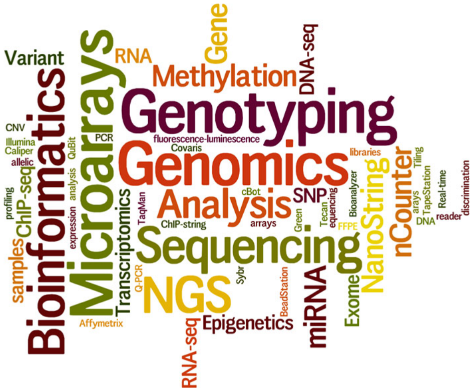
Correlation Networks

Francisco Peñagaricano

University of Florida

High throughput technologies: omics data

- **genetic variants**
- **gene expression**
- **epigenetic modifications**
- **proteins** and **metabolites**



- measuring different **phenotypic traits**

unprecedented opportunities to uncover the **genetic architecture** underlying **phenotypic variation**

Main challenge:

decipher the flow of biological information

- integrate multiple sources of biological information in order to reveal the **causal biological networks** that underlie complex traits

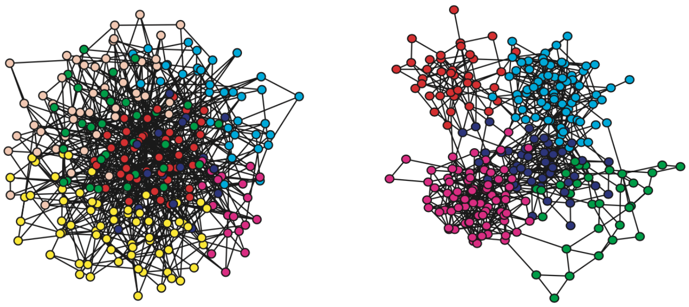
Why do we want to infer Causal Biological Networks?

- to better understand the biology of the traits
- to predict the behavior of complex systems
- to optimize management practices and breeding strategies

Biological Networks

(undirected graphs)

genes and gene products do not work in isolation; rather they are connected in complex networks



Correlation Networks

- networks constructed based on **correlation values**
- **main goal:** to explain the observed correlations between gene measurements by the presence of other genes

main challenge: how deeply correlations are resolved

FROM **correlation networks** (most simple)

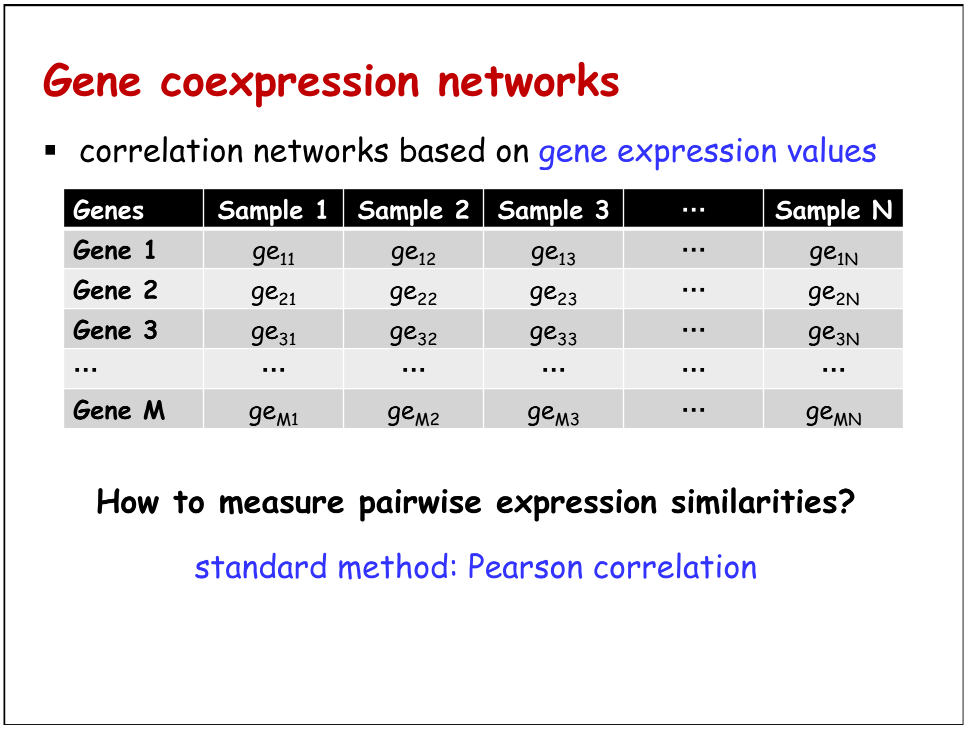
TO **Bayesian networks** (most sophisticated)

Gene coexpression networks

- correlation networks based on gene expression values

Genes	Sample 1	Sample 2	Sample 3	…	Sample N
Gene 1	ge_{11}	ge_{12}	ge_{13}	…	ge_{1N}
Gene 2	ge_{21}	ge_{22}	ge_{23}	…	ge_{2N}
Gene 3	ge_{31}	ge_{32}	ge_{33}	…	ge_{3N}
…	…	…	…	…	…
Gene M	ge_{M1}	ge_{M2}	ge_{M3}	…	ge_{MN}

gene coexpression network: it is a graph where **nodes** correspond to genes and (undirected) **edges** represent pairwise expression similarities



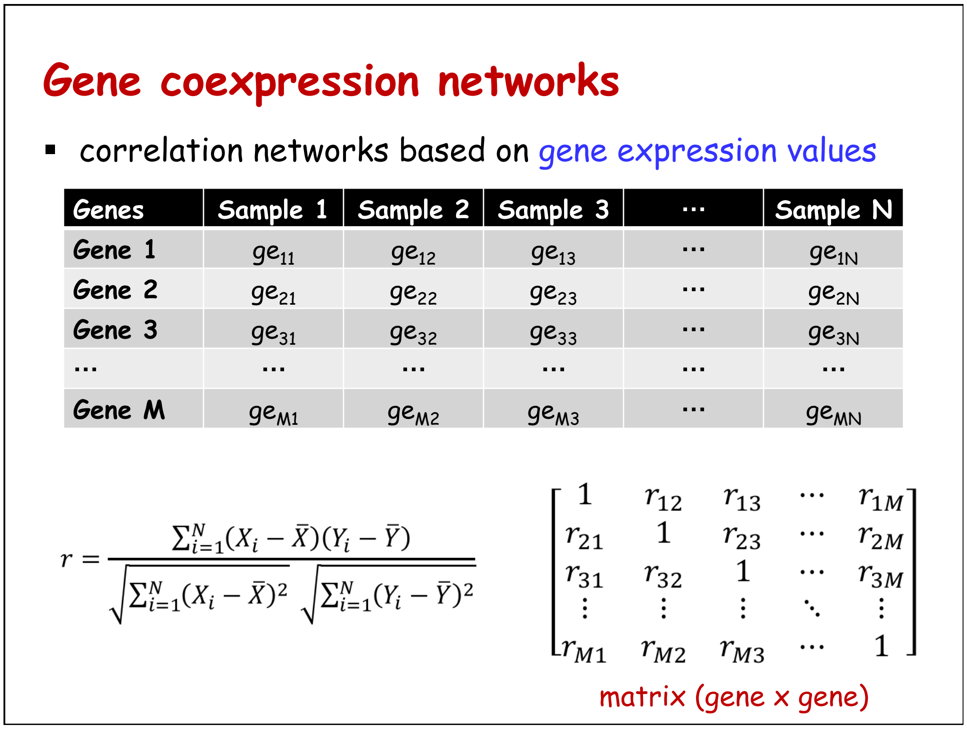
Gene coexpression networks

- correlation networks based on gene expression values

Genes	Sample 1	Sample 2	Sample 3	···	Sample N
Gene 1	ge_{11}	ge_{12}	ge_{13}	···	ge_{1N}
Gene 2	ge_{21}	ge_{22}	ge_{23}	···	ge_{2N}
Gene 3	ge_{31}	ge_{32}	ge_{33}	···	ge_{3N}
···	···	···	···	···	···
Gene M	ge_{M1}	ge_{M2}	ge_{M3}	···	ge_{MN}

How to measure pairwise expression similarities?

standard method: Pearson correlation



Gene coexpression networks

- correlation networks based on gene expression values

Genes	Sample 1	Sample 2	Sample 3	···	Sample N
Gene 1	ge_{11}	ge_{12}	ge_{13}	···	ge_{1N}
Gene 2	ge_{21}	ge_{22}	ge_{23}	···	ge_{2N}
Gene 3	ge_{31}	ge_{32}	ge_{33}	···	ge_{3N}
···	···	···	···	···	···
Gene M	ge_{M1}	ge_{M2}	ge_{M3}	···	ge_{MN}

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\ \sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

$$\begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1M} \\ r_{21} & 1 & r_{23} & \cdots & r_{2M} \\ r_{31} & r_{32} & 1 & \cdots & r_{3M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{M1} & r_{M2} & r_{M3} & \cdots & 1 \end{bmatrix}$$

matrix (gene x gene)

Gene coexpression networks

- correlation networks based on gene expression values

unweighted network

$$\begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1M} \\ r_{21} & 1 & r_{23} & \cdots & r_{2M} \\ r_{31} & r_{32} & 1 & \cdots & r_{3M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{M1} & r_{M2} & r_{M3} & \cdots & 1 \end{bmatrix}$$

Adjacency Matrix

matrix with entries are **1** or **0**

$$A = [a_{ij}]$$

rule:

$$a_{ij} = \{\mathbf{1}\ \ if\ \ r_{ij} \geq r_{threshold}; \mathbf{0}\ otherwise\}$$

$$a_{ij} = \{\mathbf{1}\ \ if\ \ r_{ij}\ is\ significant; \mathbf{0}\ otherwise\}$$

Gene coexpression networks

correlation matrix

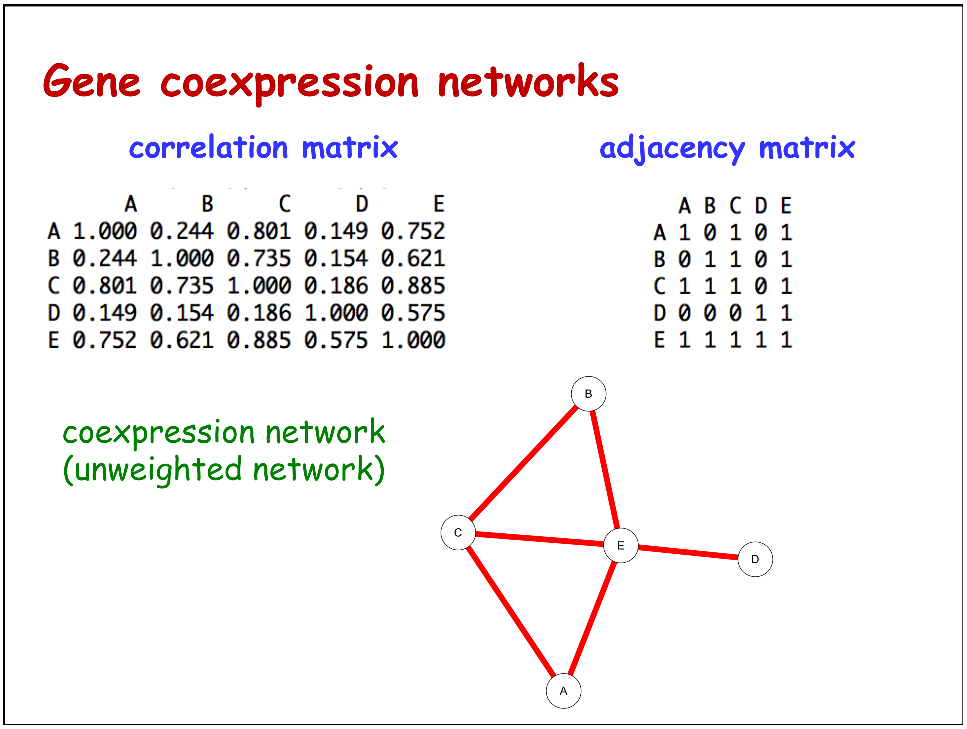
```
      A     B     C     D     E
A 1.000 0.244 0.801 0.149 0.752
B 0.244 1.000 0.735 0.154 0.621
C 0.801 0.735 1.000 0.186 0.885
D 0.149 0.154 0.186 1.000 0.575
E 0.752 0.621 0.885 0.575 1.000
```

adjacency matrix

```
  A B C D E
A 1 0 1 0 1
B 0 1 1 0 1
C 1 1 1 0 1
D 0 0 0 1 1
E 1 1 1 1 1
```

$$a_{ij} = \{\mathbf{1}\ if\ r_{ij}\ is\ significant; \mathbf{0}\ otherwise\}$$

$$T = r_{ij}\sqrt{\frac{n-2}{1-r_{ij}^2}} \qquad T \sim t_{df=n-2}$$



Gene coexpression networks

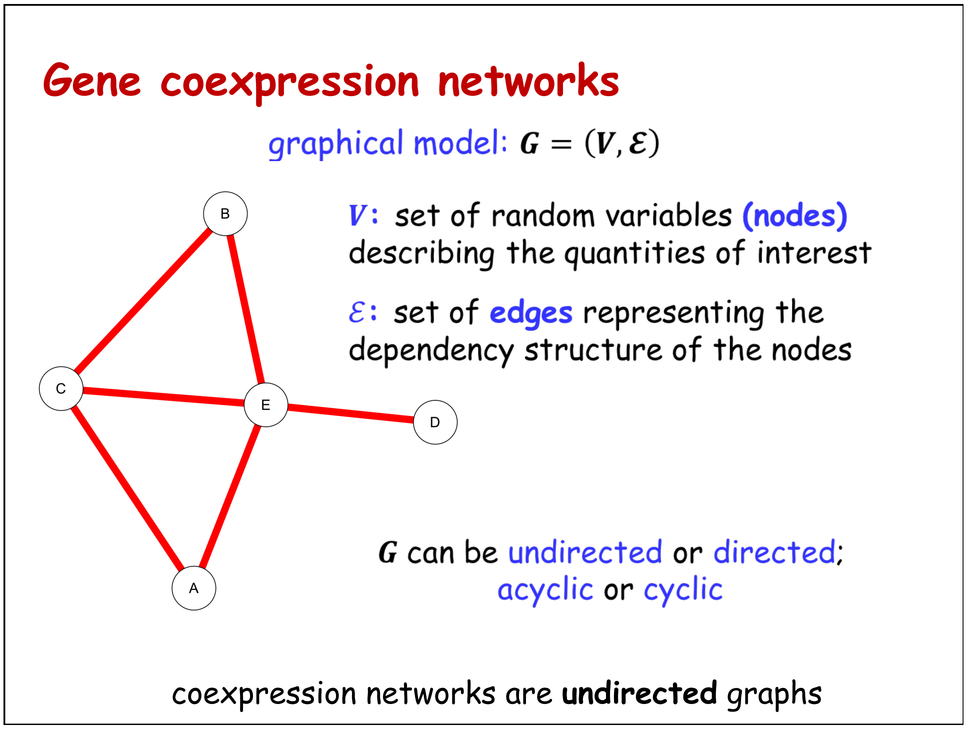
correlation matrix

	A	B	C	D	E
A	1.000	0.244	0.801	0.149	0.752
B	0.244	1.000	0.735	0.154	0.621
C	0.801	0.735	1.000	0.186	0.885
D	0.149	0.154	0.186	1.000	0.575
E	0.752	0.621	0.885	0.575	1.000

adjacency matrix

	A	B	C	D	E
A	1	0	1	0	1
B	0	1	1	0	1
C	1	1	1	0	1
D	0	0	0	1	1
E	1	1	1	1	1

coexpression network
(unweighted network)



Gene coexpression networks

graphical model: $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{\mathcal{E}})$

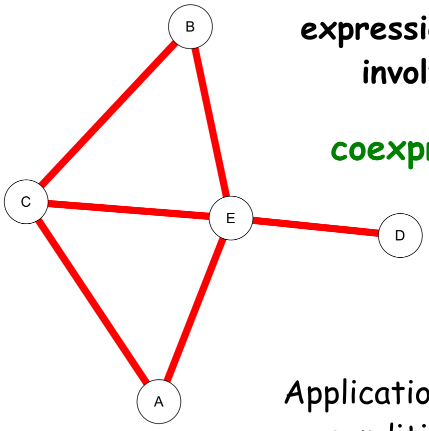
$\boldsymbol{V}$: set of random variables **(nodes)** describing the quantities of interest

$\mathcal{E}$: set of **edges** representing the dependency structure of the nodes

$\boldsymbol{G}$ can be undirected or directed;
acyclic or cyclic

coexpression networks are **undirected** graphs

Gene coexpression networks



Motivation: if two genes show **similar expression profiles**, they are probably **involved** in the **same processes**

coexpression hints coregulation

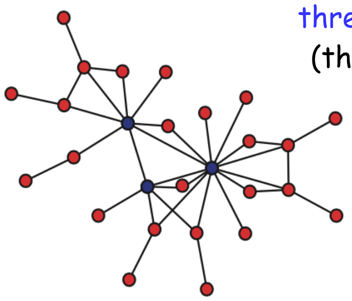
Application: **compare** networks between conditions (normal vs disease) and **identify changes** in the topology

Coexpression Network Topology

Gene connectivity: row sum of the adjacency matrix

- **number of direct neighbors** (unweighted networks)

Connectivity can be used to identify important genes



three highly connected nodes (**hubs**)
(they keep the network together)

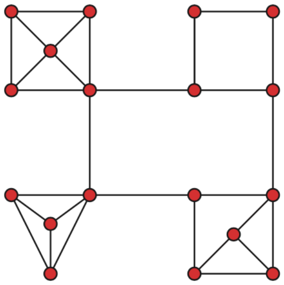
Ravasz et al. (2002) Science 297: 1551-1555

Coexpression Network Topology

Gene connectivity: row sum of the adjacency matrix

- **number of direct neighbors** (unweighted networks)

Connectivity can be used to identify important genes



four highly interconnected modules
(modules connected by a few links)

Ravasz et al. (2002) Science 297: 1551-1555

Coexpression Network Topology

modules: subset of nodes that are tightly connected

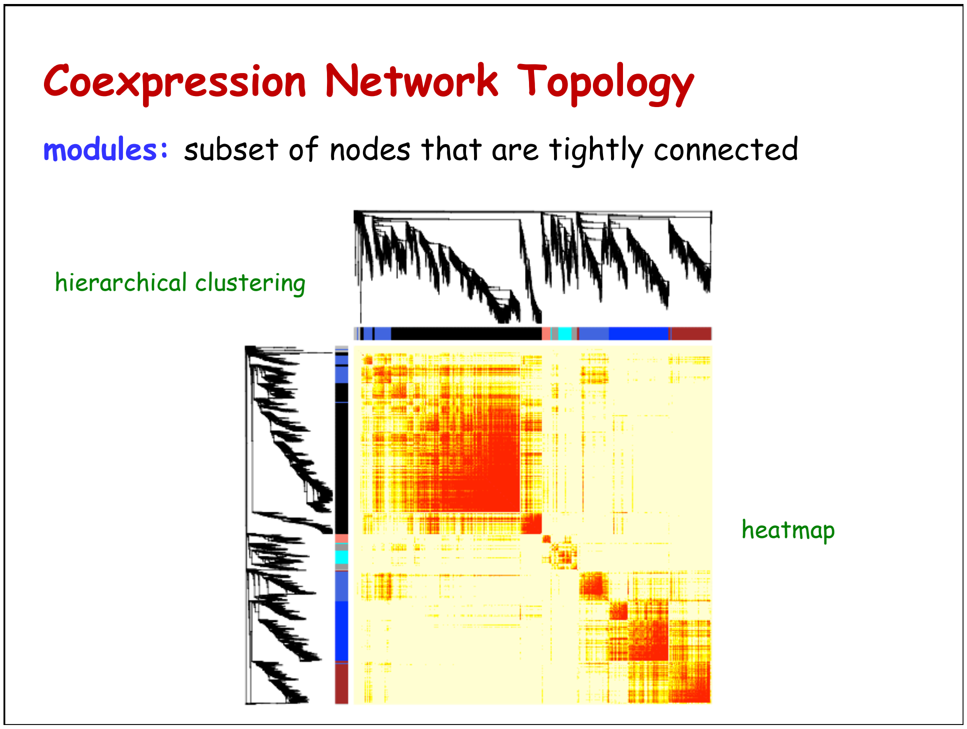
defining gene modules

1. define a dissimilarity measure between 2 genes

$$d_{ij} = 1 - abs(r_{ij})$$

2. hierarchical clustering using dissimilarity and define modules as branches of the hierarchical clustering tree

3. visualize the modules (clustering results) in a heatmap plot

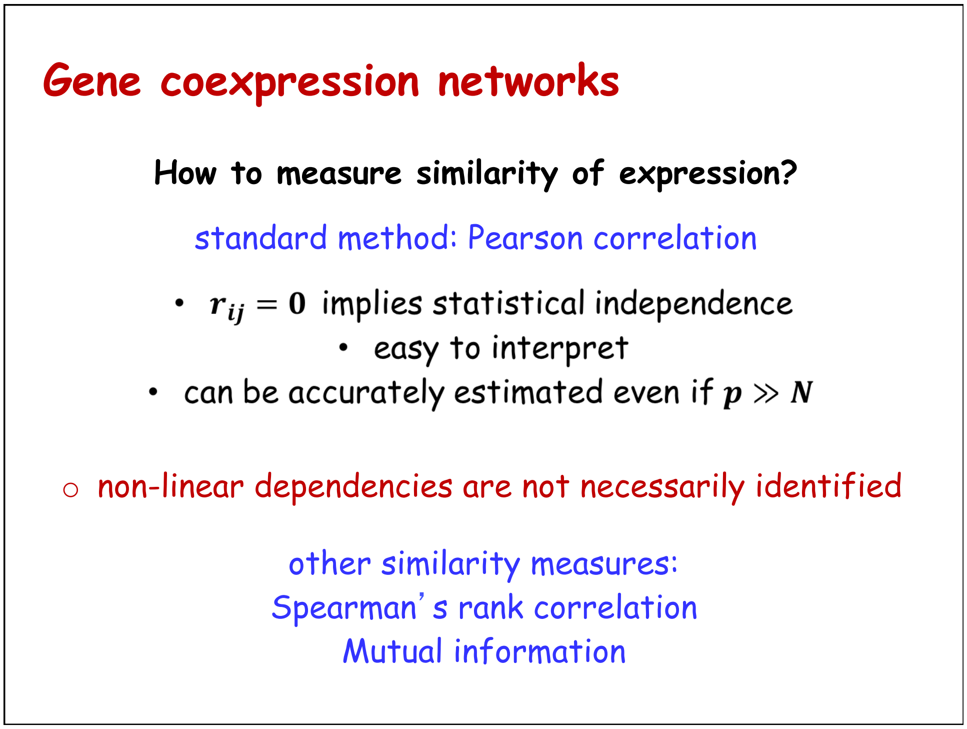


Coexpression Network Topology

modules: subset of nodes that are tightly connected

hierarchical clustering

heatmap



Gene coexpression networks

How to measure similarity of expression?

standard method: Pearson correlation

- $r_{ij} = 0$ implies statistical independence
- easy to interpret
- can be accurately estimated even if $p \gg N$

○ non-linear dependencies are not necessarily identified

other similarity measures:
Spearman's rank correlation
Mutual information

Gene coexpression networks

correlation matrix

```
      A     B     C     D     E
A 1.000 0.244 0.801 0.149 0.752
B 0.244 1.000 0.735 0.154 0.621
C 0.801 0.735 1.000 0.186 0.885
D 0.149 0.154 0.186 1.000 0.575
E 0.752 0.621 0.885 0.575 1.000
```

adjacency matrix

```
  A B C D E
A 1 0 1 0 1
B 0 1 1 0 1
C 1 1 1 0 1
D 0 0 0 1 1
E 1 1 1 1 1
```

dichotomization leads to loss of information

weighted networks: the adjacency matrix reports the connection strength between node pairs

Gene coexpression networks

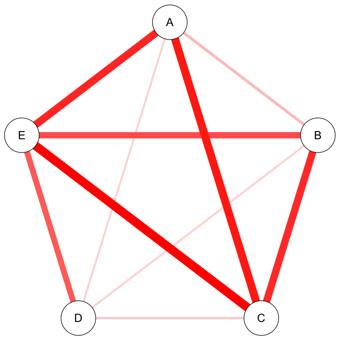
correlation matrix

```
      A     B     C     D     E
A 1.000 0.244 0.801 0.149 0.752
B 0.244 1.000 0.735 0.154 0.621
C 0.801 0.735 1.000 0.186 0.885
D 0.149 0.154 0.186 1.000 0.575
E 0.752 0.621 0.885 0.575 1.000
```

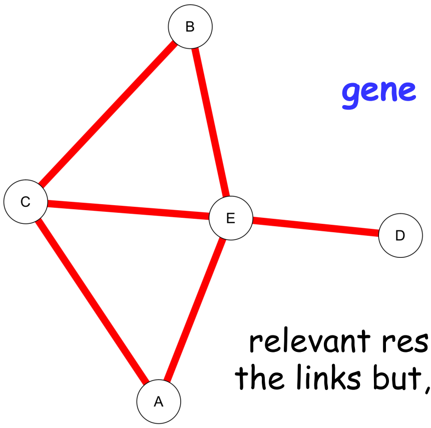
adjacency matrix

```
  A B C D E
A 1 0 1 0 1
B 0 1 1 0 1
C 1 1 1 0 1
D 0 0 0 1 1
E 1 1 1 1 1
```

coexpression network
(weighted network)



Marginal Independencies

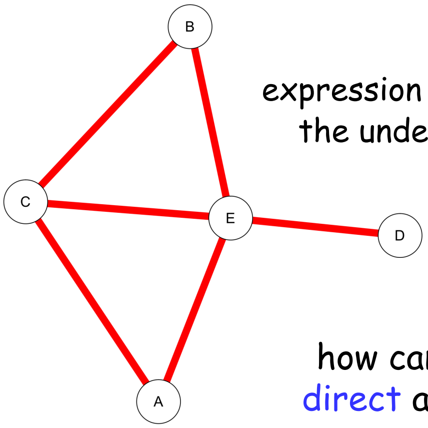


gene coexpression networks

relevant results are represented not in the links but, rather, in the missing links

- an edge indicates the possibility of direct connection
- a missing edge represents a claim of zero relationship

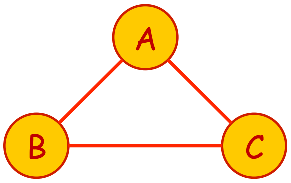
Marginal Dependencies



expression similarity tells us little about the underlying biological mechanism

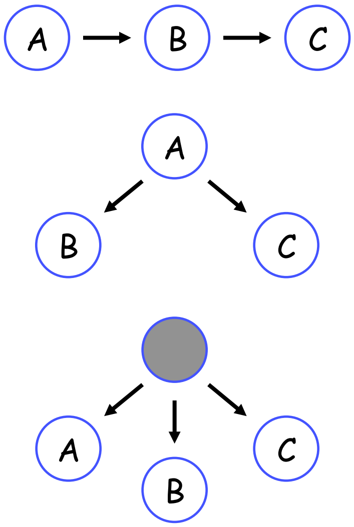
how can we distinguish between direct and indirect dependencies?

Gene Coexpression



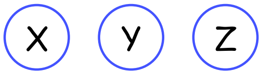
marginal dependencies

Gene Regulation



Conditional independencies

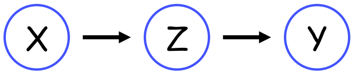
3 random variables:



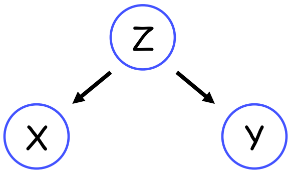
X is **conditionally independent** of Y given Z $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \mathrm{Z})$

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z)$$

knowing $\boldsymbol{Z}$, then $\boldsymbol{X}$ offers no more information about $\boldsymbol{Y}$



$(\boldsymbol{X} \perp \boldsymbol{Y} \mid \mathrm{Z})$



$(\boldsymbol{X} \perp \boldsymbol{Y} \mid \mathrm{Z})$

Gene coexpression networks

using conditional independence measures

$$(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})?$$

Is the expression of gene X **independent** of the expression of gene Y given the expression of Z?

beyond coexpression: try to recover regulatory relationships

$$\boldsymbol{X} \perp \boldsymbol{Y} \mid Z \quad \begin{cases} \boldsymbol{Z} = \emptyset & \textbf{coexpression networks} \\ \boldsymbol{Z} = & \text{single third variable} \\ \boldsymbol{Z} = & \text{all other variables except } \mathbf{X} \text{ and } \mathbf{Y} \end{cases}$$

Gene coexpression networks

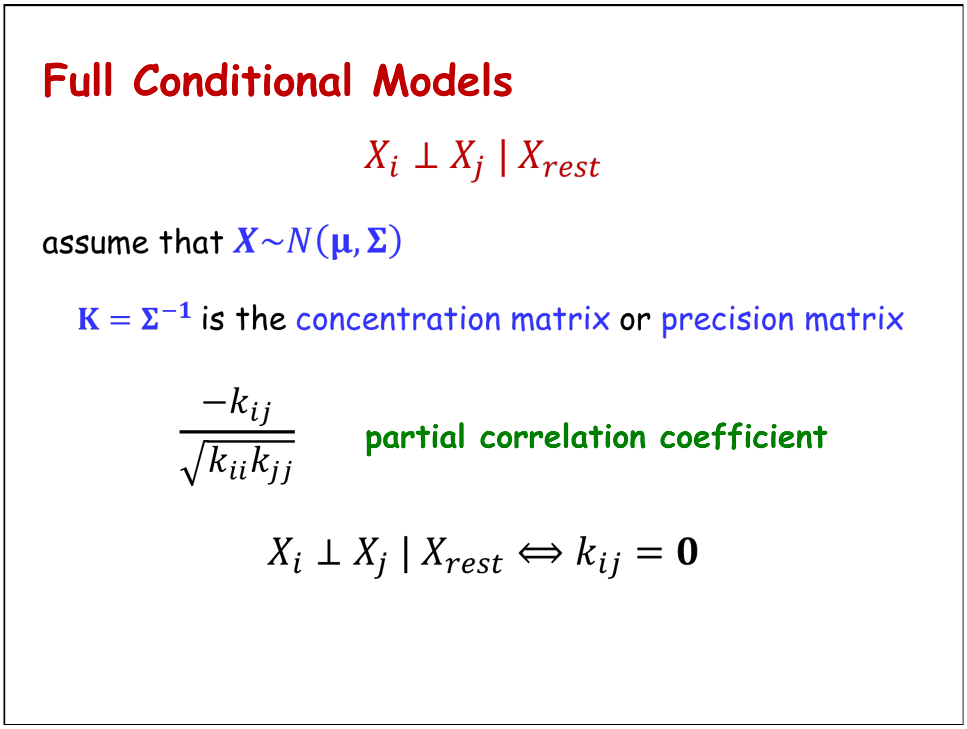
using conditional independence measures

$$X \perp Y \mid Z$$

$\boldsymbol{Z} =$ **all other variables except** $\boldsymbol{X}$ **and** $\boldsymbol{Y}$

Full Conditional Models (Markov Networks)

Can the correlation observed between X and Y be explained by **all the other genes** in the model?



Full Conditional Models

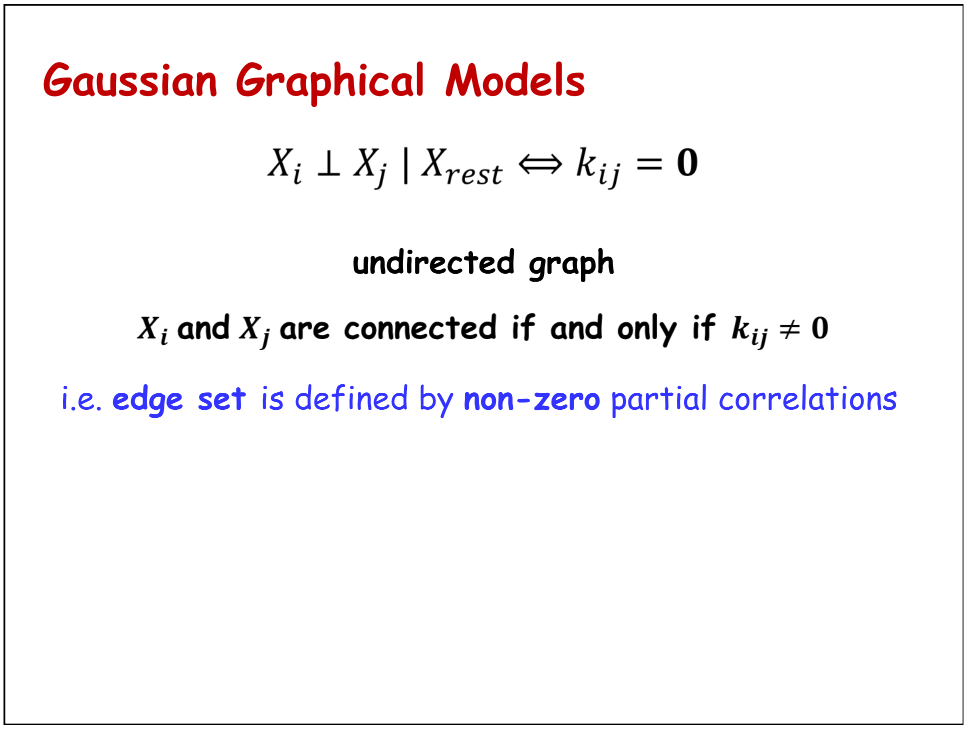
$$X_i \perp X_j \mid X_{rest}$$

assume that $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$\mathbf{K} = \boldsymbol{\Sigma}^{-1}$ is the concentration matrix or precision matrix

$$\frac{-k_{ij}}{\sqrt{k_{ii}k_{jj}}}$$ partial correlation coefficient

$$X_i \perp X_j \mid X_{rest} \Leftrightarrow k_{ij} = \mathbf{0}$$



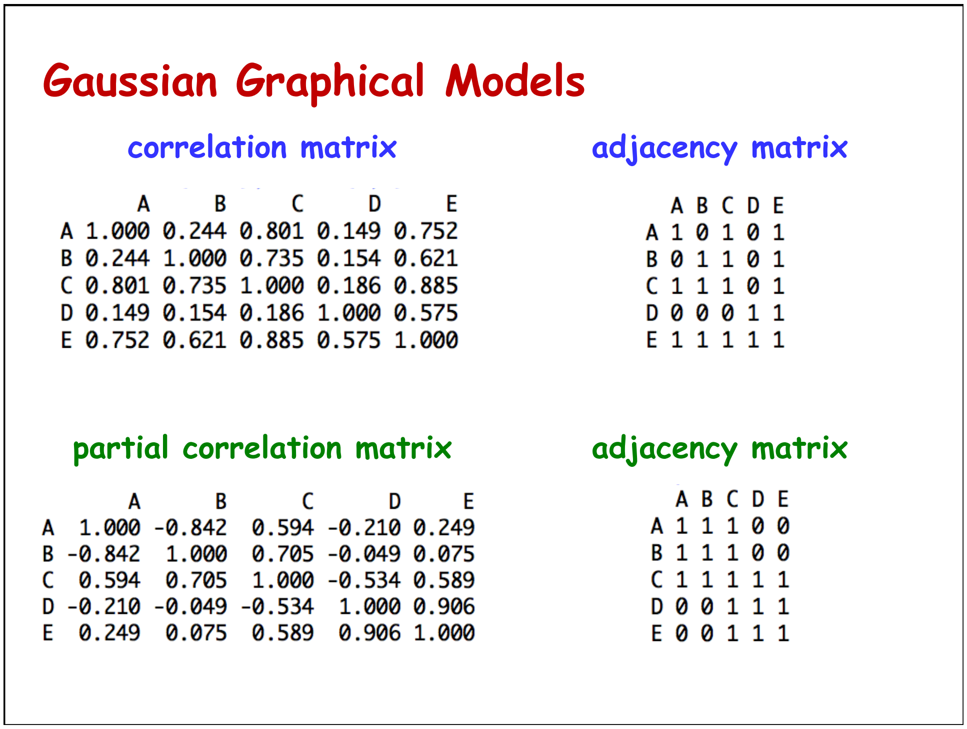
Gaussian Graphical Models

$$X_i \perp X_j \mid X_{rest} \Leftrightarrow k_{ij} = \mathbf{0}$$

undirected graph

X_i and X_j are connected if and only if $k_{ij} \neq 0$

i.e. **edge set** is defined by **non-zero** partial correlations



Gaussian Graphical Models

correlation matrix

```
      A     B     C     D     E
A 1.000 0.244 0.801 0.149 0.752
B 0.244 1.000 0.735 0.154 0.621
C 0.801 0.735 1.000 0.186 0.885
D 0.149 0.154 0.186 1.000 0.575
E 0.752 0.621 0.885 0.575 1.000
```

adjacency matrix

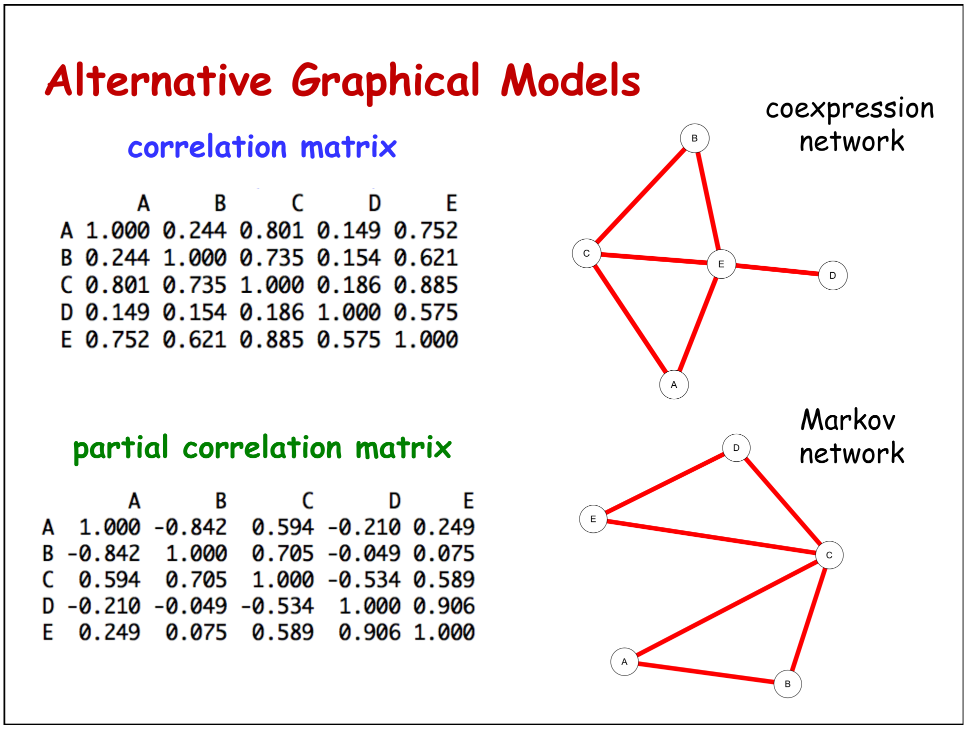
```
  A B C D E
A 1 0 1 0 1
B 0 1 1 0 1
C 1 1 1 0 1
D 0 0 0 1 1
E 1 1 1 1 1
```

partial correlation matrix

```
       A      B      C      D     E
A  1.000 -0.842  0.594 -0.210 0.249
B -0.842  1.000  0.705 -0.049 0.075
C  0.594  0.705  1.000 -0.534 0.589
D -0.210 -0.049 -0.534  1.000 0.906
E  0.249  0.075  0.589  0.906 1.000
```

adjacency matrix

```
  A B C D E
A 1 1 1 0 0
B 1 1 1 0 0
C 1 1 1 1 1
D 0 0 1 1 1
E 0 0 1 1 1
```



Alternative Graphical Models

correlation matrix

```
      A     B     C     D     E
A 1.000 0.244 0.801 0.149 0.752
B 0.244 1.000 0.735 0.154 0.621
C 0.801 0.735 1.000 0.186 0.885
D 0.149 0.154 0.186 1.000 0.575
E 0.752 0.621 0.885 0.575 1.000
```

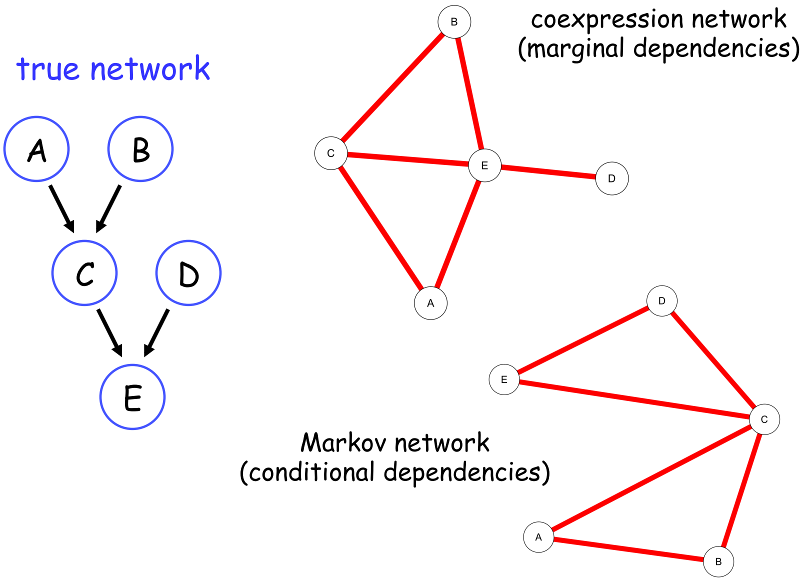
coexpression network

partial correlation matrix

```
       A      B      C      D     E
A  1.000 -0.842  0.594 -0.210 0.249
B -0.842  1.000  0.705 -0.049 0.075
C  0.594  0.705  1.000 -0.534 0.589
D -0.210 -0.049 -0.534  1.000 0.906
E  0.249  0.075  0.589  0.906 1.000
```

Markov network

Alternative Graphical Models



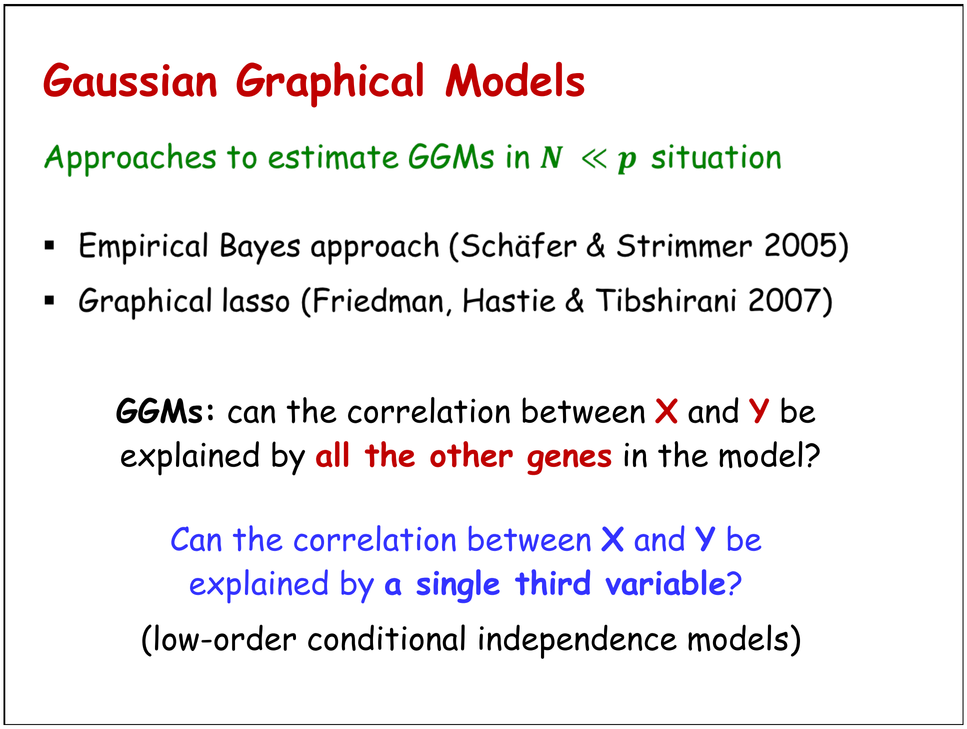
true network

coexpression network
(marginal dependencies)

Markov network
(conditional dependencies)

Gaussian Graphical Models

- full conditional relationships can only be accurately estimated if the $N \gg p$
- if $N \ll p$ then the correlation matrix does not have full rank and hence cannot be inverted
- $N \ll p$ is true for almost all genomic applications



Gaussian Graphical Models

Approaches to estimate GGMs in $N \ll p$ situation

- Empirical Bayes approach (Schäfer & Strimmer 2005)
- Graphical lasso (Friedman, Hastie & Tibshirani 2007)

GGMs: can the correlation between **X** and **Y** be explained by **all the other genes** in the model?

Can the correlation between **X** and **Y** be explained by **a single third variable**?

(low-order conditional independence models)