Statistical methods for inferring the gene regulatory networks – Part I

Lecture 1 – June 3rd, 2014
GENOME 541, Spring 2014

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

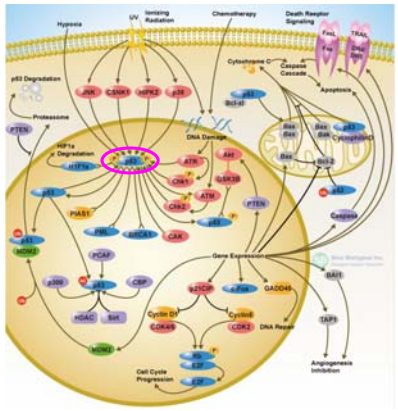
1

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.

2

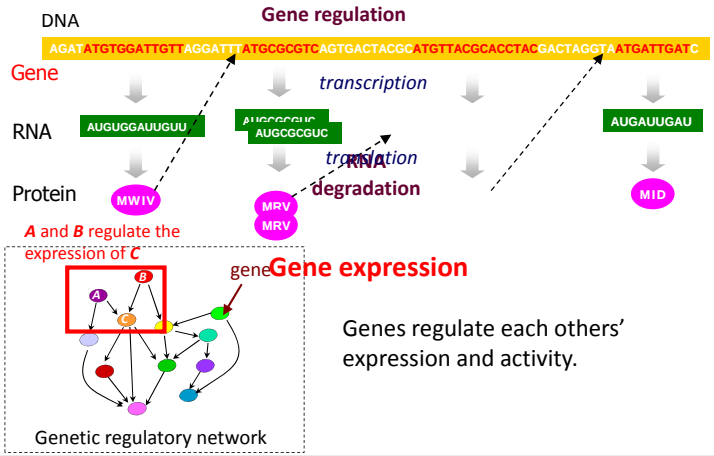
Example: P53 pathway



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

3

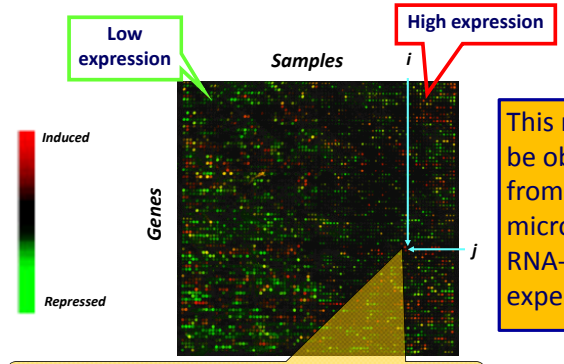
Gene regulatory network



Genes regulate each others' expression and activity.

4

We can estimate networks using observational gene expression data



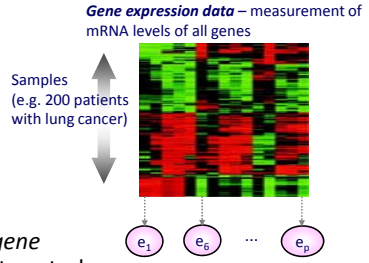
This matrix can be obtained from microarray or RNA-seq experiments

E_{ij} - RNA level of gene j in sample i

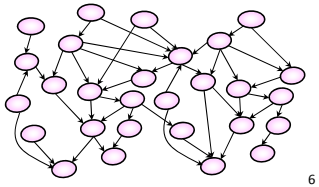
5

Learning gene regulatory networks

- **Input:**



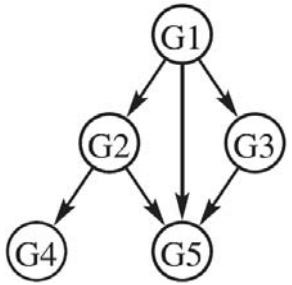
- **Goal:** Reconstruct the *gene regulatory network* that controls gene expression
- **Method:** Probabilistic graphical models to represent the regulatory network



6

Directed graphical models (BNs)

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

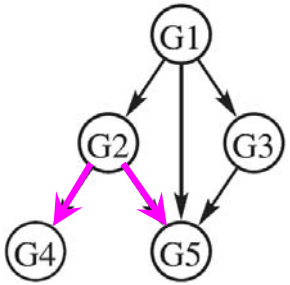


7

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.

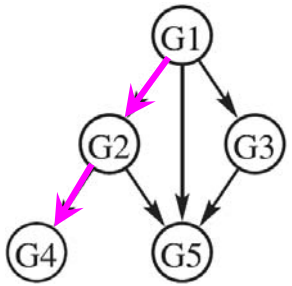
$G4 \perp G5 \mid G2$



8

Independence assumptions in BNs

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



$G4 \perp G5 \mid G2$

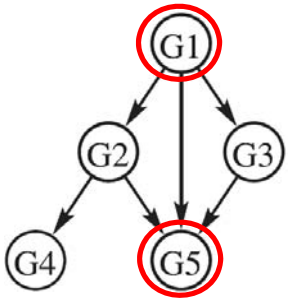
$G1 \perp G4 \mid G2$

9

Independence assumptions in BNs

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

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



10

Outline (6/3, 6/5)

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

11

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

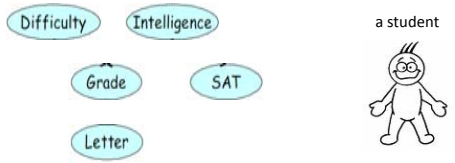
12

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
    $P(X_1, \dots, X_n)$
    $= P(X_1)\, P(X_2|X_1) P(X_3|X_1,X_2) \dots P(X_n|X_1,\dots,X_{n-1})$
- Bayes' rule
  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Probabilistic independence
  $A \perp B$ if and only if
  $P(A|B) = P(A)$ $P(A,B) = P(A)\, P(B)$

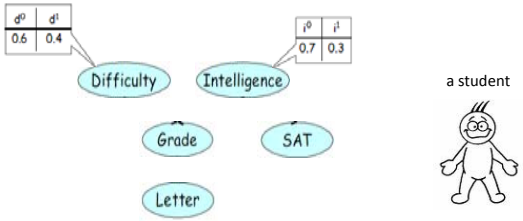
The *Student* example

- Consider the following 5 variables
  - Grade (G) Val(G) = {g^1, g^2, g^3}
  - Course difficulty (D) Val(D) = {d^1, d^0}
  - Intelligence (I) Val(I) = {i^1, i^0}
  - SAT (S) Val(S) = {s^1, s^0}
  - Quality of recommendation letter (L) Val(L) = {l^1, l^0}



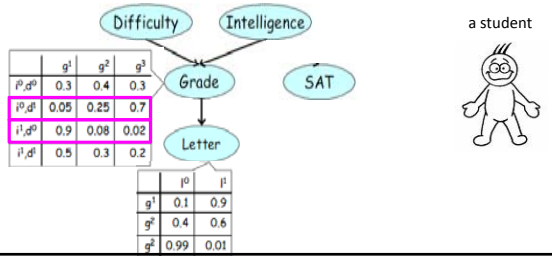
The *Student* example

- Probability distributions
  - P(D): P(D= d^0), P(D= d^1) P(I): P(I= i^0), P(I= i^1)
  - P(G): P(G= g^1), P(G= g^2), P(G= g^3) …
- Are these variables related to each other?
  - Any examples?



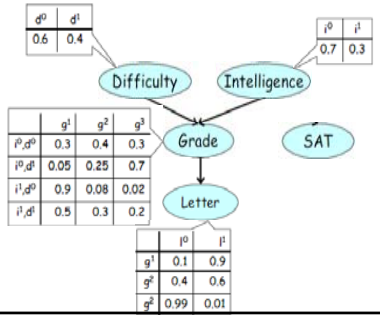
Conditional probability distributions

- Conditional probability distributions can describe the dependencies among the variables
  - Probability distribution over G depends on the values on D and I
  - Conditional distribution over G given D and I, P(G|D, I)
  - Conditional distribution over L given G, P(L|G)
    :



Statistical independence

- Some of the variables are independent to each other
  - Distribution over D does not depend on I
  - Distribution over S does not dependent on D
- *Some are conditionally independent*
  - Given G, the distribution over L does not depend on D or I



17

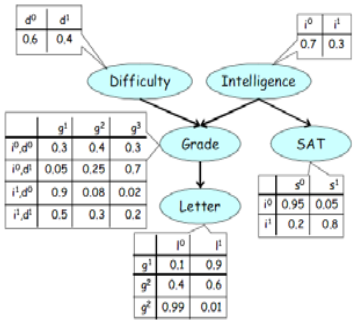
Statistical Independence

- If variables A and B are statistically independent,
  - P(A) = P(A | B)
    = P(A, B) / P(B)
  - P(A, B) = P(A) P(B)
- *Conditional* statistical independence
  - P(Letter | Grade) = P(Letter | Grade, Intelligence)
    = P(Letter | Grade, Difficulty)
    = P(Letter | Grade, Intelligence, Difficulty)
  - Given the Grade, quality of Letter does not depend on Intelligence or Difficulty

18

The *Student* network

- A network representation of the conditional independence relationships among variables



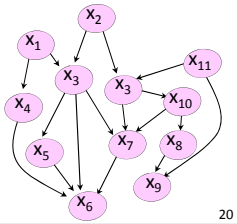
19

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 ,

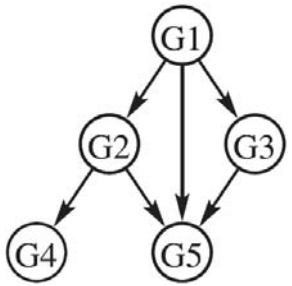
$$X_i \perp NonDesX_i \mid PaX_i$$



20

Joint probability distribution

- By the chain rule,
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) P(G3|G1, G2) P(G4|G2, G1, G3) P(G5|G1,G2,G3, G4)
- From the conditional independence relationships
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) P(G3|G1) P(G4|G2) P(G5|G1,G2,G3)



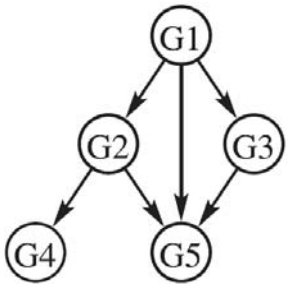
$G2 \perp G3 \mid G1$

$G1 \perp G4 \mid G2$

:

Joint probability distribution

- By the chain rule,
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) **P(G3|G1, G2)** P(G4|G2, G1, G3) P(G5|G1,G2,G3, G4)
- From the conditional independence relationships
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) **P(G3|G1)** P(G4|G2) P(G5|G1,G2,G3)



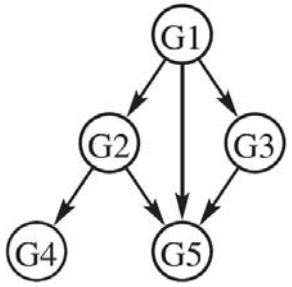
$G2 \perp G3 \mid G1$

$G1 \perp G4 \mid G2$

:

Joint probability distribution

- By the chain rule,
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) P(G3|G1, G2) **P(G4|G2, G1, G3)** P(G5|G1,G2,G3, G4)
- From the conditional independence relationships
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) P(G3|G1) **P(G4|G2)** P(G5|G1,G2,G3)



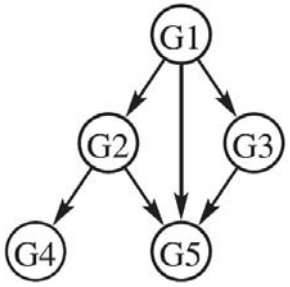
$G2 \perp G3 \mid G1$

$G1 \perp G4 \mid G2$

:

Joint probability distribution

- By the chain rule,
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) P(G3|G1, G2) P(G4|G2, G1, G3) **P(G5|G1,G2,G3, G4)**
- From the conditional independence relationships
  - P(G1,G2,G3,G4,G5) = P(G1) P(G2|G1) P(G3|G1) P(G4|G2) **P(G5|G1,G2,G3)**



$G2 \perp G3 \mid G1$

$G1 \perp G4 \mid G2$

:

JPD in Bayesian networks

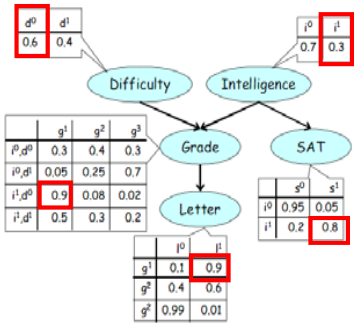
- The JPD is expressed in terms of a product of CPDs, describing each variable in terms of its parents, i.e., those variables it depends upon.

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}(x_i), \theta_i)$$

- where $\mathbf{x} = \{x_1, \ldots, x_n\}$ are the variables (nodes in the BN) and $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ denotes the model parameters, θ_i is the set of parameters describing the distribution of x_i and $\mathbf{pa}(x_i)$ denotes the parents of x_i.

Revisiting the *Student* example

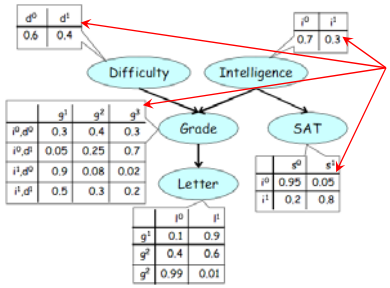
- Joint probability distribution
  - P(D,I,G,L,S) = P(D) P(I) P(G|D,I) P(S|I) P(L|G)



- What is the probability of observing {easy, intelligent, good, strong, high} ?
- P(D=easy) P(I=intelligent) P(G=good | D=easy, I=intelligent) P(S=strong | I=intelligent) P(L=strong | G=good)
  = 0.6 x 0.3 x 0.9 x 0.9 x 0.8
  = 0.1166

Parameters

- Relationship among variables can be described based on conditional probability distributions (CPDs) – P(X_i|Parents of X_i)



Conditional probability distributions (CPDs):
Probabilities inside the tables are *parameters* of the Bayesian network

- **P(X_1, …, X_n) = $\prod_i$ P(X_i|Parents of X_i)**

27

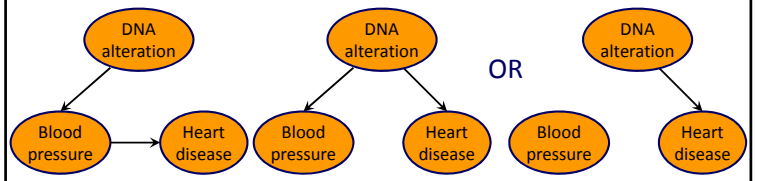
Outline

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

28

Model selection problem

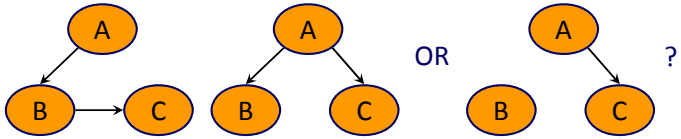
- How can we determine the Bayesian network of a certain set of variables?
- For example, how a change in a certain nucleotide in DNA (SNP), blood pressure and heart disease are related?
- There can be many possible "models"…



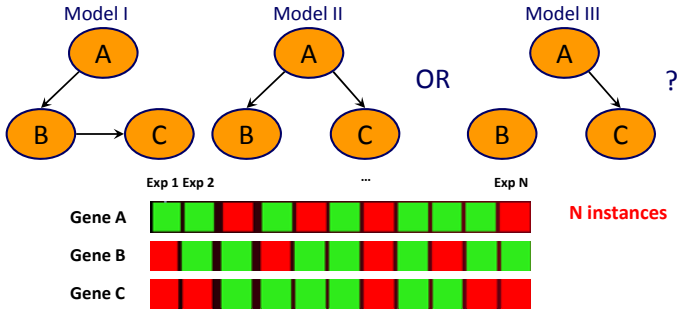
29

Model selection – another example

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



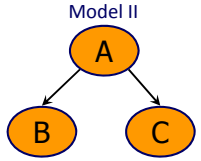
30



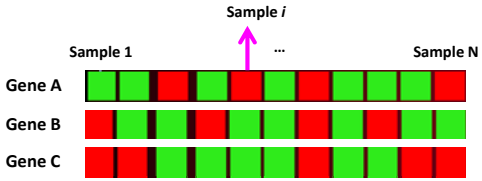
- Model selection
  - Select the model X that best explains the data
    argmax_x P(Data | model X is true)
- How to compute P(Data | model X is true)

31

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)**

32

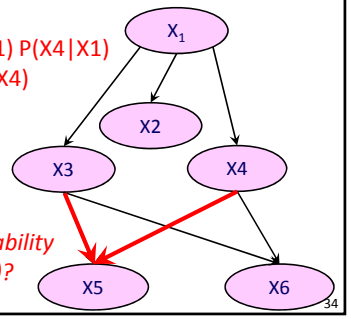
Outline

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

33

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)?

34

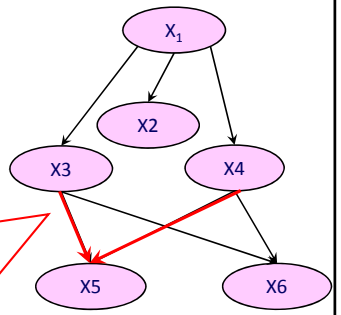
CPD for discrete expression level

- 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

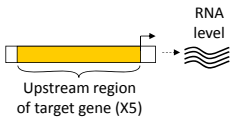
parameters



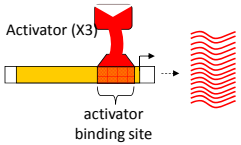
35

Context specificity of gene expression

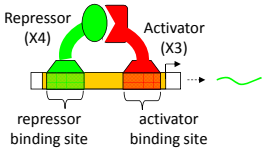
Context A
Basal expression level

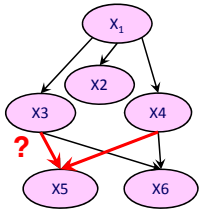


Context B
Activator induces expression



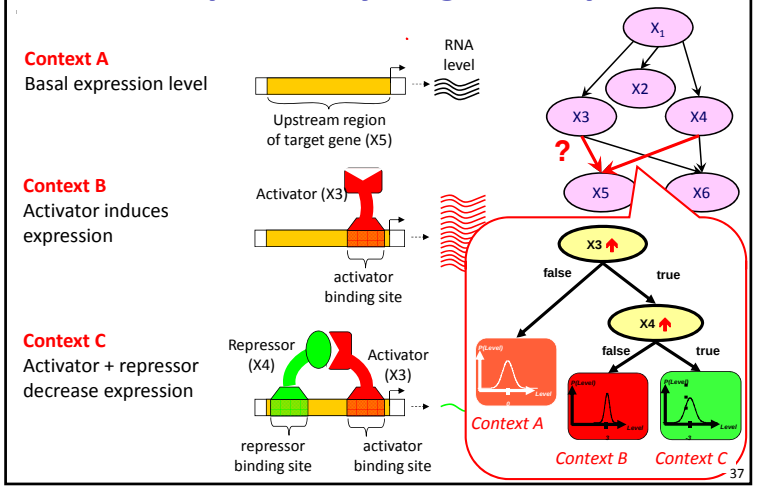
Context C
Activator + repressor decrease expression





36

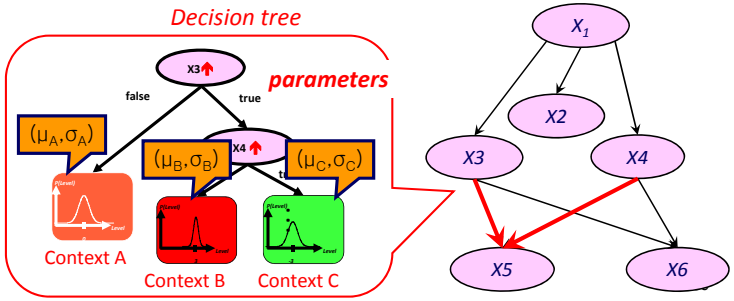
Context specificity of gene expression



37

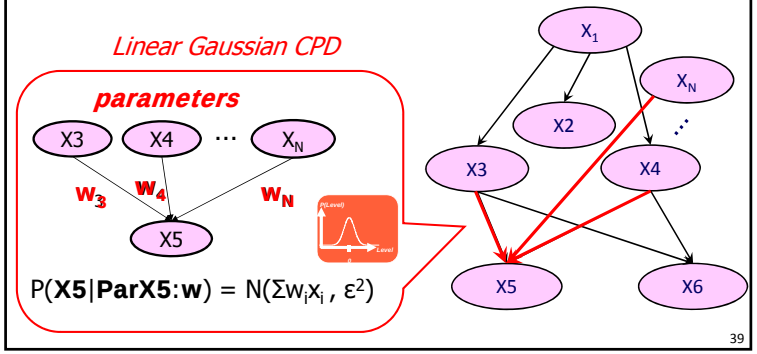
Continuous-valued expression I

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



Continuous-valued expression II

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



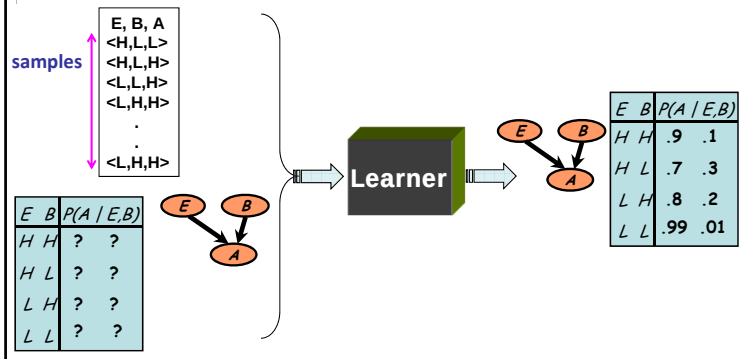
39

Outline

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

40

Known structure, complete data

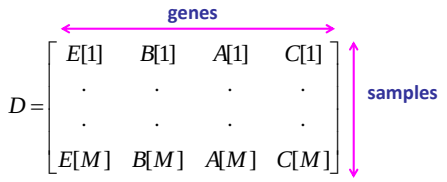


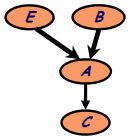
- Network structure is specified
  - Learner needs to estimate parameters
- Data does not contain missing values

41

Learning parameters

- Training data has the form:



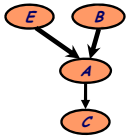


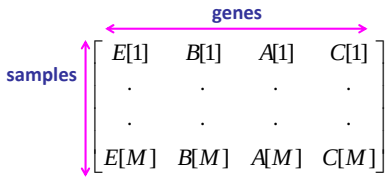
42

Likelihood function

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

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

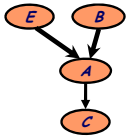




43

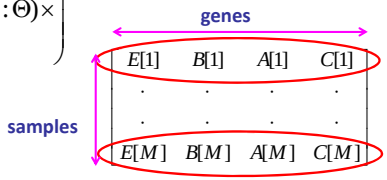
Likelihood function

- Joint distribution can be decomposed as:



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

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



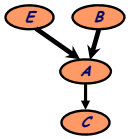
44

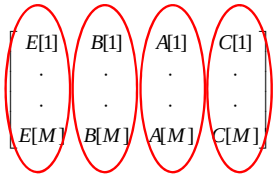
Likelihood function

- Reordering terms, we got

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

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





- Parameters can be estimated for each variable independently!

45

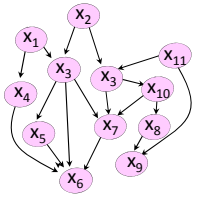
General Bayesian networks

- Generalization for any Bayesian network:

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

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

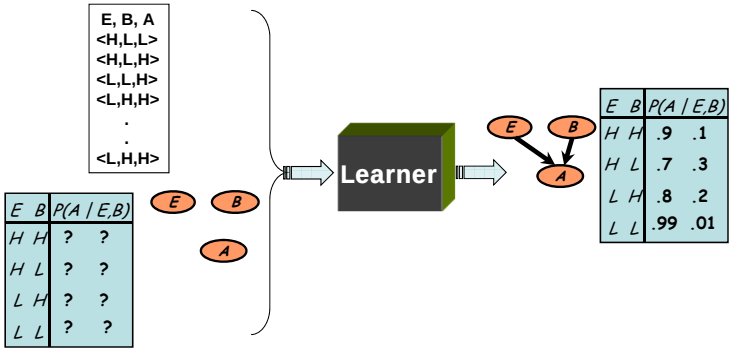
$$= \prod_i L_i(\Theta_i : D)$$



- Parameters can be estimated for each variable independently!

46

Unknown structure, complete data

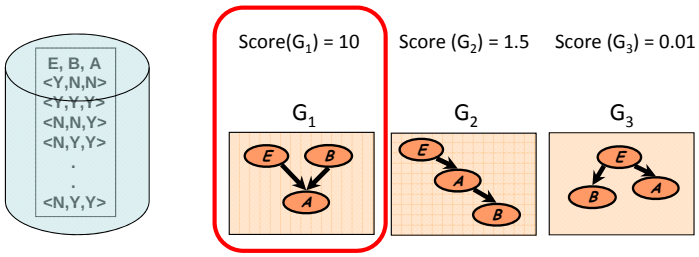


- Network structure is not specified
  - Learner needs to estimate both structure and parameters
- Data does not contain missing values

47

Score-based learning

- Define scoring function that measures how well a certain structure fits the observed data.



- Search for a structure that maximizes the score.

48