GENOME 560 Part II
Probabilistic Modeling & Learning:
Introduction to Probabilistic Models

GENOME 560

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

Why Take This Course?

- *Data* are interesting because they help us understand the world
- *Genomics*: massive amounts of data …
- Statistics is fundamental in genomics because it is integral in the **design, analysis** and **interpretation** of experiments
- This course covers the key statistical concepts and methods necessary for extracting biological insights from experimental data

Learning Goals

- 10 weeks is too short to cover all of statistics or even every specific topic that might arise in the course of your research…

- Statistical and computational methods should never be treated as "recipes" to follow!

- Instead, we should focus on
  - rigorous understanding of fundamental concepts that will provide you with the tools necessary to address routine statistical analyses
  - foundation to understand and learn more specific topics

Course Schedule

- **Syllabus:**

Date	Topic
Week 1	Introduction to probability, random variables and probability distributions, descriptive statistics, joint and conditional probability
Week 2	More probability distributions, introduction to hypothesis testing
Week 3	Parametric hypothesis testing; comparing means, comparing proportions
Week 4	Non-parametric hypothesis testing; comparing means, comparing proportions; rank-based tests; permutation testing
Week 5	More on permutation testing; resampling methods; sample size calculations

We are here.

Course Schedule

- **Syllabus:**

Date	Topic
Week 1	Introduction to probability, random variables and probability distributions, descriptive statistics, joint and conditional probability
Week 2	More probability distributions, introduction to hypothesis testing
Week 3	Parametric hypothesis testing; comparing means, comparing proportions
Week 4	Non-parametric hypothesis testing; comparing means, comparing proportions; rank-based tests; permutation testing
Week 5	More on permutation testing; resampling methods; sample size calculations

Trees

Part I covers many important trees. It’s time to shift from focusing on tress to understanding forest.

Forest

Course Schedule

- **Syllabus:**

Date	Topic
Week 1	Introduction to probability, random variables and probability distributions, descriptive statistics, joint and conditional probability
Week 2	More probability distributions, introduction to hypothesis testing
Week 3	Parametric hypothesis testing; comparing means, comparing proportions
Week 4	Non-parametric hypothesis testing; comparing means, comparing proportions; rank-based tests; permutation testing
Week 5	More on permutation testing; resampling methods; sample size calculations

Probabilistic Modeling & Learning

What is Probabilistic Model?

- A compact representation of the world
  - A set of random variables A, B, C, …
  - Probabilistic distribution over the variables P(A, B, C, …) – relationship among variables

- We can use probabilistic models to understand better about the world – e.g., *relationships* among variables

- Questions
  - Given partial data that measure the world, can infer a probabilistic model? *Learning*
  - Is the inferred model correct? How sure are we? *Model selection*

Course Schedule

- **Syllabus:**

Date	Topic
Week 1	Introduction to probability, random variables and probability distributions, descriptive statistics, joint and conditional probability
Week 2	More probability distributions, introduction to hypothesis testing
Week 3	Parametric hypothesis testing; comparing means, comparing proportions
Week 4	Non-parametric hypothesis testing; comparing means, comparing proportions; rank-based tests; permutation testing
Week 5	More on permutation testing; resampling methods; sample size calculations
Week 6	Basics of Bayesian networks; parameter estimation
Week 7	Maximum likelihood estimation (MLE), Bayesian estimation
Week 8	Linear regression, High-dimensionality, feature selection Cross-validation, model selection
Week 9	Single factor ANOVA, two-way ANOVA
Week 10	Multiple hypothesis testing

Representation

Learning

Model selection

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
  - Chapters 2.1-2.3, and 3.1-3.3

Outline

- Probability theory review
- Probabilistic models in genomics
- Bayesian networks representation
- No R exercise today

Probability Theory Review I

- Assume *random variables* A and B
  - A: Grade
  - B: Difficulty of course

- Values of A and B
  - $Val(A)=\{a^1,a^2,a^3\}$
  - $Val(B)=\{b^1,b^2\}$

- Probability distributions of A and B, P(A) and P(B)
  - P(A) consists of three probabilities: $P(A=a^1)$, $P(A=a^2)$, $P(A=a^3)$
  - P(B) consists of two probabilities: $P(B=b^1)$, $P(B=b^2)$

- Joint probability distribution P(A, B)
  - P(A, B) consists of six probabilities: $P(A=a^1, B=b^1)$, $P(A=a^1, B=b^2)$, $P(A=a^2, B=b^1)$, $P(A=a^2, B=b^2)$, $P(A=a^3, B=b^1)$, $P(A=a^3, B=b^2)$

Probability Theory Review II

- Assume random variables $Val(A)=\{a^1,a^2,a^3\}$, $Val(B)=\{b^1,b^2\}$

  $P(A), P(B)$

- Conditional probability
  - Definition $P(A|B) = \dfrac{P(A,B)}{P(B)}$

    P(A|B) consists of 6 probabilities:
    $P(A=a^1,B=b^1)$ / $P(B=b^1)$,
    $P(A=a^1,B=b^2)$ / $P(B=b^2)$, …
  - 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*
  $P(A|B) = P(A) \quad P(A,B) = P(A)\,P(B)$

Example: Probabilistic Independence

- Probabilistic independence

  $A \perp B$ *if and only if*

  $P(A|B) = P(A)$ $\quad$ $P(A,B) = P(A)\, P(B)$

- Assume *random variables* A and B
  - A: Grade
  - B: Difficulty of course

P(A,B):

P(A=A-grade, B=Difficult) < P(A=A-grade) P(B=Difficult)

P(A=C-grade, B=Difficult) > P(A=C-grade) P(B=Difficult)

- Assume *random variables* A and B
  - A: Grade
  - B: Weather in Seattle

P(A,B):

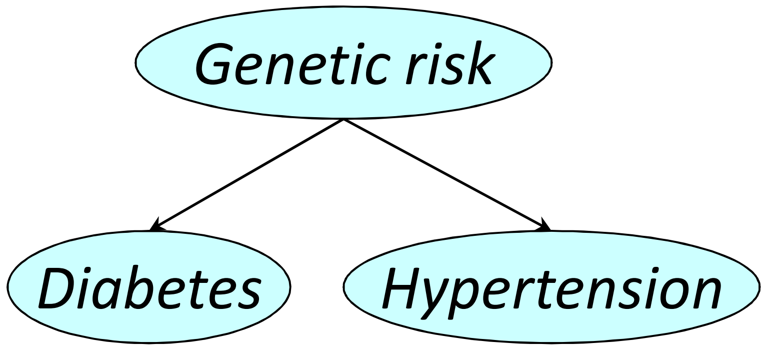
P(A=A-grade, B=Cloudy) ?

P(A=C-grade, B=Cloudy) ?

Bayesian Network 101

- Directed acyclic graph
  - Node: a random variable
  - Edge: *probabilistic dependence* of one node on another

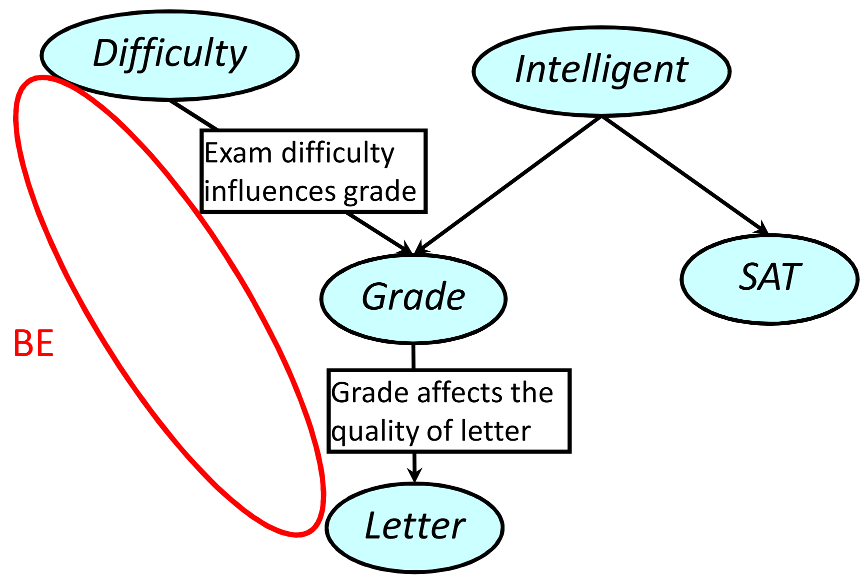
- The *Diabetes* example
  - Genetic risk (G), Diabetes (D), Hypertension (H)
  - Val (G) = $\{g^1, g^0\}$, Val (D) = $\{d^1, d^0\}$, Val (H) = $\{h^1, h^0\}$
  - P(G,D,H) = $P(G)\,P(D|G)\,P(H|D,G) = P(G)\,P(D|G)\,P(H|G)$



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



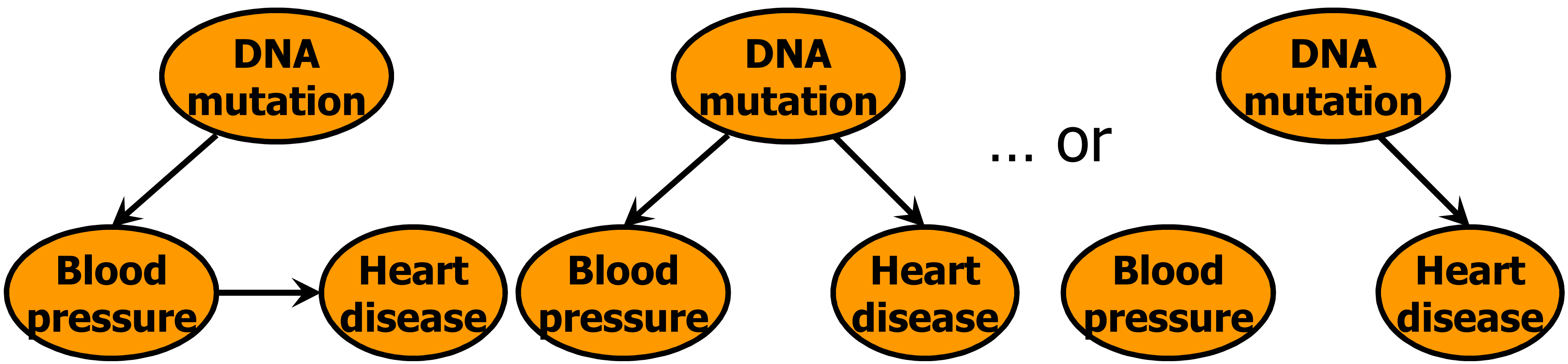
RELATIONSHIP BETWEEN DIFFICULTY AND LETTER CAN BE EXPLAINED BY GRADE

Outline

- Probability theory review
- Probabilistic models in genomics
- Bayesian networks representation
- Parameter estimation

Example 1

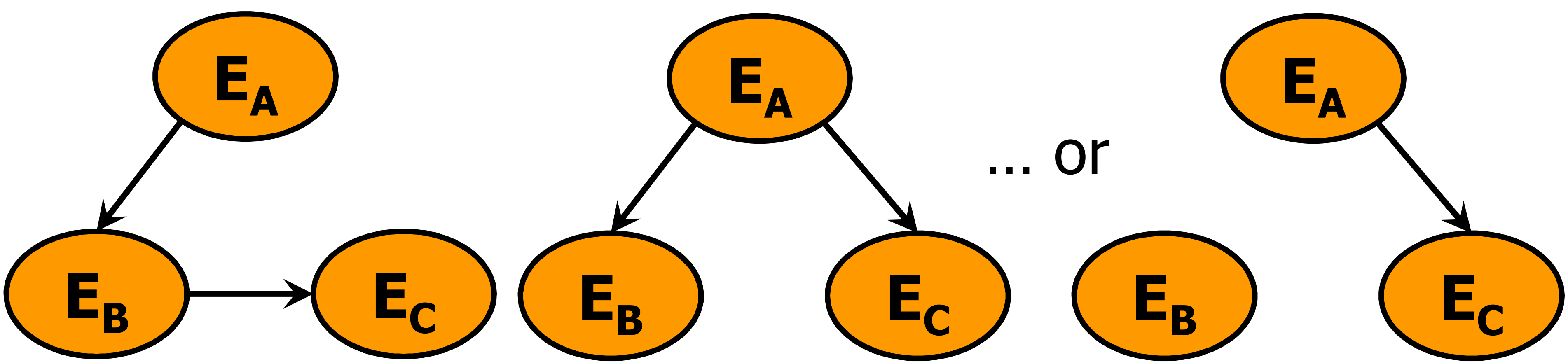
- How a certain DNA mutation, blood pressure and heart disease are related?
- There can be several “models”…



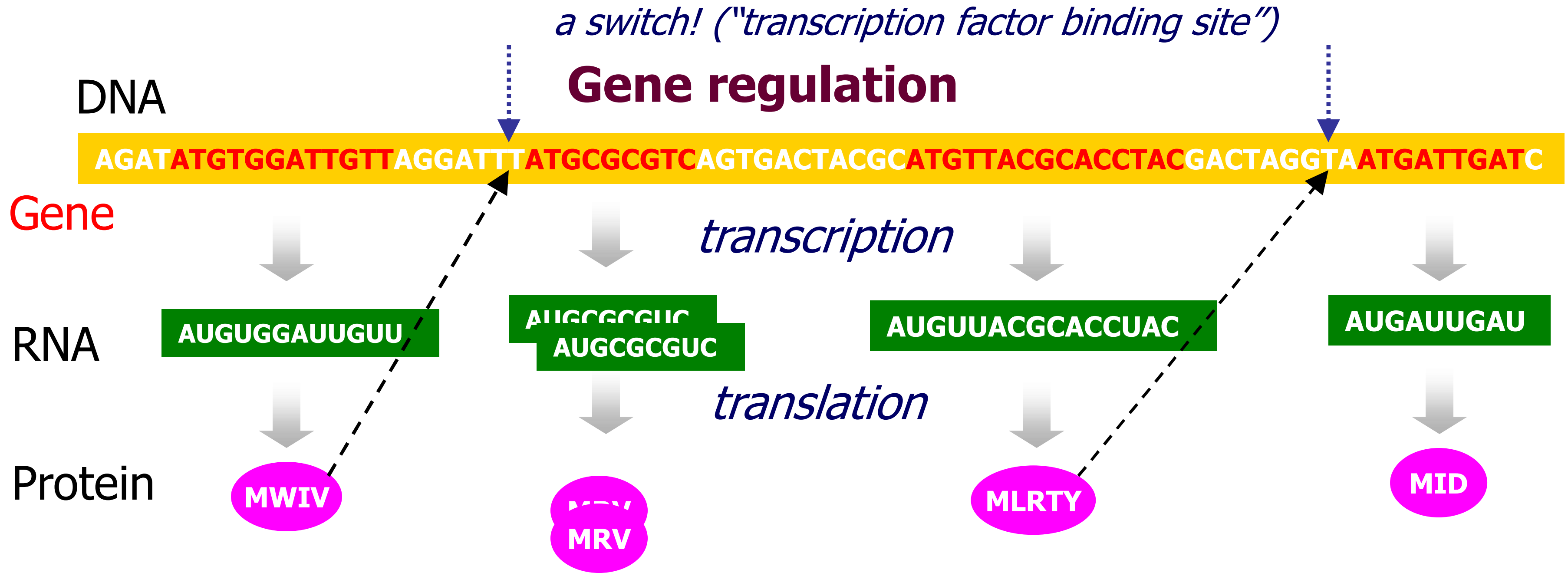
Example 2

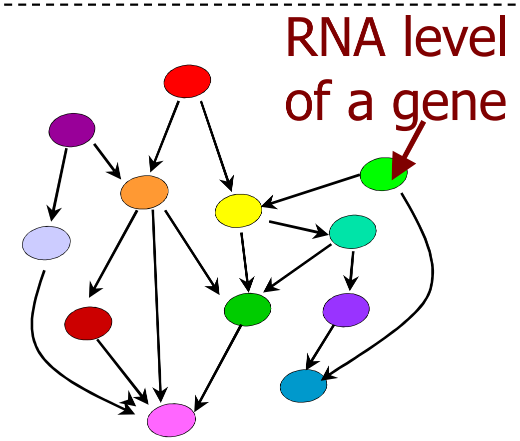
- How genes A, B and C regulate each others' expression levels (mRNA levels) ?

- There can be several models …



Gene regulatory network





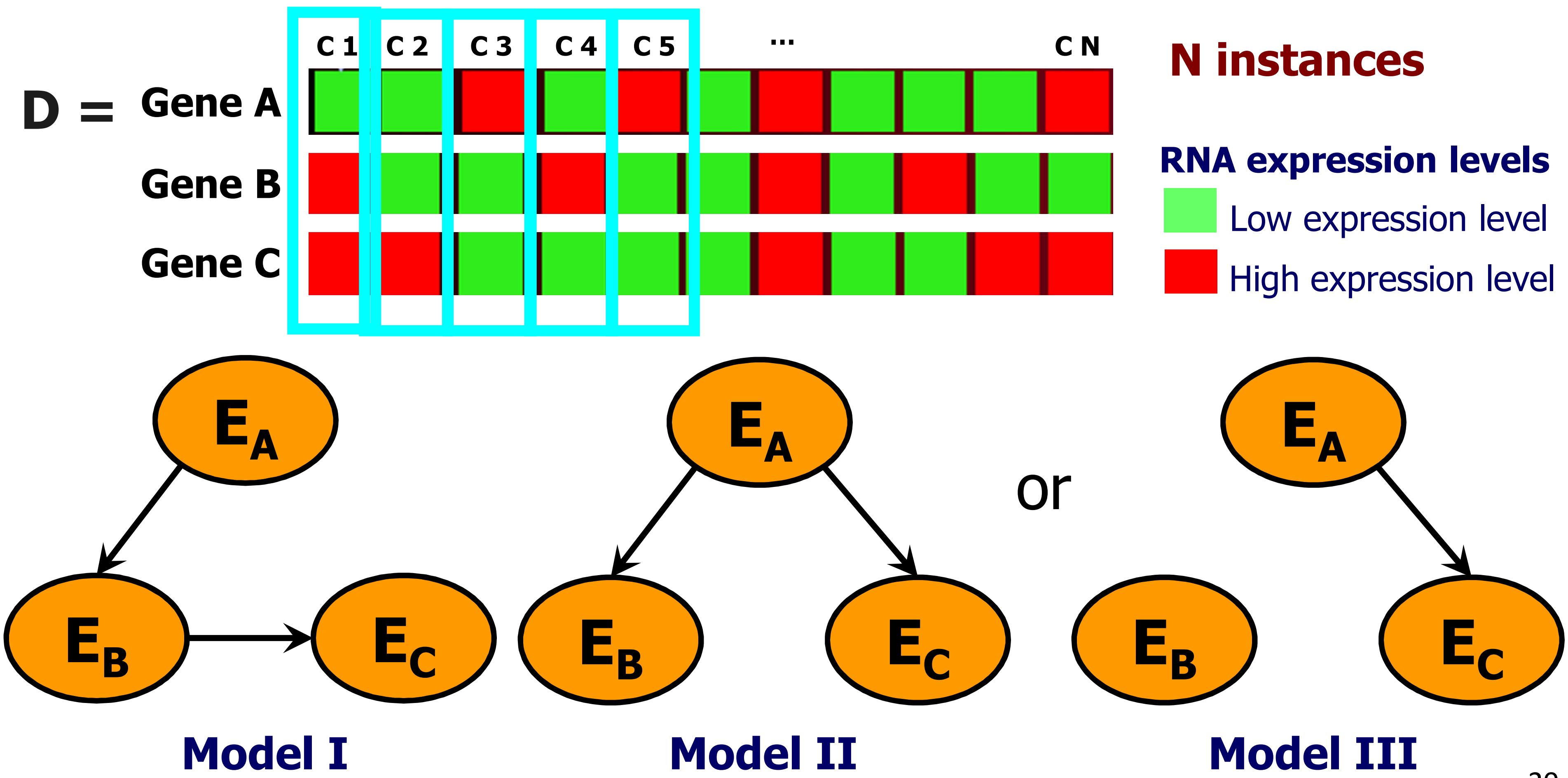
Genetic regulatory network

"Gene Expression"

Genes regulate each others' expression and activity.

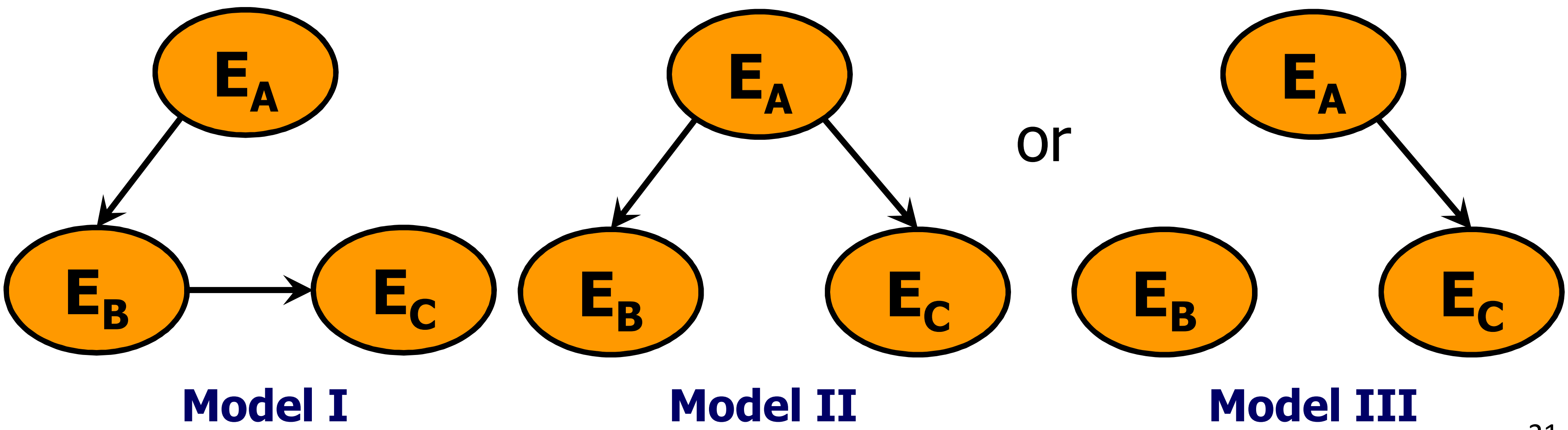
Model selection problem

- Which model do we think is the most likely?



Model selection problem

- Which model do we think is the most likely?
- Given data **D**, can we compute the following probability?
  - P (Model x is true | **D**)
  - Model selection: argmax_x P (Model x is true | **D**)
  - How to compute the probability? How about P (**D** | Model x is true)?

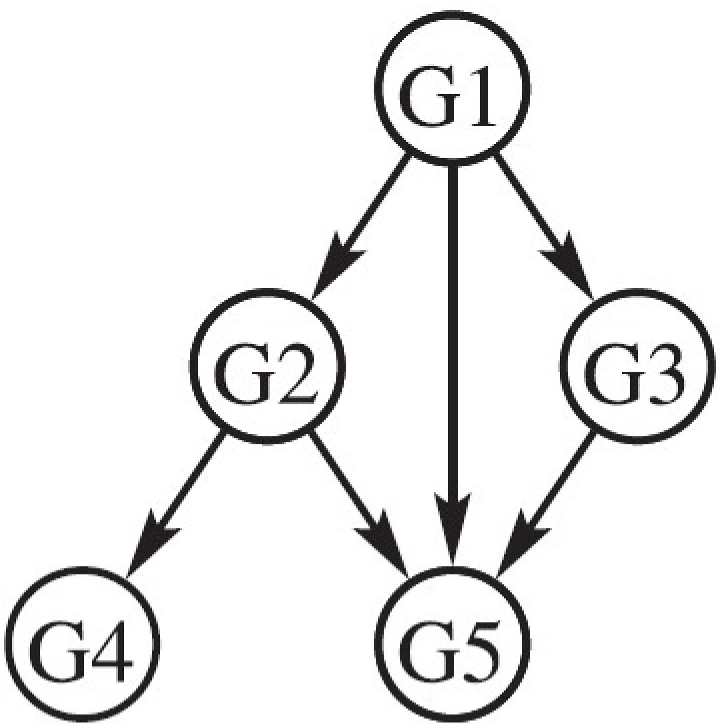


Outline

- Probability theory review
- Probabilistic models in genomics
- Bayesian networks representation
- Parameter estimation

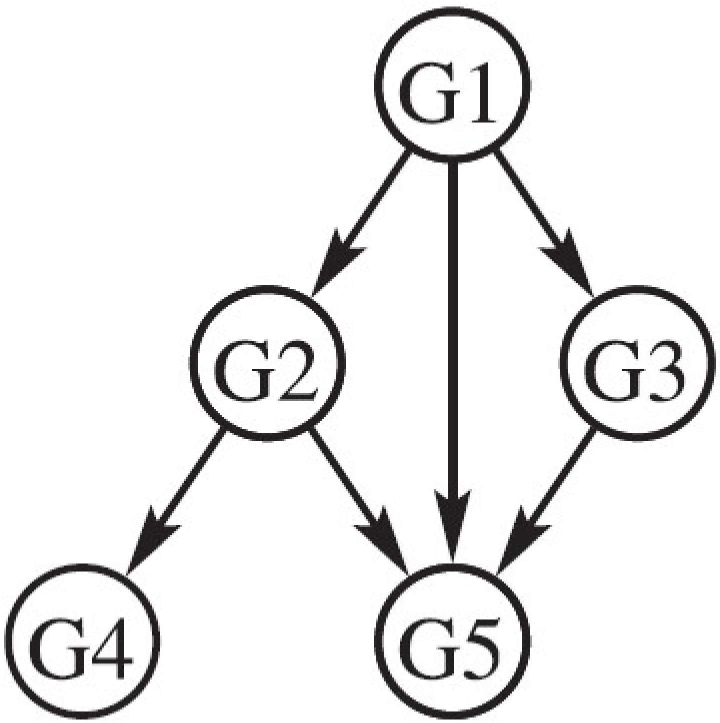
Graphical model representation

- Nodes – variables
- Edges – relationships between variables
- Bayesian network – directed acyclic graph (DAG)



Graphical model representation

- Nodes – genes
- Edges – regulatory relationships



Parameterization

- The joint probability distribution (JPD) P(G1,G2,G3,G4,G5) may be complex even for just 5 variables.

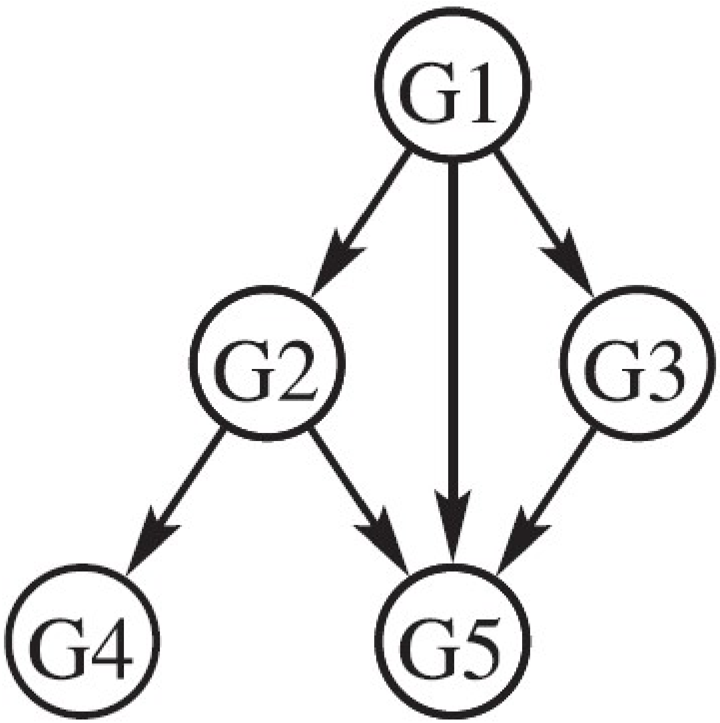
- Let's say that G's are binary.

- How many numbers do we need to fully specify JPD?
  - P(G1=1,G2=1,G3=1,G4=1,G5=0) = 0.1, …
  - $2^5 - 1$

- If G's are all independent,
  - P(G1,G2,G3,G4,G5) = P(G1)P(G2)P(G3)P(G4)P(G5)
  - Then how many numbers do we need to fully specify JPD?

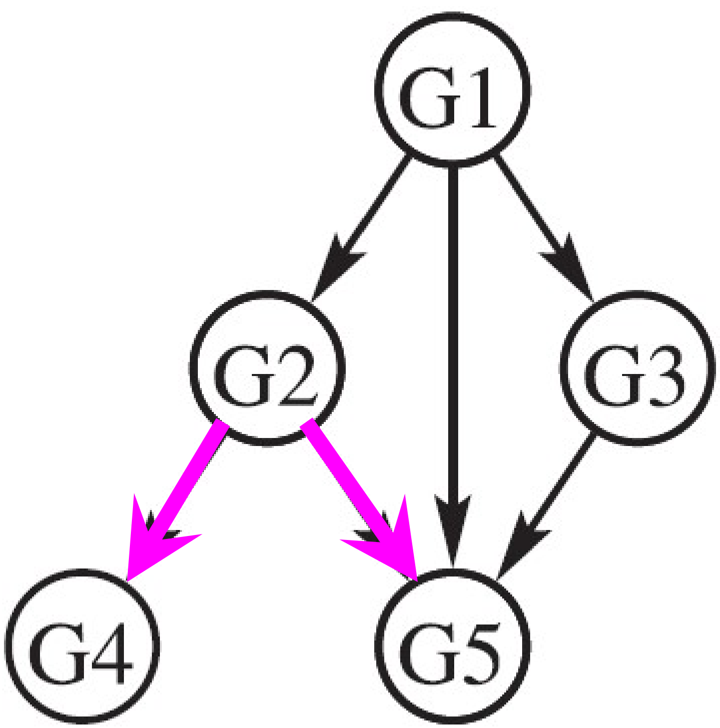
Parameterization in BNs

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



Parameterization in BNs

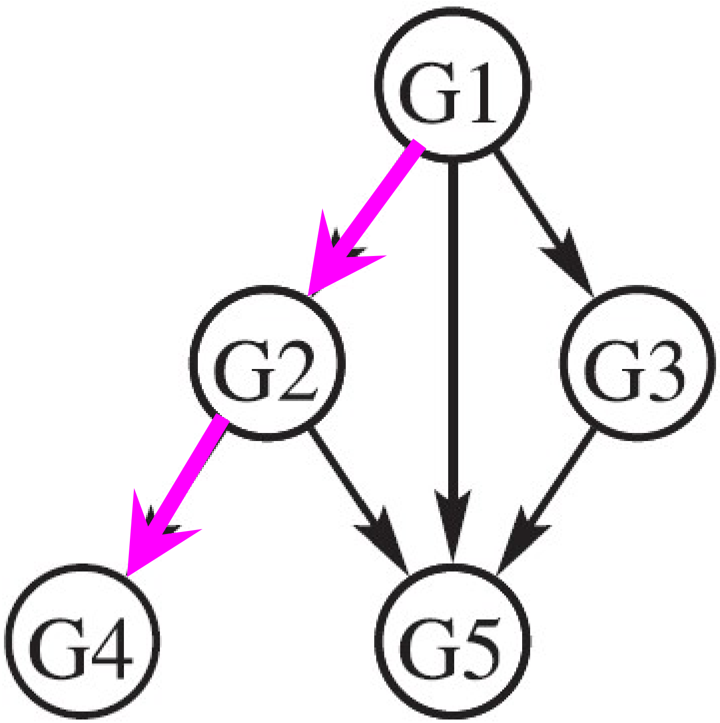
- The expression levels of G4 and G5 are related only because they share a common regulator G2.
- In mathematical term, G4 and G5 are conditionally independent given G2.



$G4 \perp G5 \mid G2$

Parameterization in BNs

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

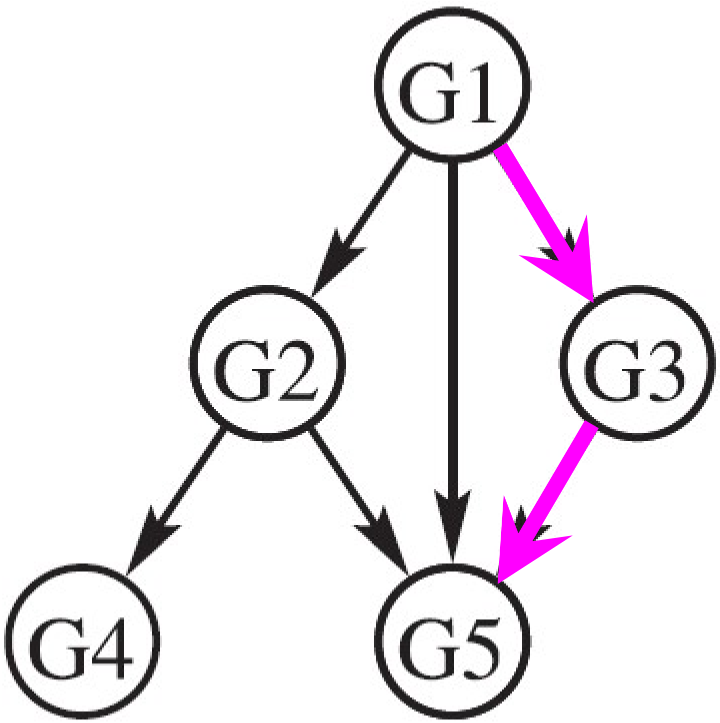


G4 $\perp$ G5 | G2

G1 $\perp$ G4 | G2

Parameterization in BNs

- The expression levels of G5 and G1 are directly related and through G2 and G3.



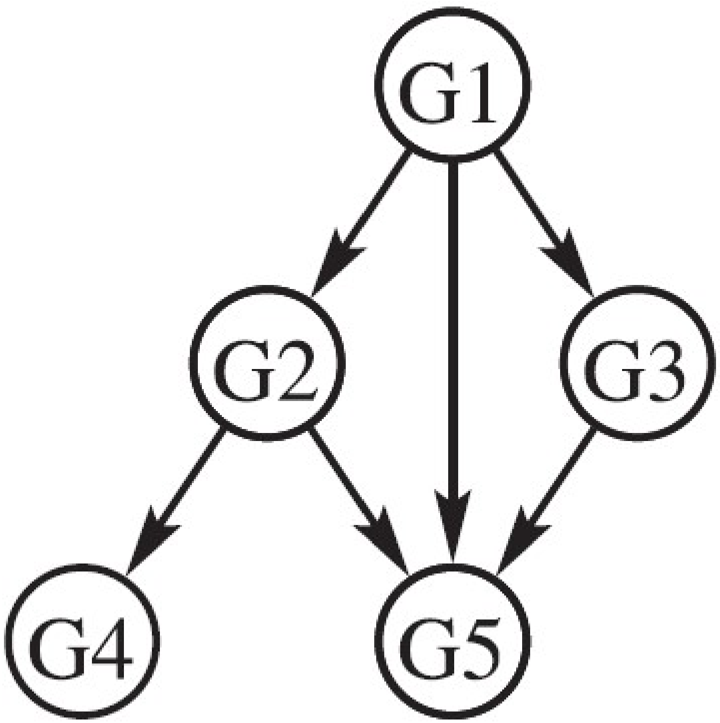
G4 ⊥ G5 | G2

G1 ⊥ G4 | G2

:

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)

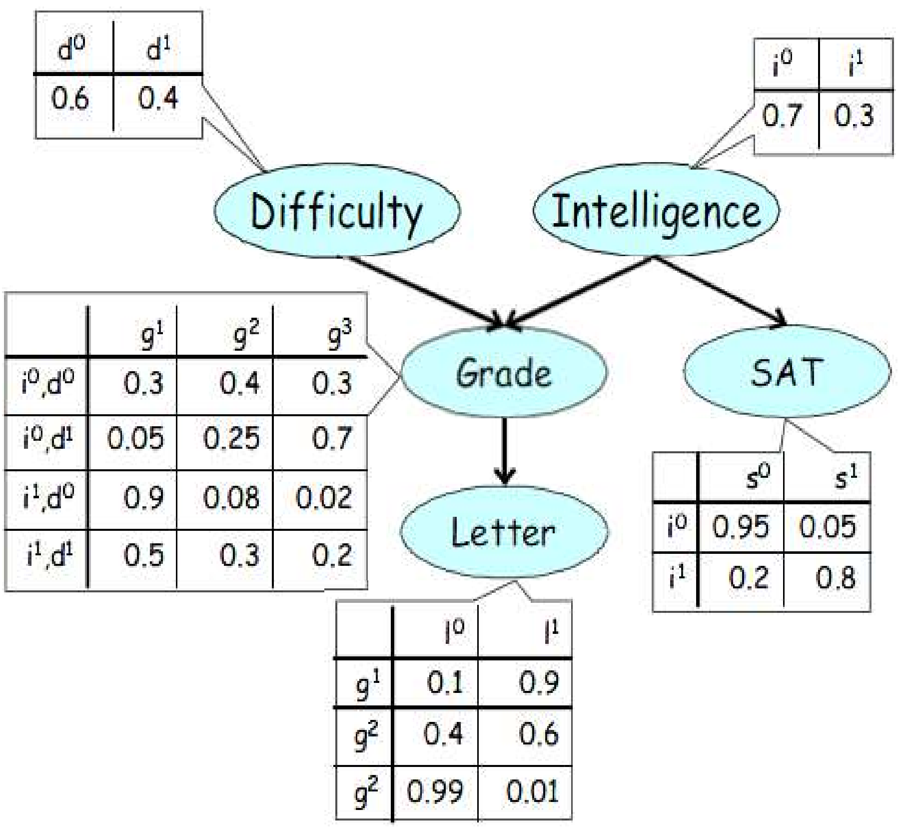


$G4 \perp G5 \mid G2$

$G1 \perp G4 \mid G2$

:

The *Student* Example



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

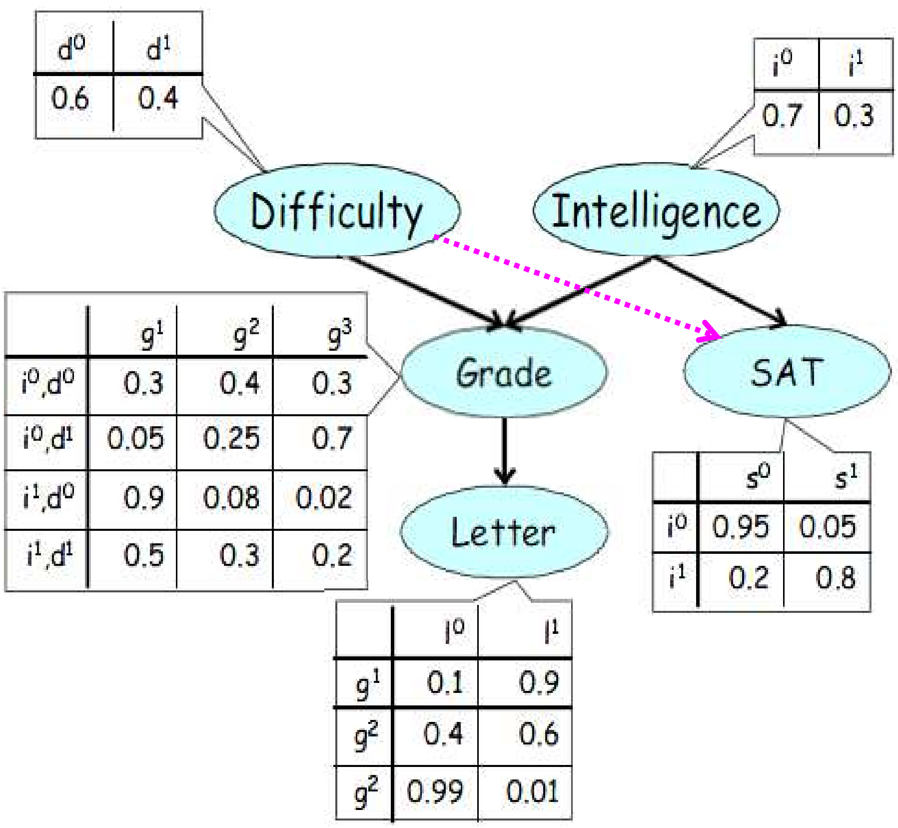
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

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

- Variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
  - Quality of Letter (L) = $\{l^0, l^1\}$
    Conditional probability distribution, P(L|G)

from Koller & Friedman (page 48 in K&F)

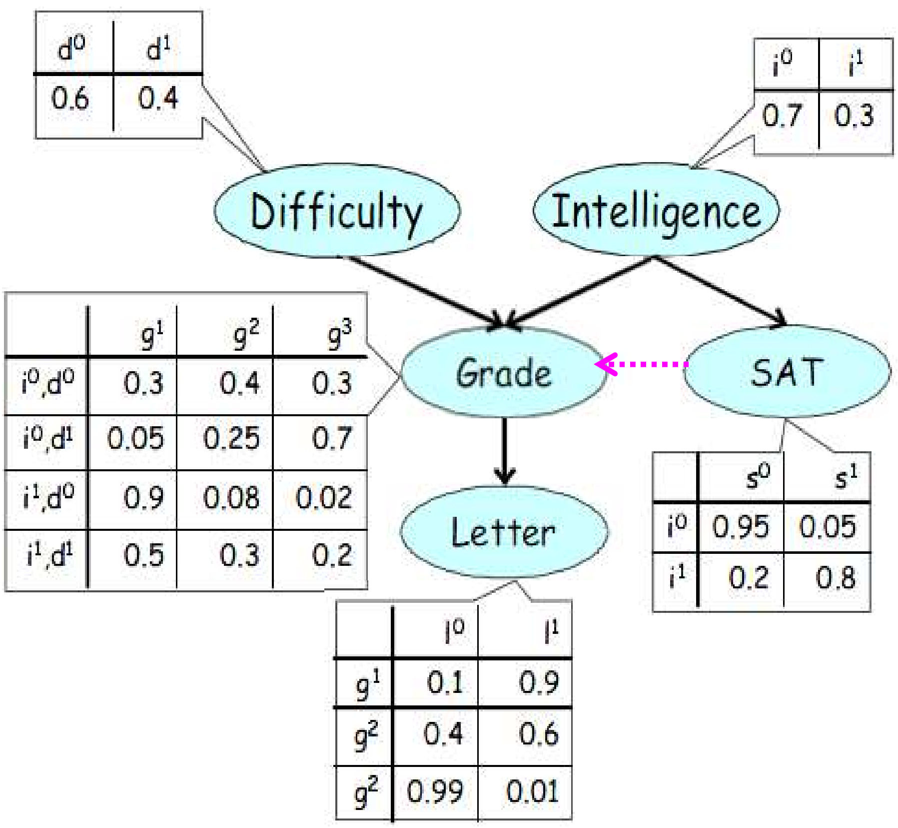
The *Student* Example



- Variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
    P(S|I,D) ?
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
  - Quality of Letter (L) = $\{l^0, l^1\}$
    Conditional probability distribution, P(L|G)

from Koller & Friedman

The *Student* Example



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

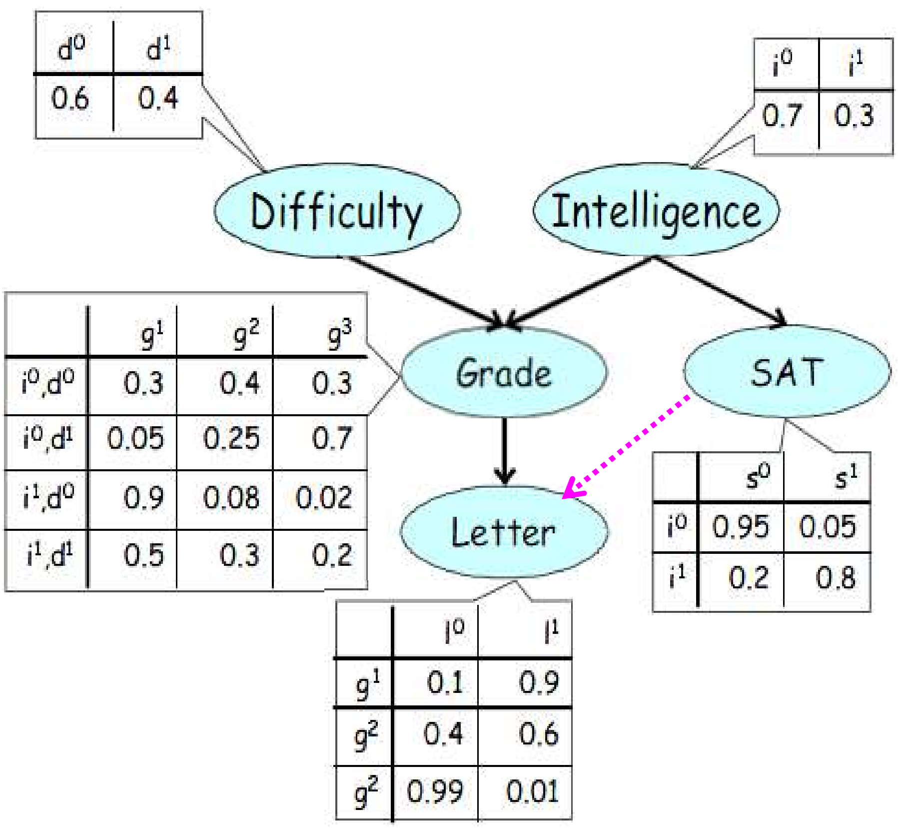
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	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

- Variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
    P(G|D,I,S) ?
  - Quality of Letter (L) = $\{l^0, l^1\}$
    Conditional probability distribution, P(L|G)

The *Student* Example



- Variables
  - Course difficulty (D) = $\{d^0, d^1\}$

    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$

    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$

    Conditional probability distribution, P(S|I)
  - Grade (G) = $\{g^1, g^2, g^3\}$

    Conditional probability distribution, P(G|D,I)
  - Quality of Letter (L) = $\{l^0, l^1\}$

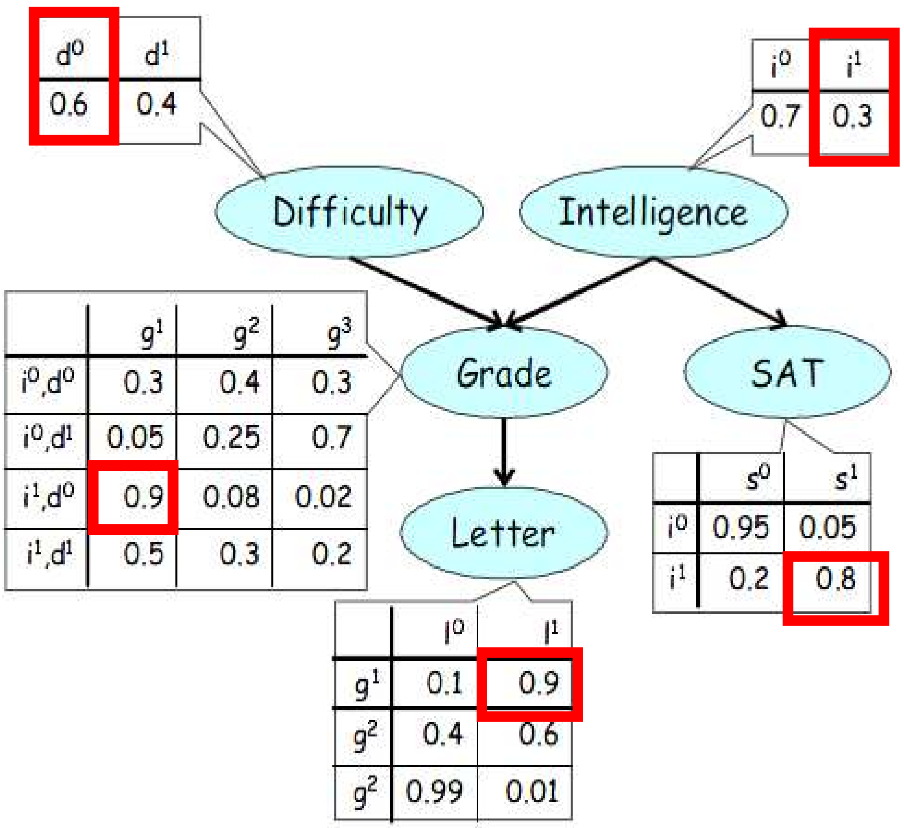
    Conditional probability distribution, P(L|G)

    P(L|G, S) ?

from Koller & Friedman

The *Student* Example

- What is the probability of observing {D=easy, I=intelligent, G=good, L=strong, S=high} ?

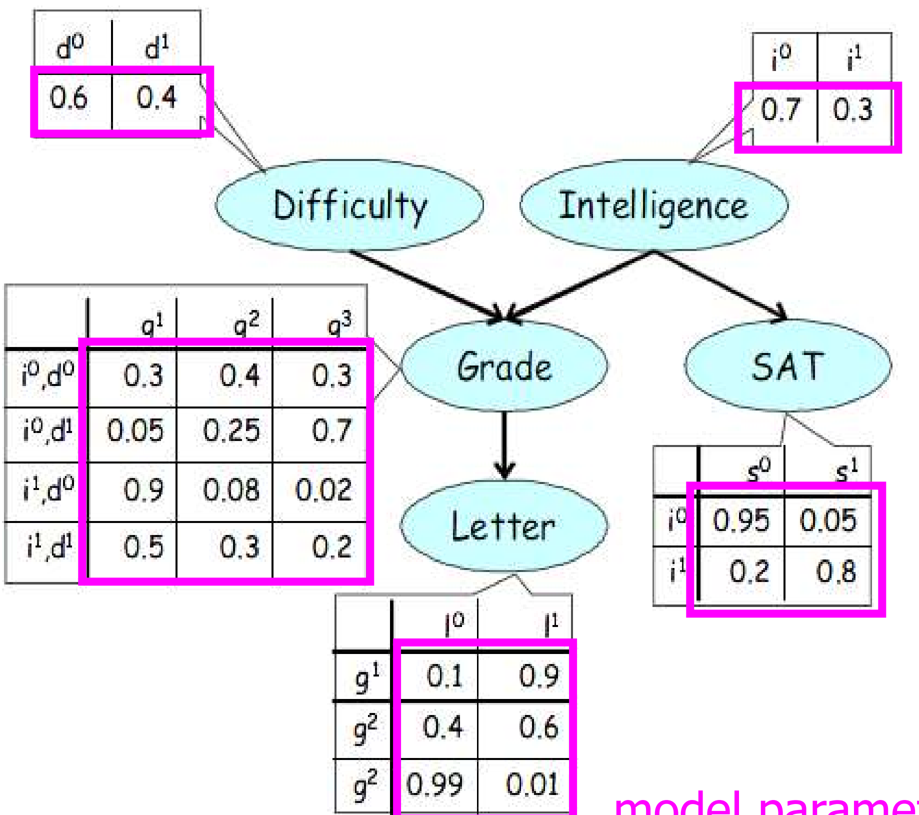


- P(D,I,G,L,S)
  = P(D) P(I) P(G|D,I) P(S|I) P(L|G)

- P(D=easy, I=intelligent, G=good, L=strong, S=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

Conditional probability tables (CPTs)

- What is the probability of observing {D=easy, I=intelligent, G=good, L=strong, S=high} ?



- P(D,I,G,L,S)

  = P(D) P(I) P(G|D,I) P(S|I) P(L|G)

- P(D=easy, I=intelligent, G=good, L=strong, S=high)

= P(D=easy) P(I=intelligent)
  P(G=good | D=easy, I=intelligent)
  P(S=strong | I=intelligent)
  P(L=strong | G=good)

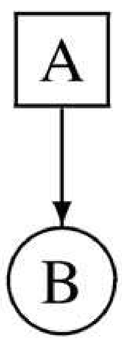
= 0.6 x 0.3 x 0.9 x 0.9 x 0.8

= 0.1166

model parameters
(can be "learned" from data!)

How about continuous variables?

- Squares – discrete nodes
- Circles – continuous nodes

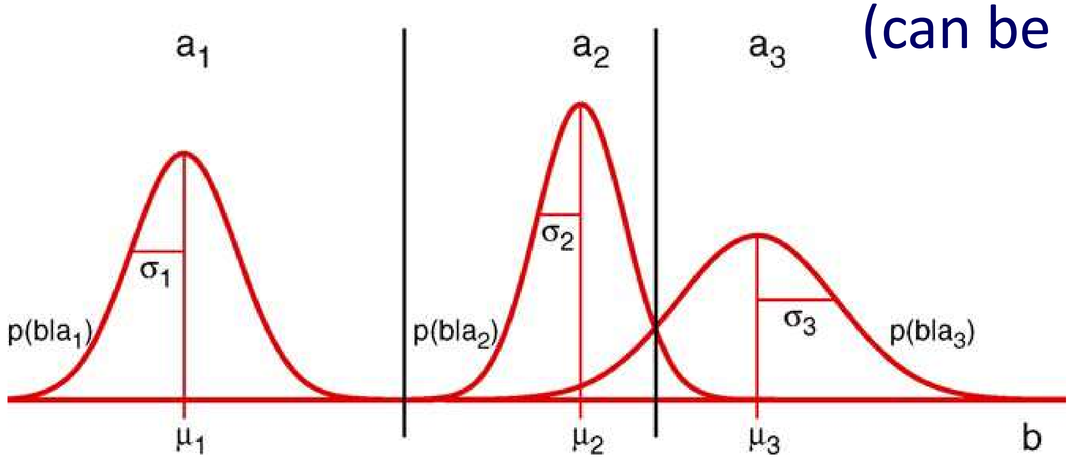


$\theta_B = p(B|A)$

	$p(B = b)$
$A = a_1$	μ_1, σ_1^2
$A = a_2$	μ_2, σ_2^2
$A = a_3$	μ_3, σ_3^2

- A: a variable with *k* =3 states
- B: a continuous node

Model parameters
(can be learned from data)



Joint probability distribution

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