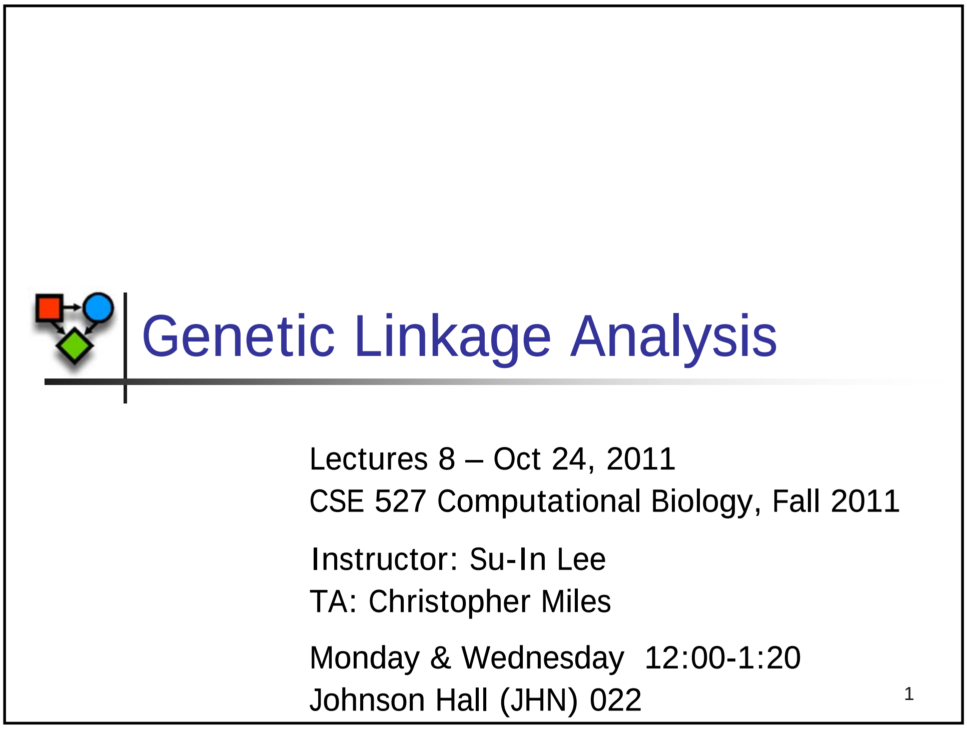
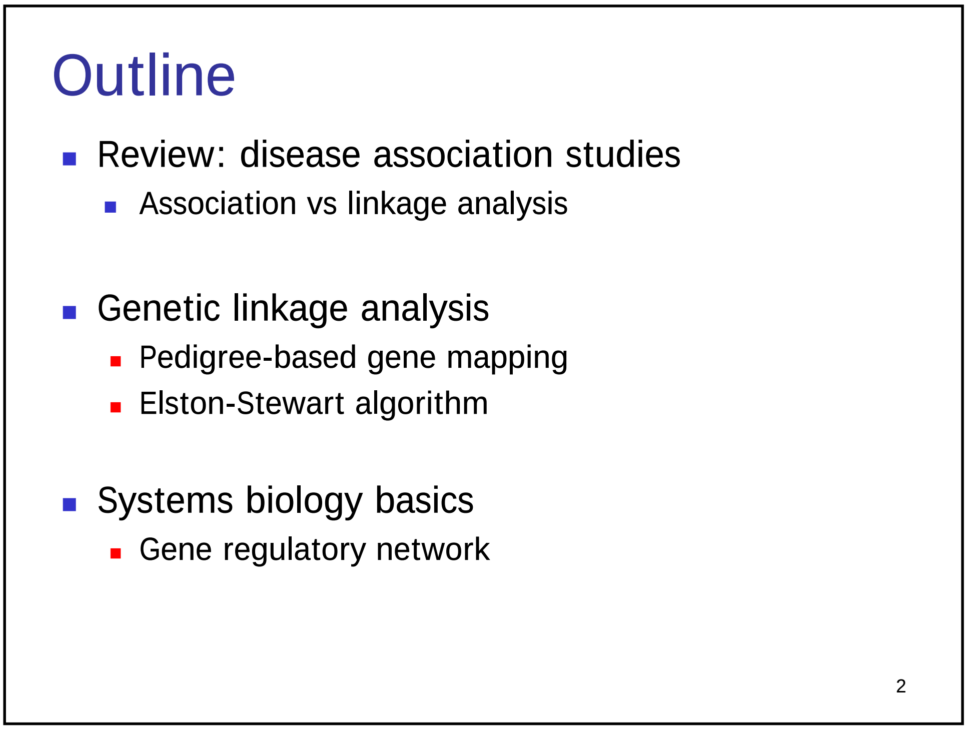


Genetic Linkage Analysis

Lectures 8 – Oct 24, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

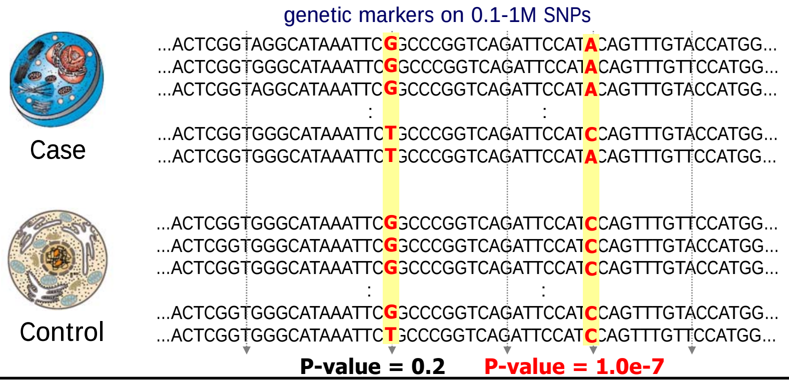


Outline

- Review: disease association studies
  - Association vs linkage analysis
- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm
- Systems biology basics
  - Gene regulatory network

Genome-Wide Association Studies

- Any disadvantages?
  - Hypothesis-free: we search the entire genome for associations rather than focusing on small candidate areas.
  - The need for extremely dense searches.
  - The massive number of statistical tests performed presents a potential for false-positive results (multiple hypothesis testing)



genetic markers on 0.1-1M SNPs

Case

```
...ACTCGGTAGGCATAAATTCGGCCCGGTCAGATTCCATACAGTTTGTACCATGG...
...ACTCGGTGGGCATAAATTCGGCCCGGTCAGATTCCATACAGTTTGTTCCATGG...
...ACTCGGTAGGCATAAATTCGGCCCGGTCAGATTCCATACAGTTTGTACCATGG...
                    :                :
...ACTCGGTGGGCATAAATTCTGCCCGGTCAGATTCCATCCAGTTTGTACCATGG...
...ACTCGGTGGGCATAAATTCTGCCCGGTCAGATTCCATACAGTTTGTTCCATGG...
```

Control

```
...ACTCGGTGGGCATAAATTCGGCCCGGTCAGATTCCATCCAGTTTGTTCCATGG...
...ACTCGGTGGGCATAAATTCGGCCCGGTCAGATTCCATCCAGTTTGTACCATGG...
...ACTCGGTGGGCATAAATTCGGCCCGGTCAGATTCCATCCAGTTTGTACCATGG...
                    :                :
...ACTCGGTGGGCATAAATTCGGCCCGGTCAGATTCCATCCAGTTTGTACCATGG...
...ACTCGGTGGGCATAAATTCTGCCCGGTCAGATTCCATCCAGTTTGTTCCATGG...
```

P-value = 0.2 **P-value = 1.0e-7**

Association vs Linkage Analysis

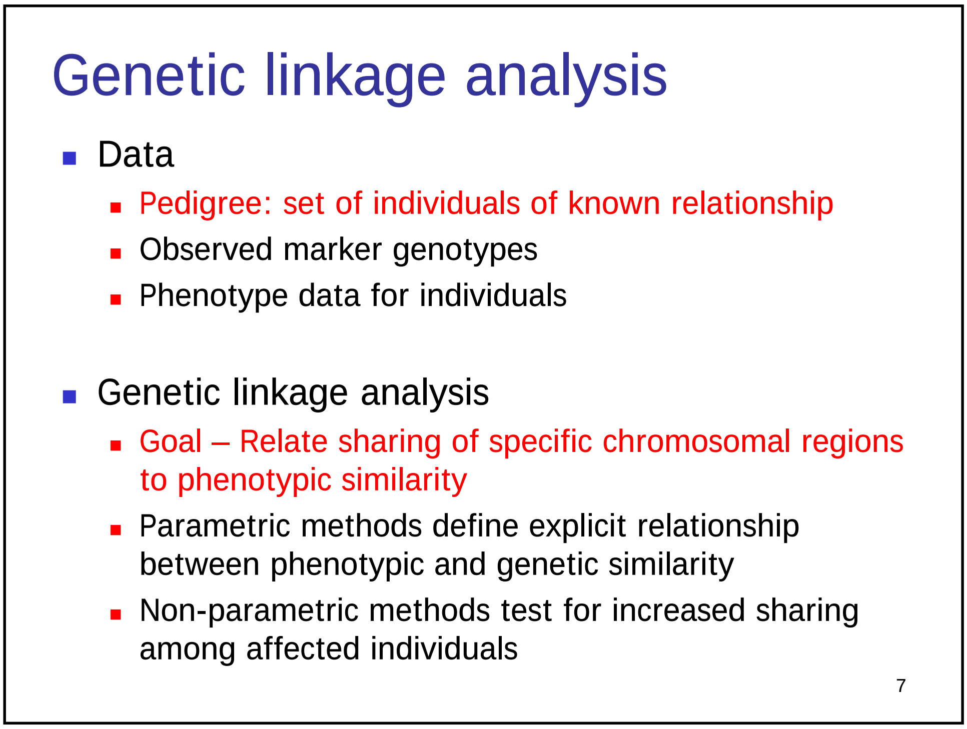
- Any disadvantages?
  - Hypothesis-free: we search the entire genome for associations rather than focusing on small candidate areas.
  - The need for extremely dense searches.
  - The massive number of statistical tests performed presents a potential for false-positive results (multiple hypothesis testing)
- Alternative strategy – Linkage analysis
  - It acts as systematic studies of variation, without needing to genotype at each region.
  - Focus on a family or families.

Basic Ideas

- Neighboring genes on the chromosome have a tendency to stick together when passed on to offspring.
- Therefore, if some disease is often passed to offspring along with specific marker-genes, we can conclude that the gene(s) responsible for the disease are located close on the chromosome to these markers.

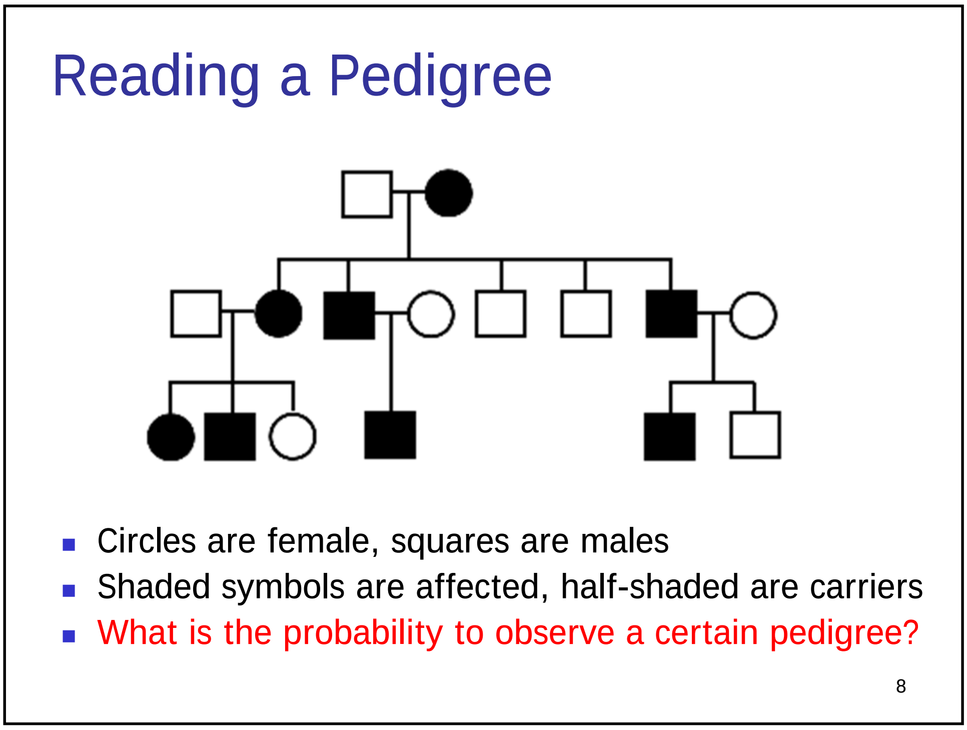
Outline

- Review: disease association studies
  - Association vs linkage analysis
- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm
- Systems biology basics
  - Gene expression data
  - Gene regulatory network



Genetic linkage analysis

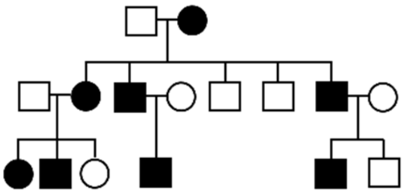
- Data
  - Pedigree: set of individuals of known relationship
  - Observed marker genotypes
  - Phenotype data for individuals
- Genetic linkage analysis
  - Goal – Relate sharing of specific chromosomal regions to phenotypic similarity
  - Parametric methods define explicit relationship between phenotypic and genetic similarity
  - Non-parametric methods test for increased sharing among affected individuals



Reading a Pedigree

- Circles are female, squares are males
- Shaded symbols are affected, half-shaded are carriers
- What is the probability to observe a certain pedigree?

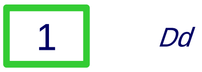
Elements of Pedigree Likelihood



- Prior probabilities
  - For founder genotypes
- Transmission probabilities
  - For offspring genotypes, given parents
- Penetrances
  - For individual phenotypes, given genotype

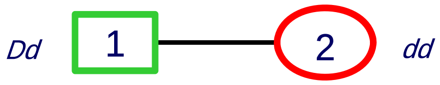
Probabilistic model for a pedigree: (1) Founder (prior) probabilities

- ***Founders*** are individuals whose parents are not in the pedigree
  - They may or may not be typed. Either way, we need to assign probabilities to their actual or possible genotypes.
  - This is usually done by assuming Hardy-Weinberg equilibrium (HWE). If the frequency of D is .01, HW says



P(father Dd) = 2 x .01 x .99

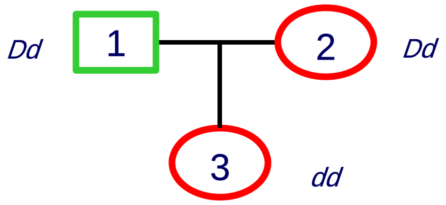
- Genotypes of founder couples are (usually) treated as independent.



P(father Dd, mother dd) = (2 x .01 x .99) x $(.99)^2$

Probabilistic model for a pedigree: (2) Transmission probabilities I

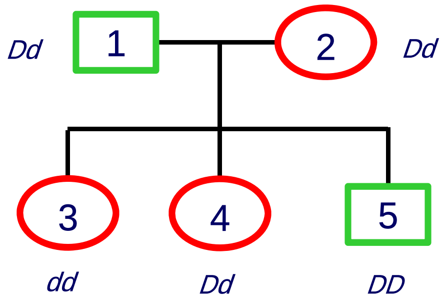
- According to Mendel’s laws, children get their genes from their parents’ genes independently:



P(children 3 dd | father Dd, mother dd) = ½ x ½

- The inheritances are independent for different children.

Probabilistic model for a pedigree: (2) Transmission probabilities II



P(3 dd, 4 Dd, 5DD | 1 Dd, 2 dd)
= (½ x ½) x (2 x ½ x ½) x (½ x ½)

- The factor 2 comes from summing over the two mutually exclusive and equiprobable ways 4 get a D and a d.

Probabilistic model for a pedigree: (3) Penetrance probabilities I

- Independent penetrance model
  - Pedigree analyses usually suppose that, given the genotype at all loci, and in some cases age and sex, the chance of having *a particular phenotype depends only on genotype at one locus*, and is independent of all other factors: genotypes at other loci, environment, genotypes and phenotypes of relative, etc

- Complete penetrance

DD ■

P(affected | DD) = 1

- Incomplete penetrance

DD □

P(affected | DD) = .8

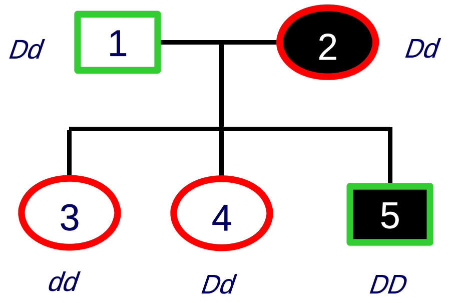
Probabilistic model for a pedigree: (3) Penetrance probabilities II

- Age & sex-dependent penetrance

DD (45) ■

P(affected | DD, male, 45 y.o.) = .6

Probabilistic model for a pedigree: Putting all together I



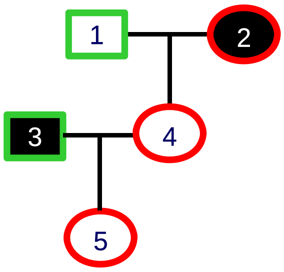
- Assumptions
  - Penetrance probabilities:
    P(affected | dd)=0.1, p(affected | Dd)=0.3, P(affected | DD)=0.8
  - Allele frequency of D is .01

- The probability of this pedigree is the product:
  - (2 x .01 x .99 x .7) x (2 x .01 x .99 x .3) x (½ x ½ x .9)
    x (2 x ½ x ½ x .7) x (½ x ½ x .8)

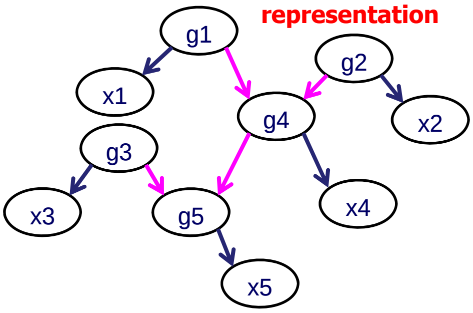
15

Elements of pedigree likelihood

A pedigree

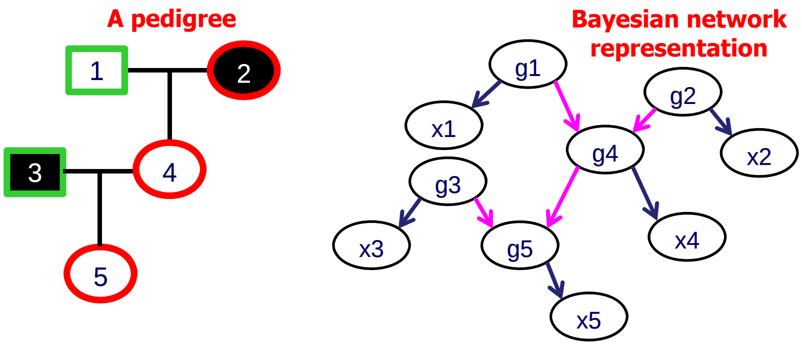


Bayesian network representation



- Prior probabilities
  - For founder genotypes e.g. P(g1), P(g2)
- Transmission probabilities
  - For offspring genotypes, given parents e.g. P(g4|g1,g2)
- Penetrance
  - For individual phenotypes, given genotype e.g. P(x1|g1)

Elements of pedigree likelihood



- Overall pedigree likelihood
  - $L = \prod_{f=\text{founders}} P(G_f) \prod_{\text{o,f,m}\}} P(G_o \mid G_f, G_m) \prod_{=\text{individuals}} P(X_i \mid G_i)$

Probability of founder genotypes

Probability of offspring given parents

Probability of phenotypes given genotypes

Probabilistic model for a pedigree: Putting all together II

- To write the likelihood of a pedigree given complete data:

$$L_C = \prod_{f=\text{founders}} P(G_f) \prod_{\{o,f,m\}} P(G_o \mid G_f, G_m) \prod_{i=\text{individuals}} P(X_i \mid G_i)$$

  - We begin by multiplying founder gene frequencies, followed by transmission probabilities of non-founders given their parents, next penetrance probabilities of all the individuals given their genotypes.

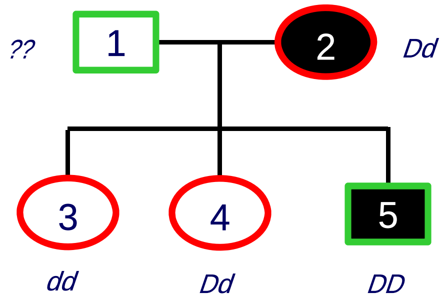
- What if there are missing or incomplete data?
  - We must sum over all mutually exclusive possibilities compatible with the observed data.

$$L = \sum_{G_1} \cdots \sum_{G_n} \prod_{f=\text{founders}} P(G_f) \prod_{\{o,f,m\}} P(G_o \mid G_f, G_m) \prod_{i=\text{individuals}} P(X_i \mid G_i)$$

All possible genotypes of individual 1

If the individual i's genotype is known to be g_i, then $G_i = \{g_i\}$

Probabilistic model for a pedigree: Putting all together II



$$L = \sum_{g_1=\{DD,Dd,dd\}} P(G_1 = g_1, G_2 = Dd, G_3 = dd, G_4 = Dd, G_5 = DD)$$

- What if there are missing or incomplete data?
  - We must sum over all mutually exclusive possibilities compatible with the observed data.

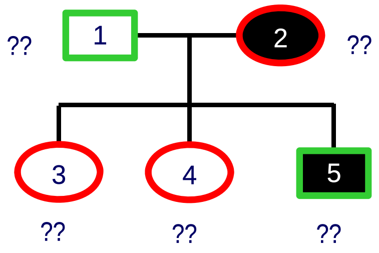
$$L = \sum_{G_1} \cdots \sum_{G_n} \prod_{f=\text{founders}} P(G_f) \prod_{\{o,f,m\}} P(G_o \mid G_f, G_m) \prod_{i=\text{individuals}} P(X_i \mid G_i)$$

Computationally ...

- To write the likelihood of a pedigree:

$$L = \sum_{G_1} \cdots \sum_{G_n} \prod_{f=\text{founders}} P(G_f) \prod_{\{o,f,m\}} P(G_o \mid G_f, G_m) \prod_{i=\text{individuals}} P(X_i \mid G_i)$$

  - Computation rises exponentially with # people *n*.
  - Computation rises exponentially with # markers
  - Challenge is summation over all possible genotypes (or haplotypes) for each individual.

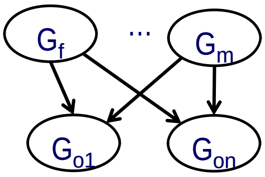


Computationally ...

- Two algorithms:
  - The general strategy of beginning with founders, then non-founders, and multiplying and summing as appropriate, has been codified in what is known as the **Elston-Stewart algorithm** for calculating probabilities over pedigrees.
  - It is one of the two widely used approaches. The other is termed the **Lander-Green algorithm** and takes a quite different approach.

Elston and Stewart's insight...

- Focus on "special pedigree" where
  - Every person is either
    - Related to someone in the previous generation
    - Marrying into the pedigree
  - No consanguineous marriages
- Process nuclear families, by fixing the genotype for one parent
  - Conditional on parental genotypes, offsprings are independent



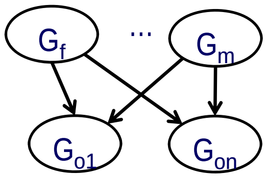
Elston and Stewart's insight...

- Conditional on parental genotypes, offsprings are independent
- Thus, avoid nested sums, and produce likelihood whose cost increases linearly with the number of offspring

$$L = \sum_{G_m}\sum_{G_f}\sum_{G_{o1}}\cdots\sum_{G_{on}} P(X_m \mid G_m)P(G_m)P(X_f \mid G_f)P(G_f)\prod_{o1\ldots on} P(X_o \mid G_o)P(G_o \mid G_m,G_f)$$

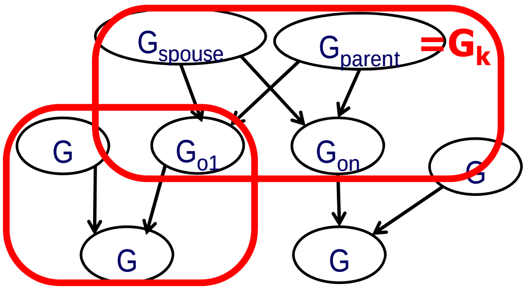
$$= \sum_{G_m} P(X_m \mid G_m)P(G_m)\sum_{G_f} P(X_f \mid G_f)P(G_f)$$

$$\prod_{o}\sum_{G_o} P(X_o \mid G_o)P(G_o \mid G_m,G_f)$$



Successive Conditional Probabilities

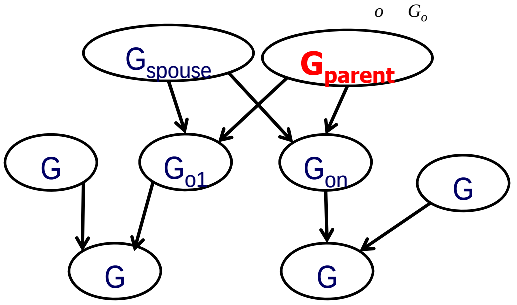
- Starting at the bottom of the pedigree...
- Calculate conditional probabilities by fixing genotypes for one parent
- Specifically, calculate H_k (G_k)
  - Probability of descendants and spouse for person k
  - Conditional on a particular genotype G_k



Formulae ...

- So for each parent, calculate

$$H_{\text{parent}}(G_{\text{parent}}) = \sum_{G_{\text{spouse}}} P(X_{\text{spouse}} \mid G_{\text{spouse}}) P(G_{\text{spouse}}) \prod_{o} \sum_{G_o} P(X_o \mid G_o) P(G_o \mid G_{\text{parent}} G_{\text{spouse}}) H_o(G_o)$$



Probability of o's spouse and descendants when it's genotype is G_o

$$H_{\text{leaf}}(G_{\text{leaf}}) = 1$$

- By convention, for individuals with no descendants

Final likelihood

- After processing all nuclear family units

- Simple sum gives the overall pedigree likelihood

$$L = \sum_{G_{\text{founder}}} P(X_{\text{founder}} \mid G_{\text{founder}}) P(G_{\text{founder}}) H_{\text{founder}}(G_{\text{founder}})$$

$$L = \sum_{G_1} \cdots \sum_{G_n} \prod_{\text{f}=\text{founders}} \text{P}(\text{G}_\text{f}) \prod_{\{\text{o,f,m}\}} \text{P}(\text{G}_\text{o} \mid \text{G}_\text{f}, \text{G}_\text{m}) \prod_{\text{i}=\text{individuals}} \text{P}(\text{X}_\text{i} \mid \text{G}_\text{i})$$

$$= \sum_{G_{\text{founder}}} P(G_{\text{founder}}) P(X_{\text{founder}} \mid G_{\text{founder}}) \boxed{\sum_{G_{\text{nonfounders}}} \prod_{\{\text{o,f,m}\}} \text{P}(\text{G}_\text{o} \mid \text{G}_\text{f}, \text{G}_\text{m}) \prod_{\text{i}=\text{nonfounders}} \text{P}(\text{X}_\text{i} \mid \text{G}_\text{i})}$$

P(X, given genotypes | G_{founder})=H_{founder} (G_{founder})

What next?

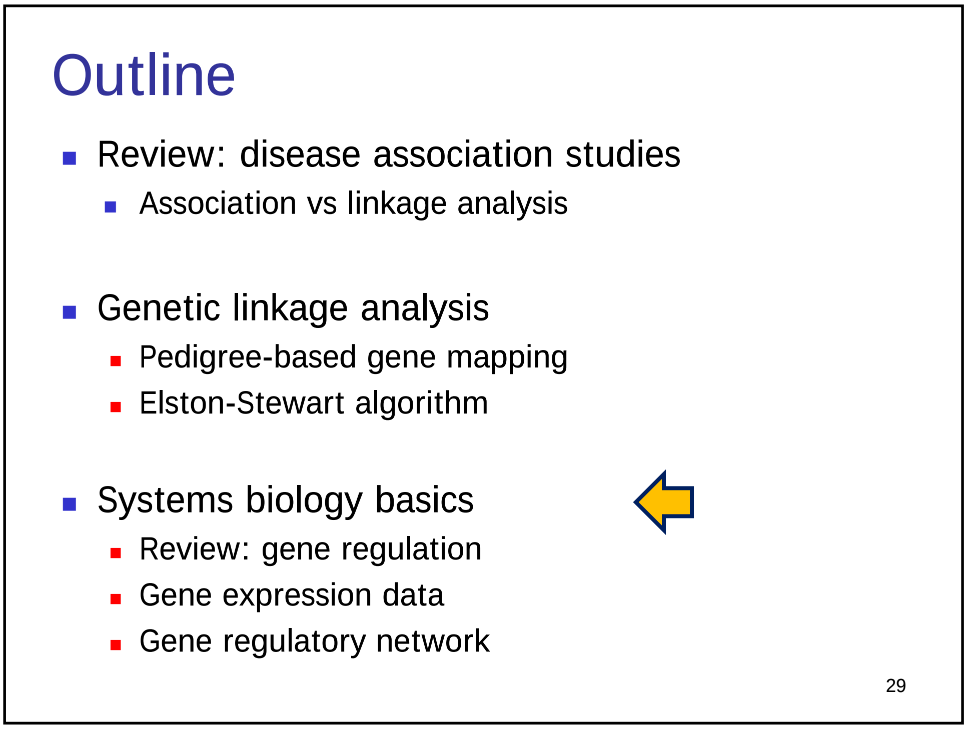
- Computation of the pedigree likelihood

- For every marker, we want to
  - Compute the pedigree likelihood for each marker and choose the marker that is closely linked to the disease gene.

Further Reading

- Part I

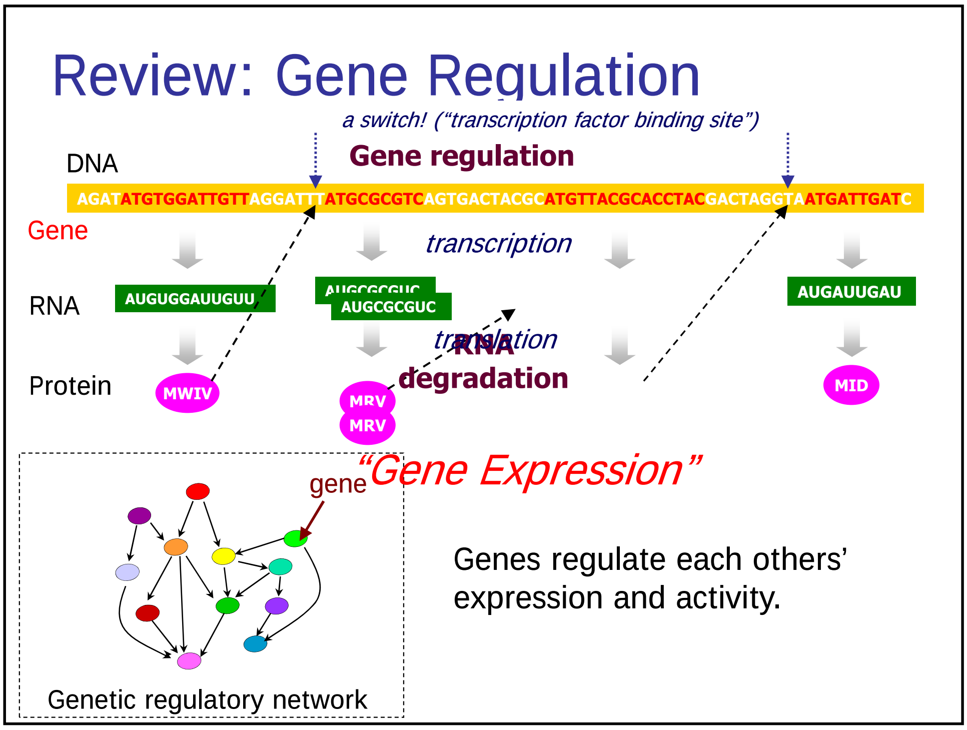
  - de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005 Nov;37(11):1217-23.
  - Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets.Nat Genet. 2006 Jun;38(6):663-7.
  - Reich, D.E. and Lander, E.S. On the allelic spectrum of human disease. *Trends Genet.*, 2001; 17, 502–510.
  - Risch N & Merikangas K, The future of genetic studies of complex human diseases. Science. 1996 Sep 13;273(5281):1516-7.
  - The International HapMap Consortium. A haplotype map of the human genome. N*ature* 2005 ; 437, 1299-1320..



Outline

- Review: disease association studies
  - Association vs linkage analysis
- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm
- Systems biology basics
  - Review: gene regulation
  - Gene expression data
  - Gene regulatory network

29

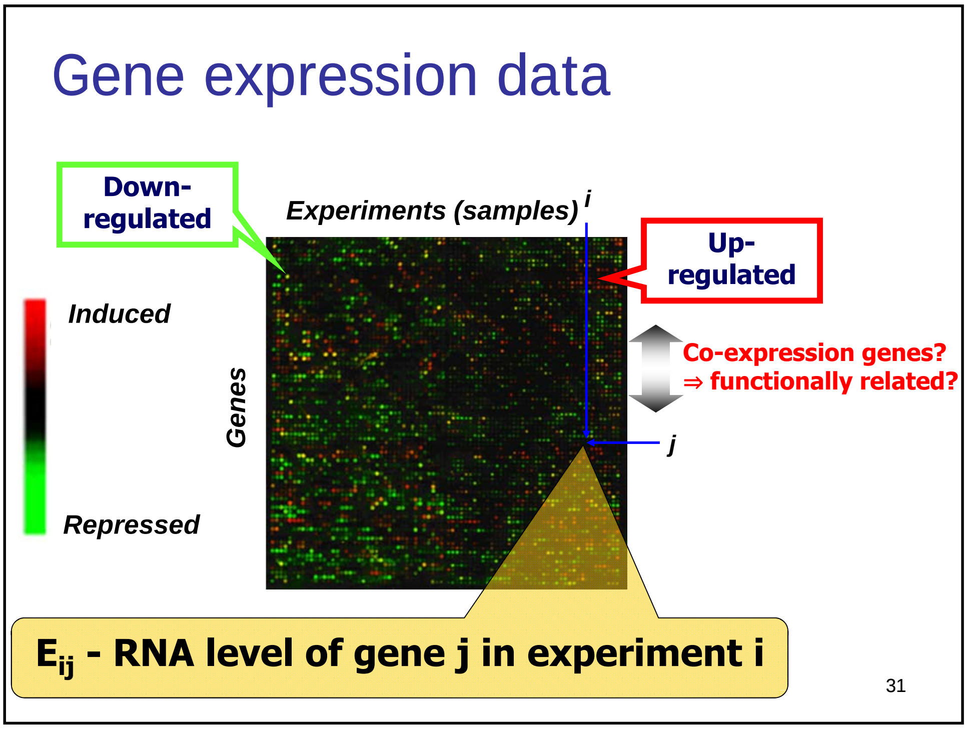


Review: Gene Regulation

a switch! (“transcription factor binding site”)
Gene regulation
DNA
AGATATGTGGATTGTTAGGATTTATGCGCGTCAGTGACTACGCATGTTACGCACCTACGACTAGGTAATGATTGATC
Gene
transcription
RNA
AUGUGGAUUGUU
AUGCGCGUC
AUGCGCGUC
AUGAUUGAU
translation
RNA
degradation
Protein
MWIV
MRV
MRV
MID
gene
Genetic regulatory network

“Gene Expression”

Genes regulate each others’ expression and activity.



Gene expression data

Down-regulated

Experiments (samples) *i*

Up-regulated

Induced

Repressed

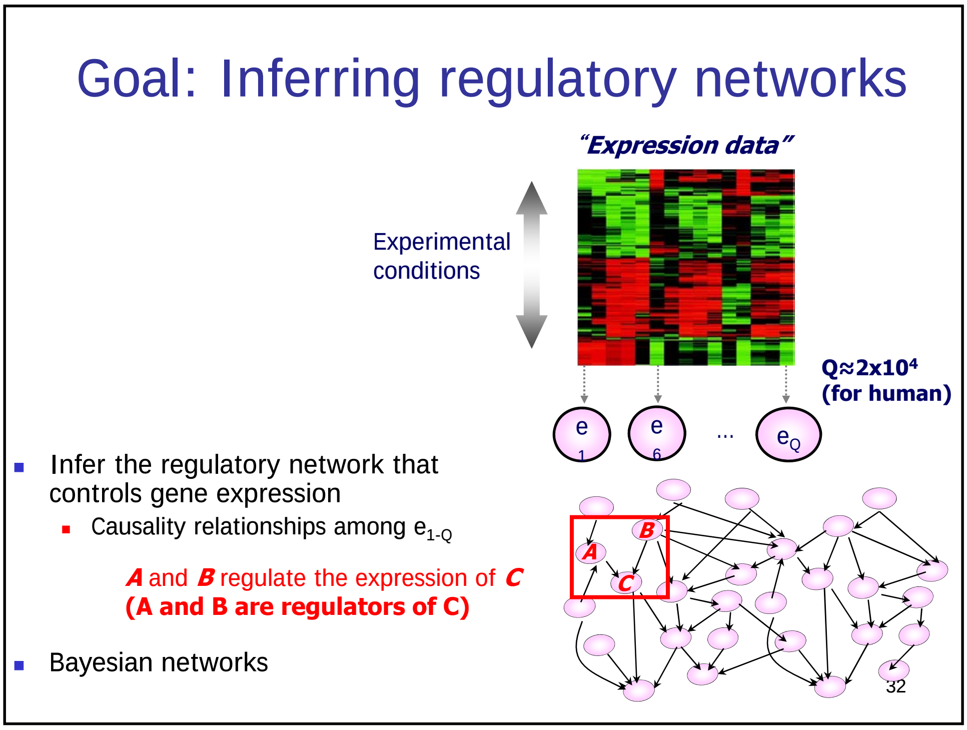
Genes

Co-expression genes?
⇒ functionally related?

j

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

31



Goal: Inferring regulatory networks

"Expression data"

Experimental conditions

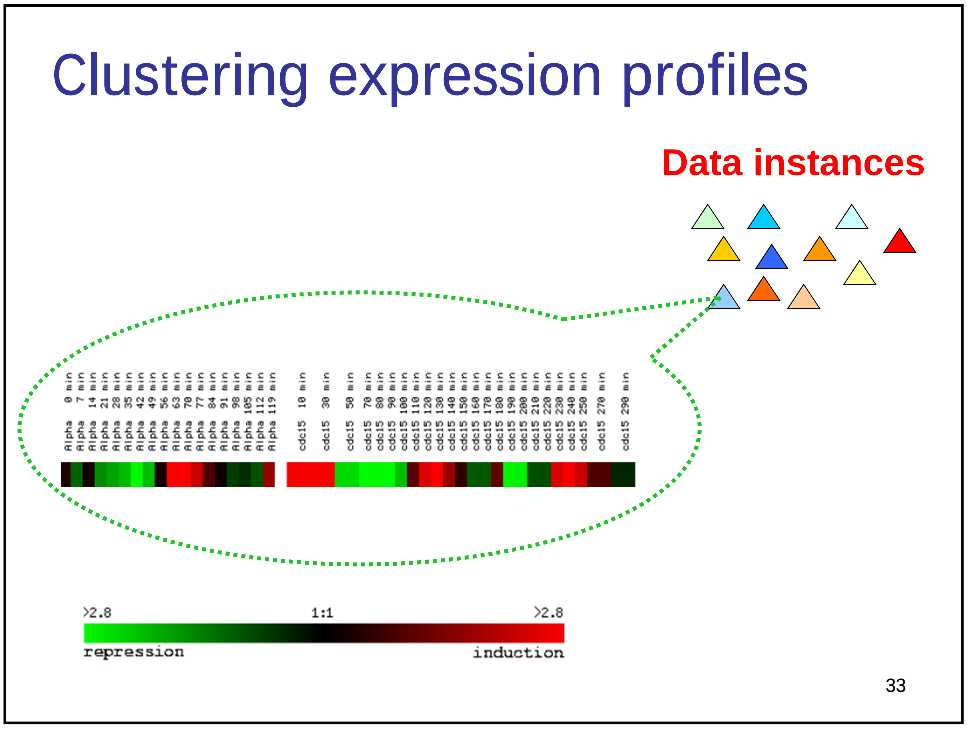
Q≈2x10⁴
(for human)

- Infer the regulatory network that controls gene expression
  - Causality relationships among $e_{1\text{-}Q}$

  A and ***B*** regulate the expression of ***C***
  (A and B are regulators of C)

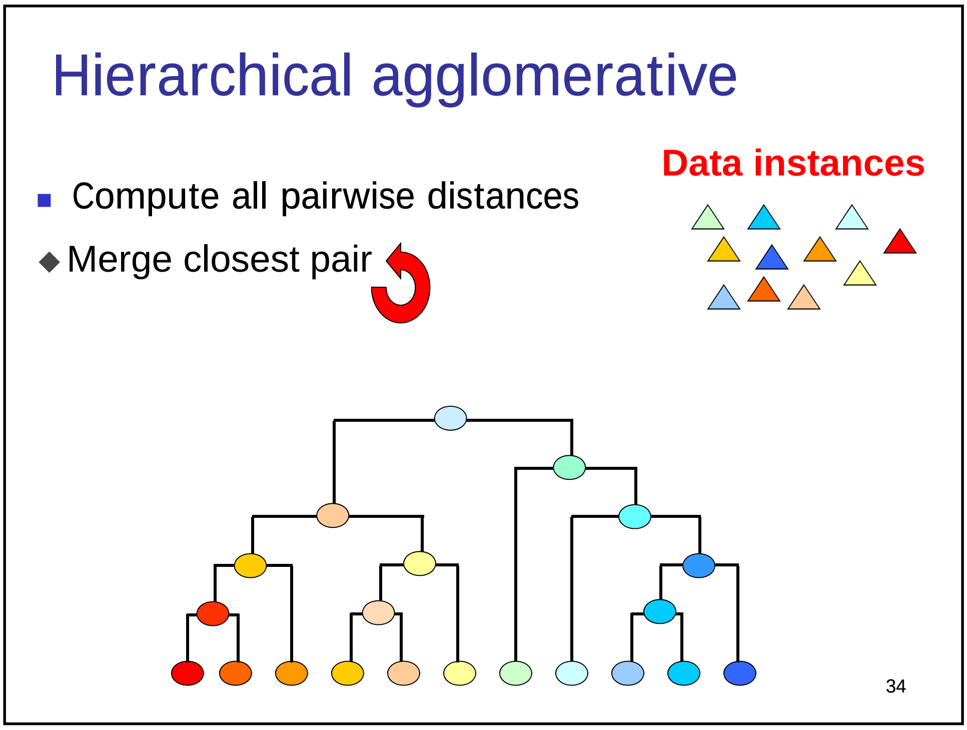
- Bayesian networks

32



Clustering expression profiles

Data instances

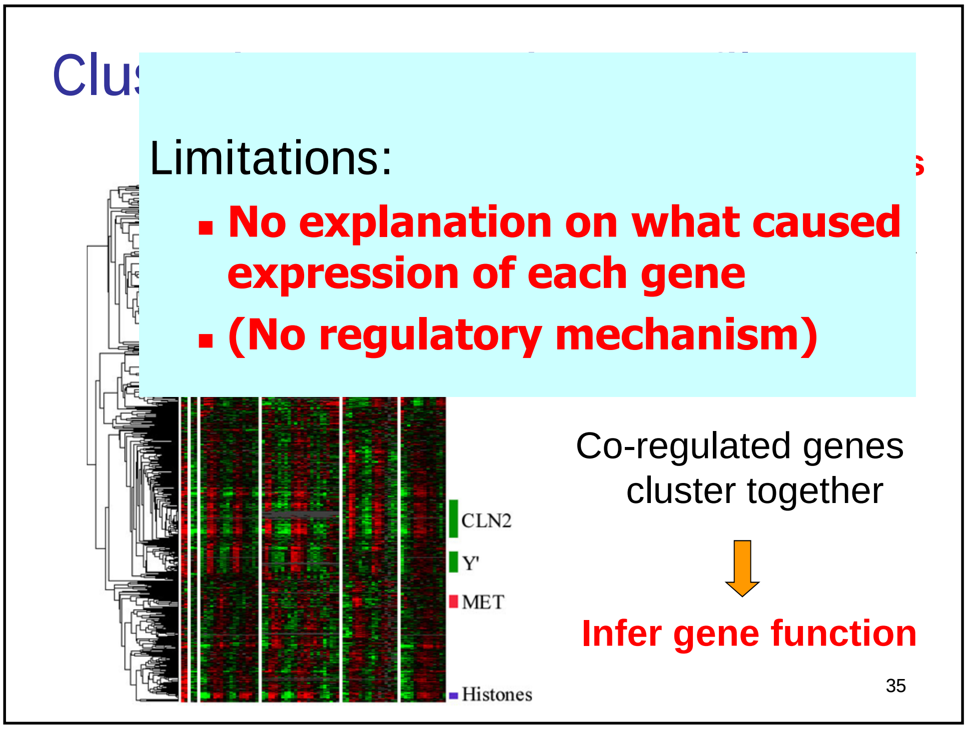


Hierarchical agglomerative

Data instances

- Compute all pairwise distances

◆ Merge closest pair



Clus

Limitations:

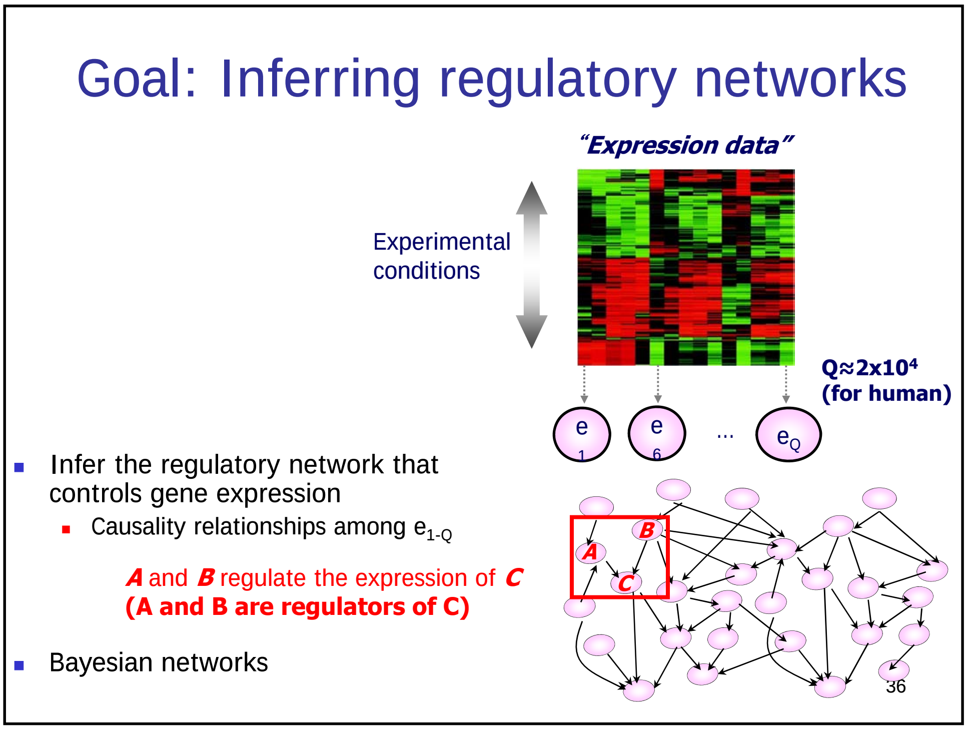
- **No explanation on what caused expression of each gene**
- **(No regulatory mechanism)**

CLN2
Y'
MET
Histones

Co-regulated genes cluster together

Infer gene function

35



Goal: Inferring regulatory networks

"Expression data"
Experimental conditions
Q≈2x10⁴ (for human)
e 1
e 6
...
eQ
A
B
C

- Infer the regulatory network that controls gene expression
  - Causality relationships among $e_{1\text{-}Q}$

  A and ***B*** regulate the expression of ***C***
  (A and B are regulators of C)

- Bayesian networks

36