# Statistical methods for inferring the gene regulatory networks – Part I

Lecture 1 – May 14th, 2013
GENOME 541, Spring 2013

Su-In Lee
GS & CSE, UW
suinlee@uw.edu

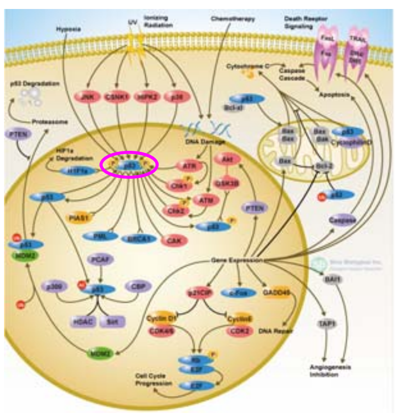


## Motivation: Why network?

- DNA, RNA, protein, and other biological molecules don't operate alone.
- Instead, they operate as part of complex *pathways* or *networks*.
- Inferring the networks from data can lead to a better understanding of disease process, evolutionary process, etc.



## Example: P53 pathway

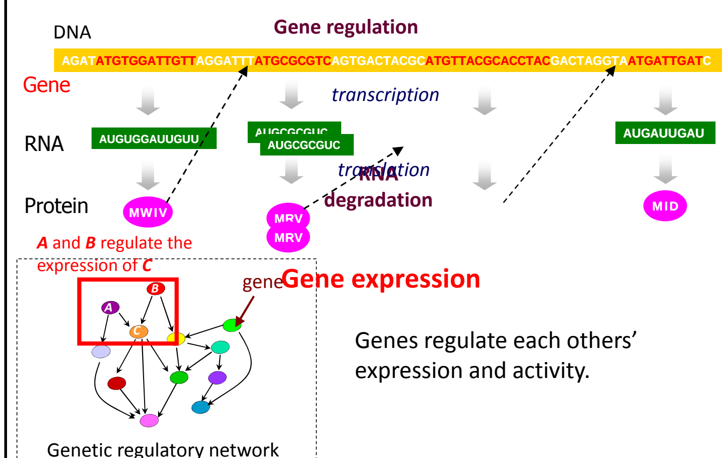


- P53
  - A transcription factor
  - A tumor suppressor protein
  - Regulates the expression of genes involved in apoptosis, inhibition of cell cycle progression and DNA repair.

This image is downloaded from Sino Biological Inc (www.sinobiological.com)

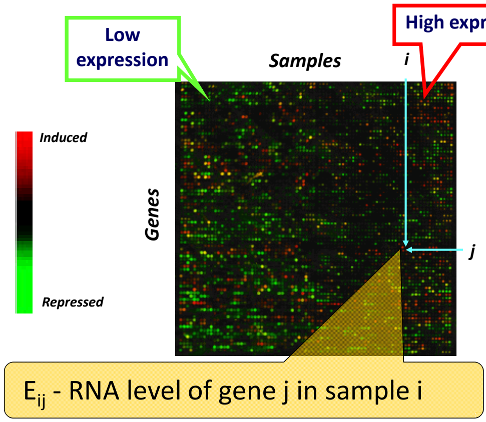


## Gene regulatory network




Genes regulate each others' expression and activity.

## We can estimate networks using observational gene expression data

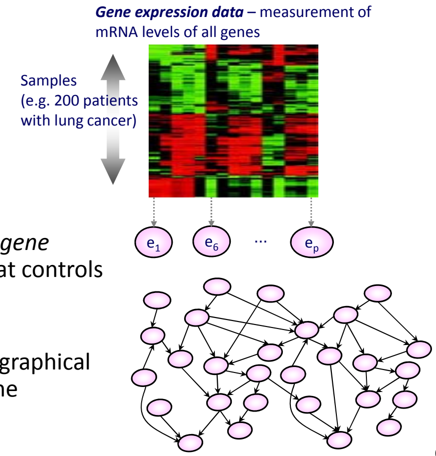



This matrix can be obtained from microarray or RNA-seq experiments

5

## Learning gene regulatory networks

- **Input:**

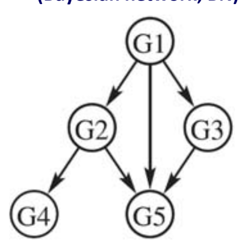



- **Goal:** Reconstruct the *gene regulatory network* that controls gene expression
- **Method:** Probabilistic graphical models to represent the regulatory network

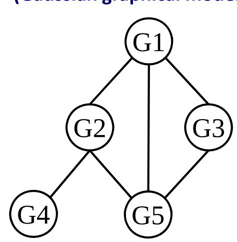
6

## Directed vs. undirected models

Directed graphical model (Bayesian network; BN)




Undirected graphical model (Gaussian graphical model)

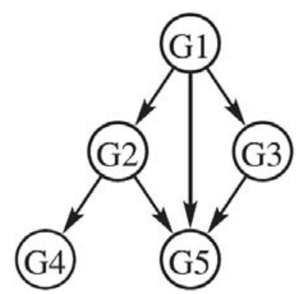



- Different conditional independence assumptions

7

## Directed graphical models (BNs)

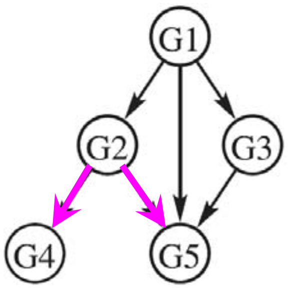
- Probability distribution for a gene expression level depends **only** on its parents (regulators) in the network




8

## Independence assumptions in BNs

- The expression levels of G4 and G5 are related only because they share a common regulator G2.
- In mathematical term, G4 and G5 are conditionally independent given G2.

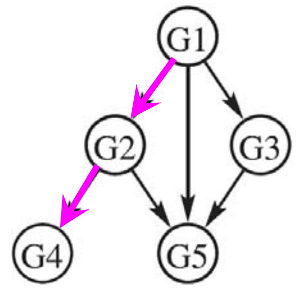



**G4 $\perp$ G5 | G2**

9

## Independence assumptions in BNs

- The expression levels of G4 and G1 are related only because of gene G2.




G4 $\perp$ G5 | G2

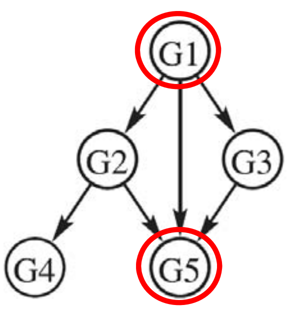
**G1 $\perp$ G4 | G2**

10

## Independence assumptions in BNs

- **Quiz:**
  - Would G5 independent of G1 given G3?

  (Would G1 and G5 are related only because of G3?)

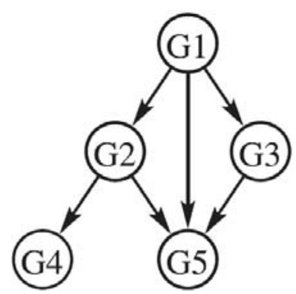



11

## Parameterization in BNs

- P(G1,G2,G3,G4,G5)

  = P(G1) P(G2|G1) P(G3|G1) P(G4|G2) P(G5|G1,G2,G3)




G4 $\perp$ G5 | G2

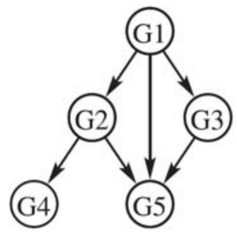
G1 $\perp$ G4 | G2

:

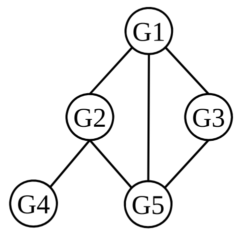
12

## Directed vs. undirected models

**Directed graphical model (Bayesian network; BN)**



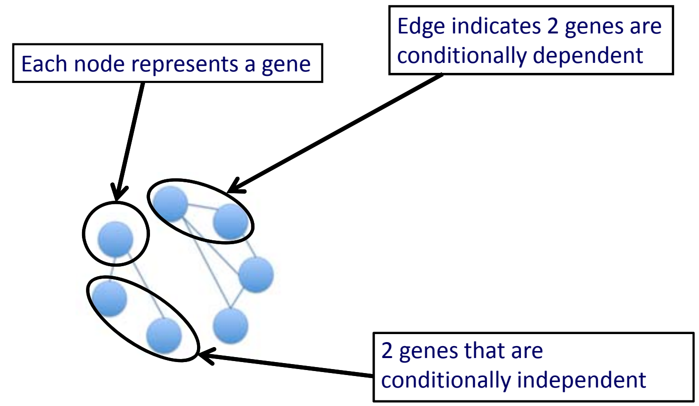

**Undirected graphical model (Gaussian graphical model)**




- Different conditional independence assumptions

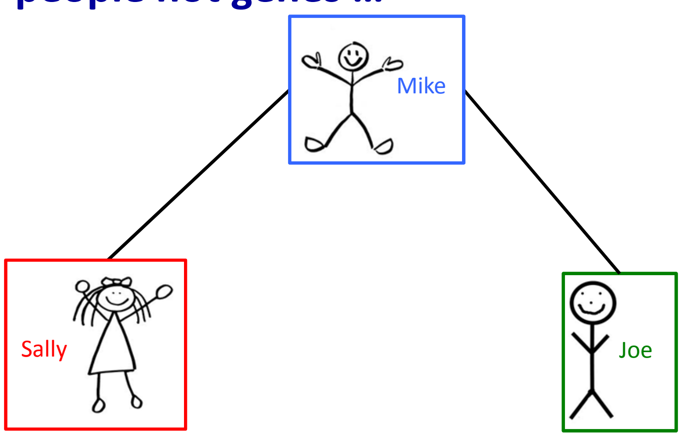
13

## In undirected graphical models …



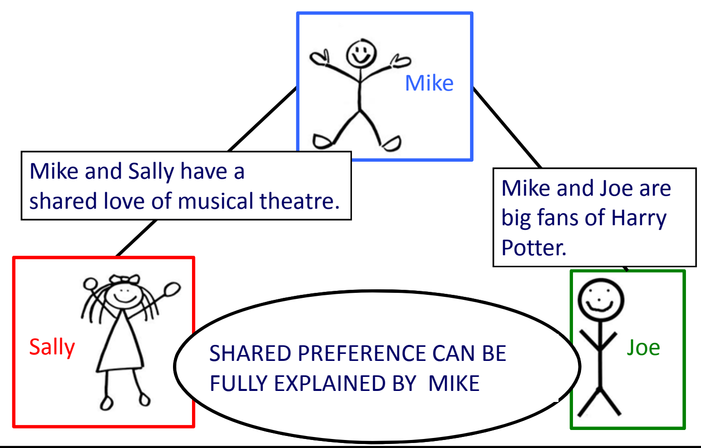

14

## An example: A network among people not genes …




15

## An example

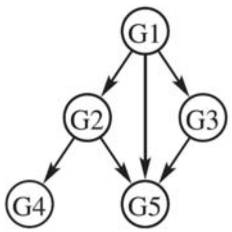



16

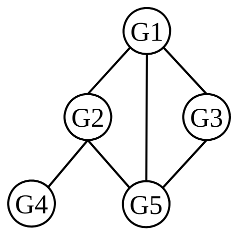
## Directed vs. undirected models

*Today*

**Directed graphical model (Bayesian network; BN)**




**Undirected graphical model (Gaussian graphical model)**




- Different conditional independence assumptions

17

## Outline (5/14, 5/16)

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms ........ *Today*
- Evaluation
- Recent probabilistic approaches to reconstructing the regulatory networks

18

## References

- A Primer on Learning in Bayesian Networks for Computational Biology
  - Chris Needhan et al. PLoS Computational Biology, 2007
- Probabilistic Graphical Models: Principles and Techniques
  - Daphne Koller and Nir Friedman, MIT Press 2009

19

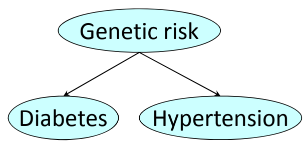
## Probability theory review

- Assume random variables $Val(A)=\{a^1,a^2,a^3\}$, $Val(B)=\{b^1,b^2\}$
  $P(A)$, $P(B)$
- Conditional probability
  - Definition $P(A|B) = \frac{P(A,B)}{P(B)}$
  - Chain rule $P(X_1, \ldots, X_n)$
    $= P(X_1)\, P(X_2|X_1) P(X_3|X_1,X_2) \ldots P(X_n|X_1,\ldots,X_{n-1})$
- Bayes' rule
  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Probabilistic independence
  $A \perp B$ if and only if
  $P(A|B) = P(A)$ $P(A,B) = P(A)\, P(B)$

20

# Bayesian network 101

- Directed acyclic graph
  - Node: a random variable
  - Edge: *direct* influence of one node on another
- The *Diabetes* example
  - Genetic risk (G), Diabetes (D), Hypertension (H)
  - Val (G) = $\{g^1, g^0\}$, Val (D) = $\{d^1, d^0\}$, Val (H) = $\{h^1, h^0\}$
  - P(G,D,H) = P(G) P(D|G) P(H|G)



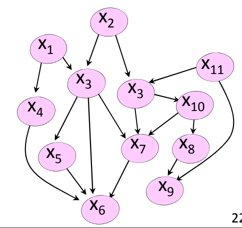

21

# Bayesian network semantics

- A Bayesian network structure *G* is a DAG whose nodes represent random variables $X_1, \ldots, X_p$.
  - $PaX_i$: parents of $X_i$ in G
  - $NonDesX_i$: variables in G that are not descendants of $X_i$.
- Local Markov assumptions
  - *G* encodes the following set of conditional independence assumptions:

For each variable $X_i$ ,

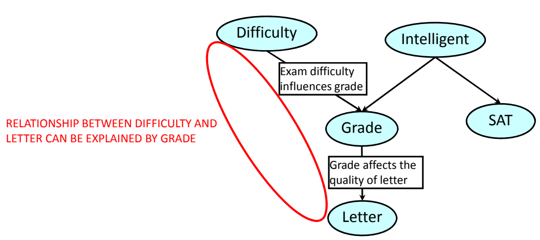
$$X_i \perp NonDesX_i \mid PaX_i$$




22

# The *Student* Example

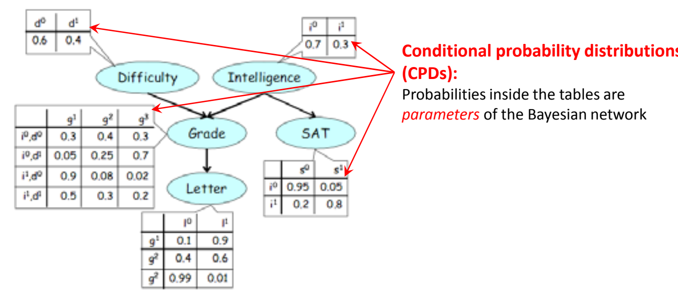
- Variables
  - Course difficulty (D), Val(D) = {easy, hard}
  - Quality of the rec. letter (L) , Val(L) = {strong, weak}
  - Intelligence (I), Val(I) = $\{i^1, i^0\}$
  - SAT (S) , Val (S) = $\{s^1, s^0\}$
  - Grade (G) , Val (G) = $\{g^1, g^2, g^3\}$
- Bayesian network G




23

# Parameters

- Relationship among variables can be described based on conditional probability distributions (CPDs) – $P(X_i \mid \text{Parents of } X_i)$




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0,d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0,d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1,d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1,d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

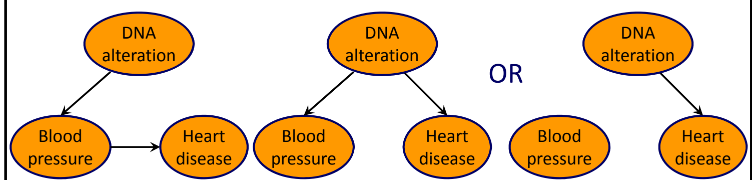
| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^2$ | 0.99 | 0.01 |

- $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \text{Parents of } X_i)$

24

## Model selection problem

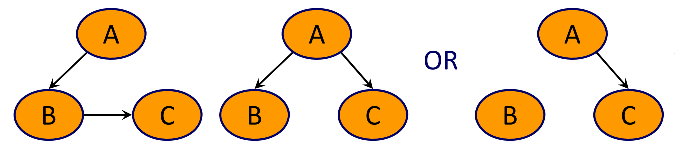
- How can we determine the Bayesian network of a certain set of variables?
- For example, how a change in a certain nucleotide in DNA (SNP), blood pressure and heart disease are related?
- There can be many possible "models"…



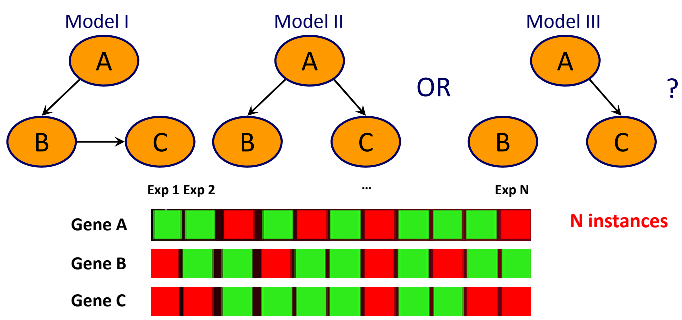

25

## Model selection – another example

- How genes A, B and C regulate each other's expression levels (mRNA levels) ?
- There can be many possible models…



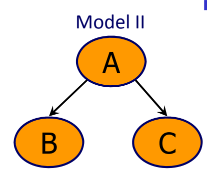

26



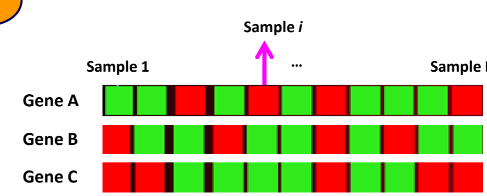

- Model selection
  - Select the model X that best explains the data
    $\text{argmax}_x$ P( Data | model X is true )
- How to compute P( Data | model X is true )

27

## Computing P(Data | model II is true)




- P(A,B,C | model II is true)= ?
  - P(A)P(B|A)P(C|A)
  - For sample i,
    P(A=high)P(B=low|A=high)P(C=low|A=high)




- **P(Data | model II is true) = $\prod_i$ P( [A,B,C] in sample *i* | model II )**

28

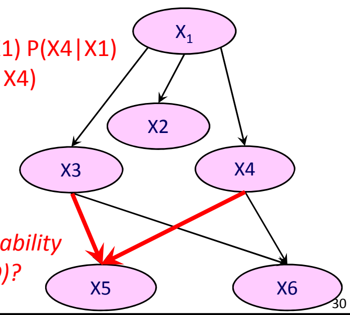
# Outline

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
- Evaluation
- Recent probabilistic approaches to reconstructing the regulatory networks

29

# Regulatory network

- Bayesian network representation
  - Xi: expression level of gene i
  - Val(Xi): continuous
- Joint distribution
  - P(**X**) = P(X1) P(X2|X1) P(X3|X1) P(X4|X1) P(X5|X3, X4) P(X6|X3, X4)
- Interpretation
  - Conditional independence




*Conditional probability distribution (CPD)?*

30

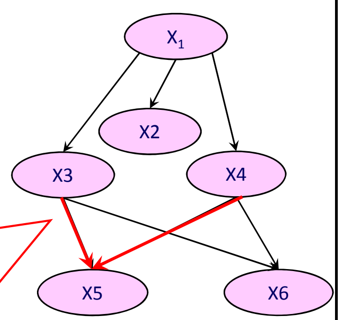
# CPD for discrete expression level

- After discretizing the expression levels to "high" and "low"…
  - Parameters – probability values in every entry

*Table CPD*

| | X5=high | X5=low |
|---|---|---|
| X3=high, X4=high | 0.3 | 0.7 |
| X3=high, X4=low | 0.95 | 0.05 |
| X3=low, X4=high | 0.1 | 0.9 |
| X3=low, X4=low | 0.2 | 0.8 |

***parameters***



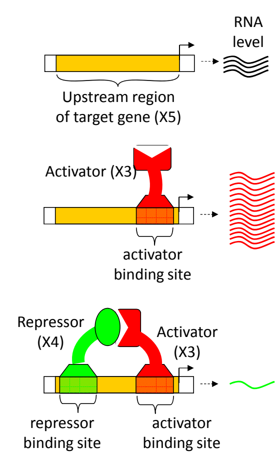

31

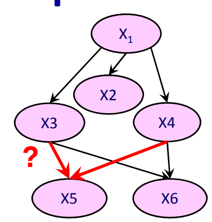
# Context specificity of gene expression

**Context A**
Basal expression level

**Context B**
Activator induces expression

**Context C**
Activator + repressor decrease expression



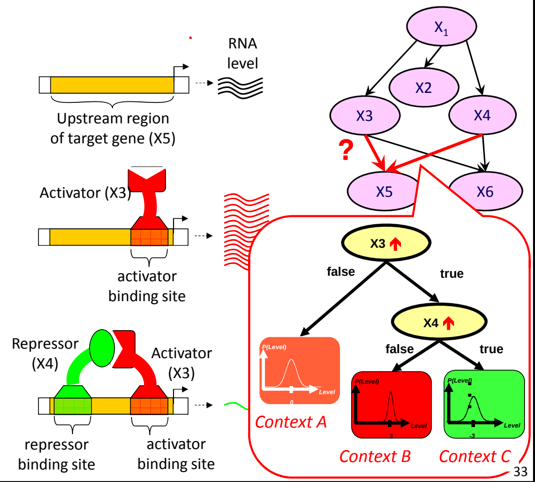

32

# Context specificity of gene expression

**Context A**
Basal expression level

**Context B**
Activator induces expression

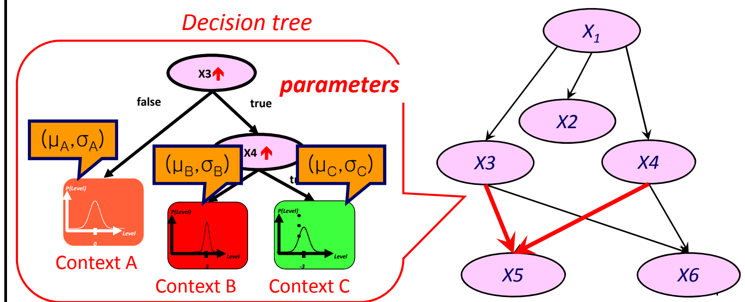
**Context C**
Activator + repressor decrease expression




33

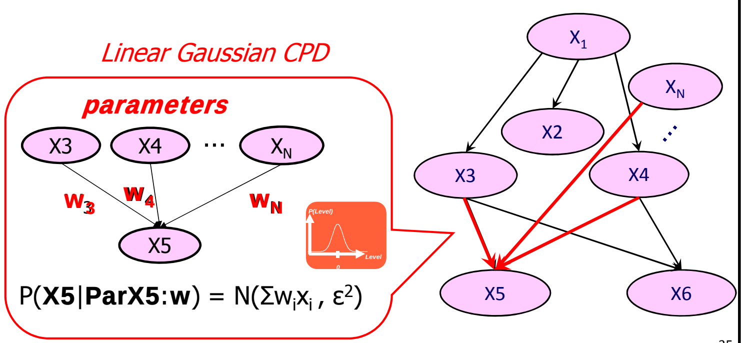
# Continuous-valued expression I

- Tree conditional probability distributions (CPD)
  - Parameters – mean (μ) & variance ($\sigma^2$) of the normal distribution in each context
  - Represents combinatorial and context-specific regulation




# Continuous-valued expression II

- Linear Gaussian CPD
  - Parameters – weights $w_1$,…,$w_N$ associated with the parents (regulators)



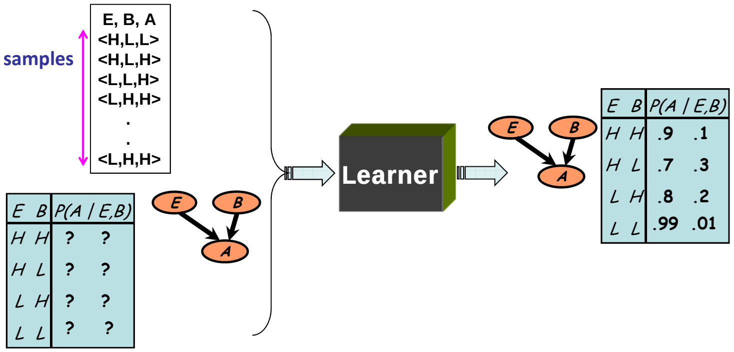

35

# Outline

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
  - Parameter learning
  - Structure learning
  - Structure discovery
- Evaluation
- Recent probabilistic approaches to reconstructing the regulatory networks

36

## Known structure, complete data

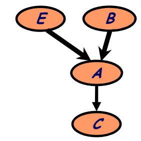



- Network structure is specified
  - Learner needs to estimate parameters
- Data does not contain missing values



## Learning parameters

- Training data has the form:




genes

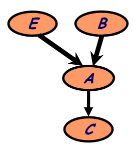
$$D = \begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

samples



## Likelihood function

- Assume i.i.d. samples
- Likelihood function is defined as:




$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

genes

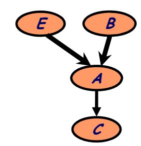
samples

$$\begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$



## Likelihood function

- Joint distribution can be decomposed as:




$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

$$= \prod_m \begin{pmatrix} P(E[m] : \Theta) \times \\ P(B[m] : \Theta) \times \\ P(A[m] \mid B[m], E[m] : \Theta) \times \\ P(C[m] \mid A[m] : \Theta) \end{pmatrix}$$

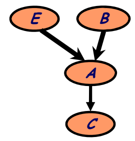
genes

samples

$$\begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

## Likelihood function

- Reordering terms, we got




$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

$$= \left( \begin{array}{l} \prod_m P(E[m] : \Theta_E) \times \\ \prod_m P(B[m] : \Theta_B) \times \\ \prod_m P(A[m] \mid B[m], E[m] : \Theta_{A|B,E}) \times \\ \prod_m P(C[m] \mid A[m] : \Theta_{C|A}) \end{array} \right)$$

$$\begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

- Parameters can be estimated for each variable independently!

41

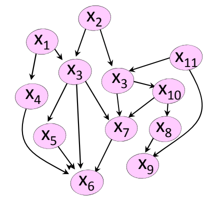
## General Bayesian networks

- Generalization for any Bayesian network:

$$L(\Theta : D) = \prod_m P(x_1[m], \ldots, x_n[m] : \Theta)$$

$$= \prod_m \prod_i P(x_i[m] \mid Pa_i[m] : \Theta_i)$$

$$= \prod_i L_i(\Theta_i : D)$$



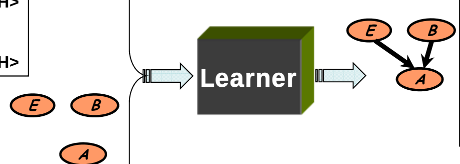

- Parameters can be estimated for each variable independently!

42

## Unknown structure, complete data

E, B, A
<H,L,L>
<H,L,H>
<L,L,H>
<L,H,H>
.
.
<L,H,H>

| E | B | P(A \| E,B) | |
|---|---|---|---|
| H | H | ? | ? |
| H | L | ? | ? |
| L | H | ? | ? |
| L | L | ? | ? |




| E | B | P(A \| E,B) | |
|---|---|---|---|
| H | H | .9 | .1 |
| H | L | .7 | .3 |
| L | H | .8 | .2 |
| L | L | .99 | .01 |

- Network structure is not specified
  - Learner needs to estimate both structure and parameters
- Data does not contain missing values

43