Undirected network reconstruction - part 3

Wessel van Wieringen
w.n.van.wieringen@vu.nl

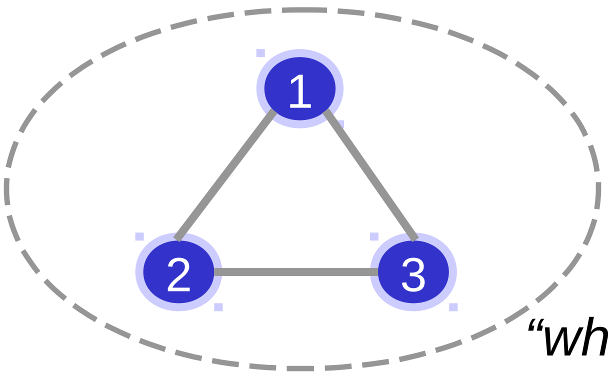
Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

vrije Universiteit

VU medisch centrum

Partial correlation

Multi-gene pathways comprise of more than two genes, and assume no gene “lives” outside the pathway.



pathway =
“whole universe”

Network reconstruction:
→ bivariate normal: no correlation → dependence.
→ suggests study of correlations in multivariate normal.

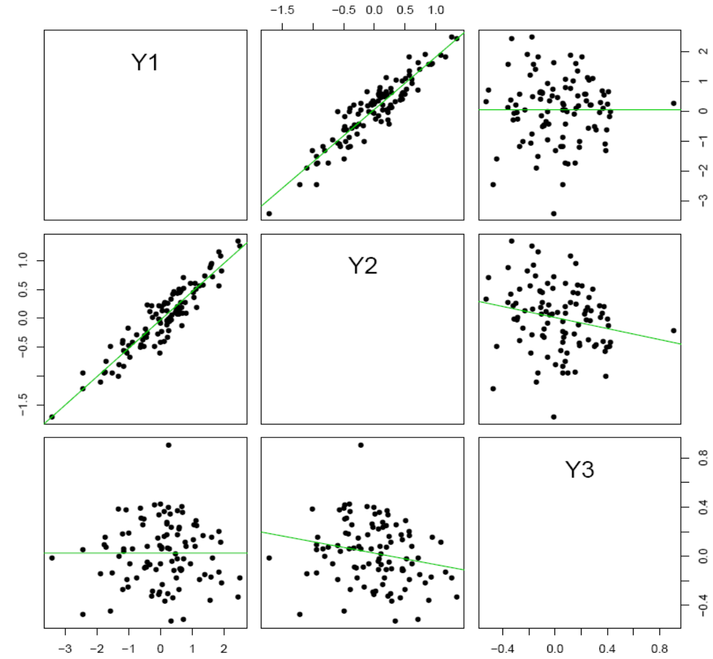
But … correlation ignores the other variables, and thus assesses the marginal dependence between two variables.

Partial correlation

Example

Consider a 3-gene pathway. Assume the genes' expression levels follow a linear system:

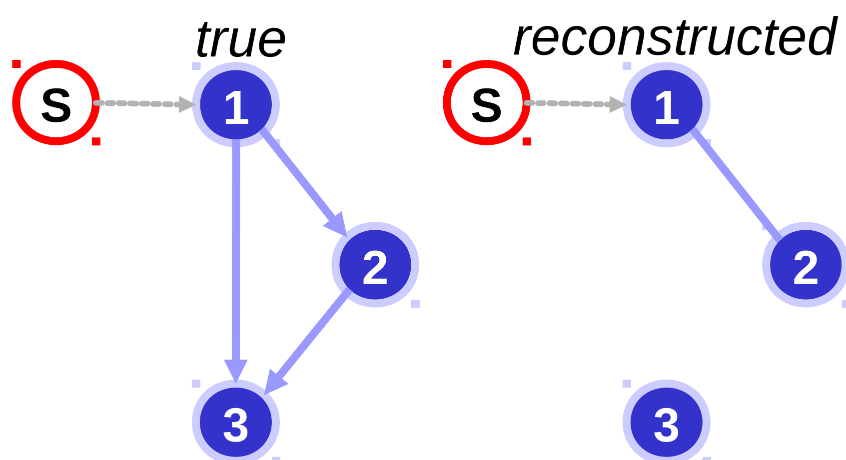
$$\begin{cases} Y_1 &= S + \varepsilon_1 \\ Y_2 &= \frac{1}{2} Y_1 + \varepsilon_2 \\ Y_3 &= \frac{1}{4} Y_1 - \frac{1}{2} Y_2 + \varepsilon_3 \end{cases}$$



with the signal *S* and errors independent and normal.

Estimated correlation matrix:

	Y1	Y2	Y3
Y1	1.000	0.930	0.000
Y2	0.930	1.000	-0.211
Y3	0.000	-0.211	1.000



Multi-gene pathways

So far

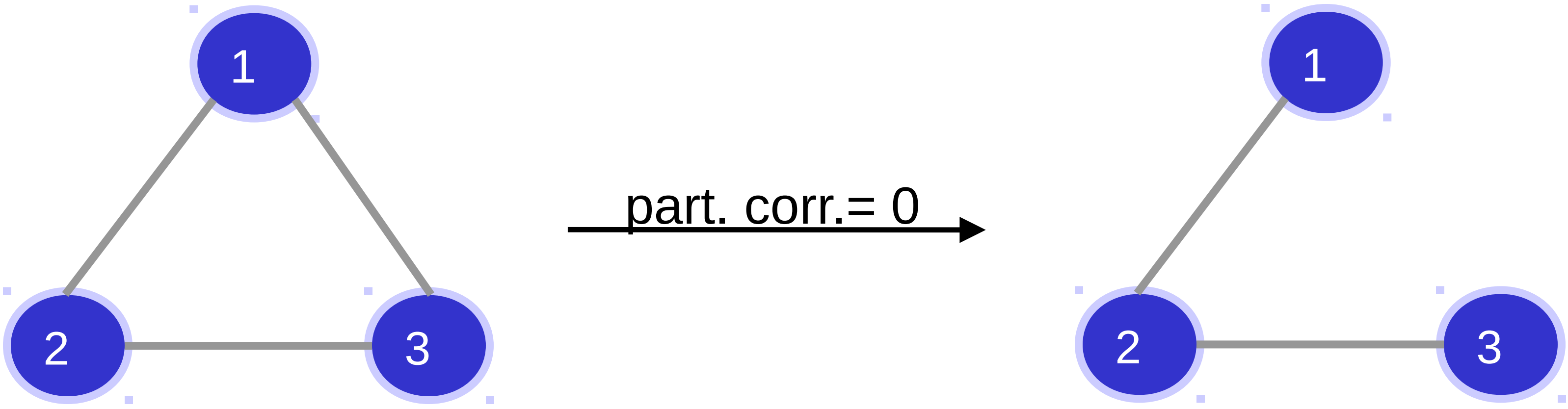
CIG reconstruction based on correlation often fails as correlation only looks at marginal independence.

Up ahead

Conditional independence → *partial correlation*.

Why?

A zero partial correlation → edge absent in the CIG.



Partial correlation

Partial correlation

A *partial correlation coefficient* quantifies the correlation between two variables when conditioning on other variables.

The partial correlation coefficient between Y_a and Y_b conditional on variables $\mathbf{Y}_c$ is defined as:

$$\rho(Y_a, Y_b \mid \mathbf{Y}_c) = \frac{\mathrm{Cov}(Y_a, Y_b \mid \mathbf{Y}_c)}{\sqrt{\mathrm{Var}(Y_a \mid \mathbf{Y}_c)}\sqrt{\mathrm{Var}(Y_b \mid \mathbf{Y}_c)}}$$

The number of variables conditioned on is the *order*.

The partial *correlation* above thus measures the *linear* dependence between Y_a and Y_b conditional on $\mathbf{Y}_c$.

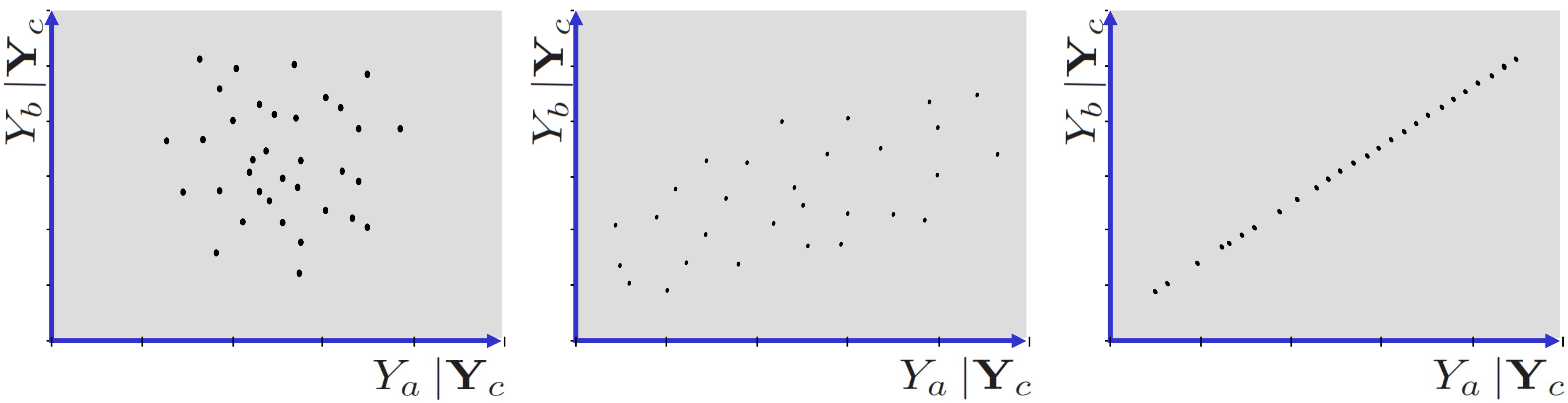
Partial correlation

The partial correlation is normalized and thus:

$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) \in [-1, 1]$$

with:

$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) = 0$ $\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) = 0.2$ $\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) = 1$



Partial correlation

Interpretation

Let Y_1, Y_2, Y_3 be random variables. Then, $\rho(Y_1, Y_2 \mid Y_3) \approx$ amount of information in Y_1 on Y_2 after removal of all information on either of them contained in Y_3.

$$\rho(Y_1, Y_2 \mid Y_3) = 0$$

```
Call:
lm(formula = Y1 ~ 0 + Y2 + Y3)

Coefficients:
   Estimate Pr(>|t|)
Y2 -0.01444    0.638
Y3  1.01584   <2e-16 ***
```

Y_2 adds nothing to Y_3 in explaining variation in Y_1.

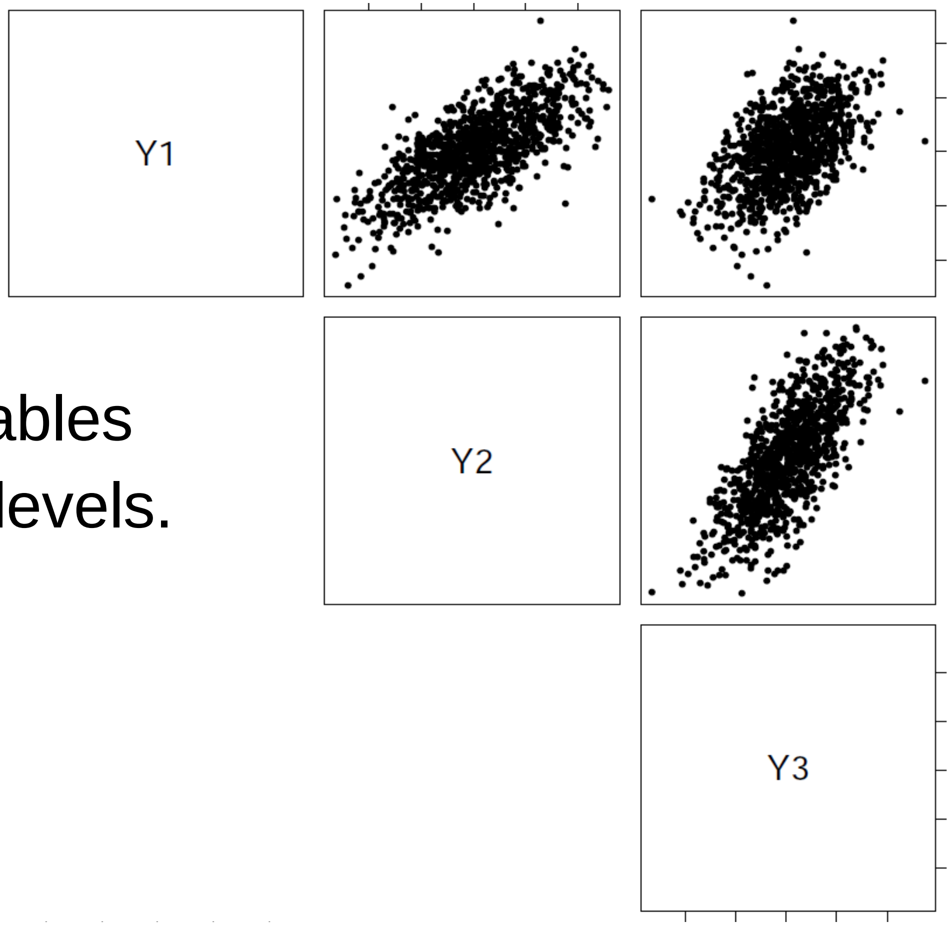
$$\rho(Y_1, Y_2 \mid Y_3) \neq 0$$

```
Call:
lm(formula = Y1 ~ 0 + Y2 + Y3)

Coefficients:
   Estimate Pr(>|t|)
Y2  0.24869 2.95e-15 ***
Y3  0.96542  < 2e-16 ***
```

Y_2 does add to Y_3 in explaining variation in Y_1.

Partial correlation

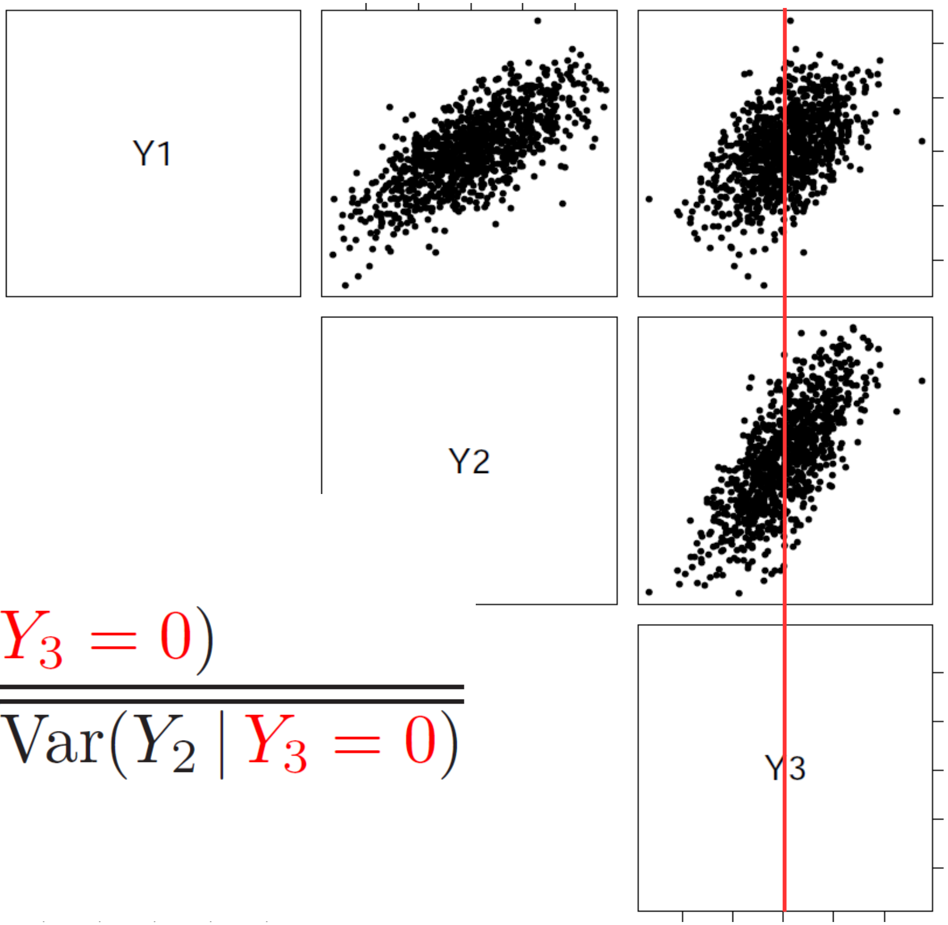


Consider three genes.
Let Y_1, Y_2, Y_3 be random variables representing their expression levels.

$$\begin{cases} Y_1 &= Y_2 + \varepsilon_1 \\ Y_2 &= \varepsilon_2 \\ Y_3 &= Y_2 + \varepsilon_3 \end{cases}$$

Question
What about the partial correlations?
→ $\rho(Y_1, Y_2 \mid Y_3) = 0$ or $\rho(Y_1, Y_2 \mid Y_3) \neq 0$?
→ $\rho(Y_1, Y_3 \mid Y_2) = 0$ or $\rho(Y_1, Y_3 \mid Y_2) \neq 0$?

Partial correlation



Question

Consider $\rho(Y_1, Y_2 \mid Y_3)$.

Wish to know e.g.:

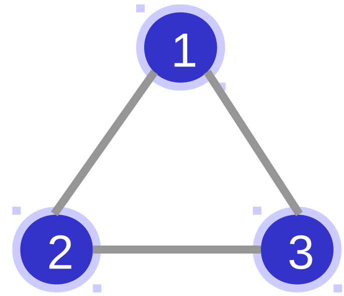
$$\mathrm{Cor}(Y_1, Y_2 \mid Y_3 = 0) = \frac{\mathrm{Cov}(Y_1, Y_2 \mid Y_3 = 0)}{\sqrt{\mathrm{Var}(Y_1 \mid Y_3 = 0)}\sqrt{\mathrm{Var}(Y_2 \mid Y_3 = 0)}}$$

Effect of conditioning:

$$\begin{cases} Y_1 = Y_2 + \varepsilon_1 \\ Y_2 = \varepsilon_2 \\ 0 = Y_2 + \varepsilon_3 \end{cases} \xrightarrow{\text{fix error}_3} \varepsilon_3 = -Y_2$$

Setting $Y_3 = 0$ does not affect relation between Y_1 and Y_2!

Partial correlation



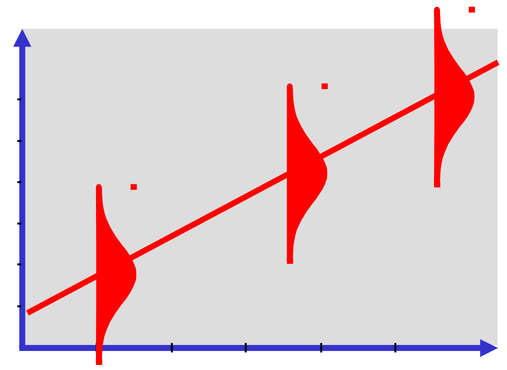
How to condition on another gene?
Condition gene 1 on gene 3
within a three-gene pathway.

Recall

$$Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i$$

is equivalent to:

$$Y_i \,|\, \mathbf{X}_{i,*} \sim N(\mathbf{X}_{i,*}\boldsymbol{\beta}, \sigma^2)$$

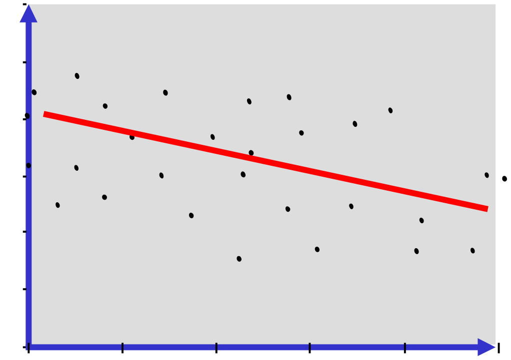


Regress gene 1 on gene 3:

$$Y_{i,1} = \beta_0 + \beta_1 Y_{i,3} + \varepsilon_{i,1}$$

and “obtain”:

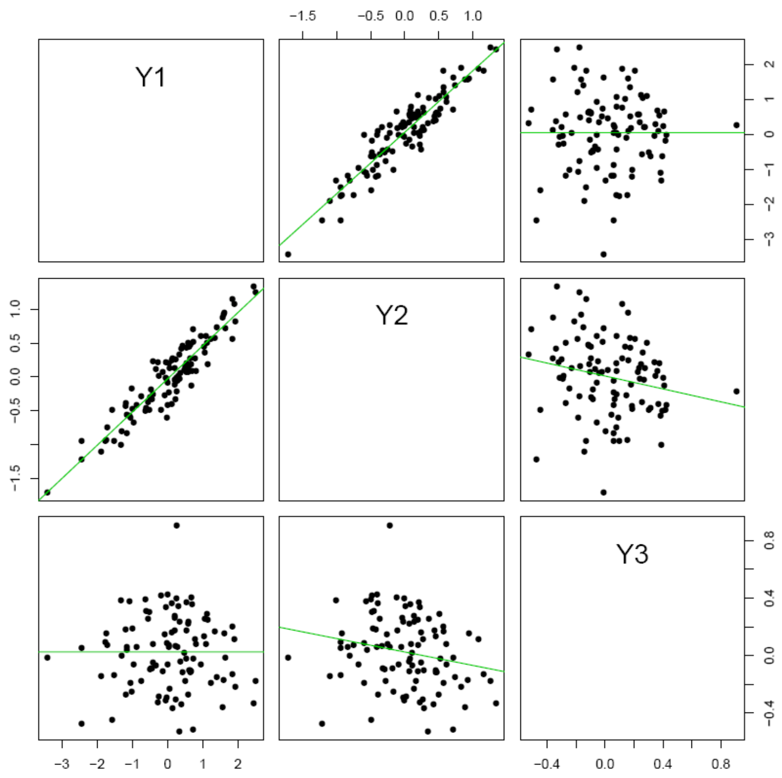
$$Y_{i,1} \,|\, Y_{i,3} \sim \mathcal{N}(\beta_0 + \beta_1 Y_{i,3}, \sigma^2)$$

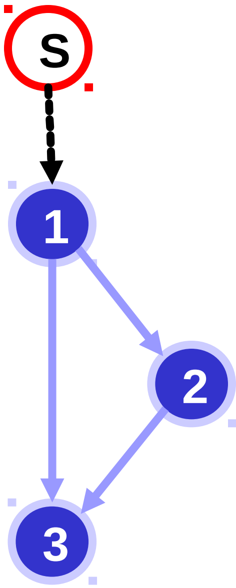


Partial correlation

Example

Recall the pathway of 3 genes.

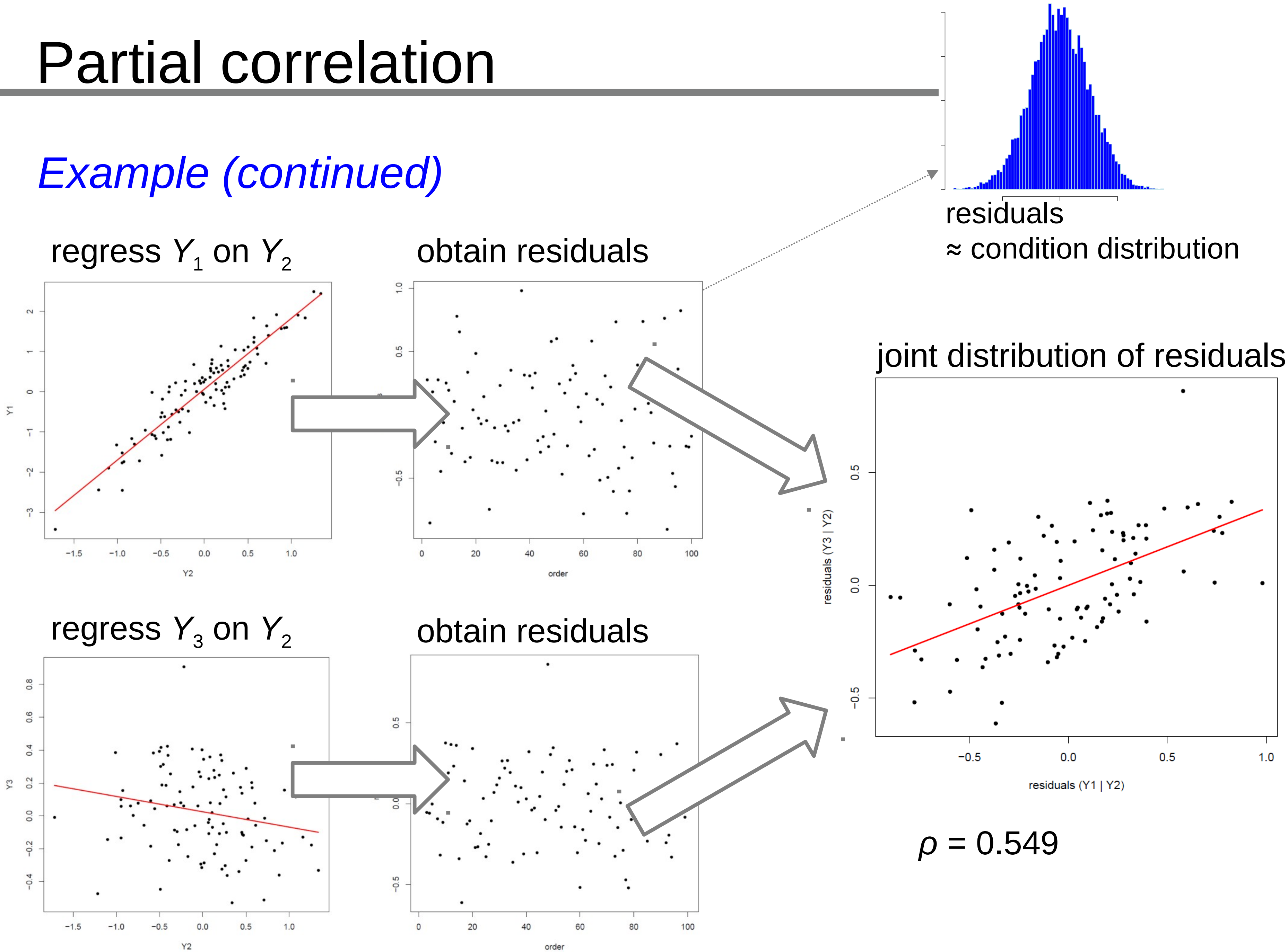




Calculate partial correlation between Y_1 and Y_3.

Partial correlation

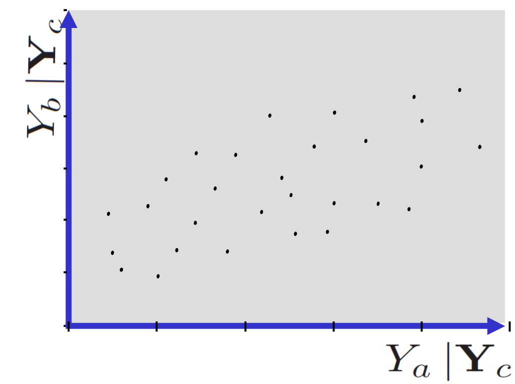
Example (continued)



$\rho = 0.549$

Partial correlation

The partial correlation between Y_a and Y_b conditional on $\mathbf{Y}_c$ is the correlation between the residuals of Y_a and Y_b after regressing them on $\mathbf{Y}_c$.



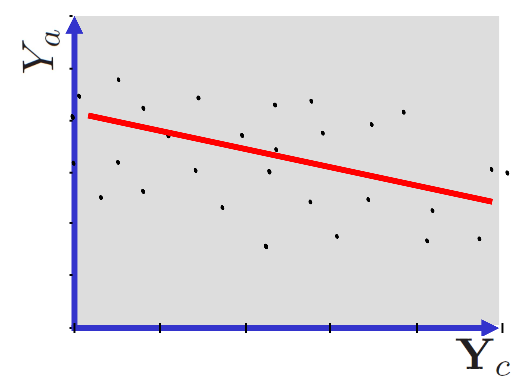
Estimate partial correlation by:

$$\hat{\rho}(Y_a, Y_b \,|\, \mathbf{Y}_c) = \hat{\rho}(\hat{\varepsilon}_{Y_a}, \hat{\varepsilon}_{Y_b})$$

where

$$\hat{\varepsilon}_{Y_a} = Y_a - E(Y_a \,|\, \mathbf{Y}_c)$$

the residual obtained when regressing Y_a on $\mathbf{Y}_c$.



Partial correlation

Distribution

The Fisher transformed partial correlation coefficient:

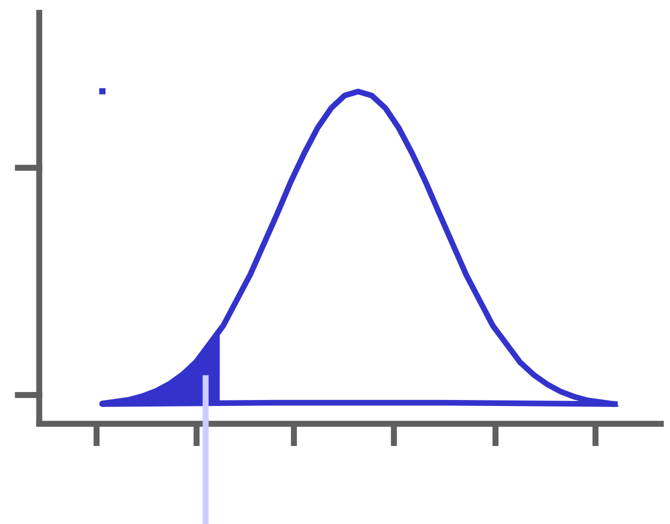
$$F(r) = \frac{1}{2}\log[(1+r)/(1-r)] = \operatorname{arctanh}(r)$$

follows asymptotically a normal distribution:

$$F(r) \sim \mathcal{N}[F(\rho), (n-k-2)^{-1}]$$

where *k* is the number of variables conditioned upon.

Can now to test H_0: $\rho(\cdot|\cdot) = 0$.



p-value = P (X ≤ x)

Partial correlation

Question

Consider three random variable Y_1, Y_2, and Y_3, representing expression levels of the three genes.

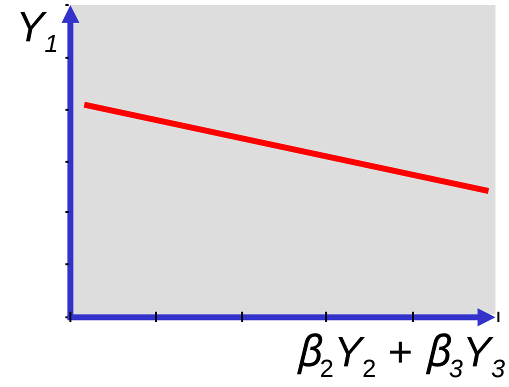
The expression levels are not linearly dependent:

$$|\rho(Y_1, Y_2)| \neq 1, |\rho(Y_1, Y_3)| \neq 1, |\rho(Y_2, Y_3)| \neq 1$$

When regressing the first gene on the other two, i.e.:

$$Y_1 = \beta_2 Y_2 + \beta_3 Y_3 + \varepsilon_1$$

with coefficient of determination $R^2 = 1$.



What is $\rho(Y_1, Y_2 \mid Y_3)$?

Partial correlation

Alternatively, partial correlation can be calculated by covariance matrix inversion. Hereto we need:

Inverse variance lemma (Whittaker, 1991)

Let **X** and **Y** be *p* and *q*-dimensional random variables. The inverse of the partitioned variance *Var*(**X**, **Y**) is given by:

$$\{\mathrm{Var}[(\mathbf{X}, \mathbf{Y})]\}^{-1} = \begin{pmatrix} \mathrm{Var}(\mathbf{X}) & \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) \\ \mathrm{Cov}(\mathbf{Y}, \mathbf{X}) & \mathrm{Var}(\mathbf{Y}) \end{pmatrix}^{-1} = \begin{pmatrix} * & * \\ * & \underline{[\mathrm{Var}(\mathbf{Y} \,|\, \mathbf{X})]^{-1}} \end{pmatrix}$$

Take-away: inversion ≈ (reciprocal of) conditioning!

Partial correlation

Corollary (Whittaker, 1991)

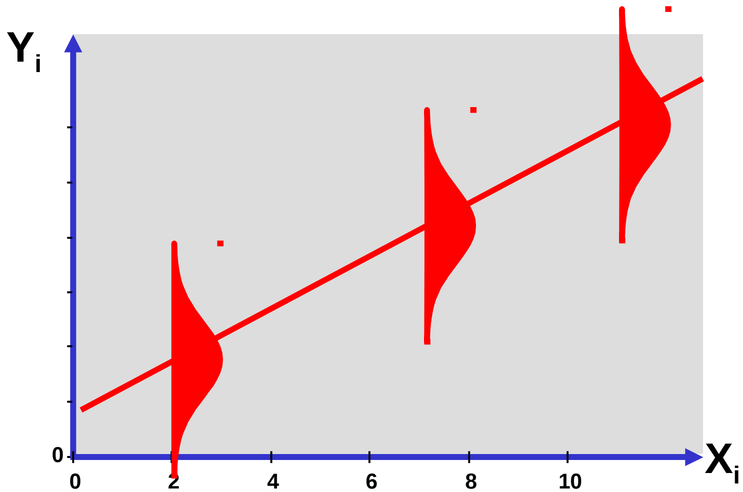
Each diagonal element of the inverse variance matrix is the reciprocal of a partial variance:

$$[(\boldsymbol{\Sigma})^{-1}]_{jj} = 1/[\underbrace{\mathrm{Var}(Y_j \mid \mathbf{Y}_{\mathcal{V}\backslash j})}_{\text{partial variance}}]$$

Familiar quantity from regression analysis. Let:

$$Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$$

Then: $\mathrm{Var}(Y \mid \mathbf{X}) = \sigma^2$



Partial correlation

Example

Trivariate normally distributed data:

```
> # prelimary
> set.seed(1); library(mvtnorm);
> Sigma <- matrix(0.4, 3, 3); diag(Sigma) <- 1;

> # draw and center data
> Y <- rmvnorm(100, sigma=Sigma)
> Y <- sweep(Y, 2, apply(Y, 2, mean))
```

regression

```
> # residuals and their variance
> errorHat <- residuals(lm(Y[,1] ~ 0 + Y[,2] + Y[,3]))
> mean(errorHat^2)
[1] 0.915961
```

reciprocal

inverse covariance

```
> # sample covariance and inverse
> S <- t(Y) %*% Y / 100
> solve(S)
           [,1]       [,2]       [,3]
[1,]  1.0917495 -0.3157168 -0.3764417
[2,] -0.3157168  1.6184612 -0.4741769
[3,] -0.3764417 -0.4741769  1.6385053
```

Partial correlation

Corollary (Whittaker, 1991)
Each off-diagonal element of the inverse variance matrix (scaled to have a unit diagonal) is the negative of the partial correlation between the two corresponding variables, conditioned on all remaining variables.

In formula, this gives:

$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) \quad = \quad \frac{-\left(\mathbf{\Sigma}^{-1}\right)_{a,b}}{\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{a,a}}\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{b,b}}}$$

Partial correlation estimate by plugging covariance estimate.

Partial correlation

From covariance to partial correlation

inversion

covariance matrix

	Y1	Y2	Y3	Y4	Y5
Y1	1.52	-0.61	-0.32	-0.07	-0.53
Y2	*	1.35	-0.57	-0.64	0.39
Y3	*	*	1.39	1.08	-0.08
Y4	*	*	*	1.82	-0.20
Y5	*	*	*	*	0.97

precision matrix

	Y1	Y2	Y3	Y4	Y5
Y1	1.21	0.66	0.60	-0.03	0.43
Y2	*	1.40	0.62	0.13	-0.12
Y3	*	*	1.71	-0.77	0.06
Y4	*	*	*	1.06	0.08
Y5	*	*	*	*	1.34

standarization

off-diagonal minus one

partial correlation matrix

	Y1	Y2	Y3	Y4	Y5
Y1	1.00	-0.51	-0.42	0.03	-0.34
Y2	*	1.00	-0.40	-0.11	0.09
Y3	*	*	1.00	0.57	-0.04
Y4	*	*	*	1.00	-0.07
Y5	*	*	*	*	1.00

standardized precision matrix

	Y1	Y2	Y3	Y4	Y5
Y1	1.00	0.51	0.42	-0.03	0.34
Y2	*	1.00	0.40	0.11	-0.09
Y3	*	*	1.00	-0.57	0.04
Y4	*	*	*	1.00	0.07
Y5	*	*	*	*	1.00

Partial correlation

Example

Verify the corollary. Consider a 3x3 correlation matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

Its inverse is given by:

$$\frac{1}{\det(\mathbf{\Sigma})} \begin{pmatrix} 1-\rho_{23}^2 & \boxed{\rho_{13}\,\rho_{23}-\rho_{12}} & \rho_{12}\,\rho_{23}-\rho_{13} \\ \rho_{13}\,\rho_{23}-\rho_{12} & 1-\rho_{13}^2 & \rho_{12}\,\rho_{13}-\rho_{23} \\ \rho_{12}\,\rho_{23}-\rho_{13} & \rho_{12}\,\rho_{13}-\rho_{23} & 1-\rho_{12}^2 \end{pmatrix}$$

proportional to partial correlation

Partial correlation

Example (continued)

The inverse is (up a factor) identical to:

$$
\begin{aligned}
\mathrm{Var}(\mathbf{X} \mid Z) &= \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XZ}\,\mathbf{\Sigma}_{ZZ}^{-1}\,\mathbf{\Sigma}_{ZX} \\
&= \begin{pmatrix} 1 & \boxed{\rho_{12}} \\ \rho_{12} & 1 \end{pmatrix} - \begin{pmatrix} \rho_{13}^2 & \boxed{\rho_{13}\,\rho_{23}} \\ \rho_{13}\,\rho_{23} & \rho_{23}^2 \end{pmatrix}
\end{aligned}
$$

proportional to $\rho(Y_1, Y_2 \mid Y_3)$

The factor cancels out in the partial correlation:

$$
\rho_{12.3} = \frac{-\left(\mathbf{\Sigma}^{-1}\right)_{12}}{\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{11}}\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{22}}} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}
$$

Partial correlation

Let $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be partitioned as:

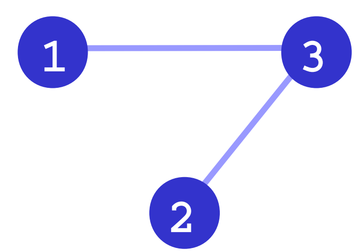
$$\begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \\ \mathbf{Y}_c \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{bb} & \boldsymbol{\Sigma}_{bc} \\ \boldsymbol{\Sigma}_{ca} & \boldsymbol{\Sigma}_{cb} & \boldsymbol{\Sigma}_{cc} \end{pmatrix} \right)$$

Then:

$$\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b | \mathbf{Y}_c \quad \Longleftrightarrow \quad \boldsymbol{\Omega}_{ab} = (\boldsymbol{\Sigma}^{-1})_{ab} = \mathbf{0}.$$

Simple criterion for (conditional) pairwise independence:

$$(\boldsymbol{\Omega})_{1,2} = 0 \iff (\boldsymbol{\Sigma}^{-1})_{1,2} = 0$$
$$\iff Y_1 \perp\!\!\!\perp Y_2 \,|\, Y_3, \ldots, Y_p \iff$$

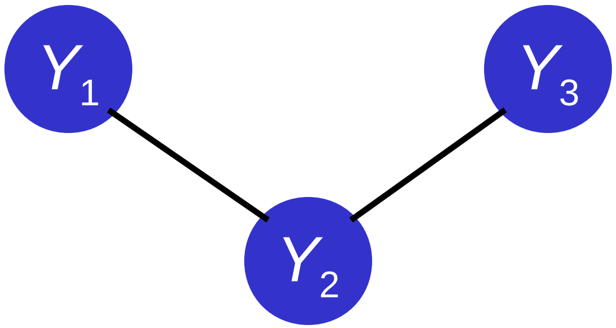


Whittaker (1990).

Partial correlation

Example (continued)

Assume: $(\mathbf{\Omega})_{13} = (\mathbf{\Sigma}^{-1})_{13} = 0$



Corollary: Y_1 and Y_3 independent, conditionally on Y_2.

Proposition: joint density function of Y_1, Y_2, and Y_3 factorizes.

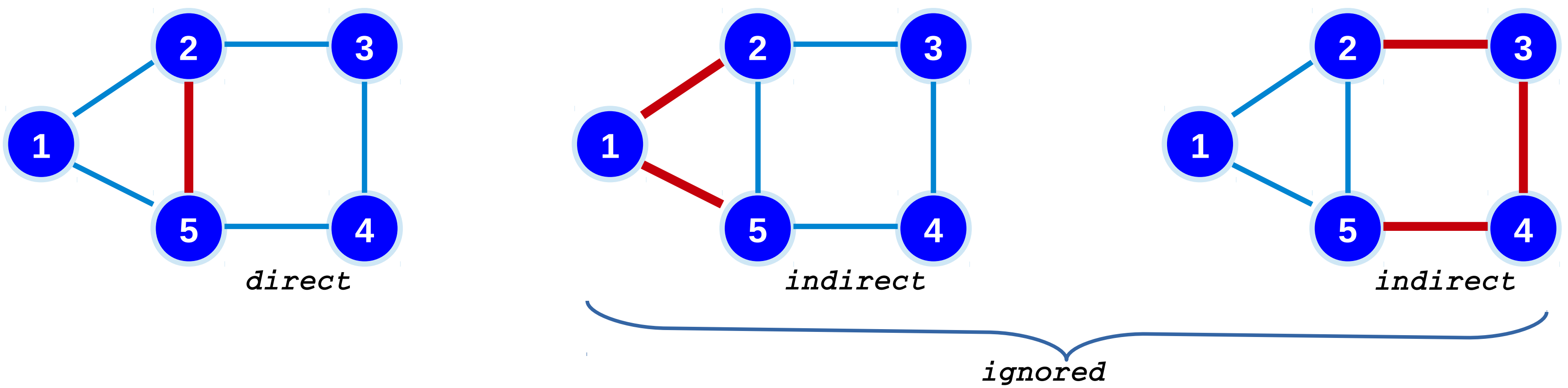
Question

Confirm factorization.

Partial correlation

Elements of P measure the *direct* relation between two nodes while excluding effects of others.

ω_{25} : direct association between nodes 2 and 5:



Standardization yields the *partial correlations*, e.g.:

$$\rho(Y_a, Y_b \mid \mathbf{Y}_c) = \frac{\mathrm{Cov}(Y_a, Y_b \mid \mathbf{Y}_c)}{\sqrt{\mathrm{Var}(Y_a \mid \mathbf{Y}_c)}\sqrt{\mathrm{Var}(Y_b \mid \mathbf{Y}_c)}}$$

= *linear* dependence between Y_a and Y_b conditional on Y_c.

Partial correlation

Roadmap

estimation

data

	Y1	Y2	Y3	Y4	Y5
i=1	-0.34	-0.62	1.51	-2.07	-0.98
i=2	1.78	1.53	-0.39	-0.47	0.28
i=3	0.91	-1.33	1.39	1.07	-0.76
i=4	0.03	0.84	-0.70	1.27	-0.18
...	...	...	...	...	...

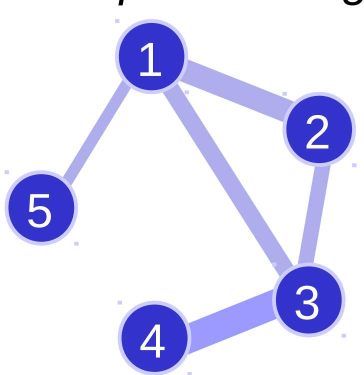
covariance matrix

	Y1	Y2	Y3	Y4	Y5
Y1	1.52	-0.61	-0.32	-0.07	-0.53
Y2	*	1.35	-0.57	-0.64	0.39
Y3	*	*	1.39	1.08	-0.08
Y4	*	*	*	1.82	-0.20
Y5	*	*	*	*	1.34

inversion + ...

test for zeros

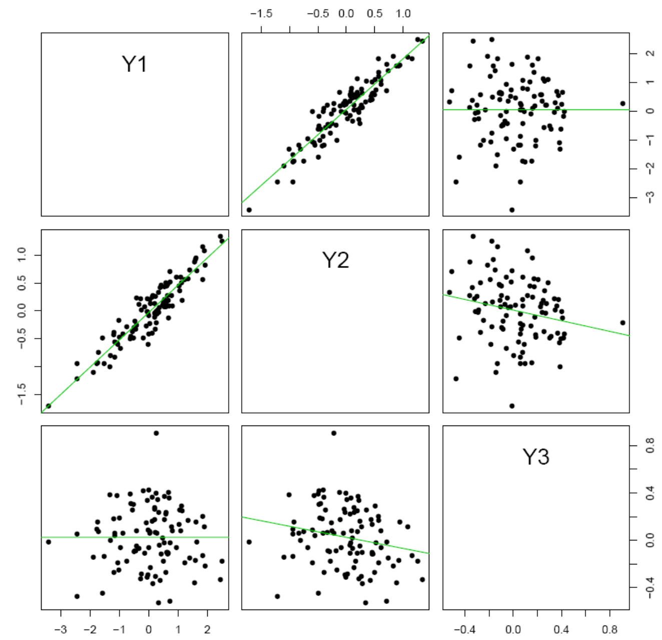
cond. independence graph



partial correlation matrix

	Y1	Y2	Y3	Y4	Y5
Y1	1.00	-0.51	-0.42	0.03	-0.34
Y2	*	1.00	-0.40	-0.11	0.09
Y3	*	*	1.00	0.57	-0.04
Y4	*	*	*	1.00	-0.07
Y5	*	*	*	*	1.00

Partial correlation

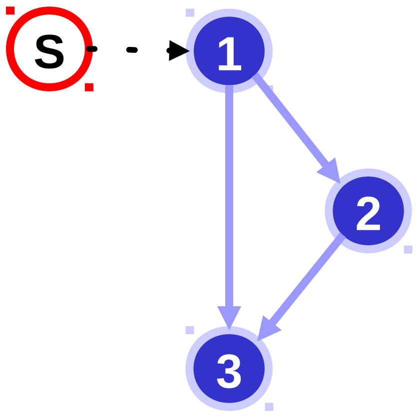


Example (continued)

Estimated partial correlation matrix:

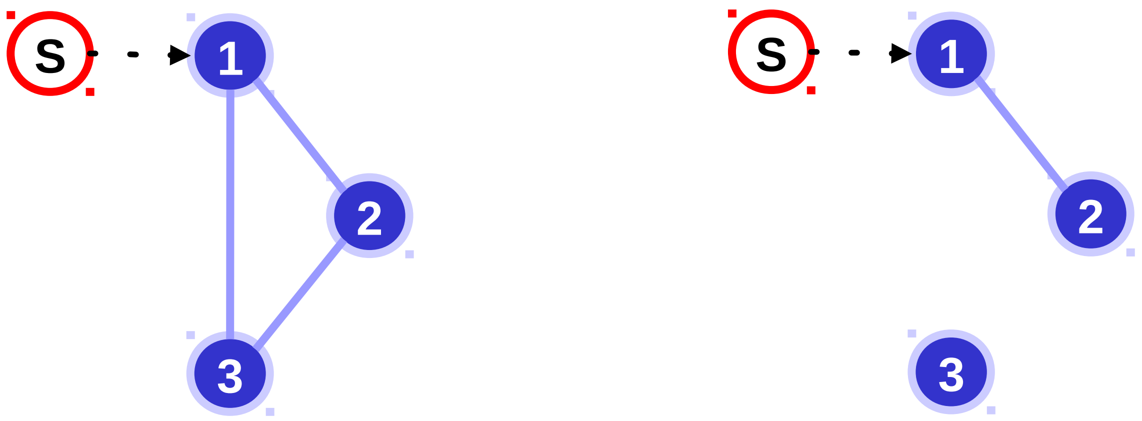
	Y1	Y2	Y3
Y1	1.000	0.952	0.549
Y2	0.952	1.000	-0.576
Y3	0.549	-0.576	1.000

true network



reconstructed (partial correlations)

reconstructed (marginal correlations)



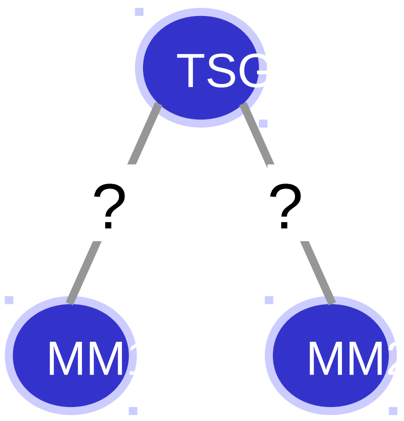
Partial correlation

Cancer research example

Y : gene expression measurements of a tumor
suppressor gene

X_1 : gene expression of methylation marker 1

X_2 : gene expression of methylation marker 2



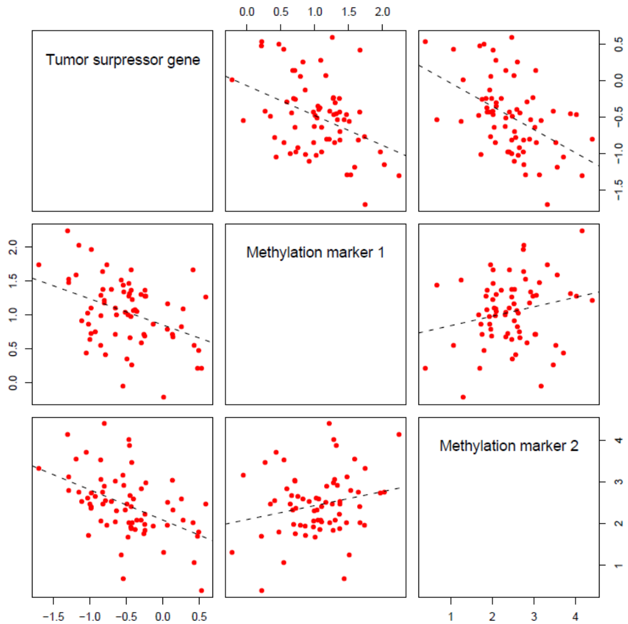
Question

Do the methylation markers (MMs) influence the expression of the tumor suppressor gene (TSG)?

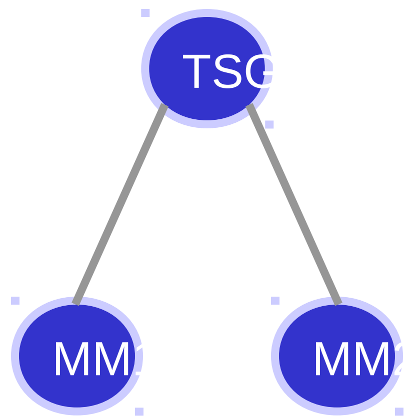
Partial correlation

Cancer research example

Finally, the partial correlation method clearly indicates there is no conditional correlation between MM1 and MM2,.




```
> Sigma <- var(cbind(TSG, MM1, MM2))
> invSigma <- solve(Sigma)
> partCorMat <- cov2cor(invSigma)
> round(partCorMat, d=3)
      [,1]   [,2]   [,3]
[1,] 1.000  0.342  0.441
[2,] 0.342  1.000 -0.032
[3,] 0.441 -0.032  1.000
```



Partial correlation

Conclusion

→ Partial correlation measures correlation between two variables while taking others into account.

→ Partial correlations are readily obtained from the standardized inverse of the covariance matrix.

→ Zero partial correlations indicate conditional independence.

→ The conditional independence graph can be reconstructed using partial correlations.

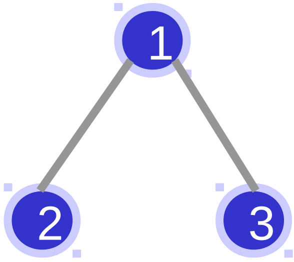
Partial correlation vs. regression

Multi-gene pathway & regression

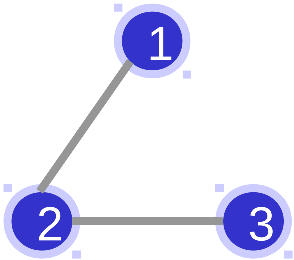
Regression analysis

Regress the expression data of each gene on that of all other genes.

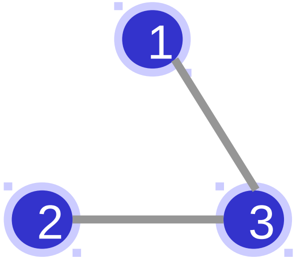
$$Y_1 = b_{01} \qquad\qquad + b_{21}Y_2 + b_{31}Y_3 + e_1$$



$$Y_2 = b_{02} + b_{12}Y_1 \qquad\qquad + b_{32}Y_3 + e_2$$



$$Y_3 = b_{03} + b_{13}Y_1 + b_{23}Y_2 \qquad\qquad + e_3$$



Partial corr. vs. regression

So far

→ Partial correlation is closely related to regression: confer its calculation.

→ Recall the two-gene pathway. In this bivariate case ρ and β are 1-1 related. In particular, $\rho = 0 \leftrightarrow \beta = 0$. Independence between the two genes of the pathway can be assessed by either ρ and β.

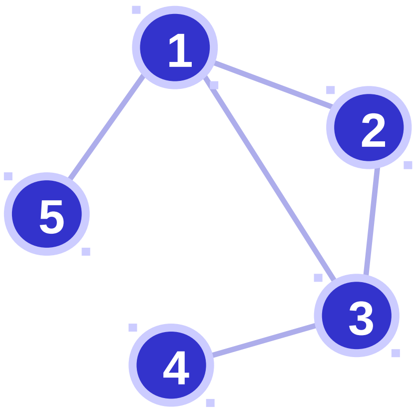
Question

Does a similar relation between ρ and β hold in the multi-gene pathways?

Partial correlation vs. regression

Example

A pathway comprising of five genes



Expression data distributed as:

$$\mathbf{Y}_i \quad \sim \quad \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

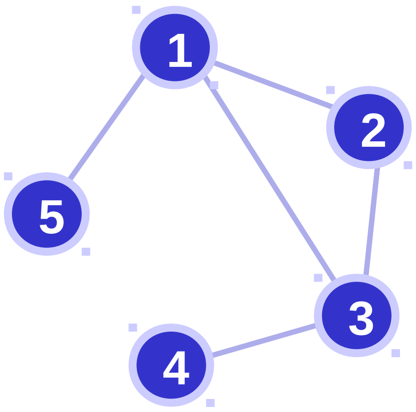
with:

$$\mathbf{\Sigma}^{-1} \quad = \quad \begin{pmatrix} 1.00 & -0.50 & -0.50 & 0.00 & 0.50 \\ -0.50 & 1.00 & 0.50 & 0.00 & 0.00 \\ -0.50 & 0.50 & 1.00 & 0.50 & 0.00 \\ 0.00 & 0.00 & 0.50 & 1.00 & 0.00 \\ 0.50 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}$$

Partial correlation vs. regression

Example (continued)

The partial correlation matrix suggests that Y_1 and Y_4 are conditionally independent.



Confirmed by the regression approach?

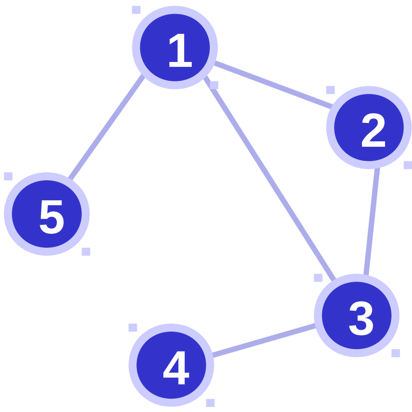
```
> summary(lm(Y[,1] ~ 0 + Y[,2]+ Y[,3] + Y[,4] + Y[,5]))

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
Y[, 2]  0.45838    0.02781  16.482   <2e-16 ***
Y[, 3]  0.50208    0.02698  18.611   <2e-16 ***
Y[, 4]  0.03506    0.03065   1.144    0.253
Y[, 5] -0.47669    0.02805 -16.995   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Partial correlation vs. regression

Example (continued)

The estimated regression coefficients are closely related to the partial correlation coefficients.




```
> summary(lm(Y[,1] ~ 0 + Y[,2]+ Y[,3] + Y[,4] + Y[,5]))

Coefficients:
        Estimate
Y[, 2]   0.45838
Y[, 3]   0.50208
Y[, 4]   0.03506
Y[, 5]  -0.47669
---
Signif. codes:
```

$$\Sigma^{-1} = \begin{pmatrix} 1.00 & -0.50 & -0.50 & 0.00 & 0.50 \\ -0.50 & 1.00 & 0.50 & 0.00 & 0.00 \\ -0.50 & 0.50 & 1.00 & 0.50 & 0.00 \\ 0.00 & 0.00 & 0.50 & 1.00 & 0.00 \\ 0.50 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}$$

Partial correlation vs. regression

From ρ(·|·) to β

An explicit relationship between the regression and partial correlation coefficients exists.

Hereto formulate the simultaneous-regression model:

$$
\begin{aligned}
Y_{i,1} &= \qquad\qquad\qquad \beta_{12} Y_{i,2} + \ldots + \beta_{1p} Y_{i,p} + \varepsilon_{i,1} \\
Y_{i,2} &= \beta_{21} Y_{i,1} \qquad\qquad + \ldots + \beta_{2p} Y_{i,p} + \varepsilon_{i,2} \\
\ldots &= \ldots \\
Y_{i,p} &= \beta_{p1} Y_{i,1} + \beta_{p2} Y_{i,2} + \ldots + \qquad\qquad \varepsilon_{i,p}
\end{aligned}
$$

Each $Y_{i,j}$ is regressed on all other $Y_{i,j}$'s.

Partial correlation vs. regression

From ρ(·|·) to β

It turns out that:

$$\beta_{12} = \rho(Y_1, Y_2 \,|\, Y_3, \ldots)\sqrt{(\mathbf{\Sigma}^{-1})_{11}/(\mathbf{\Sigma}^{-1})_{22}}$$

and its reverse:

$$\rho(Y_1, Y_2 \,|\, Y_3, \ldots) = \operatorname{sign}(\beta_{12})\sqrt{\beta_{12}\beta_{21}}$$

Conclusion

Thus:

$$\rho(Y_1, Y_2 \,|\, Y_3, \ldots) \overset{1-1}{\Longleftrightarrow} \beta_{12}$$

In particular:

$$\beta_{12} = 0 \iff Y_1 \perp\!\!\!\perp Y_2 \,|\, Y_3, \ldots, Y_p$$

All nice …
… but to what end?

All nice ...

Example

Reconstruct the topology of the TP53 signaling pathway.

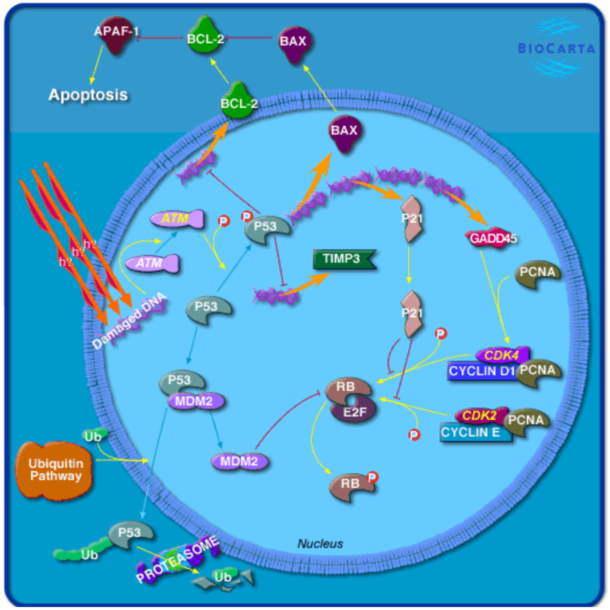
Available

→ Genes that comprise the TP53 pathway (from Biocarta)

→ Gene expression data of breast cancer samples (from Bioconductor)

Goal

→ identify gene-gene interaction



BioCarta: p53 signalling pathway

All nice ...

Extract pathway data

Replace UNT by e.g. VDX

Replace 150 by e.g. 190 for ErbB2 signalling pathway

```
# packages and Sources
library(Biobase)
library(rags2ridges)
library(graphite)
library(breastCancerUNT)

# load biocarta pathway
biocaPathway <- biocarta[[150]]
biocaPathway <- convertIdentifiers(biocaPathway, "entrez")
entrezIDs <- nodes(biocaPathway)

# load expression data of pathway genes
data(unt)
unt <- unt[match(entrezIDs, as.character(levels(fData(unt)[,5])
              [fData(unt)[,5]])),]
unt <- unt[, which(pData(unt)[,8] == 1)]
gNames <- fData(unt)[,3]
Y <- t(exprs(unt))
```

All nice ...

Pathway reconstruction

```
# specify number edges to select
# probabilistic selection only for large pathways
top <- 14

# reconstruct netwerk
lambdaOpt <- optPenaltyLOOCVauto(Y, 0.001, 100)
estP <- ridgeS(covML(Y), lambdaOpt)
sparseP <- sparsify(estP, "top", top=top,
                    output="heavy")$sparsePrecision
colnames(sparseP) <- rownames(sparseP) <- gNames

# plot inferred pathway
Ugraph(sparseP, lay=layout.circle, type="fancy")
```

Try
`lay=layout.random,`
`lay=layout.auto`

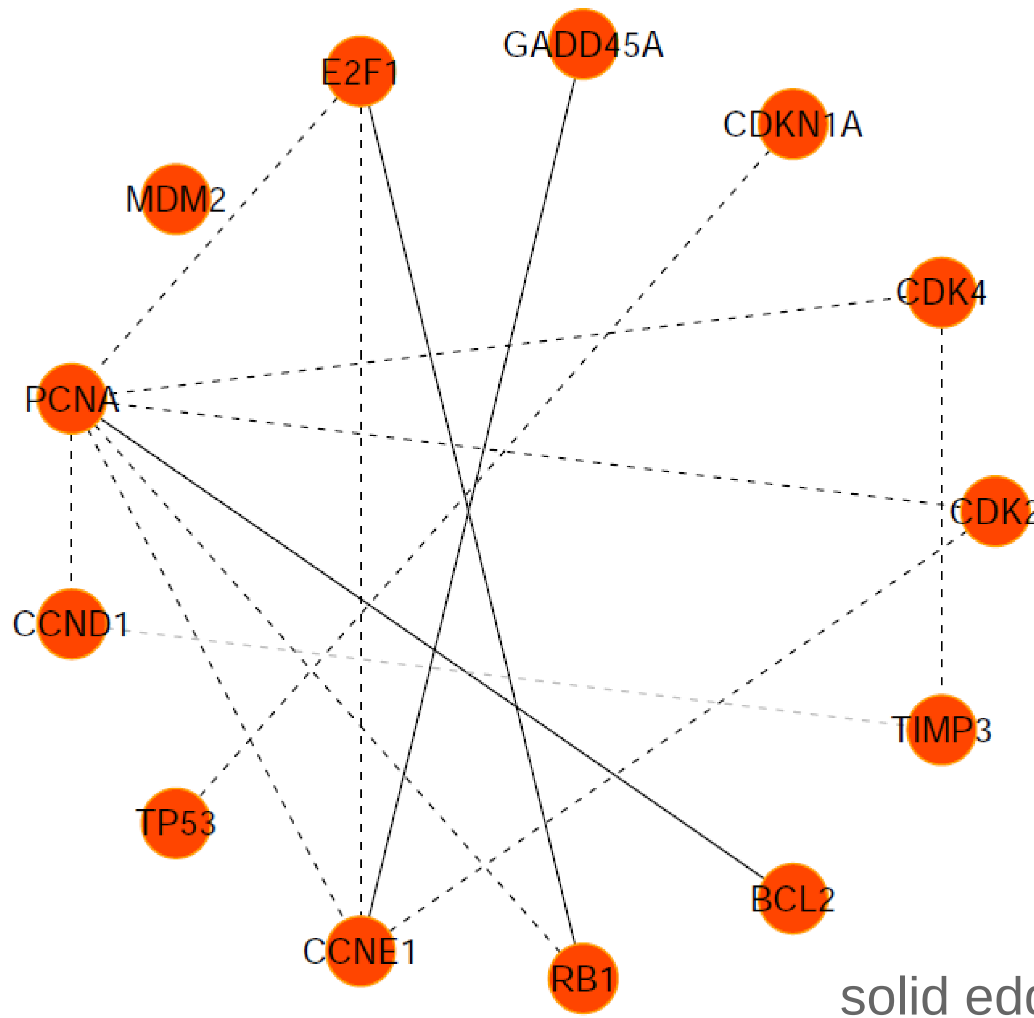
Try
`type="weighted",`
`type="plain"`

All nice ...

Pathway reconstruction

Visualization is important!

Question
Which genes interact? E.g. do genes RB1 and E2F1 interact?



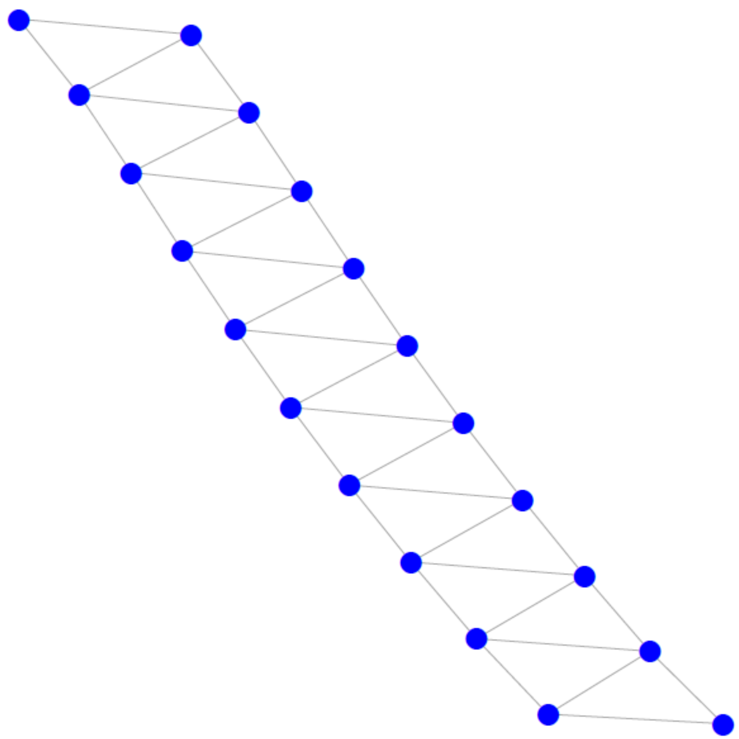
solid edge: > 0
dashed edge: < 0

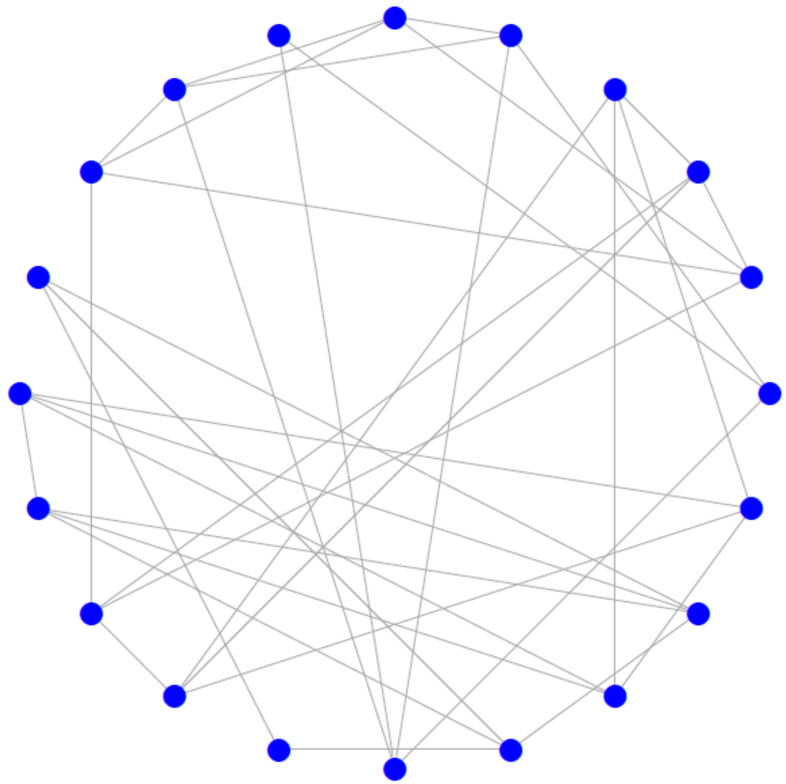
All nice ...

Pathway reconstruction

Visualization is important!

Two instances of same network:



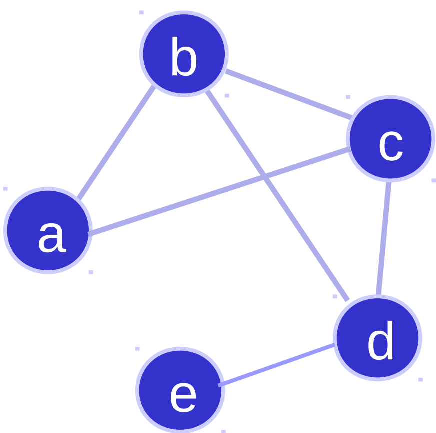


All nice ...

Comparison with Biocarta

Comparison is not visual, but via the adjacency matrix.

Pathway topology



Adjacency matrix

to

from	*a*	*b*	*c*	*d*	*e*
a	0	1	1	0	0
b	1	0	1	1	0
c	1	1	0	1	0
d	0	1	1	0	1
e	0	0	0	1	0

This adjacency matrix is symmetric, as the pathway topology is undirected.

All nice ...

Comparison with Biocarta

```
# inferred adjancency matrix
inferAdj <- adjacentMat(sparseP)

# biocarta adjancency matrix
biocaEdges <- edges(biocaPathway)[,1:2]
biocaEdges <- matrix(unlist(lapply(c(biocaEdges[,1],
                biocaEdges[,2]), function(X, Y){
                which(X == Y) }, Y=entrezIDs)), ncol=2,
                byrow=FALSE)
biocaAdj <- 0 * inferAdj
biocaAdj[biocaEdges] <- 1
biocaAdj[cbind(biocaEdges[,2], biocaEdges[,1])] <- 1

# compare adjacency matrix
table(biocaAdj[upper.tri(biocaAdj)],
inferAdj[upper.tri(inferAdj)])
```

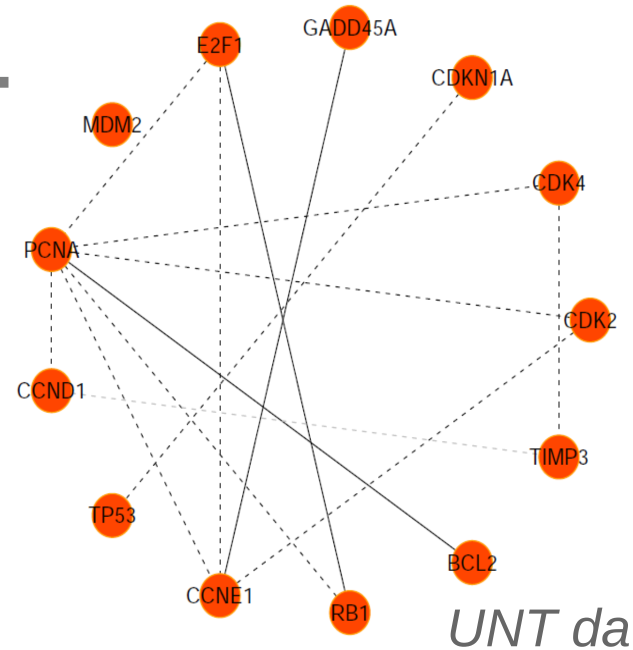
All nice ...

Comparison with Biocarta

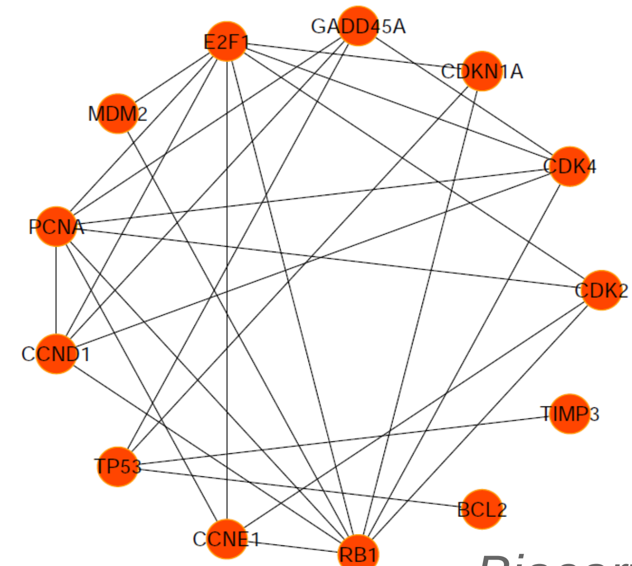
Contingency table of nonredundant elements of both adjacency matrices:

		UNT	
Biocarta		0	1
	0	46	4
	1	18	10

E.g. ten overlapping edges.



UNT data



Biocarta

All nice ...

Node analysis

BIOINFORMATICS **ORIGINAL PAPER**

Systems biology

Global topological features of cancer proteins interactome

Pall F. Jonsson and Paul A. Bates*

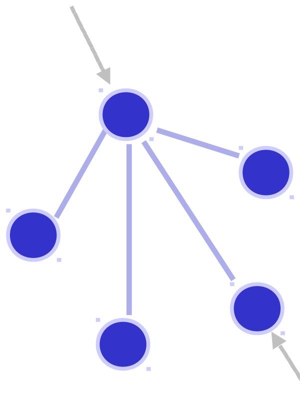
“The most striking property of cancer proteins is the increased frequency of interactions they participate in. This observation indicates an underlying evolutionary pressure to which cancer genes, as genes of central importance, are subjected.”

All nice ...

Node analysis

Hub ≈ many connections.

gene with
high degree



gene with
low degree

Question: role of the hub?

Hypothesis
Hubs are disease genes.

Infer network and compare to *census of human cancer genes** from:

Hypothesis not confirmed.

*Futreal *et al*. (2004), *Nature Reviews Cancer*.

All nice ...

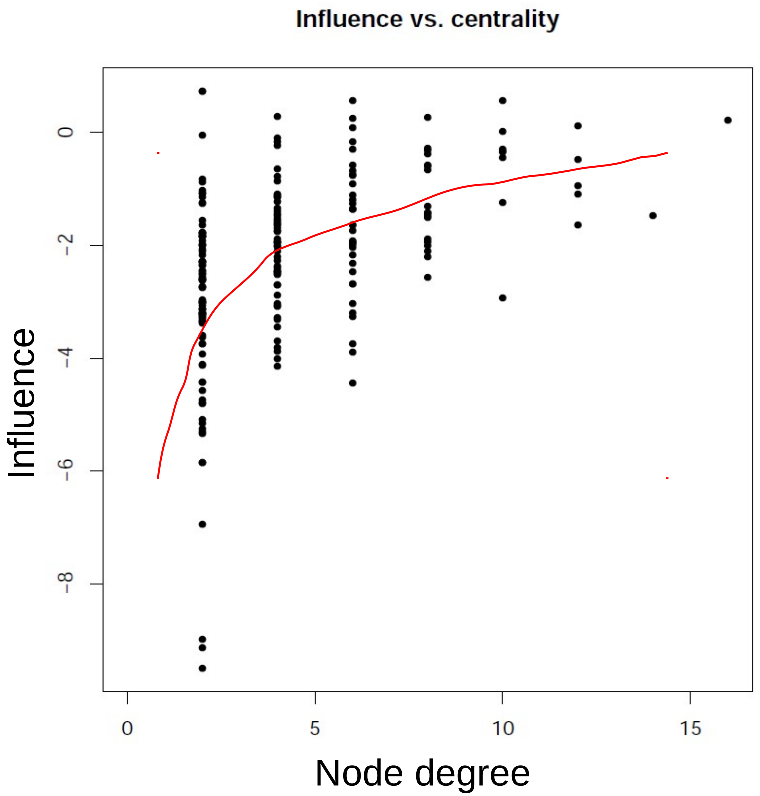
Node analysis

Measure influence between gene and rest by *mutual information*:

$$\begin{aligned}\mathcal{I}(\mathbf{Y}_{\setminus j}; Y_j) &= \mathcal{H}(\mathbf{Y}_{\setminus j}) - \mathcal{H}(\mathbf{Y}_{\setminus j} \,|\, Y_j) \\ &= \log\{|\mathrm{Var}(\mathbf{Y}_{\setminus j})|\} \\ &\quad - \log\{|\mathrm{Var}(\mathbf{Y}_{\setminus j} \,|\, Y_j)|\}\end{aligned}$$

Measure of information shared between two random variables.

Hypothesis
Hubs are influential

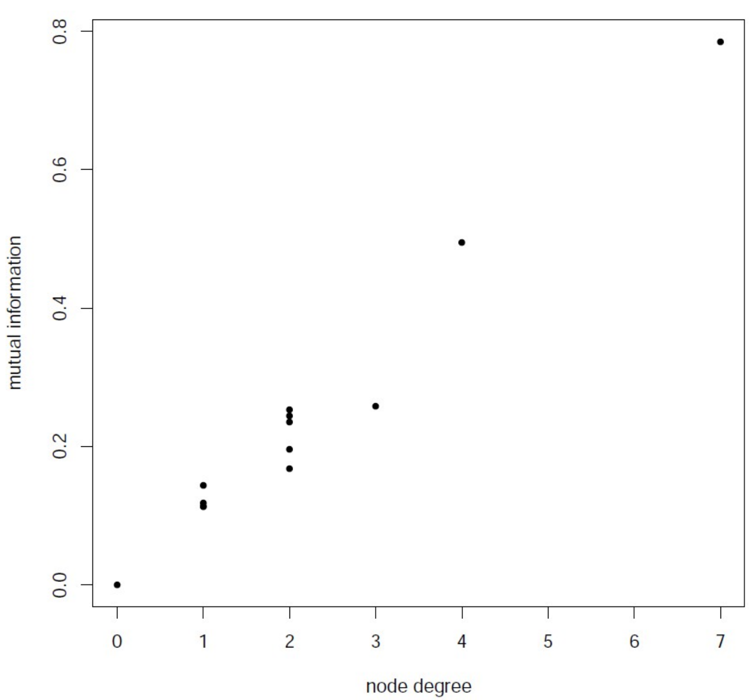


All nice ...

Node analysis

```
# calculate network statistics per gene
nodeStats <- GGMnetworkStats(sparseP, as.table=TRUE)
nodeStats[1:2,]
     degree betweenness  closeness nNeg nPos mutualInfo     ...
CDK2      2         0.0 0.01785714    2    0  0.2524956     ...
CDK4      2         3.5 0.01785714    2    0  0.2348174     ...
```

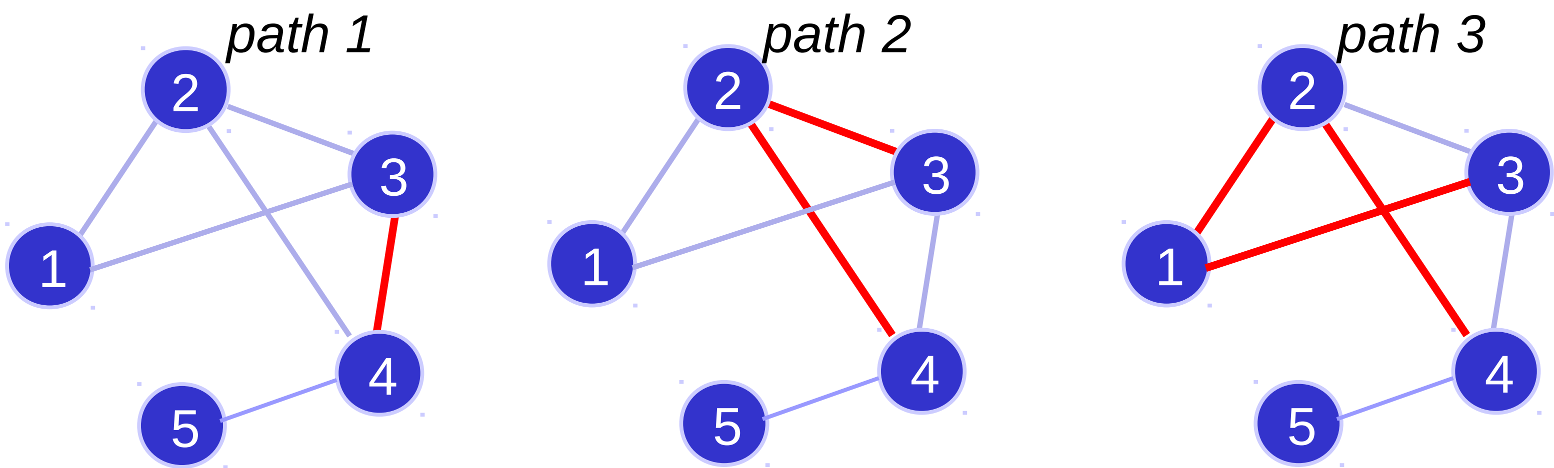
```
# degree vs. MI
plot(nodeStats[,1],
   nodeStats[,6]))
```



All nice ...

Path analysis

Understand the covariance between genes 3 and 4 by decomposition into the paths that propagate signals between these genes.



Which path contributes most to covariance?

All nice ...

Path analysis

The covariance between two nodes can be decomposed into the contributions of the paths connecting these nodes.

The covariance between nodes j_1 and j_2 equals:

$$(\mathbf{\Sigma})_{j_1,j_2} \quad = \sum_{P \in \mathcal{P}_{j_1 j_2}} (-1)^{r+1} \frac{\det(\mathbf{\Omega}_{\setminus P, \setminus P})}{\det(\mathbf{\Omega})} \prod_{s=2}^{r} (\mathbf{\Omega})_{p_{s-1}, p_s}$$

where $\mathcal{P}_{j_1 j_2}$ the set of all paths from j_1 to j_2 and

$$P \quad = \quad \{(p_1 = j_1, p_2), (p_2, p_3), \ldots, (p_{r-1}, p_r = j_2)\}$$

a path of length r from j_1 to j_2.

All nice ...

Path analysis

Also try e.g.
TP53 (node 6) and
MDM2 (node 9)

```
# E2F1 and RB1
node1 <- 5
node2 <- 11
pathStats <- GGMpathStats(sparseP, node1, node2, prune=FALSE,
                          nodecol="red", VBcolor="orange")


Covariance between node pair : -0.0103
----------------------------------------
              path length contribution
 1           5--11      1     -0.02735
 2        5--7--11      2      0.01243
 3    5--10--7--11      3      0.00392
 4 5--10--1--7--11      4      0.00071
----------------------------------------
Sum path contributions       : -0.0103
```

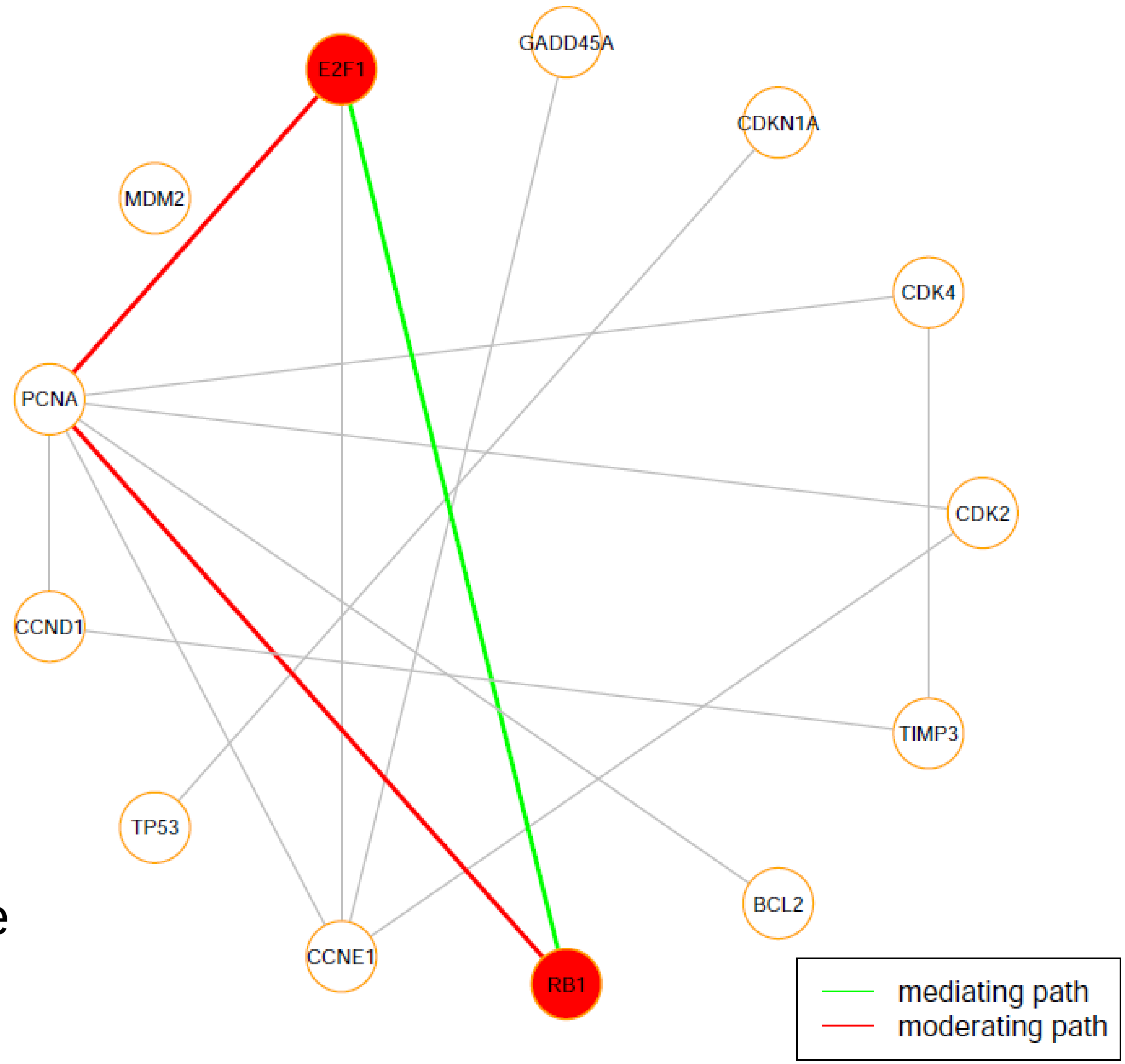
All nice ...

Path analysis

Top mediating and moderating paths are plotted.

Mediating path:
contribution has same sign as observed covariance

Moderating path:
contribution has opposite sign as observed covariance



Towards a causal graph

Causality

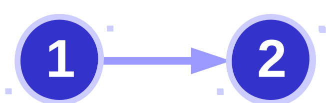
Causal relation

A causes B if a change in A may lead to one in B.

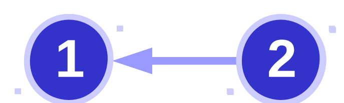
A *causal graph* depicts the causal relations among variables.
Some classic examples:

no cause

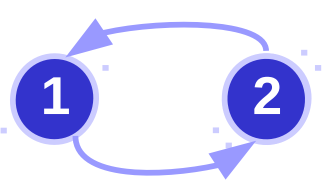
cause



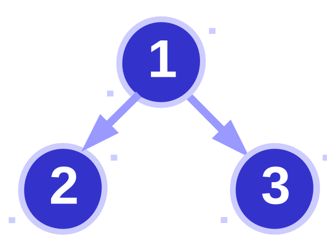
cause



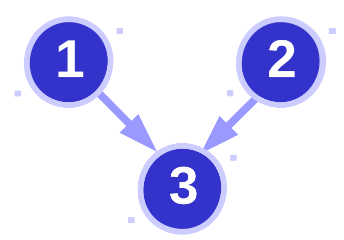
mutual cause



common cause



common effect



 : a direct causal relationship

DAG

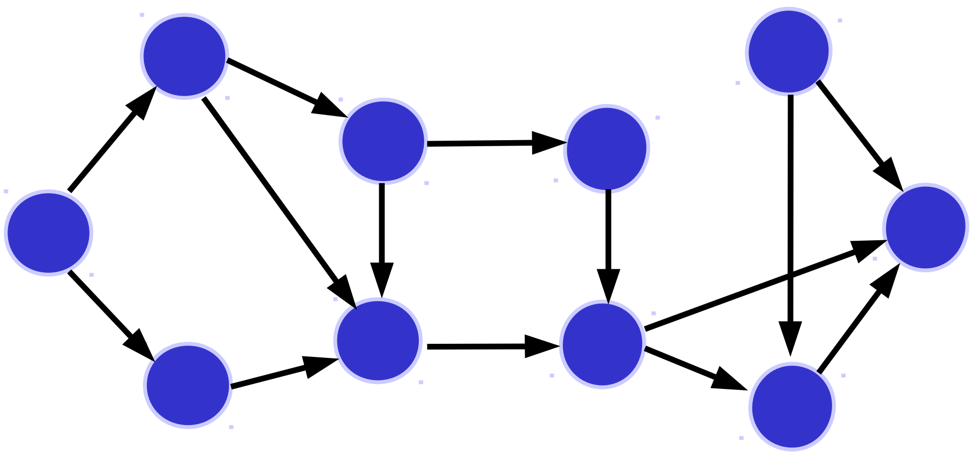
Directed acyclic graph (DAG)

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that contains:

→ directed edges only: if $(j, j') \in \mathcal{E}$ then $(j', j) \notin \mathcal{E}$.

→ no directed cycles.

The graph $\mathcal{G}$ contains a *directed cycle* if $(j_k, j_{k+1}) \in \mathcal{E}$ for $k = 1, \ldots, K-1$ with $j_1 = j_K$.



The phylogenetic tree is another example.

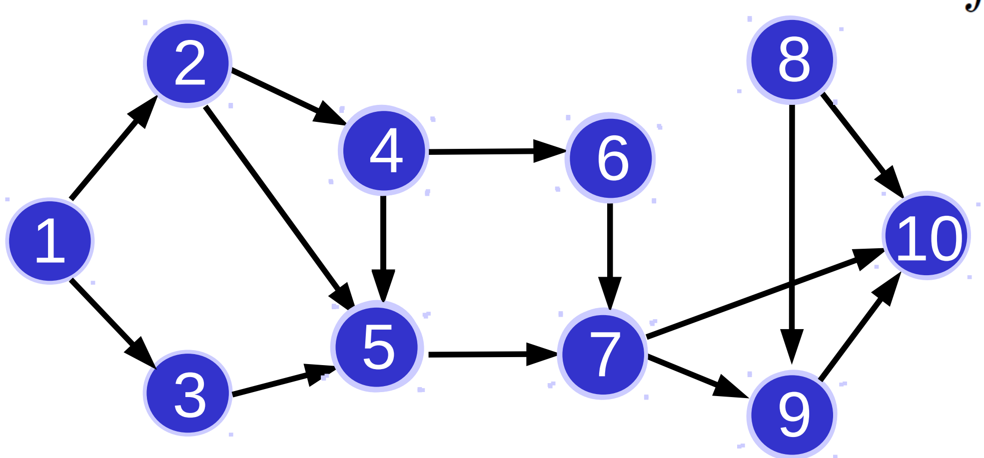
Bayesian network

A Bayesian network is a graphical model that describes the behaviour of a random variable **Y** through a DAG and accompanying density:

$$f_{\mathbf{Y}}(y_1, \ldots, y_p) = \prod_{j=1}^{p} f(y_j \,|\, \text{all } y_{j'} \text{ s.t. } j' \in pa(j))$$

where pa(j) denotes the parents of j (according to the DAG).

For instance, in:



$$\begin{aligned} f_{\mathbf{Y}}(y_1, \ldots, y_{10}) \\ = \quad & f(y_2 \,|\, y_1) f(y_3 \,|\, y_1) \\ & f(y_4 \,|\, y_2) f(y_5 \,|\, y_2, y_3, y_4) \\ & \times \ldots \end{aligned}$$

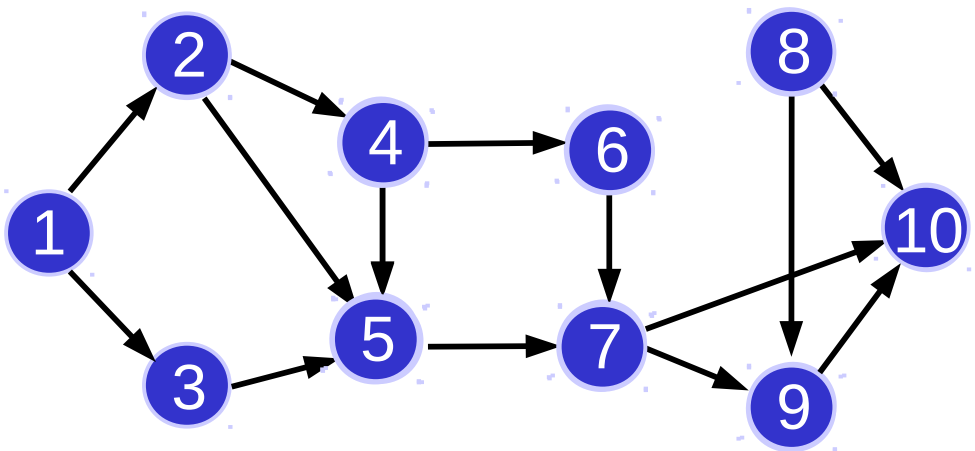
Markov again

A Bayesian network satisfies the local Markov property:

$$Y_j \perp\!\!\!\perp \mathbf{Y}_{nd(j)} \,|\, \mathbf{Y}_{pa(j)}$$

where nd(j) denotes the non-decendants of j (i.e. there is no directed path from j to these nodes with the DAG).

For instance, in:

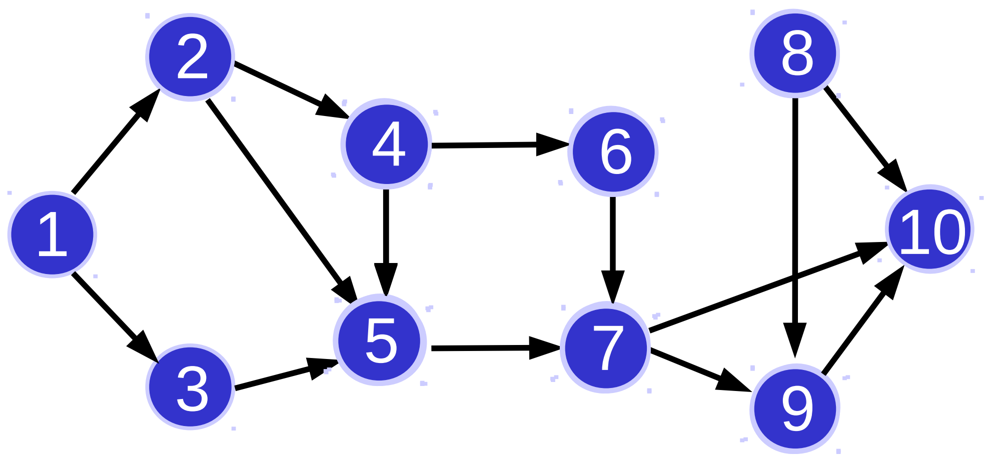


$$Y_4 \perp\!\!\!\perp Y_1, Y_3 \,|\, Y_2$$

Learning a Bayesian network

Would the causal graph be known and be a DAG,
it can be fitted by a system of regression equations.

For instance, to fit:



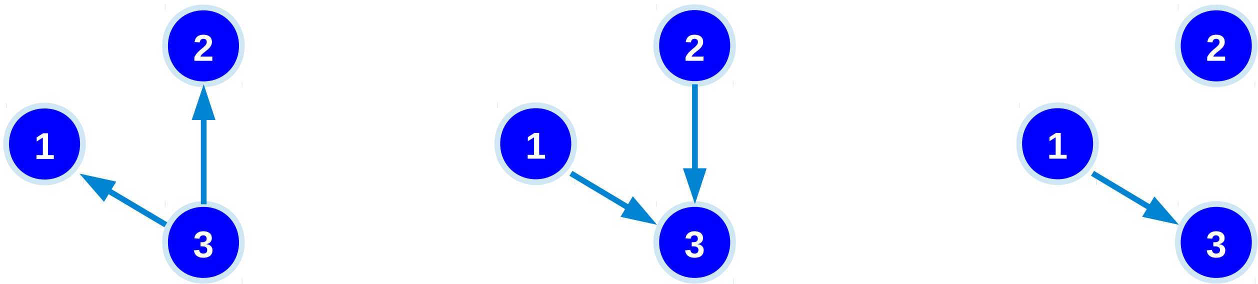
→ Regression Y_3 on Y_1,

→ ...

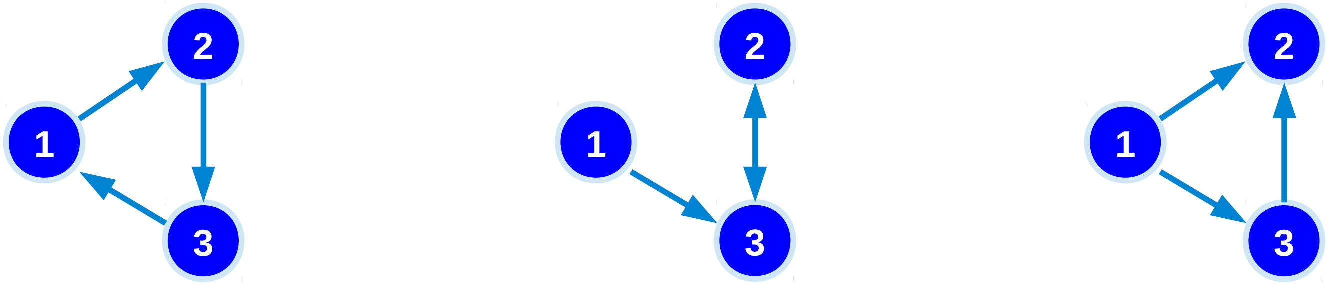
→ Regress on Y_{10} on Y_7, Y_8 and Y_9

Directionality

Assume causal structure has directed tree structure, e.g.:



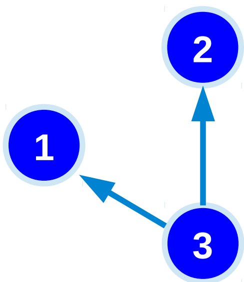
but not, e.g.:



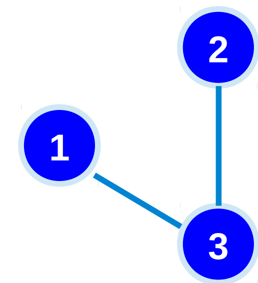
(contain loops: forbidden by assumption)

Directionality

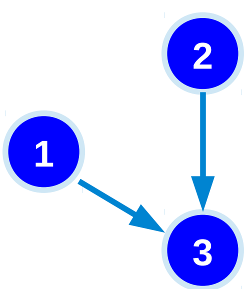
Causal structure induces factorization and graph:



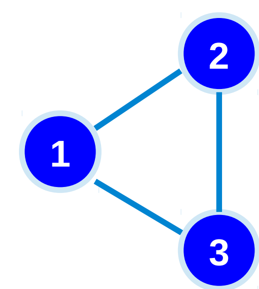
$$
\begin{aligned}
&f(y_1, y_2, y_3) \\
&\quad = f(y_3) f(y_1, y_2 \mid y_3) \\
&\quad = f(y_3) f(y_1 \mid y_3) f(y_2 \mid y_3)
\end{aligned}
$$



whereas:



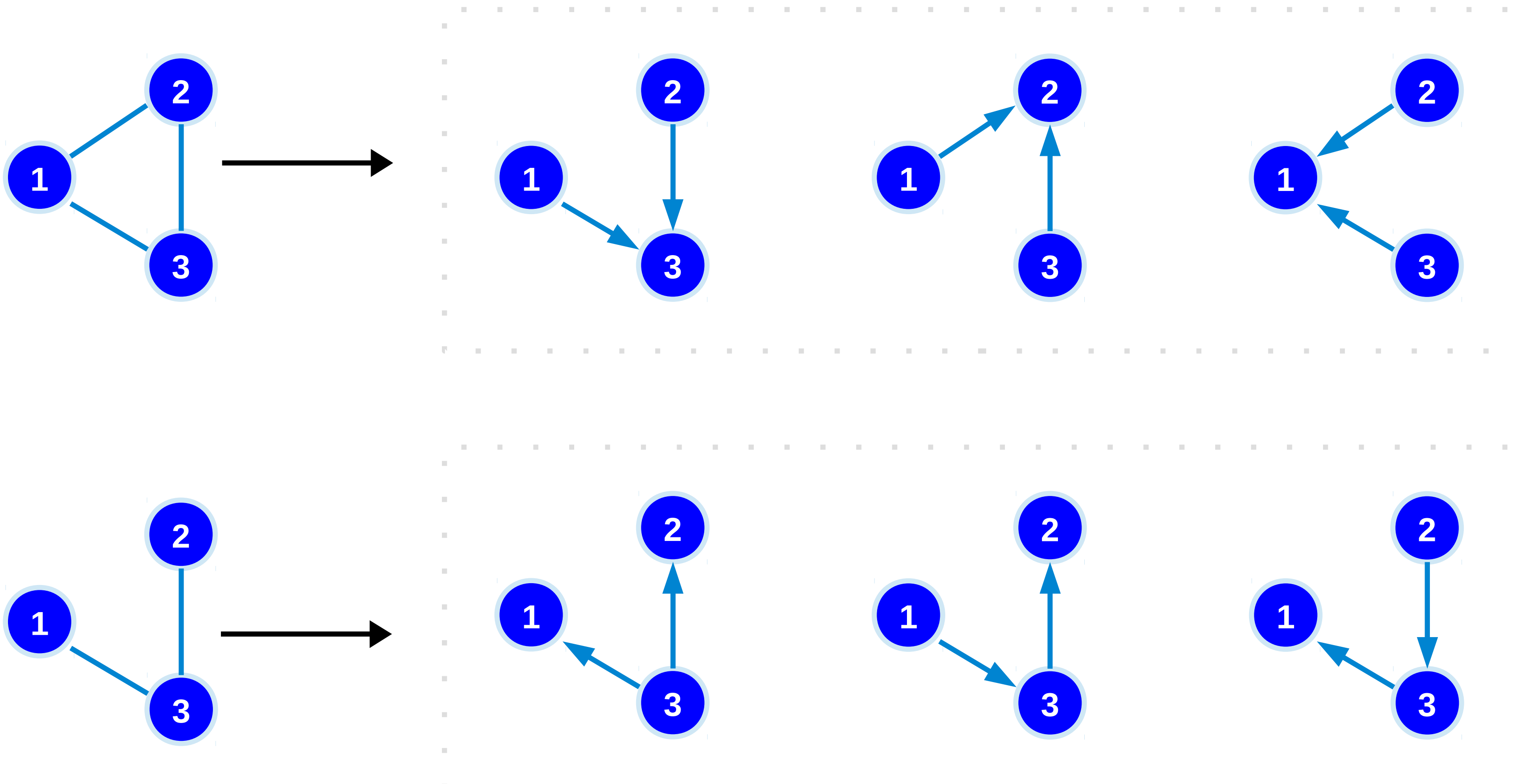
$$
\begin{aligned}
&f(y_1, y_2, y_3) \\
&\quad = f(y_1) f(y_2) f(y_3 \mid y_1, y_2) \\
&\quad \neq g(y_1, y_3) h(y_2, y_3)
\end{aligned}
$$



Directionality

Observed undirected networks

Possible underlying directed networks



Interpretation pitfall revisited

(*or*: the case for integration)

Interpretation pitfall revisited

Reconstruct the ErbB signalling pathway

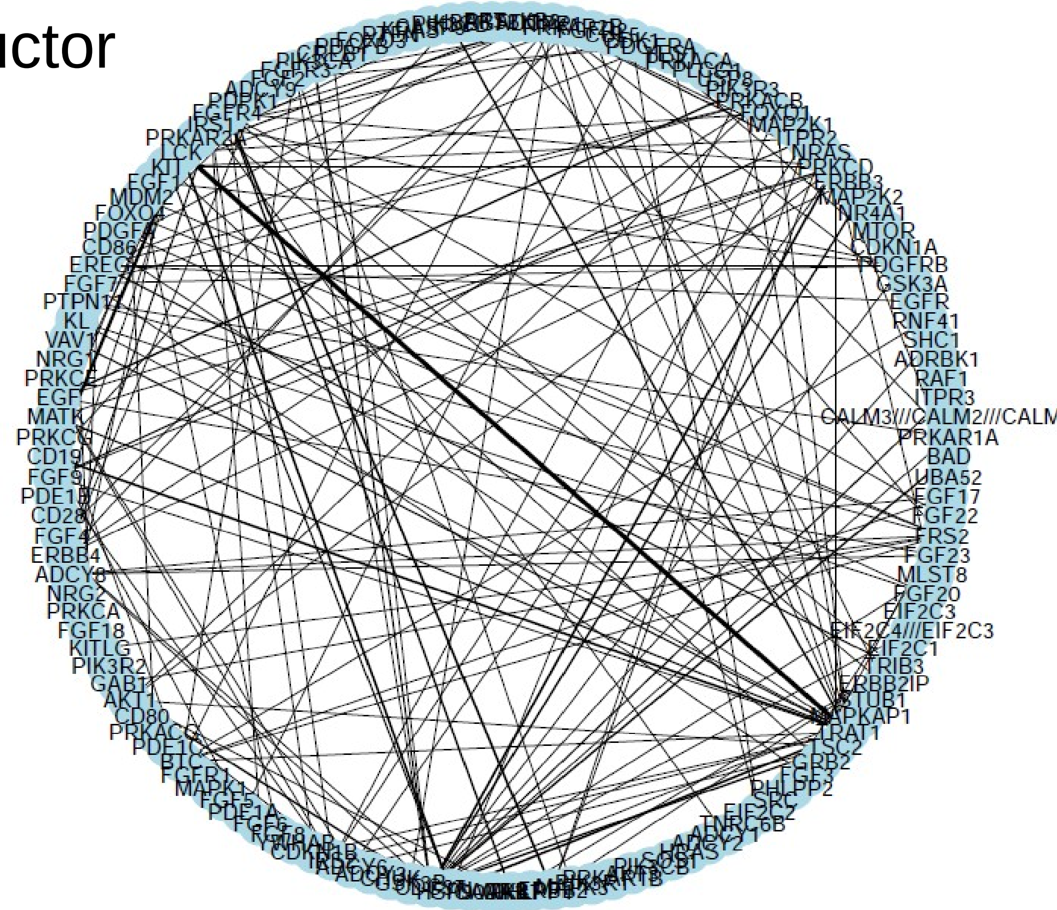
Available

Off the shelf from Bioconductor

4 breast cancer data sets

Reconstruction

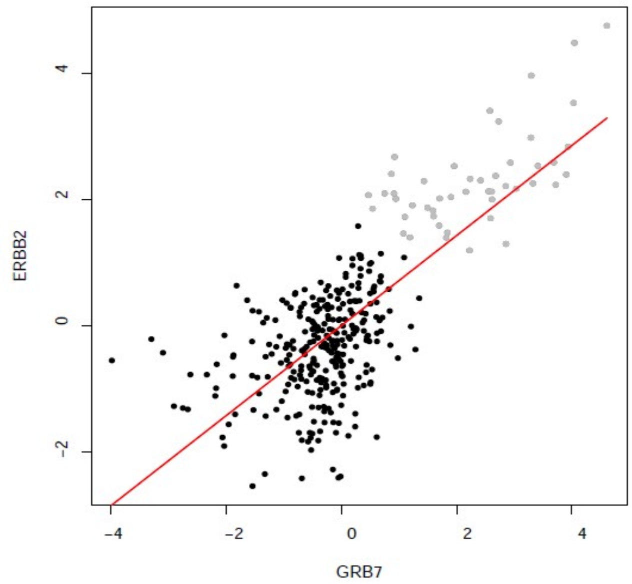
(Fit GGM with ridge penalty to pathway data. Post-hoc sparsification by local FDR procedure.)



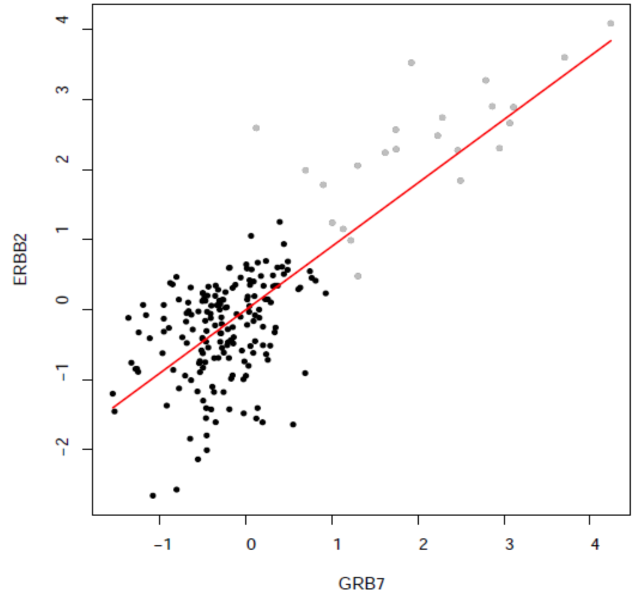
Interpretation pitfall revisited

ErbB2 - GRB7 edge

→ Top ranking edge
→ ErbB2 often amplified
→ GRB7 maps to ErbB2 amplicon
→ ErbB2 and GRB7 co-expressed



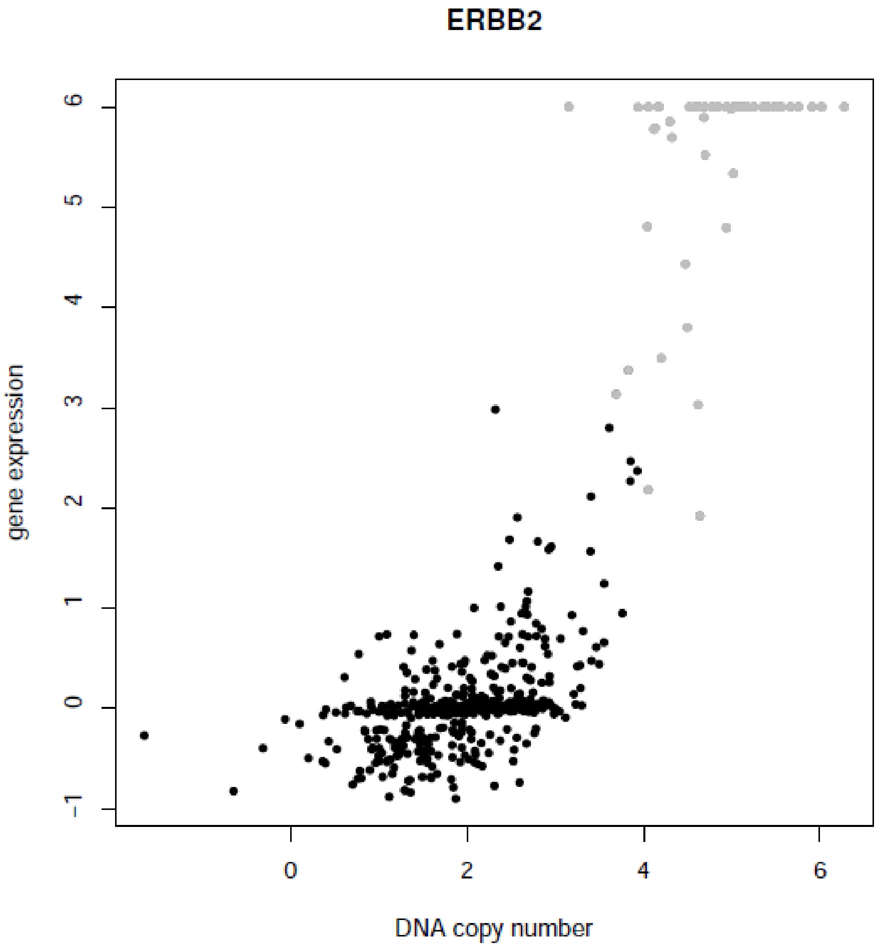
Breast cancer data set	Marginal correlation	Partial correlation
VDX	0.733	0.624
MAINZ	0.772	0.668
TRANSBIG	0.795	0.767
UPP	0.866	0.815

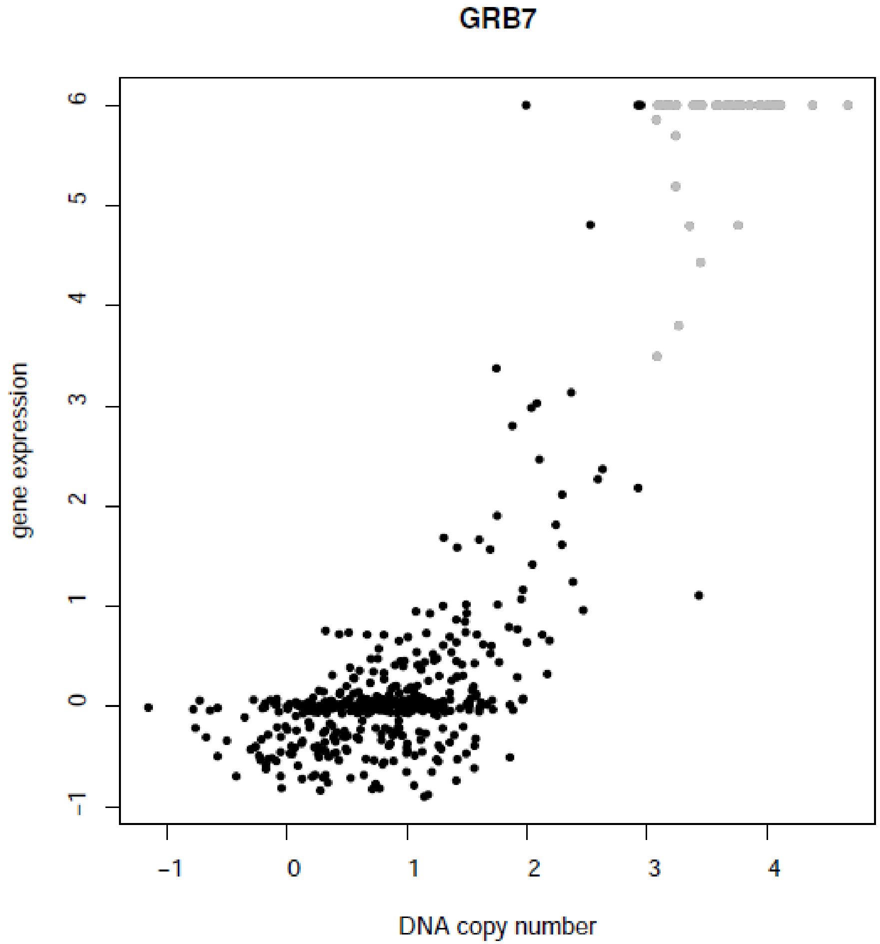


Interpretation pitfall revisited

ErbB2 - GRB7 edge

TCGA breast cancer data: copy number vs. expression

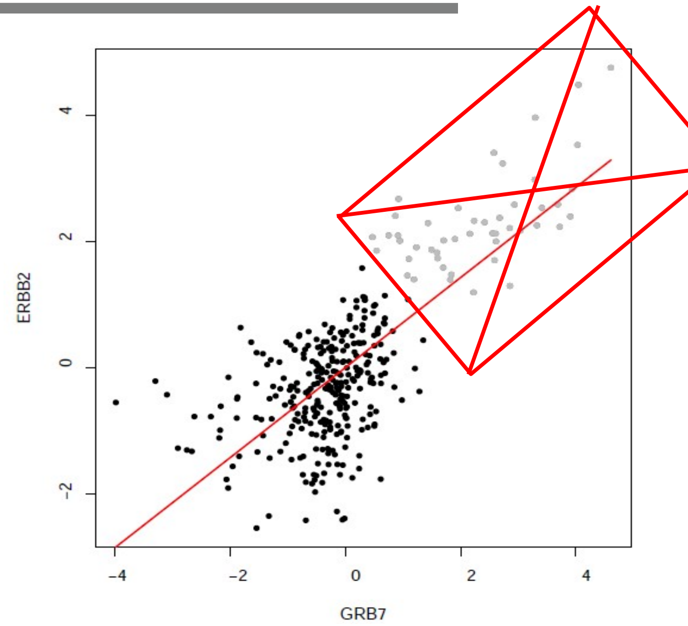




Interpretation pitfall revisited

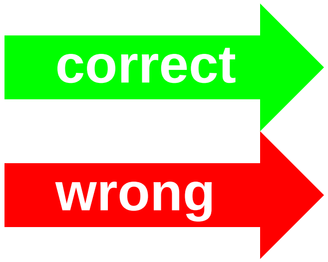
ErbB2 - GRB7 edge

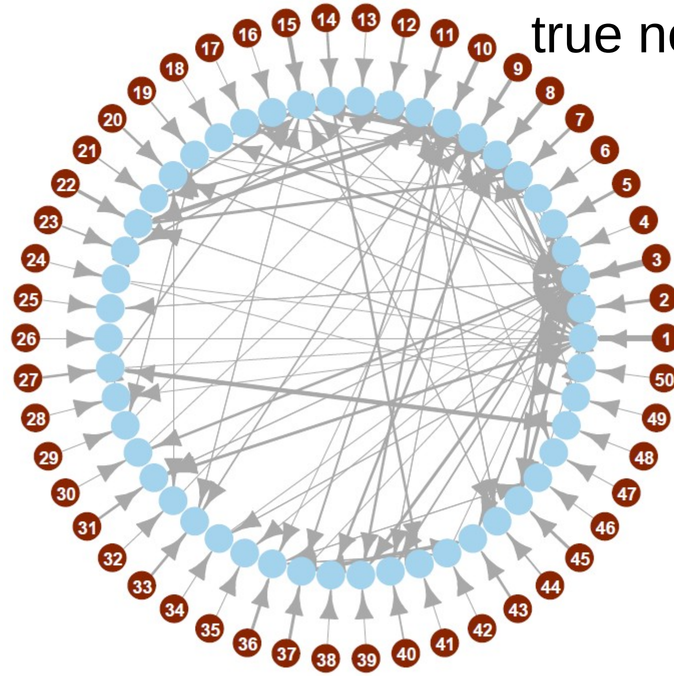
→ Remove “amplified samples”
→ Reconstruct CIG of pathway
→ Amplification contributes to edge strength



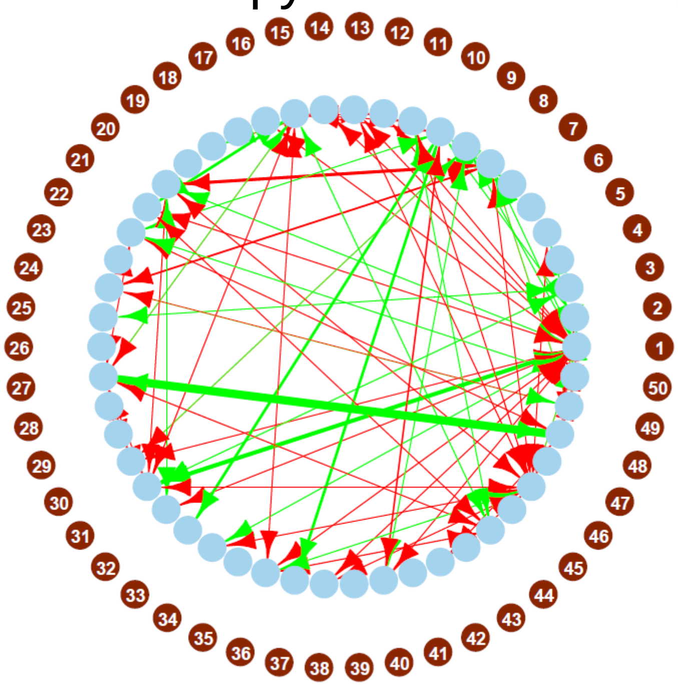
Breast cancer data set	Marginal correlation	Partial correlation	Marg. cor. ampl. rem.	Part. cor. ampl. rem.	Rank
VDX	0.733	0.624	0.340	0.238	69 (out of 9453)
MAINZ	0.772	0.668	0.357	0.293	521 (out of 9453)
TRANSBIG	0.795	0.767	0.428	0.314	457 (out of 9453)
UPP	0.866	0.815	0.370	0.305	237 (out of 11628)

Interpretation pitfall revisited

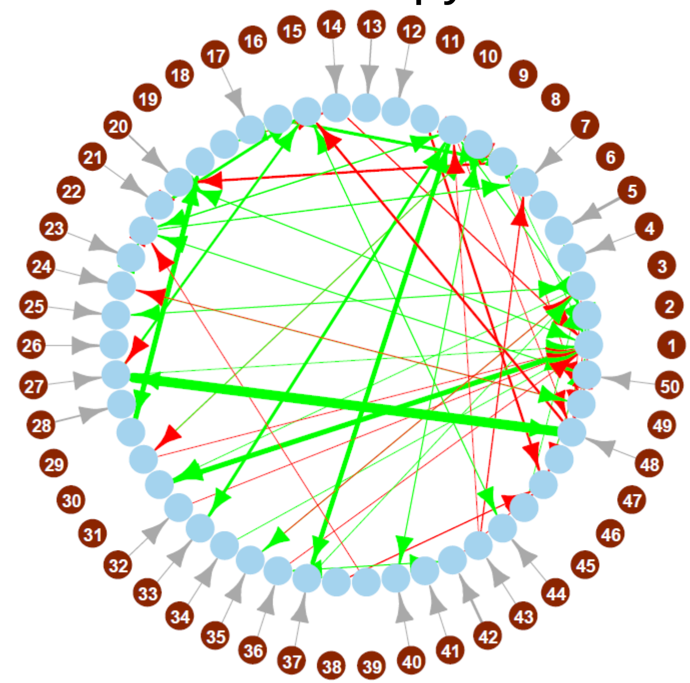



true network

Reconstruction
without copy number

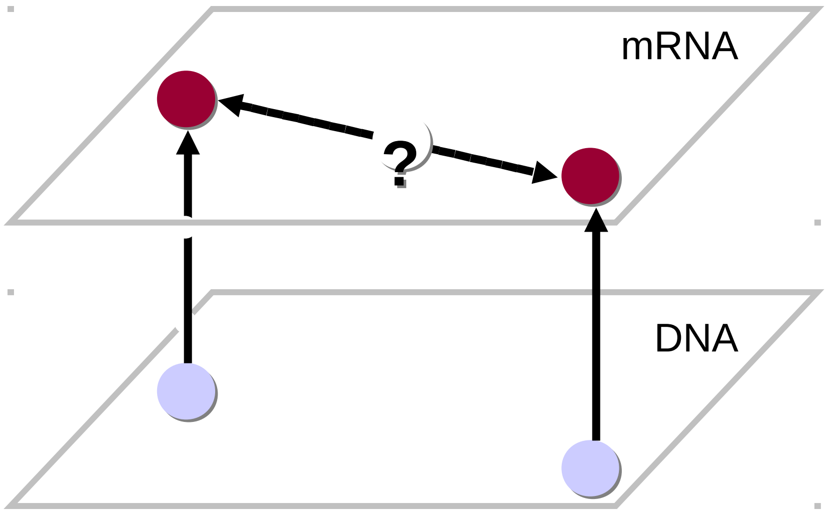


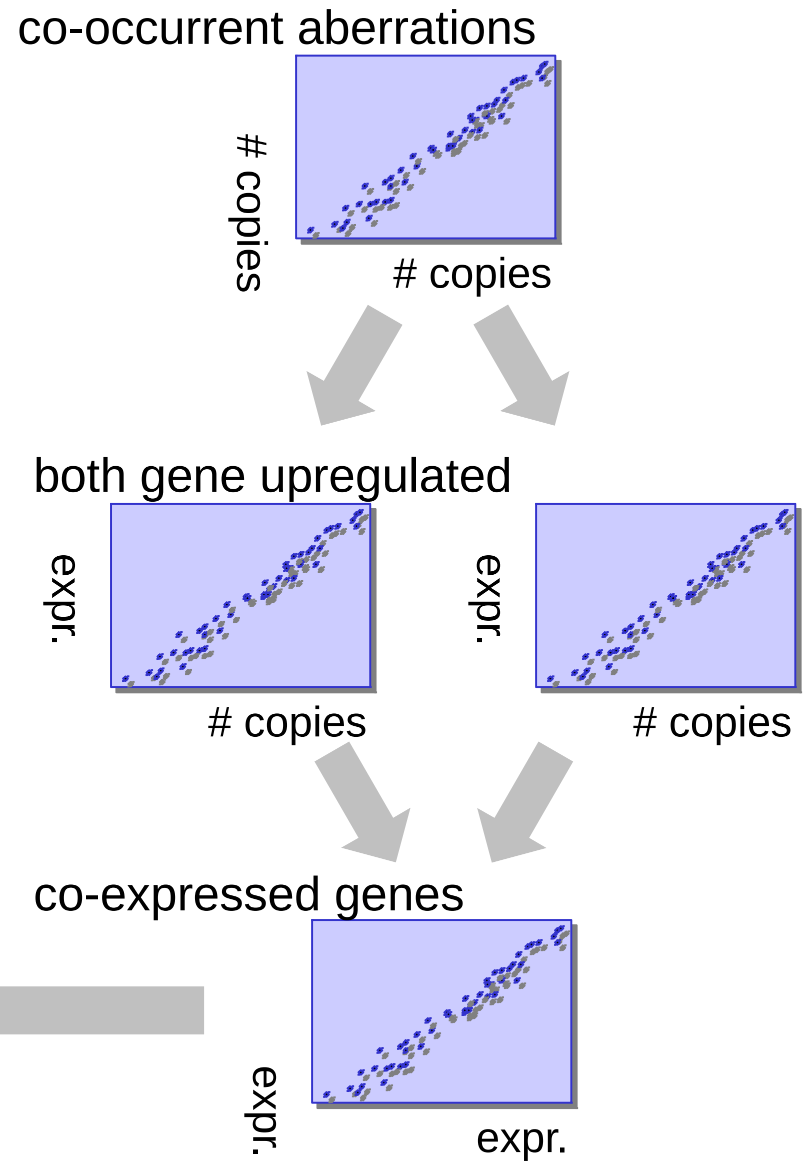
Reconstruction
with copy number



Interpretation pitfall revisited

Simple case: two genes



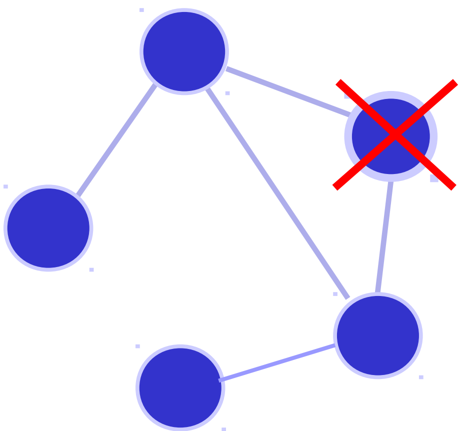


co-expression ≠ co-regulated

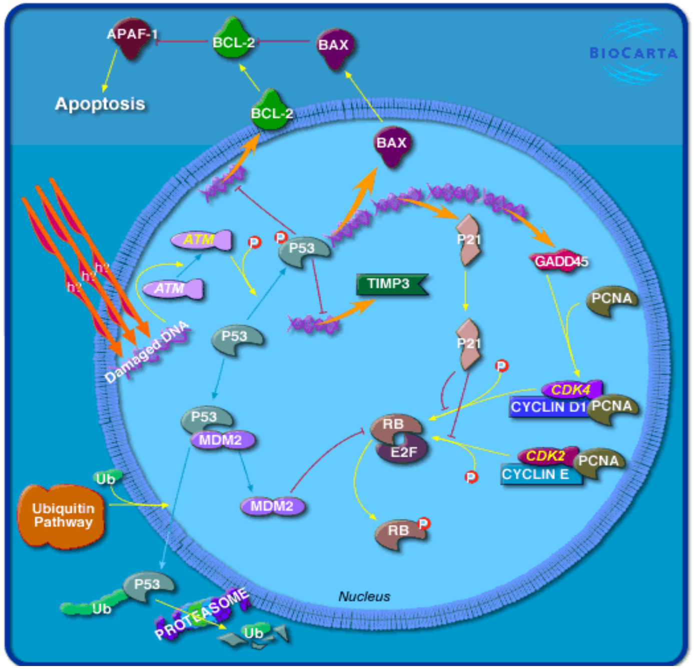
Further topics

Further topics

Effect of changes in the regulatory system.



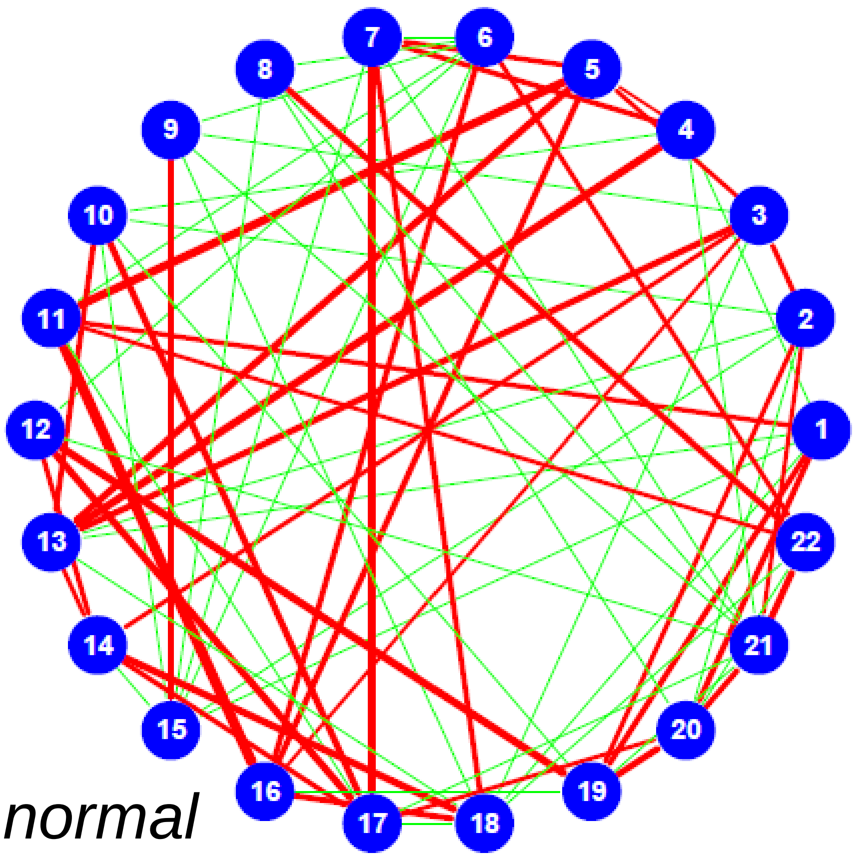
interpretation



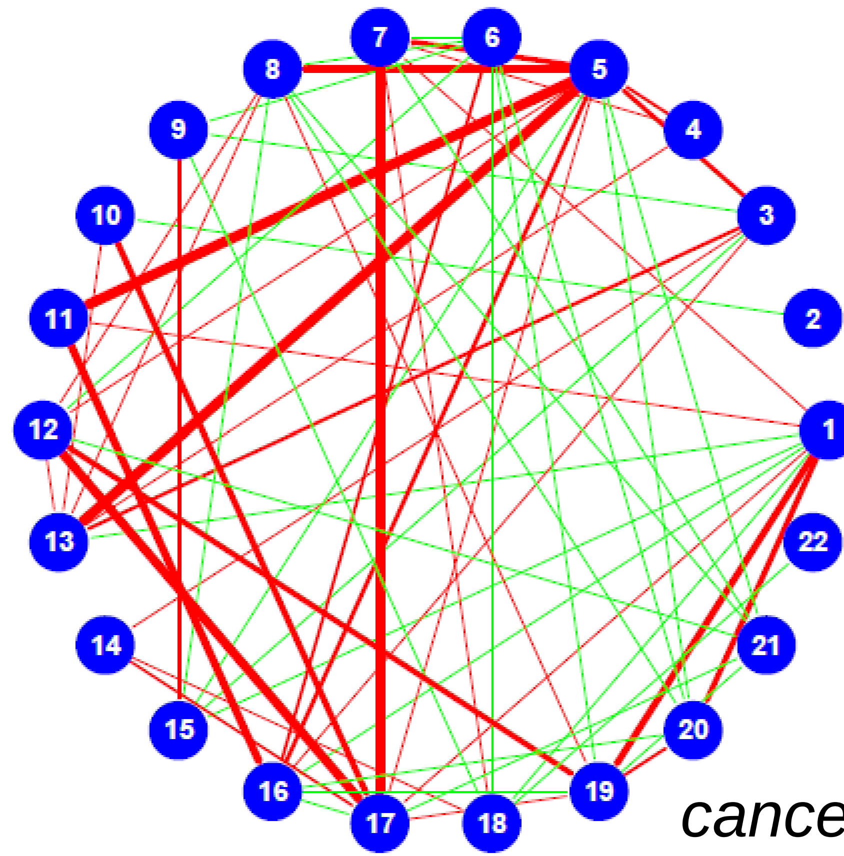
knock-out:
model predicts effect

Further topics

Network differences between two conditions



normal

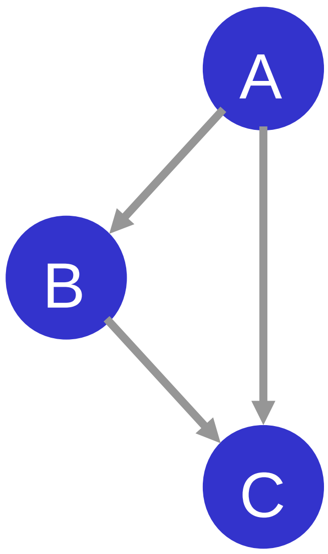


cancer

Further topics

Dynamic networks

Feedforward loop



Feedforward loop (unrolled)