam, the Netherlands
March 10, 2017

Contributors

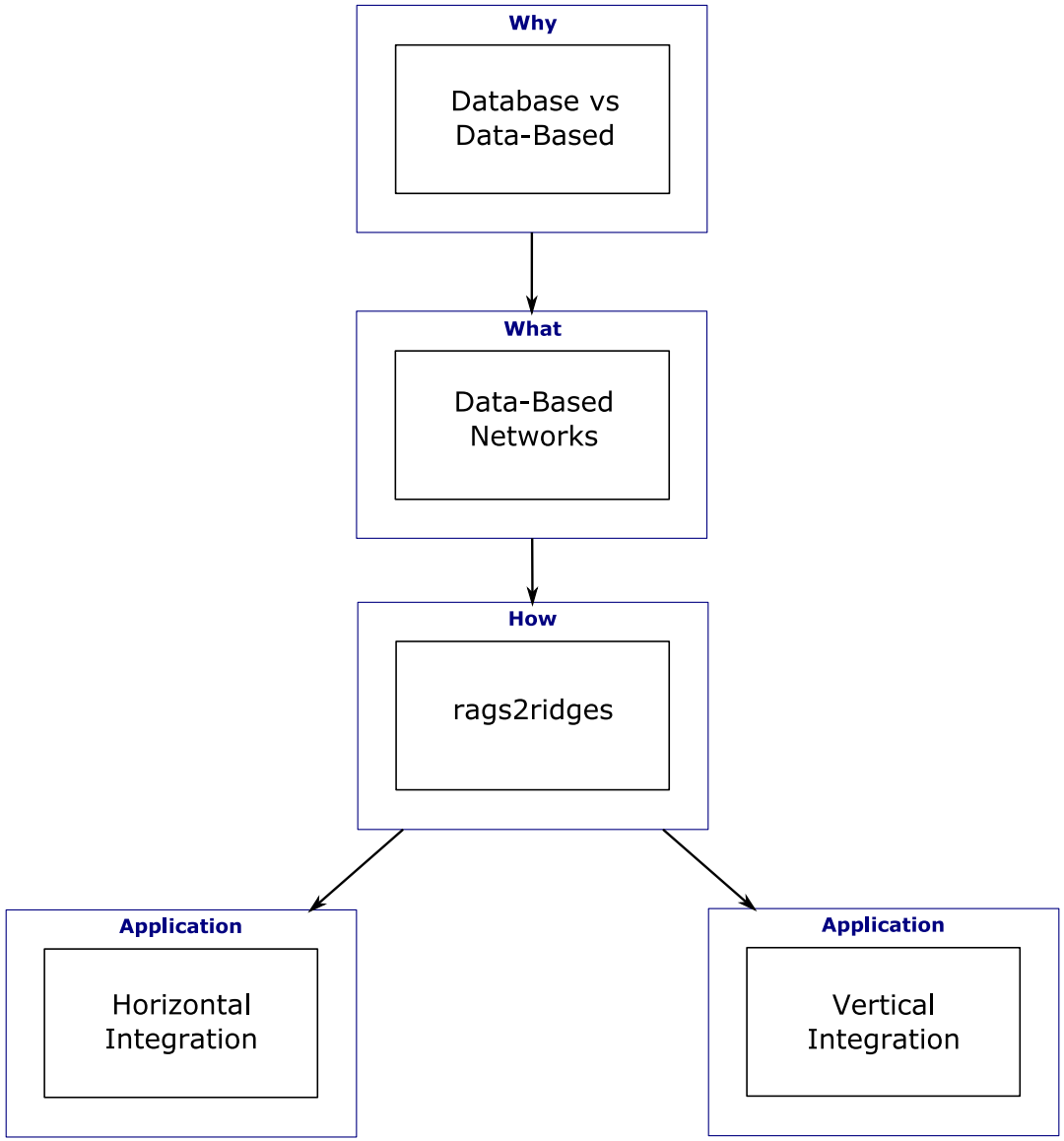
Anders E. Bilgrau
Novo Nordisk
Dept. of Mathematical Sciences, Aalborg University

Wessel N. van Wieringen
Dept. of Epimiology & Biostatistics, VUMC
Dept. of Mathematics, VU University Amsterdam

Mark A. van de Wiel
Dept. of Epimiology & Biostatistics, VUMC
Dept. of Mathematics, VU University Amsterdam


hot stuff!

Overview


Why
Database vs
Data-Based
What
Data-Based
Networks
How
rags2ridges
Application
Horizontal
Integration
Application
Vertical
Integration

Pathway Databases

Metabolic pathway databases

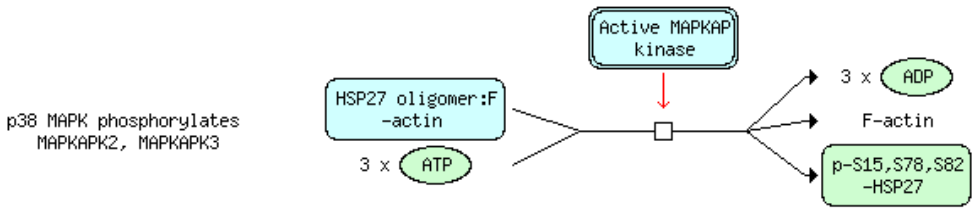
- MetaCyc
- HumanCyc
- BioCyc

Signaling pathway databases

- KEGG
- BioCarta
- Reactome
- Biomodels
- Human signaling network
- Ingenuity
- PID
- BioPP

Manual Curation

1.2.50 p-MAPK2/3 phosphorylates HSP27 (BlackBoxEvent)



Authors
Garapati, P V, 2013-08-30.

Reviewers
Welsh, Michael, Berger, Philipp, Ballmer-Hofer, Kurt, 2014-05-12.

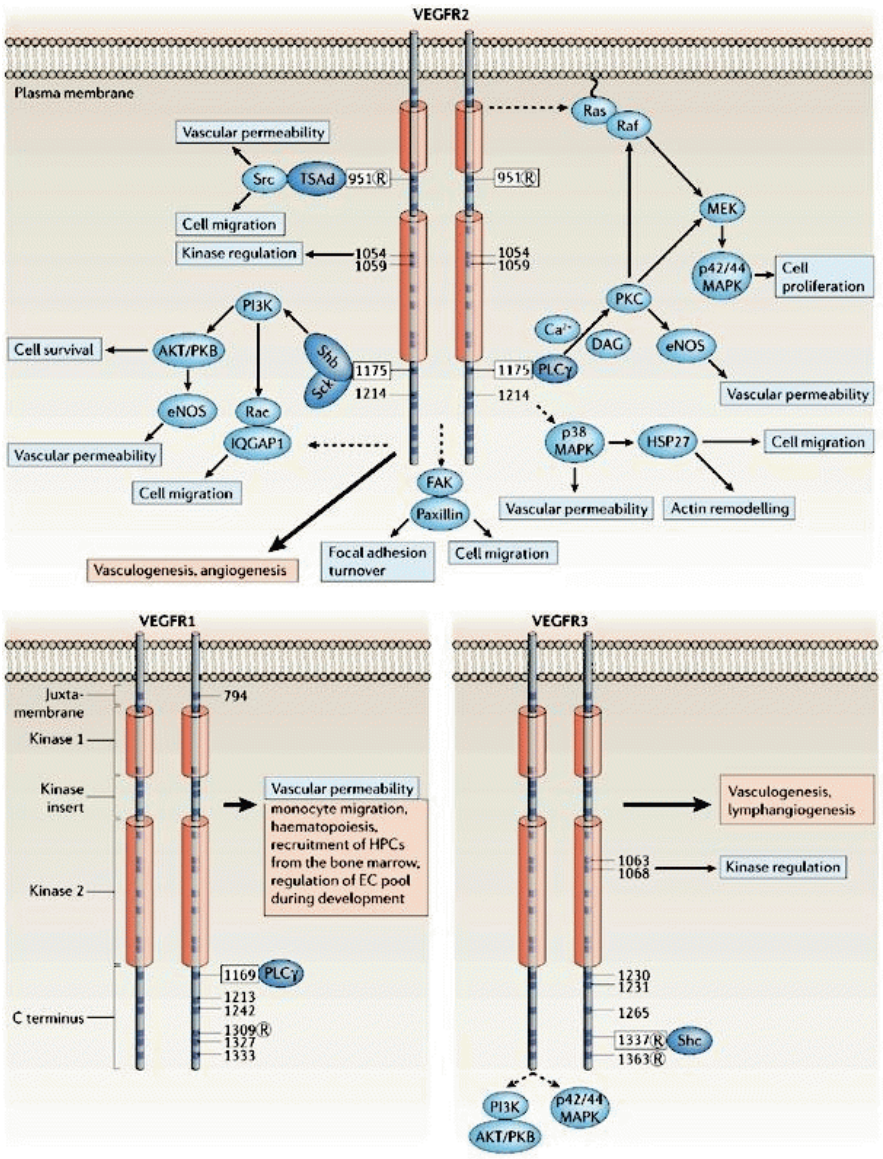
Activated MAP kinase-activated protein kinase (MAPK/MAPKAPK) 2 and 3 in turn phosphorylate heat shock protein beta 1 (HSPB1, HSP27).

References

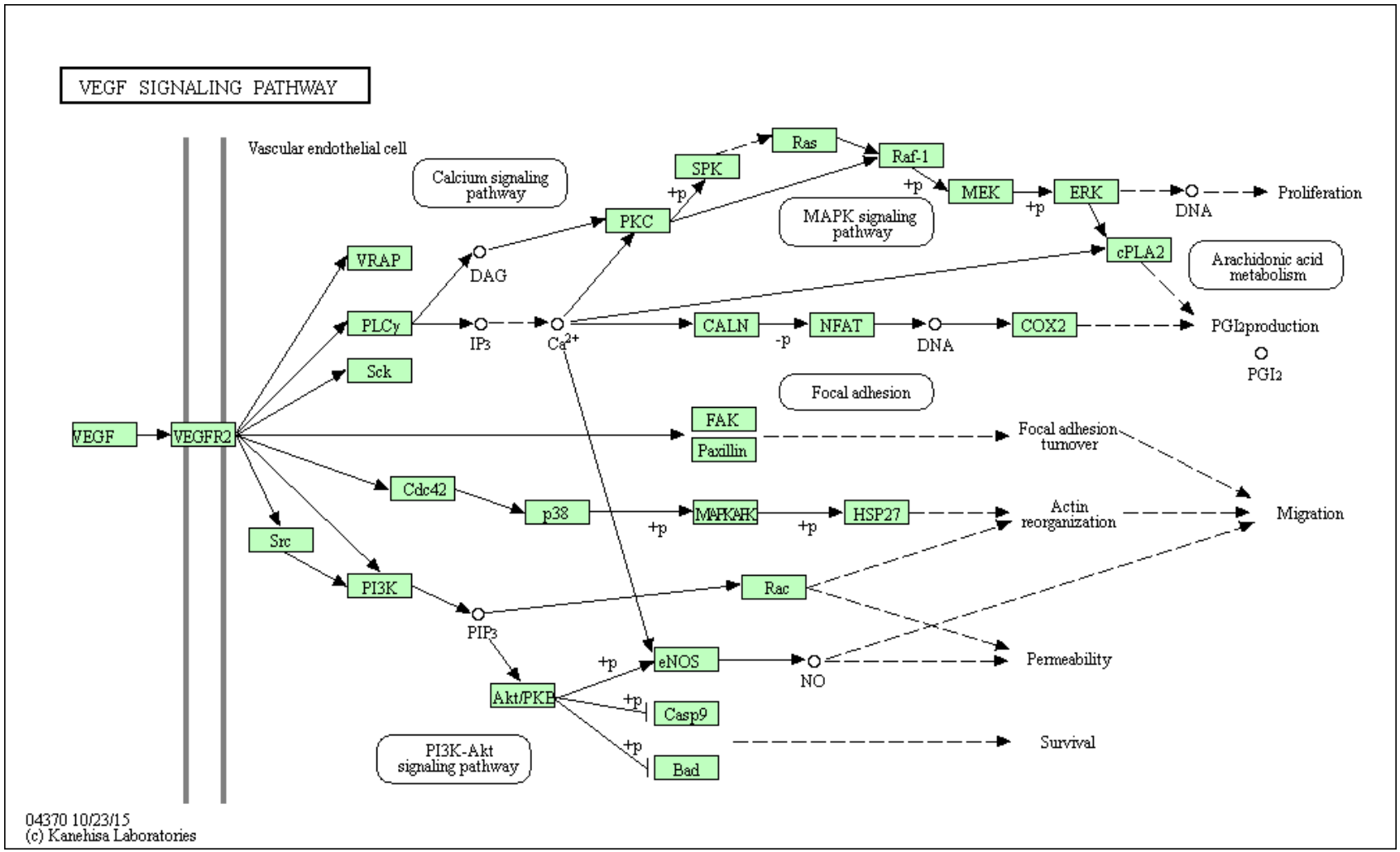
Landry J, Huot J, "Regulation of actin dynamics by stress-activated protein kinase 2 (SAPK2)-dependent phosphorylation of heat-shock protein of 27 kDa (Hsp27)", Biochem. Soc. Symp., 64, 1999, 79-89.

Lavoie JN, Lambert H, Hickey E, Weber LA, Landry J, "Modulation of cellular thermoresistance and actin filament stability accompanies phosphorylation-induced changes in the oligomeric structure of heat shock protein 27", Mol. Cell. Biol., 15, 1995, 505-16.

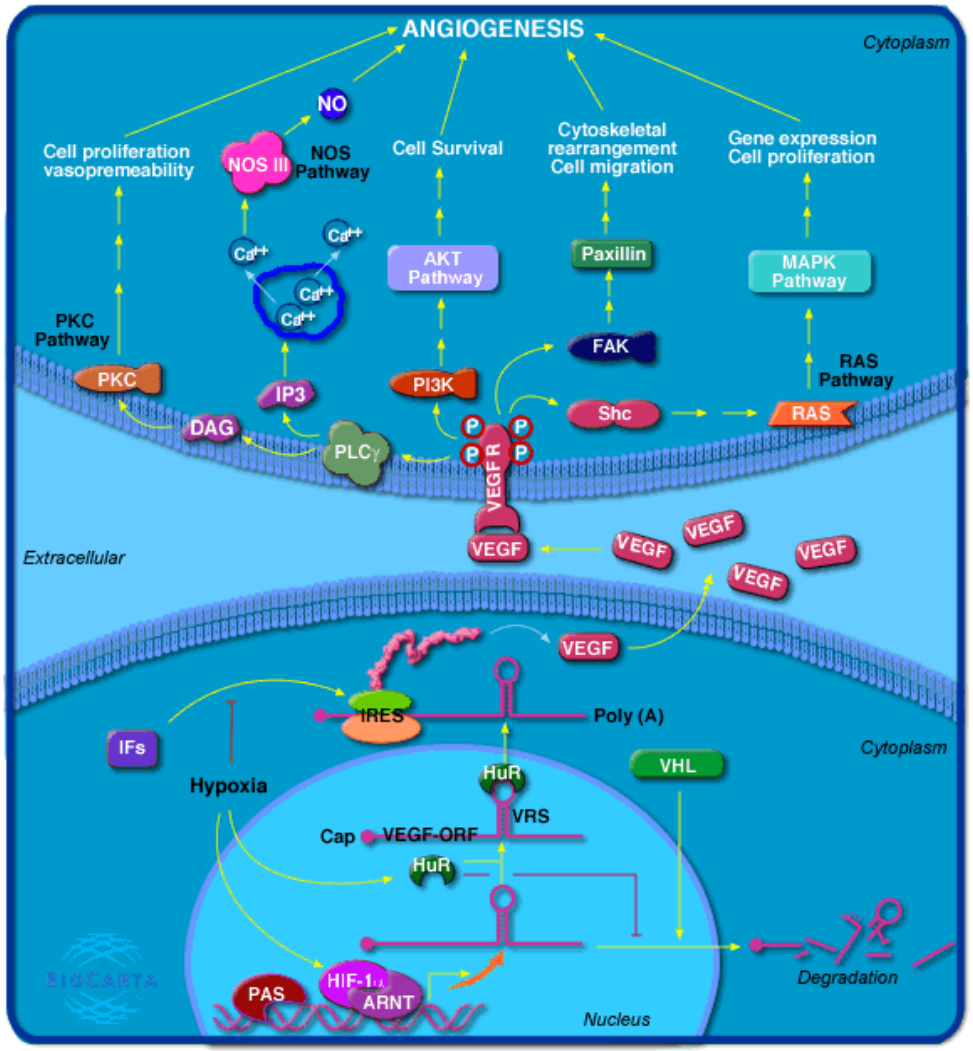
VEGF Signaling According to REACTOME



VEGF Signaling According to KEGG



VEGF Signaling According to BIOCARTA



Graphs

Representation

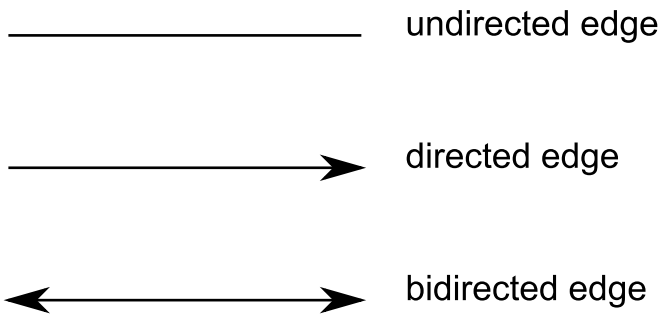
Pathways are represented by a *graph* (or *network*)

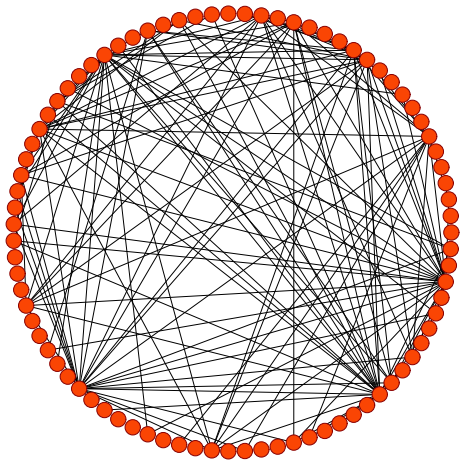
Vertices

- *Node* or *vertex* represents molecular feature

Edges

Edge or *arrow* represents some functional relation



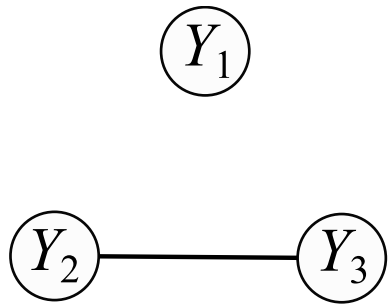


Correlation networks

Example

Three variables: Y_1, Y_2, and Y_3

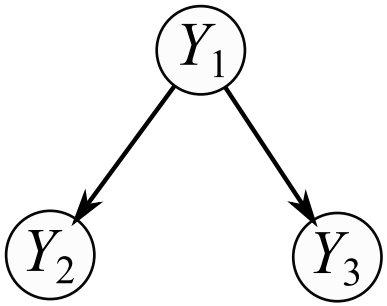
$$\text{cor}(Y_1, Y_2) = 0$$
$$\text{cor}(Y_1, Y_3) = 0$$
$$\text{cor}(Y_2, Y_3) \neq 0$$



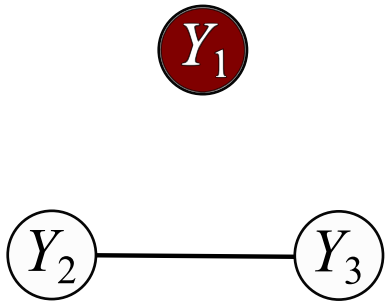
Marginal dependence

Undirected edge represents marginal dependence

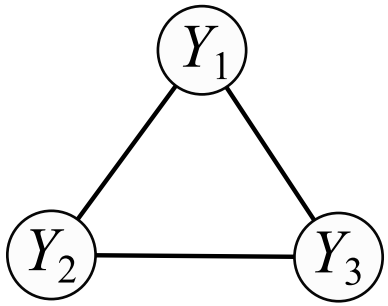
Interpretational danger



True mechanism



Not observing Y_1:
Spurious association



Observing Y_1:
Saturated graph

Conditional dependence

Partial correlation

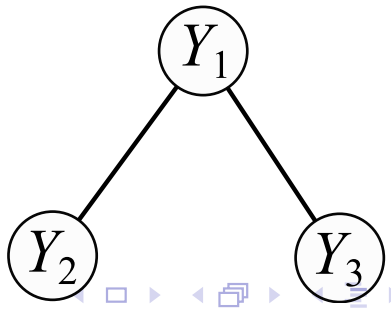
Measures degree of association between two random variables when controlling for third variables

Conditioned correlation

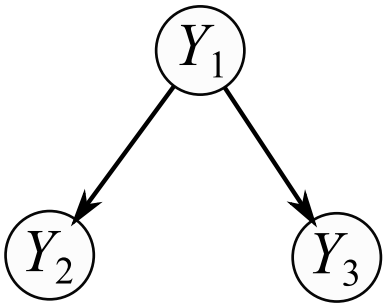
$$\text{cor}(Y_1, Y_2|Y_3)$$
$$\text{cor}(Y_1, Y_3|Y_2)$$
$$\text{cor}(Y_2, Y_3|Y_1)$$

If, e.g., $\text{cor}(Y_2, Y_3|Y_1) = 0$, we say Y_2 and Y_3 are independent *given* Y_1

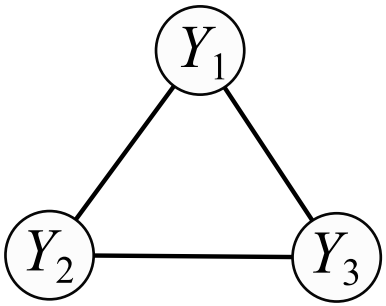
$$\text{cor}(Y_1, Y_2|Y_3) \neq 0$$
$$\text{cor}(Y_1, Y_3|Y_2) \neq 0$$
$$\text{cor}(Y_2, Y_3|Y_1) = 0$$



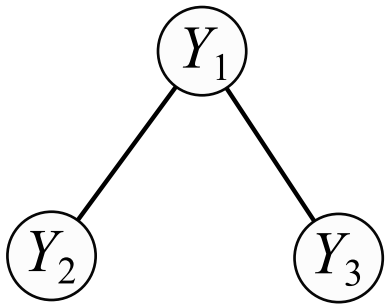
Partial Correlation Networks



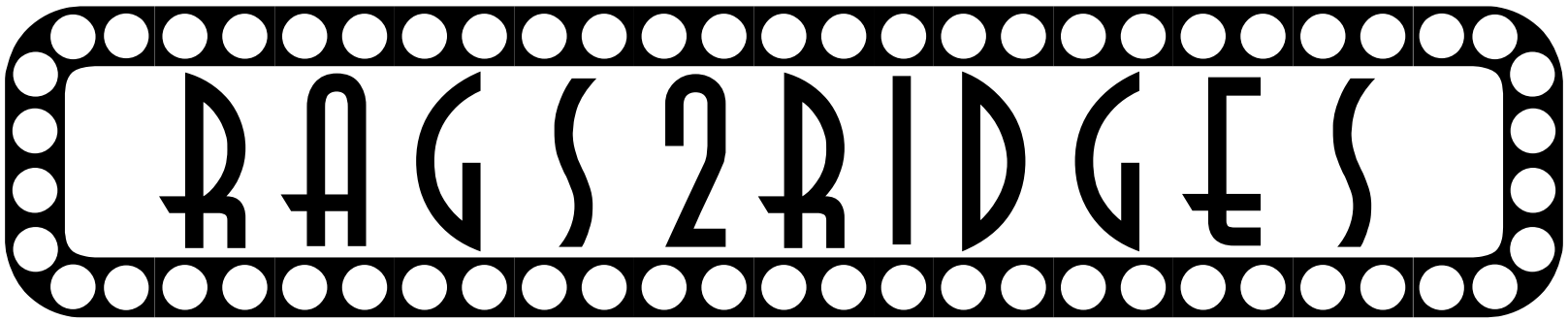
True mechanism



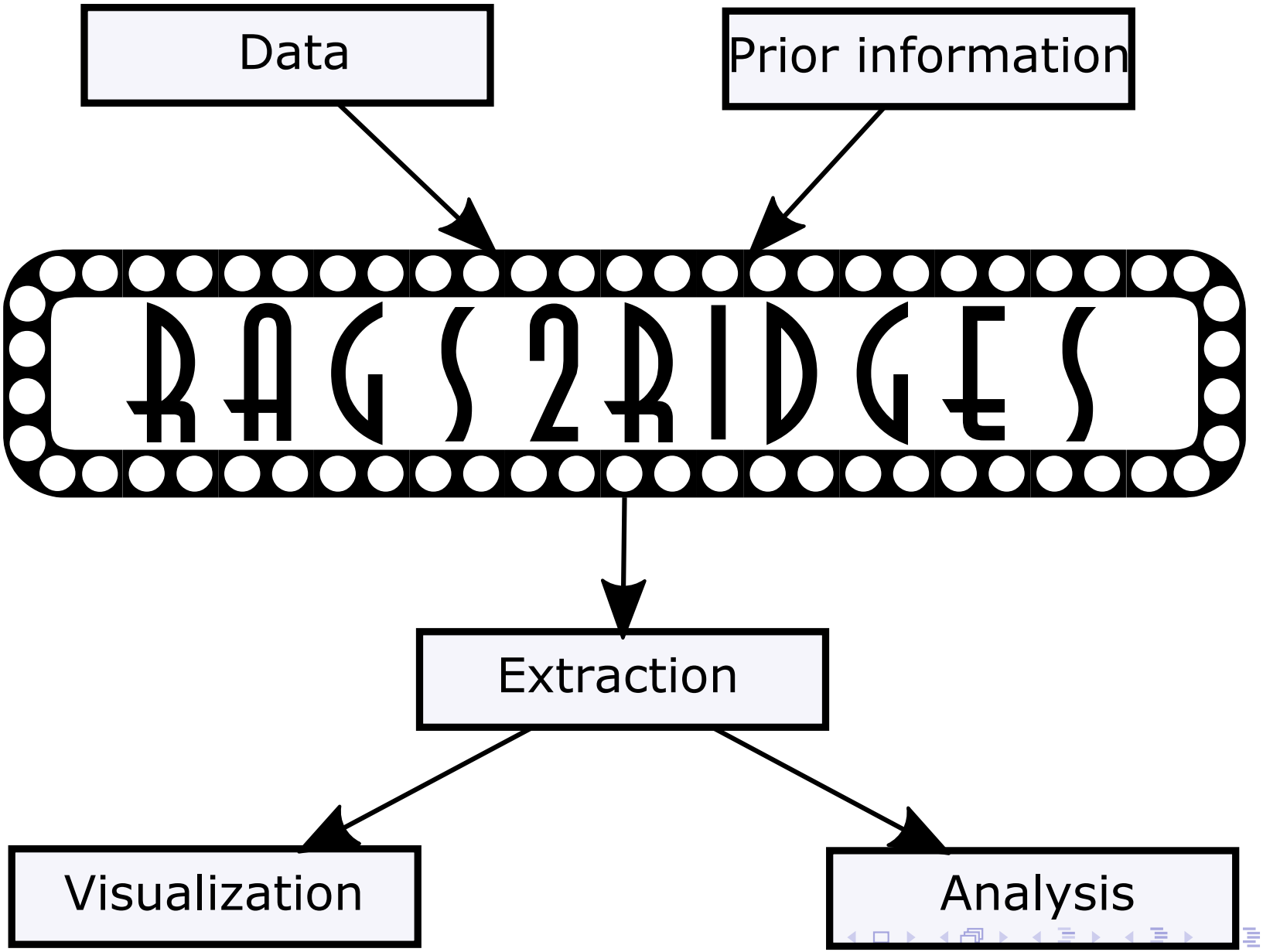
Correlation graph



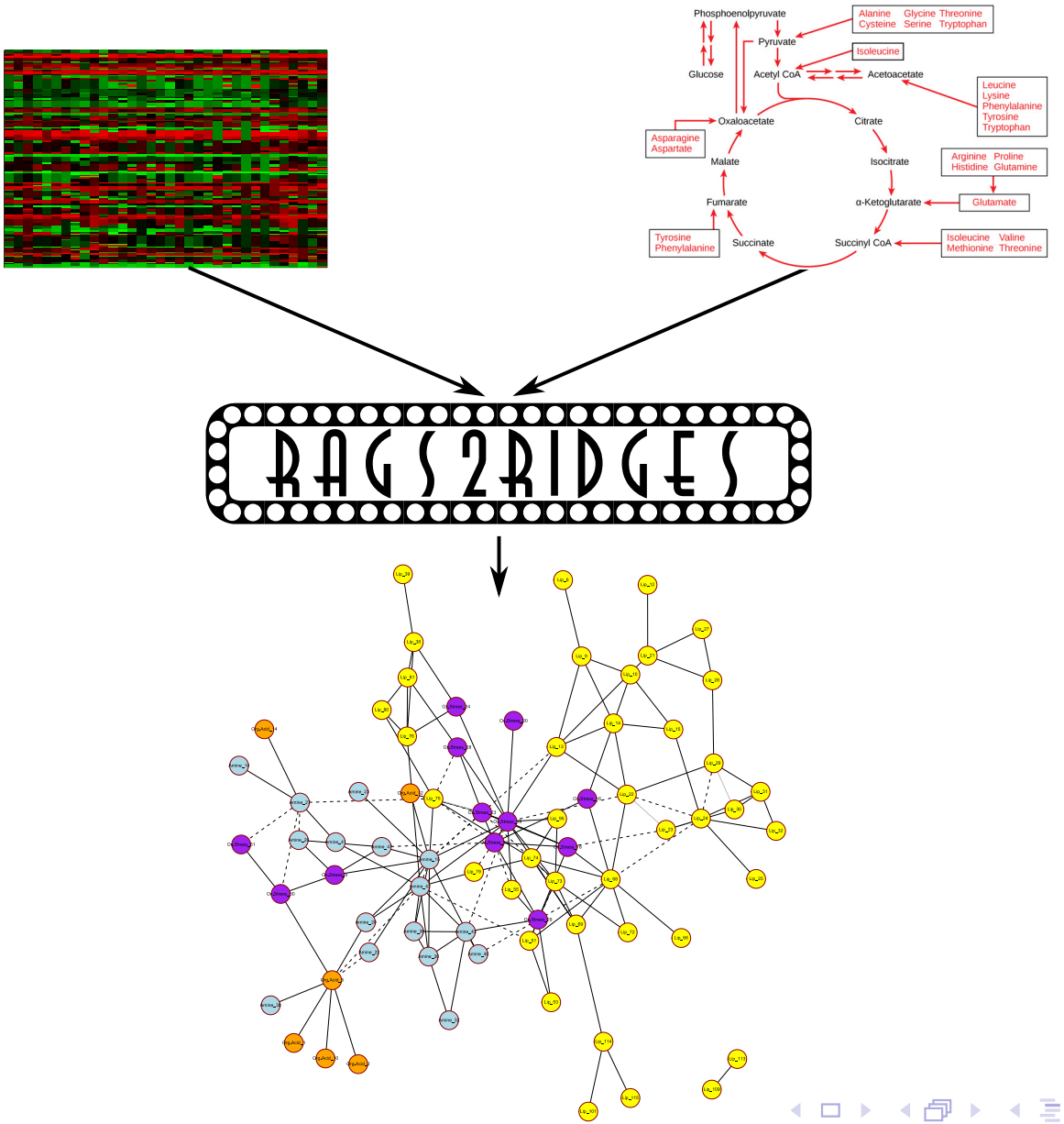
Conditional independence graph


RAGS2RIDGES

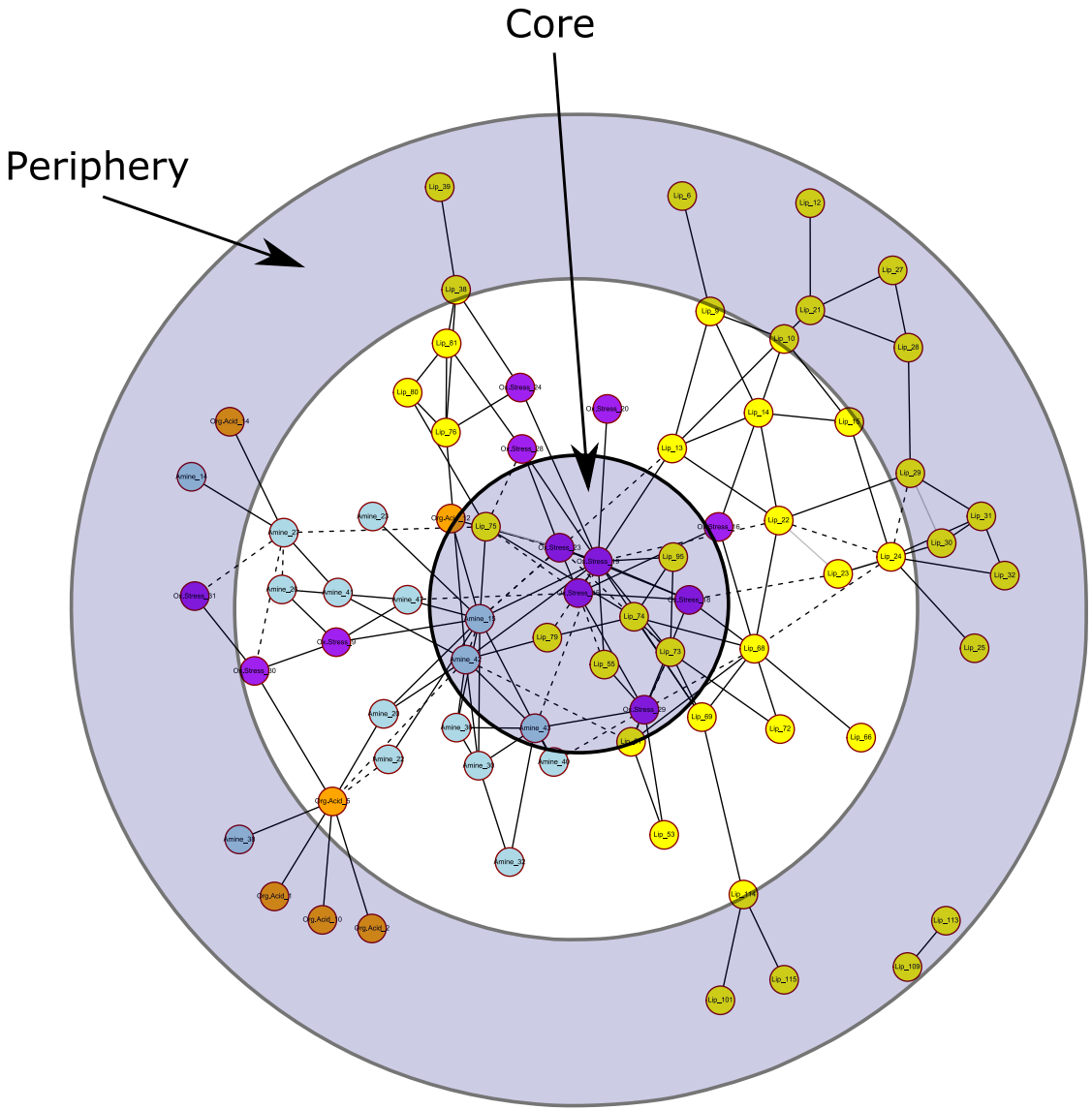
One-Stop-Go


Data
Prior information
RAGS2RIDGES
Extraction
Visualization
Analysis

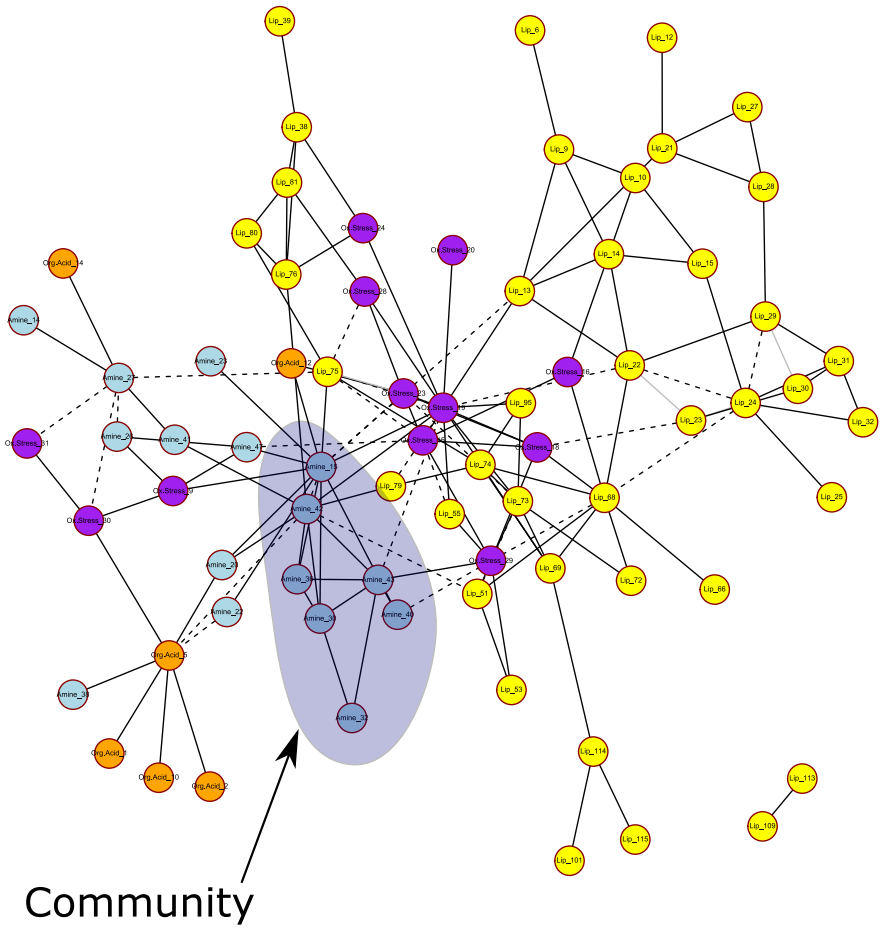
One-Stop-Go


Phosphoenolpyruvate
Alanine Glycine Threonine
Cysteine Serine Tryptophan
Pyruvate
Isoleucine
Glucose
Acetyl CoA
Acetoacetate
Leucine
Lysine
Phenylalanine
Tyrosine
Tryptophan
Oxaloacetate
Citrate
Asparagine
Aspartate
Malate
Isocitrate
Arginine Proline
Histidine Glutamine
Fumarate
α-Ketoglutarate
Glutamate
Tyrosine
Phenylalanine
Succinate
Succinyl CoA
Isoleucine Valine
Methionine Threonine
RAGS2RIDGES

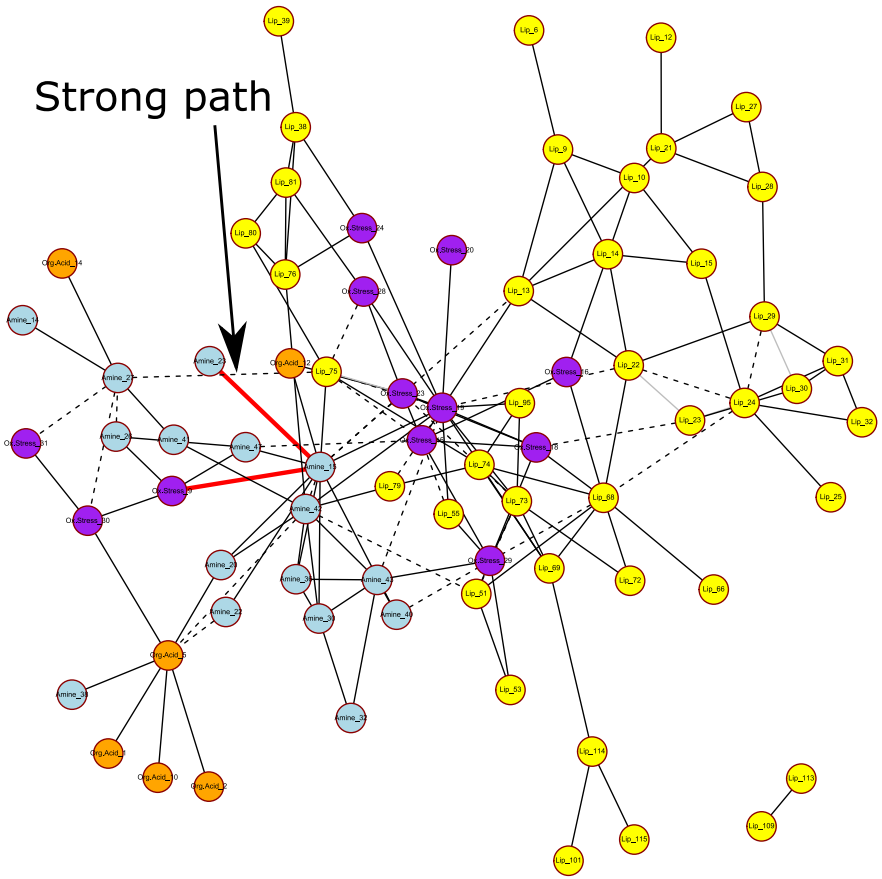
Analysis: Global Level



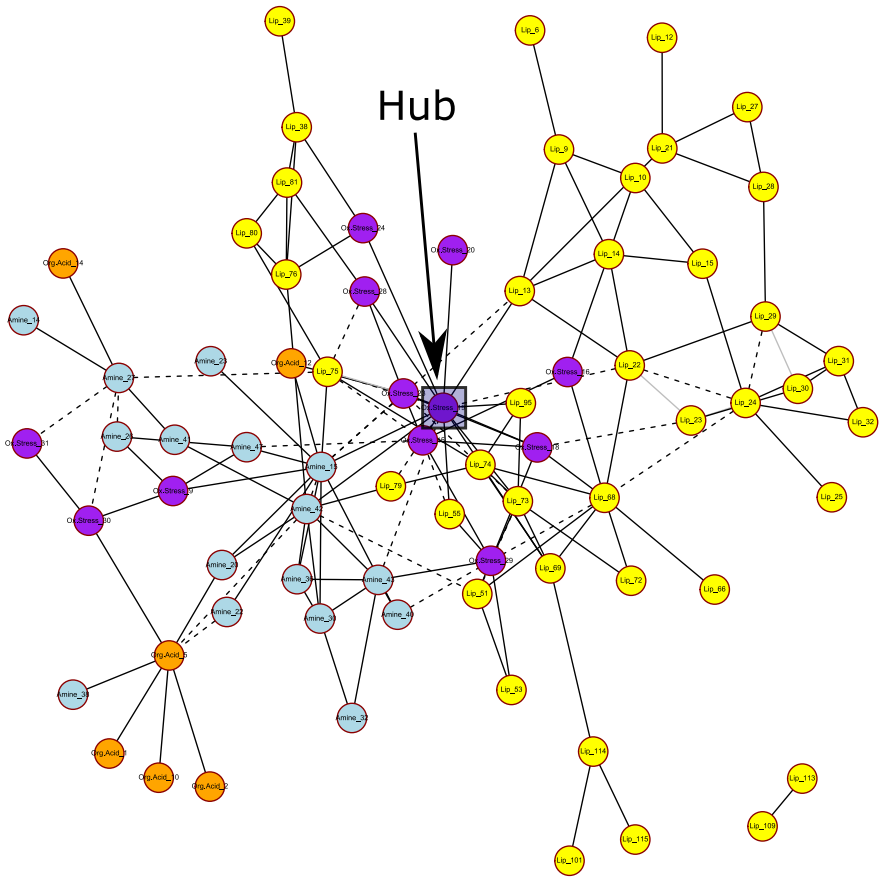
Analysis: Group Level



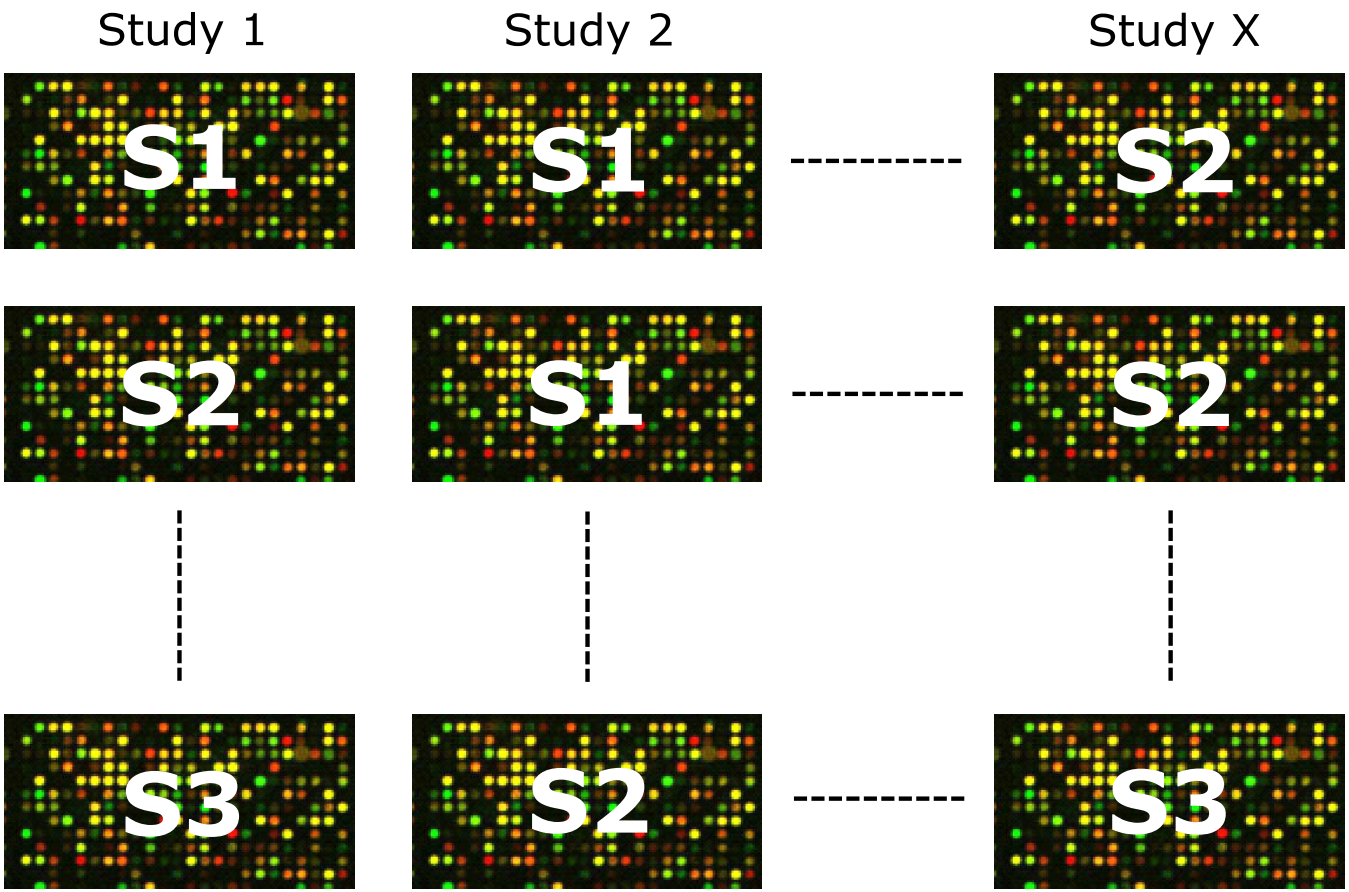
Analysis: Path Level



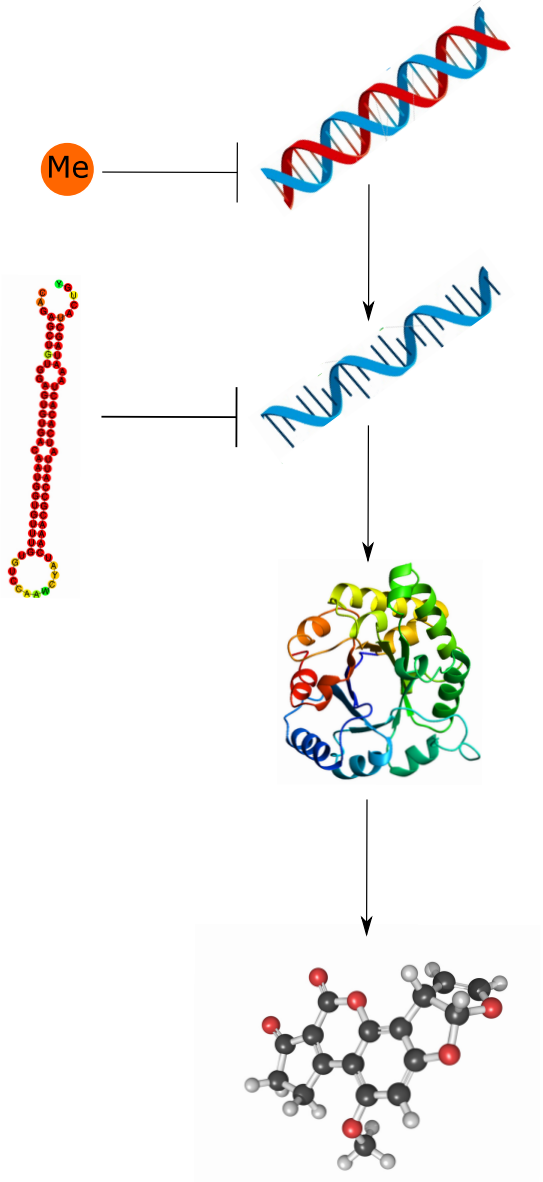
Analysis: Node Level



Integration: Horizontal



Integration: Vertical



Background

DLBCL

- Diffuse large B-cell lymphomas
- A non-Hodgkin type of blood cancer

DLBCL subtypes

At least two major genetic subtypes of tumors:

- ABC: activated B-cells
- GCB: germinal centre B-cells
- III: cannot be classified as either ABC or GCB

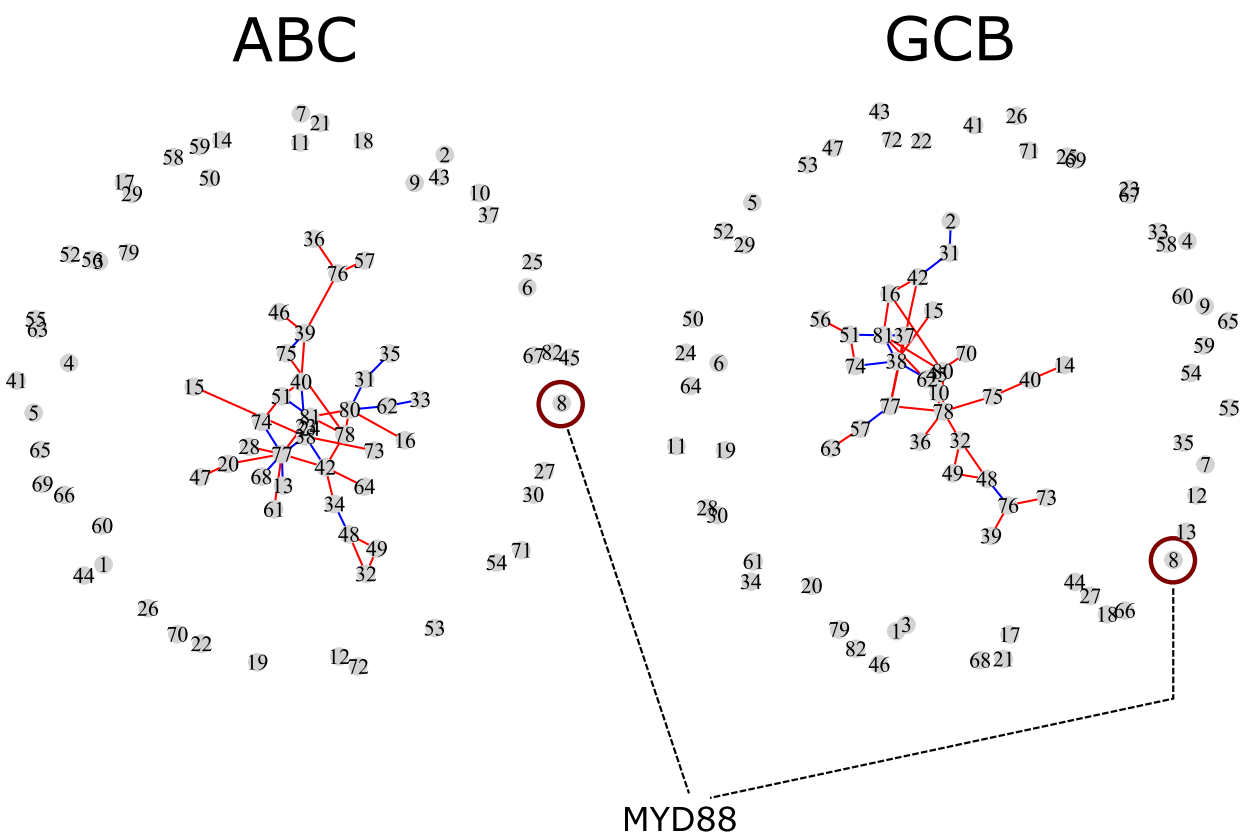
NF-κB signaling pathway

- Responsible for control of cell survival
- Known to be deregulated in DLBCL
- Hallmark distinguishing *poor prognostic* ABC from *good prognostic* GCB

A non-integrative analysis

Data

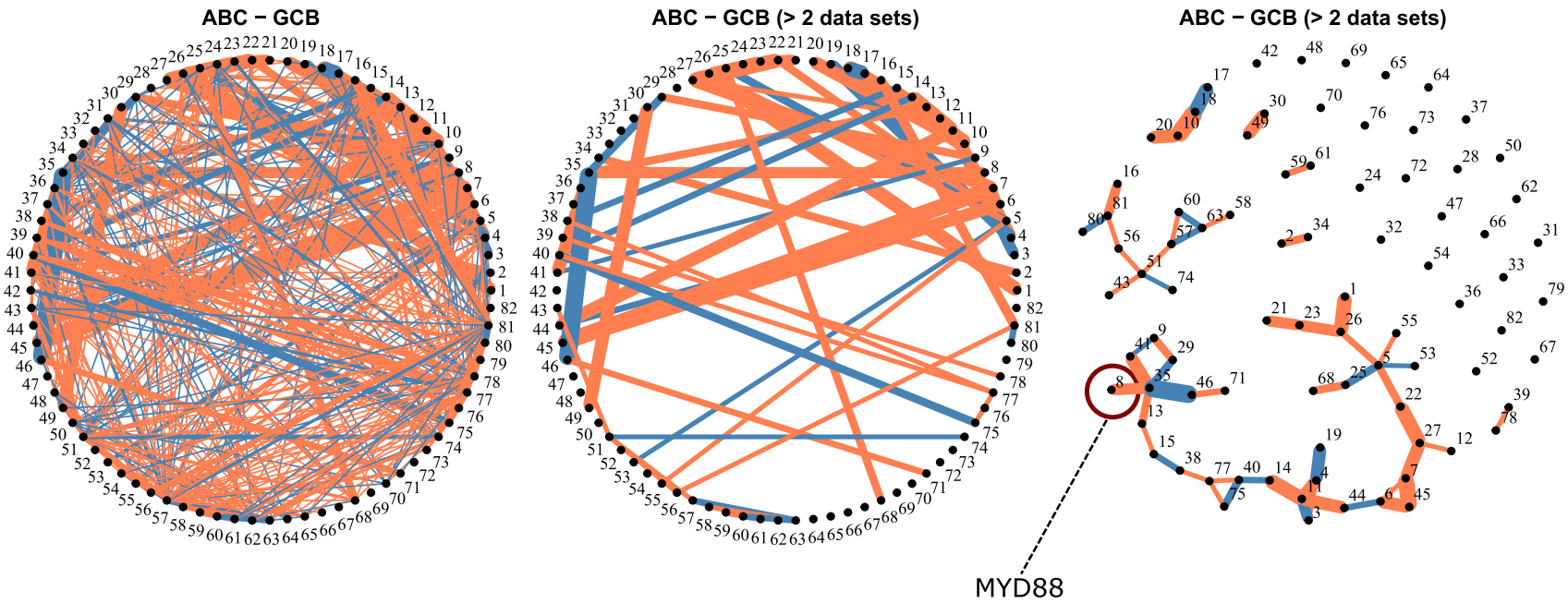
- $n = 89$ DLBCL tumor samples
- ABC ($n_1 = 31$), III ($n_2 = 13$), and GCB ($n_3 = 45$)
- $p = 82$ (KEGG)



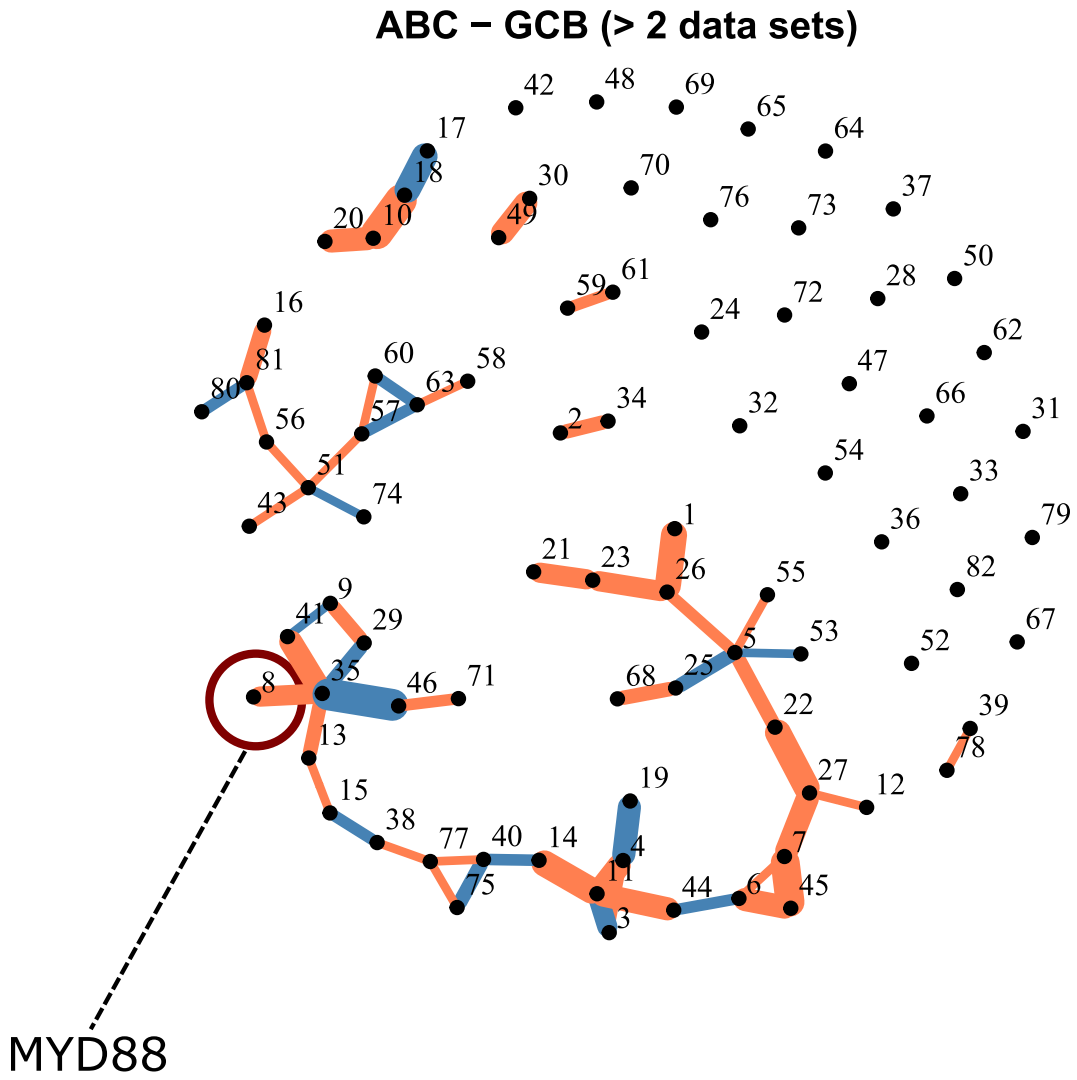
Integrative analysis: Data

	ABC		Type III		GBC		
	g	n_g	g	n_g	g	n_g	$\sum n_g$
Pilot data							
GSE11318		74		71		27	172
Data set							
GSE56315	1	31	2	13	3	45	89
GSE19246	4	51	5	30	6	96	177
GSE12195	7	40	8	18	9	78	136
GSE22895	10	31	11	21	12	49	101
GSE31312	13	146	14	97	15	224	467
GSE10846.CHOP	16	64	17	28	18	89	181
GSE10846.RCHOP	19	75	20	42	21	116	233
GSE34171.hgu133plus2	22	23	23	15	24	52	90
GSE34171.hgu133AplusB	25	18	26	17	27	43	78
GSE22470	28	86	29	43	30	142	271
GSE4475	31	73	32	20	33	128	221
$\sum n_g$		638		344		1062	2044

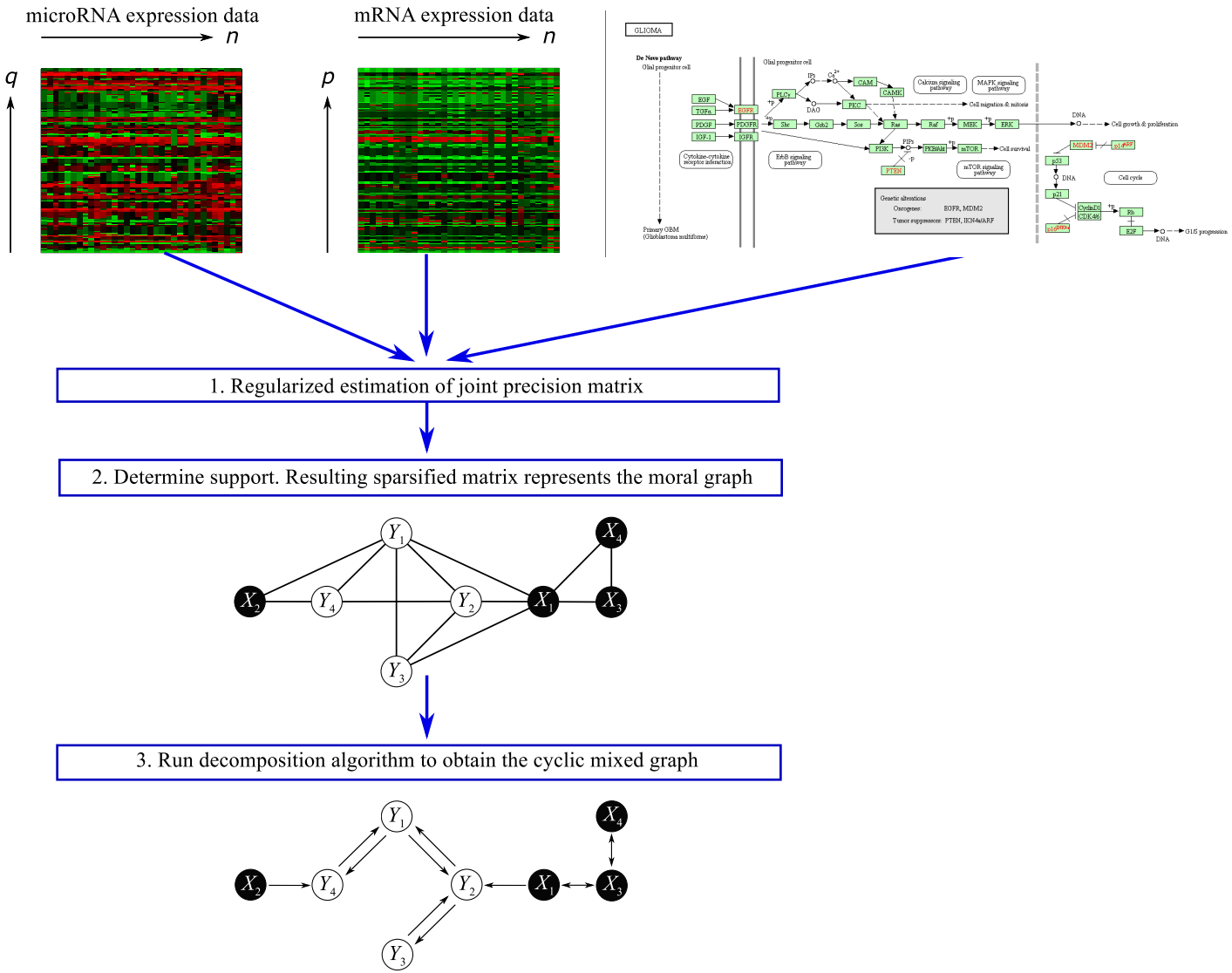
Integrative analysis: Results



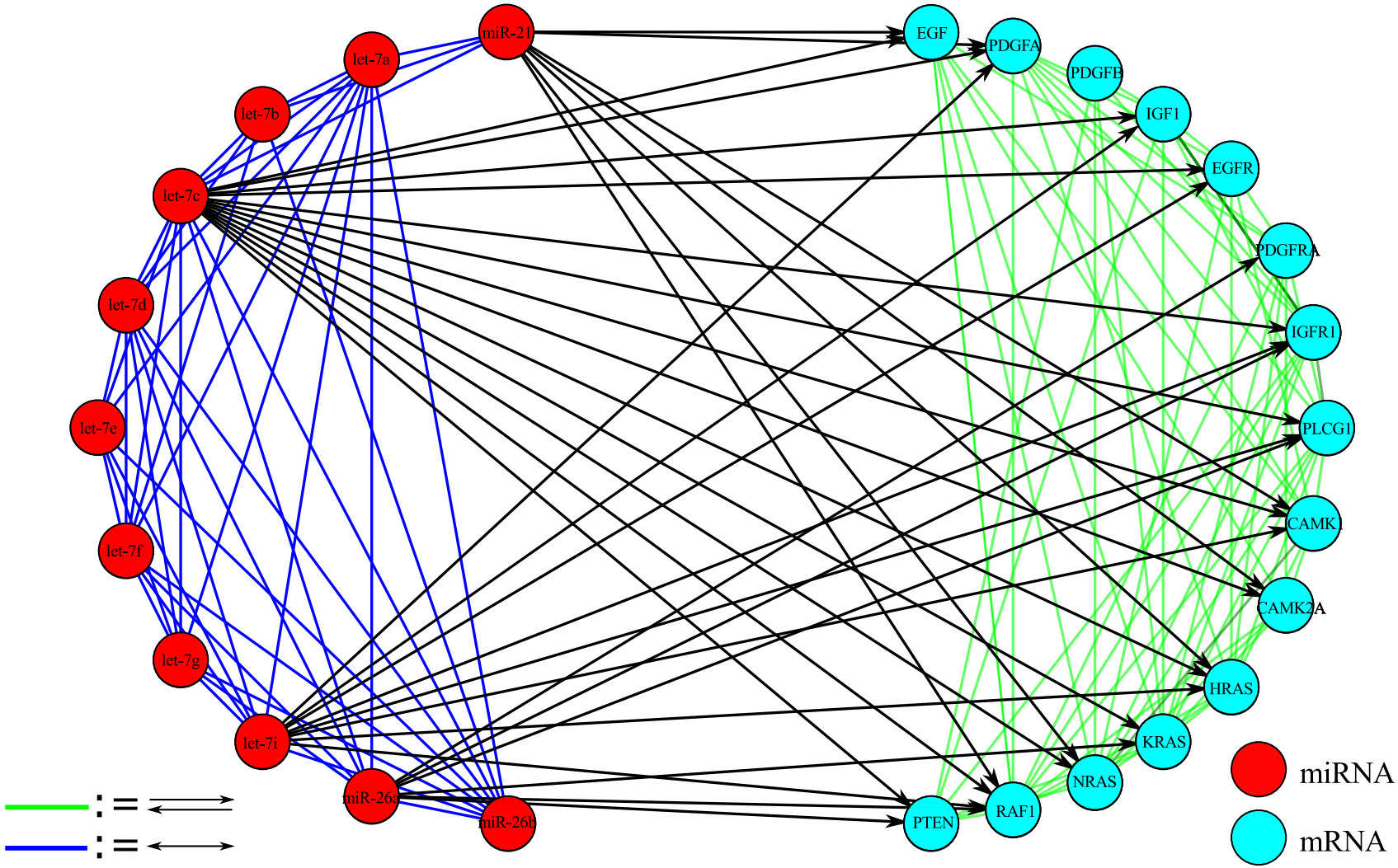
Integrative analysis: Results



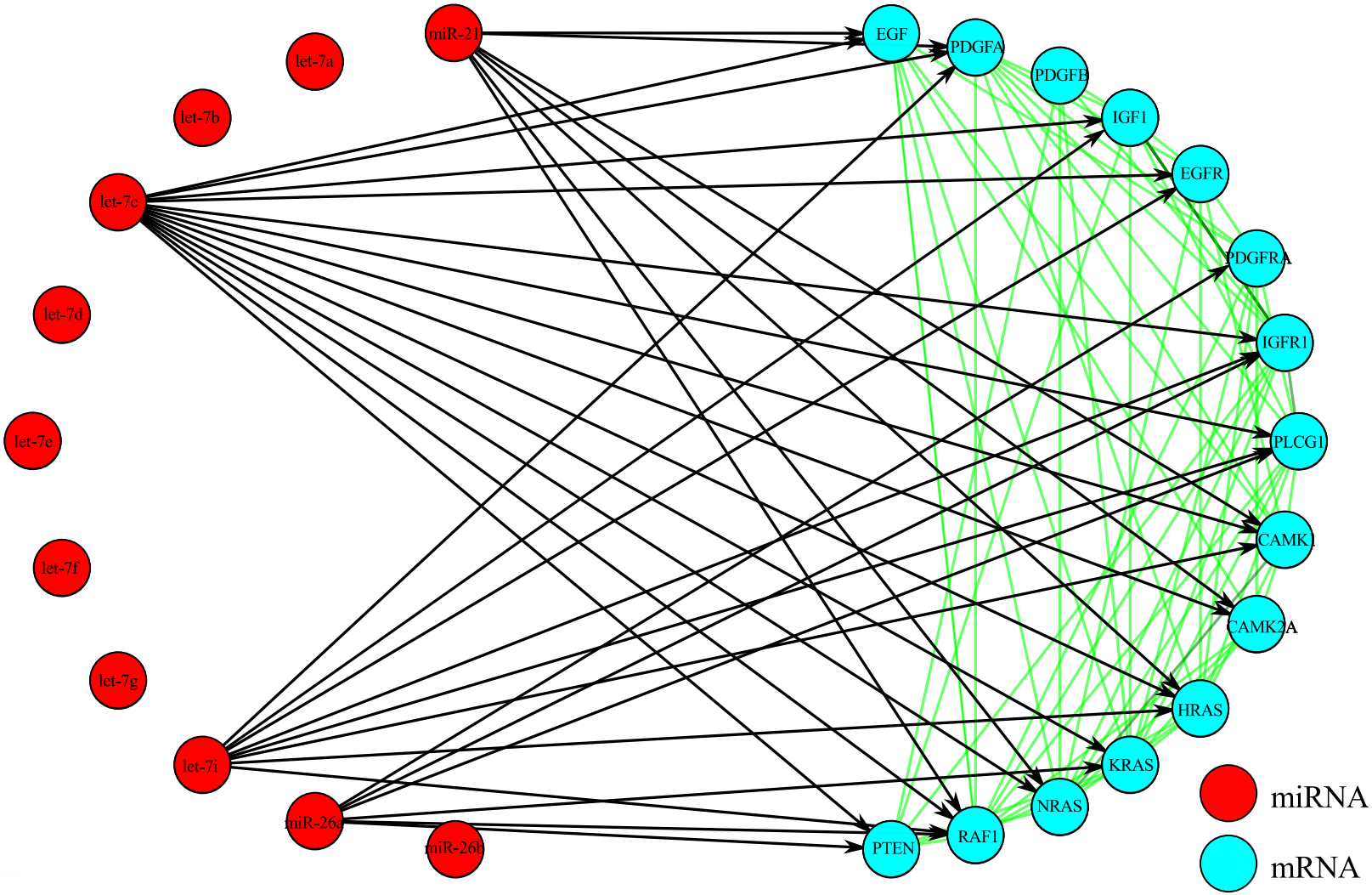
Background


microRNA expression data
n
q
mRNA expression data
n
p
GLIOMA
1. Regularized estimation of joint precision matrix
2. Determine support. Resulting sparsified matrix represents the moral graph
3. Run decomposition algorithm to obtain the cyclic mixed graph

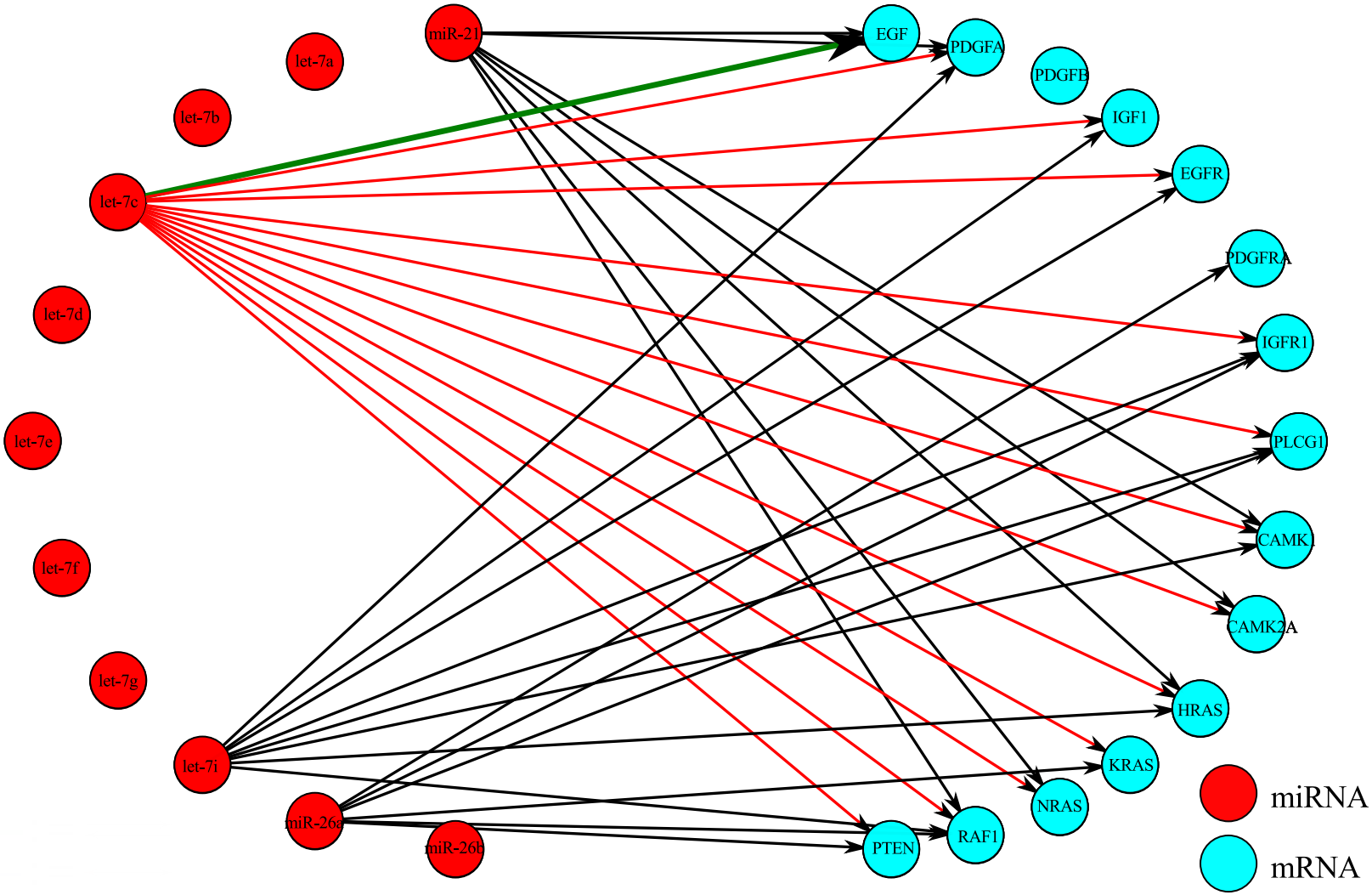
Full DCMG


miR-21
let-7a
let-7b
let-7c
let-7d
let-7e
let-7f
let-7g
let-7i
miR-26a
miR-26b
EGF
PDGFA
PDGFB
IGF1
EGFR
PDGFRA
IGFR1
PLCG1
CAMK1
CAMK2A
HRAS
KRAS
NRAS
RAF1
PTEN
miRNA
mRNA

Endogenous Relations



Exogenous Shocks



Manual/Download

- Peeters, C.F.W., Bilgrau, A.E., & van Wieringen, W.N. (2016). "`rags2ridges`: Ridge Estimation of Precision Matrices from High-Dimensional Data". R package, version 2.1.1 URL: https://cran.r-project.org/package=rags2ridges.

Theory/Methodology

- Bilgrau*, A.E., Peeters*, C.F.W., Eriksen, P.S., Bøgsted, M., & van Wieringen, W.N. (2015). "Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes". arXiv:1509.07982v1 [stat.ME].
- Peeters, C.F.W., van Wieringen, W.N., & van de Wiel, M.A. (in preparation). "Directed Cyclic Mixed Graph Modeling for High-Dimensional Genomic Data Integration".
- van Wieringen, W.N. & Peeters, C.F.W. (2016). "Ridge Estimation of Inverse Covariance Matrices from High-Dimensional Data". Computational Statistics & Data Analysis, 103: 284-303. arXiv:1403.0904v3 [stat.ME].

Software

- Peeters, C.F.W., van de Wiel, M.A., & van Wieringen, W.N. (2016) "The Spectral Condition Number Plot for Regularization Parameter Determination". arXiv:1608.04123v1 [stat.CO].
- van Wieringen, W.N. & Peeters, C.F.W. (2015). "Application of a New Ridge Estimator of the Inverse Covariance Matrix to the Reconstruction of Gene-Gene Interaction Networks". In: di Serio, C., Lio, P., Nonis, A., and Tagliaferri, R. (Eds.) 'Computational Intelligence Methods for Bioinformatics and Biostatistics'. Lecture Notes in Computer Science, vol. 8623. Springer, pp. 170–179.

Explaining the inverse

The scalar inverse

- Let a denote a number (excluding 0)
- The inverse is then the number b such that $a \times b = 1$
- Clearly, $b = \frac{1}{a}$

Matrix

A matrix is a generalization of a number, an array of numbers

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$$

Explaining the inverse

The Matrix Inverse

Consider the matrix $\mathbf{A}$. Its inverse $\mathbf{B} = \mathbf{A}^{-1}$ is defined such that

$$\mathbf{AB} = \mathbf{I},$$

where

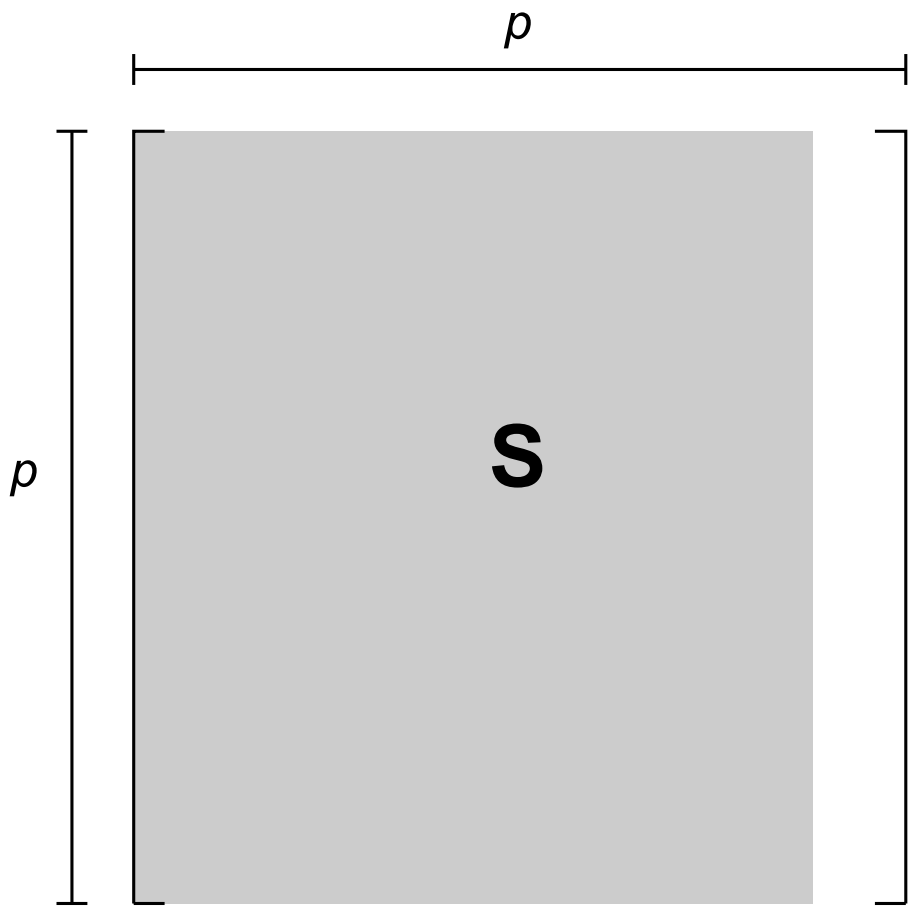
$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Solution

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{Q}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{Q}^{-1} \end{bmatrix},$$

with $\mathbf{Q} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ denoting the Schur complement.

Singularity



Ridge estimation

Maximize

$$\underbrace{\ln|\boldsymbol{\Omega}| - \text{tr}(\mathbf{S}\boldsymbol{\Omega})}_{\log-\text{likelihood}} - \underbrace{\frac{\lambda}{2}\|\boldsymbol{\Omega} - \mathbf{T}\|_2^2}_{\ell_2-\text{penalty}}$$

- $\mathbf{T}$ denotes a p.d. symmetric target matrix
- $\lambda \in (0, \infty)$ denotes a penalty parameter

Analytic penalized ML estimator

$$\hat{\boldsymbol{\Omega}}(\lambda) = \left\{ \left[\lambda \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda\mathbf{T}) \right\}^{-1}$$

Properties

Behavior

i. $\hat{\boldsymbol{\Omega}}(\lambda) \succ 0$, for all $\lambda \in (0, \infty)$;

ii. $\lim_{\lambda \to 0^+} \hat{\boldsymbol{\Omega}}(\lambda) = \mathbf{S}^{-1}$ if $p < n$;

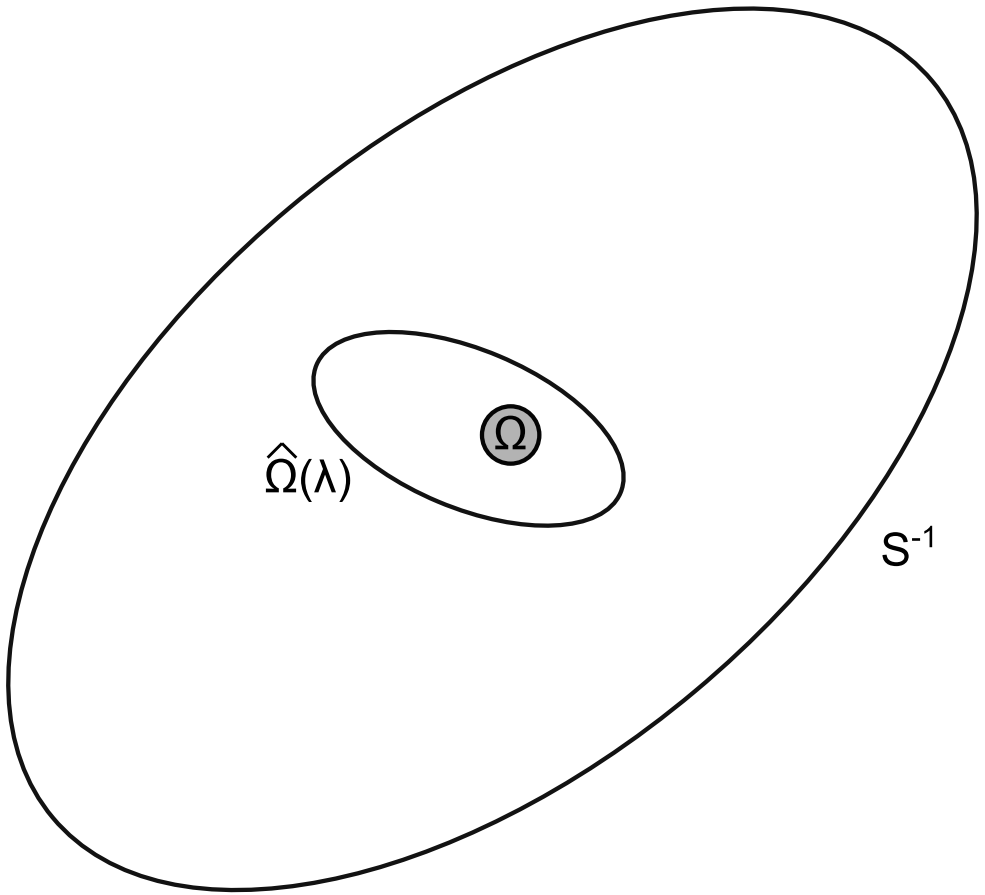
iii. $\lim_{\lambda \to \infty} \hat{\boldsymbol{\Omega}}(\lambda) = \mathbf{T}$.

Consistency

i. $\lim_{n \to \infty} \mathbb{E}\left[\hat{\boldsymbol{\Omega}}_n(\lambda_n)\right] \longrightarrow \lim_{n \to \infty} \mathbb{E}\left(\mathbf{S}_n^{-1}\right) = \boldsymbol{\Omega}$;

ii. $\lim_{n \to \infty} \mathbb{E}\left(\|\hat{\boldsymbol{\Omega}}_n(\lambda_n) - \boldsymbol{\Omega}\|_F^2\right) = 0$.

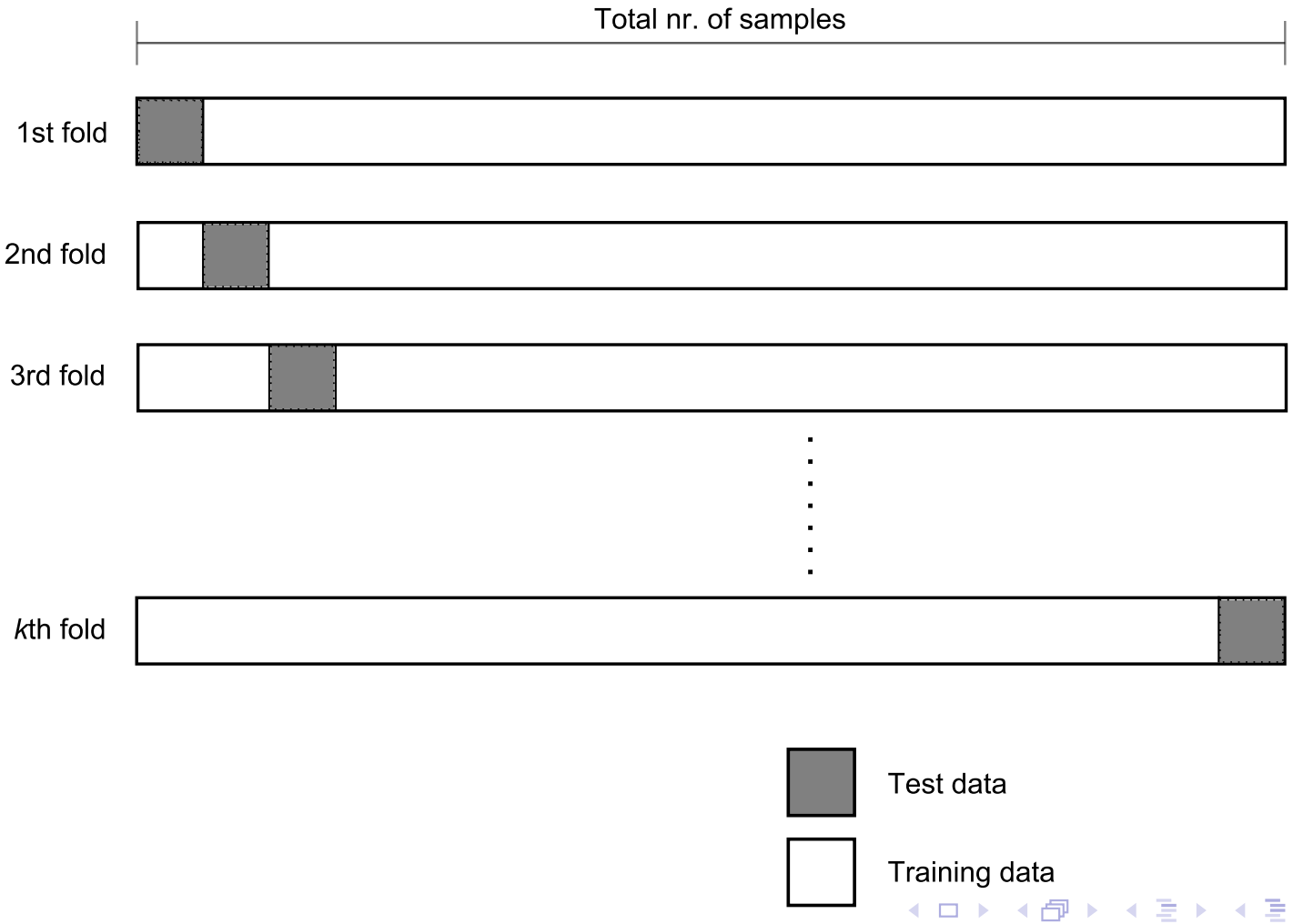
Visual explanation



Choosing the penalty value

K-fold cross-validation (CV)

Single iteration of *K*-fold CV



Choosing the penalty value

K-fold CV score

$$\varphi^K(\lambda) = \sum_{k=1}^{K} n_k \left\{ -\ln |\hat{\boldsymbol{\Omega}}(\lambda)_{-k}| + \text{tr}[\hat{\boldsymbol{\Omega}}(\lambda)_{-k}\mathbf{S}_k] \right\},$$

n_k is the size of subset k, for $k = 1, \ldots, K$ disjoint subsets;
$\mathbf{S}_k$ denotes the sample covariance matrix on kth test set;
$\hat{\boldsymbol{\Omega}}(\lambda)_{-k}$ denotes the estimated regularized precision matrix on kth training set

Highest predictive accuracy

Choose $n_k = 1$, such that $K = n$ (known as leave-one-out CV - LOOCV)

Problem

K-fold CV is computationally demanding for large p and/or large K

Solution

Computationally efficient approximate LOOCV score

Support determination

Scaling

$\hat{\mathbf{P}}(\lambda)$: Regularized precision estimate scaled to partial correlation form

Assume

Nonredundant off-diagonal partial correlation coefficients (indexed by $j < j'$) follow a mixture distribution:

$$f\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right\} = \eta_0 f_0\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}; \kappa\right\} + (1-\eta_0) f_{\mathcal{E}}\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right\}$$

- $\eta_0 \in [0,1]$ is the mixture weight
- $f_0\{\cdot\}$ denotes the distribution of a null-edge
- $f_{\mathcal{E}}\{\cdot\}$ denotes the distribution of a present edge

Determine

$$P\left(Y_j \neq Y_{j'} | [\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right)$$

Situation

Data

- G classes of $(n_g \times p)$-dimensional data
- Classes defined by data sets and/or (subtypes of) diseases

Assumption

Precision matrices of constituent classes chiefly share the same structure but potentially differ in a number of locations of interest

Desire

Integrative or meta-analytic Gaussian graphical modeling

Targeted fused ridge estimation: General Formulation

Maximize

$$\underbrace{\mathcal{L}\left(\{\boldsymbol{\Omega}_g\};\{\mathbf{S}_g\}\right)}_{\text{log-likelihood}} - \sum_g \underbrace{\frac{\lambda_{gg}}{2}\left\|\boldsymbol{\Omega}_g - \mathbf{T}_g\right\|_F^2}_{\text{ridge-penalty}} - \sum_{g_1,g_2} \underbrace{\frac{\lambda_{g_1g_2}}{4}\left\|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\right\|_F^2}_{\text{fusion-penalty}}$$

- $\mathbf{T}_g$ indicate class-specific target matrices
- $\lambda_{gg} \in (0, \infty)$ denote class-specific ridge penalty parameters
- $\lambda_{g_1g_2} \in [0, \infty)$ denote pair-specific fusion penalty parameters, $\lambda_{g_1g_2} = \lambda_{g_2g_1}$

Penalty matrix

All penalties can be collected into a non-negative symmetric matrix $\boldsymbol{\Lambda} = [\lambda_{g_1g_2}]$

Targeted fused ridge estimation

Maximizing argument for class g_0

$$\hat{\boldsymbol{\Omega}}_{g_0}\left(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_g\}_{g \neq g_0}\right) = \left\{ \left[\bar{\lambda}_{g_0} \mathbf{I}_p + \frac{1}{4}\left(\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0} \mathbf{T}_{g_0}\right)^2 \right]^{1/2} + \frac{1}{2}\left(\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0} \mathbf{T}_{g_0}\right) \right\}^{-1},$$

where

$$\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} - \sum_{g \neq g_0} \frac{\lambda_{g g_0}}{n_{g_0}} (\boldsymbol{\Omega}_g - \mathbf{T}_g), \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\sum_g \lambda_{g g_0}}{n_{g_0}}$$

Properties

Behavior

i. $\hat{\boldsymbol{\Omega}}_g \succ \mathbf{0}$ for all $\lambda_{gg} \in (0, \infty)$;

ii. $\lim_{\lambda_{gg} \to 0^+} \hat{\boldsymbol{\Omega}}_g = \mathbf{S}_g^{-1}$ if $\sum_{g' \neq g} \lambda_{gg'} = 0$ and $p \leq n_g$;

iii. $\lim_{\lambda_{gg} \to \infty} \hat{\boldsymbol{\Omega}}_g = \mathbf{T}_g$ if $\lambda_{gg'} < \infty$ for all $g' \neq g$;

iv. $\lim_{\lambda_{g_1 g_2} \to \infty} (\hat{\boldsymbol{\Omega}}_{g_1} - \mathbf{T}_{g_1}) = \lim_{\lambda_{g_1 g_2} \to \infty} (\hat{\boldsymbol{\Omega}}_{g_2} - \mathbf{T}_{g_2})$ if $\lambda_{g_1' g_2'} < \infty$ for all $\{g_1', g_2'\} \neq \{g_1, g_2\}$.

Block coordinate ascent

Input:
Sufficient data: $(\mathbf{S}_1, n_1), \ldots, (\mathbf{S}_G, n_G)$
Penalty matrix: $\boldsymbol{\Lambda}$
Convergence criterion: $\varepsilon > 0$
Output:
Estimates: $\hat{\boldsymbol{\Omega}}_1, \ldots, \hat{\boldsymbol{\Omega}}_G$
procedure RIDGEP.FUSED$(\mathbf{S}_1, \ldots, \mathbf{S}_G, n_1, \ldots, n_G, \boldsymbol{\Lambda}, \varepsilon)$
    Initialize: $\hat{\boldsymbol{\Omega}}_g^{(0)}$ for all g.
    for $c = 1, 2, 3, \ldots$ **do**
        for $g = 1, 2, \ldots, G$ **do**
            Update $\hat{\boldsymbol{\Omega}}_g^{(c)} := \hat{\boldsymbol{\Omega}}_g\left(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Omega}}_1^{(c)}, \ldots, \hat{\boldsymbol{\Omega}}_{g-1}^{(c)}, \hat{\boldsymbol{\Omega}}_{g+1}^{(c-1)}, \ldots, \hat{\boldsymbol{\Omega}}_G^{(c-1)}\right)$
        end for
        if $\max_g \left\{ \frac{\|\hat{\boldsymbol{\Omega}}_g^{(c)} - \hat{\boldsymbol{\Omega}}_g^{(c-1)}\|_F^2}{\|\hat{\boldsymbol{\Omega}}_g^{(c)}\|_F^2} \right\} < \varepsilon$ **then**
            return $\left(\hat{\boldsymbol{\Omega}}_1^{(c)}, \ldots, \hat{\boldsymbol{\Omega}}_G^{(c)}\right)$
        end if
    end for
end procedure