# Bayesian (Belief) Network Models,

2/10/03 & 2/12/03

# Outline of This Lecture

1. Overview of the model
2. Bayes Probability and Rules of Inference
   - Conditional Probabilities
   - Priors and posteriors
   - Joint distributions
3. Dependencies and Independencies
4. Bayesian Networks, Markov Assumption
5. Inference
6. Complexity of Representations: exponential vs. polynomial
7. (Friedman et al., 2000)
8. Equivalence Classes of Bayesian Networks
9. Learning Bayesian Networks

# Why Bayesian Networks

- Bayesian Nets are graphical (as in graph) representations of precise statistical relationships between entities
- They combine two very well developed scientific areas: Probability +Graph Theory
- Bayesian Nets are graphs where the nodes are random variables and the edges are directed causal relationships between them, A→B
- They are very high level qualitative models, making them a good match for gene networks modeling

# Bayesian Networks

**(1) An annotated <u>directed acyclic graph</u> G(V,E), where the nodes are random variables $X_i$,**
**(2) conditional distributions $P(X_i \mid \text{ancestors}(X_i))$ defined for each $X_i$.**

A Bayesian network uniquely specifies a joint distribution:

$$p(X) = \prod_{i=1}^{n} p(X_i \mid \text{ancestors}(X_i))$$

From the joint distribution one can do inferences, and choose likely causalities

# Learning the Network

- Given data we would like to come up with Bayesian Network(s) that fit that data well
- Algorithms exist that can do this efficiently (though the optimal ones are NP-complete)
- We'll discuss this later in this lecture...

# Choosing the Best Bayesian Network: Model Discrimination

- Many Bayesian Networks may model given data well
- In addition to the data fitting part, here we need to discriminate between the many models that fit the data
- Scoring function: Bayesian Likelihood
- More on this later...

# General Properties

- Fixed Topology (doesn’t change with time)
- Nodes: Random Variables
- Edges: Causal relationships
- DAGs
- Allow testing inferences from the model and the data

# 1. Bayes Probability and Rules of Inference

# Bayes Logic

- Given our knowledge that an event may have been the result of two or more causes occurring, what is the probability it occurred as a result of a particular cause?
- We would like to predict the unobserved, using our knowledge, i.e. assumptions, about things

# Conditional Probabilities

If two events, A and B are independent:

$$P(AB)=P(A)P(B)$$

If they are not independent:

$$P(B|A)=P(AB)/P(A)$$

or

$$P(AB)=P(B|A)*P(A)$$

# Bayes Formula

- *P(AB)=P(B|A)*P(A)*

Joint distributions

- *P(AB)=P(A|B)*P(B)*

Thus *P(B|A)*P(A)=P(A|B)*P(B)* and

$$\boxed{P(B|A)=P(A|B)*P(B)/P(A)}$$

If $B_i$, $i=1,\dots,n$ are mutually exclusive events, then

Possible causes

Prior probabilities (assumptions)

Posterior

$$P(B_i\,|\,A)=\frac{P(A\,|\,B_i)\cdot P(B_i)}{\sum_{j=1}^{n} P(A\,|\,B_j)\cdot P(B_j)}$$

Observation

# Joint Probability

- The probability of all events: $P(AB)=P(A)*P(B|A)$ *or* $P(ABCD)=P(A)*P(B|A)*P(C|AB)*P(D|ABC)$

- For n variables it takes $2^n$ terms to write it out!

# Conditional Independencies

- Recall P(AB)=P(A)*P(B) A is independent of B
- Conditional Independency: A is independent of B, given C

  P(A;B|C) =P(A|C)*P(B|C)

**Markov Assumption:** Each variable is independent of its non-descendents, given its parents

Bayesian Networks implicitly encode the Markov assumption. The joint probability becomes:

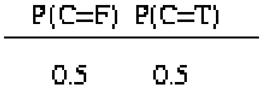
$$p(X) = \prod_{i=1}^{n} p(X_i \mid \text{ancestors}(X_i))$$

joint conditionals

Notice that if the ancestors (fan in) are bound by k, the complexity of this joint becomes $n2^{k+1}$

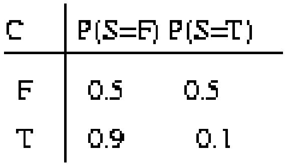
# Example

KP Murphy’s web site:
http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html

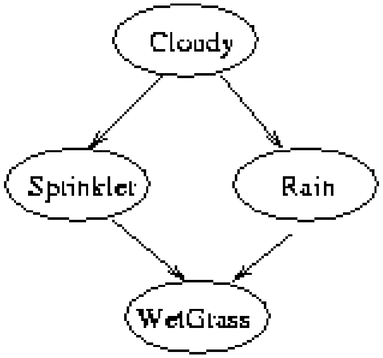


| P(C=F) | P(C=T) |
|---|---|
| 0.5 | 0.5 |

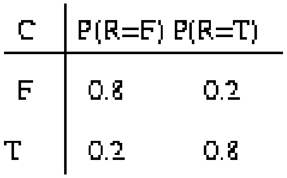
P(S|C)



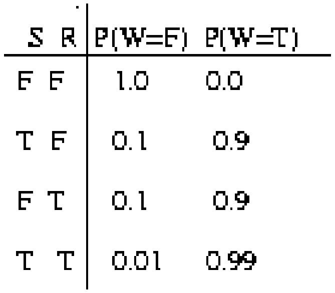
| C | P(S=F) | P(S=T) |
|---|---|---|
| F | 0.5 | 0.5 |
| T | 0.9 | 0.1 |



P(R|C)



| C | P(R=F) | P(R=T) |
|---|---|---|
| F | 0.8 | 0.2 |
| T | 0.2 | 0.8 |



| S | R | P(W=F) | P(W=T) |
|---|---|---|---|
| F | F | 1.0 | 0.0 |
| T | F | 0.1 | 0.9 |
| F | T | 0.1 | 0.9 |
| T | T | 0.01 | 0.99 |

P(W|S,R)

**Joint:** *P(C,R,S,W)=P(C)*P(R|C)*P(S|C,R)*P(W|C,R,S)*

or *P(C,R,S,W)=P(C)*P(S|C)*P(R|C)*P(W|S,R)*

**Independencies**: ***I(S;R|C), I(R;S|C)***

**Dependencies:** *P(S|C), P(R|C), P(W|S,R)*

# Bayesian Inference

Which event is more likely, wet grass observed and it is because of

- sprinkler:

  $P(S=1|W=1)=P(S=1,W=1)/P(W=1)=0.430$

- rain:

  $P(R=1|W=1)=P(R=1,W=1)/P(W=1)=0.708$

<u>Algorithms exist that can answer such questions given the Bayesian Network</u>

# BN, Second Lecture

1. Applying Bayesian Networks to Microarray Data
2. Learning Causal Patterns: Causal Markov Assumption
3. Gene Expression Data

- time-series data (Friedman et al., 2000)
  - partial models
  - estimating statistical confidence
  - efficient learning algorithms
  - discretization
  - experimental results
- perturbation data (Pe'er et al., 2001)
  - ideal interventions
  - feature identification
  - reconstructing significant sub-networks
  - analysis of results

# Bayesian Networks and Expression Data

- Friedman et al., 2000
- Learned <u>pronounced features</u> of <u>equivalence classes</u> of Bayesian Networks from time-series measurements of microarray data
- Data set used: Spellman et al., 1998
  - Objective: Cell cycling genes
  - Yeast genome microarrays (6177 genes)
  - 76 observations at different time-points

# Equivalence Classes of Bayesian Networks

- A Bayesian Network G implies a set of independencies, I(G), in addition to the ones following from Markov assumption
- Two Bayesian Networks that have the same set of independencies are equivalent
- Example G: $X \rightarrow Y$ and G':$X \leftarrow Y$ are equivalent, since $I(G)=I(G')=\emptyset$

# Equivalence Classes

- v-structure: two directed edges converging into the same node, i.e. X→Z←Y
- Thm: Two graphs are equivalent iff their DAGs have the same underlying directed graphs and the same v-structures
- Graphs in an equivalence class can be represented simply by Partially Directed Graph, PDAG where
  - a directed edge, X→Y implies all members of the equivalence class contain that directed edge
  - an undirected edge, X—Y implies that some DAGs in the class contain X→Y and others X←Y.
- Given a DAG, a PDAG can be constructed efficiently

# Learning Bayesian Networks

- Problem: Given a training set D=($x_1$,$x_2$,...,$x_n$) of independent instances of the random variables ($X_1$,$X_2$,...,$X_n$), find a network G (or equivalence class of networks) that best matches D.
- A commonly used scoring function is the Bayesian Score which has some very nice properties.
- Finding G that maximizes the Bayesian Score is NP-hard; heuristics are used that perform well in practice.

# Scoring Bayesian Networks

- The Scoring Measure is based on Bayesian considerations
- The measure is the posterior probability of the graph, given the data:

$$S(G{:}D)=\log P(G|D)=\log P(D|G)+\log P(G)+C$$

- *P(D|G)* averages the probability of the data over all possible parametric assignments to G (i.e. all posible cond. probabilities)
- What is important: Good choice of priors makes this scoring metric S(G:D) have nice properties:
  - graphs that capture the exact properties of the network very likely score higher than ones that do not
  - Score is decomposable

# Optimizing S(G:D)

- Once the priors are specified, and the data is given, the Bayesian Network is learned, i.e. the network with the highest score is chosen
- But Maxiizing this scoring function is an NP-hard problem
- Heuristics: local search in the form of arc reversals yield good results

# Closer to the Goal: Causal Networks

1. We want "<u>A is a cause for B</u>"
2. We have "<u>B independent of non-descendants given A</u>"

- So, we want to get from the second to the first, i.e. from Bayesian to stronger, causal networks
- Difference between Causal and Bayesian Networks: X→Y and X←Y are equivalent Bayesian Nets, but very different causally
- Causal Networks can be interpreted as Bayesian if we make another assumption
- <u>Causal Markov Assumption</u>: given the values of a variable's immediate causes, it is independent of its earlier causes (Example: Genetic Pedigree)
- Rule of thumb: In a PDAG equivalence class, X→Y can be interpreted as a causal link

# Putting it All Together: Expression Data Analysis

- Random Variables denote expression levels of genes
- Closed biological system
- The result is a joint probability distribution over all random variables
- The joint can be used to answer queries:
  - Does the gene depend on the experimental conditions?
  - Is this dependence direct or not?
  - If it is indirect, which genes mediate the dependence?

# Putting it all Together: Issues

In learning such a model the following issues come up:

1. interpreting the results: what do they mean?
2. algorithmic complexities in learning from the data
3. choice of local probability models

# Dimensionality Curse

- Again, we are hurt by having many more genes than observations (6200 vs. 20)
- Instead of trying to learn a model that explains the whole data the authors attempt to characterize features common to high-scoring models
- The intuition is that preserved features in many high-scoring networks are biologically important
- They call these models Partial Models

# Partial Models

- We do this because the data set has a small number of samples (compare to under-determined in linear models!)
- A small data set is not sufficient to determine that a single model is the “right” one (no one model dominates!)
- Idea: Compare highest scoring models for features common to all of them
- Simple features considered: pair-wise relations

# Partial Model Features

- Markov Relations
  - Is Y in the Markov Blanket of X?
  - Markov Blanket is the minimal set of variables that shield X from the rest of the variables in the model
  - Formally, X is independent from the rest of the network given the blanket
  - It can be shown that X and Y are either directly linked or share parenthood of a node
  - In biological context, a MR indicates that X and Y are related in some joint process
- Order Relations
  - Is X an ancestor of Y in all networks of a given class?
  - An indication of causality!

# 1. Are the Features Trustworthy?

- To what extent does the data support a given feature?
- The authors develop a measure of confidence in features as the likelihood that a given feature is actually true
- Confidence is estimated by generating slightly “perturbed” versions of the original data set and learning from them
- Thus, any false positives should disappear if the features are truly strong
- This is the case in their experiments

# 2. Learning Algorithms Complexity

- The solution space for all these problems is huge: super-exponential
- Thus some additional simplification is needed
- Assumption: Number of parents of a node is limited
- Trick: Initial guesses for the parents of a node are genes whose temporal curves cluster well

# 3. Local Probability Models

- The authors experimented with two different models: multinomial and linear gaussian
- These models are chosen for mathematical convenience
- Pros et cons:
  - Former needs discretization of the data. Gene expression levels are {-1,0,1}. Can capture combinatorial effects
  - Latter can take continuous data, but can only detect linear or close to linear dependencies

# Data and Methods

- Data set used: Spellman et al., 1998
  - Objective: Cell cycling genes
  - Yeast genome microarrays (6177 genes)
  - 76 observations at different time-points
- They ended up using 800 genes (250 for some experiments)
- Learned features with both the multinomial and the linear gaussian probability models
- They used no prior knowledge, only the data

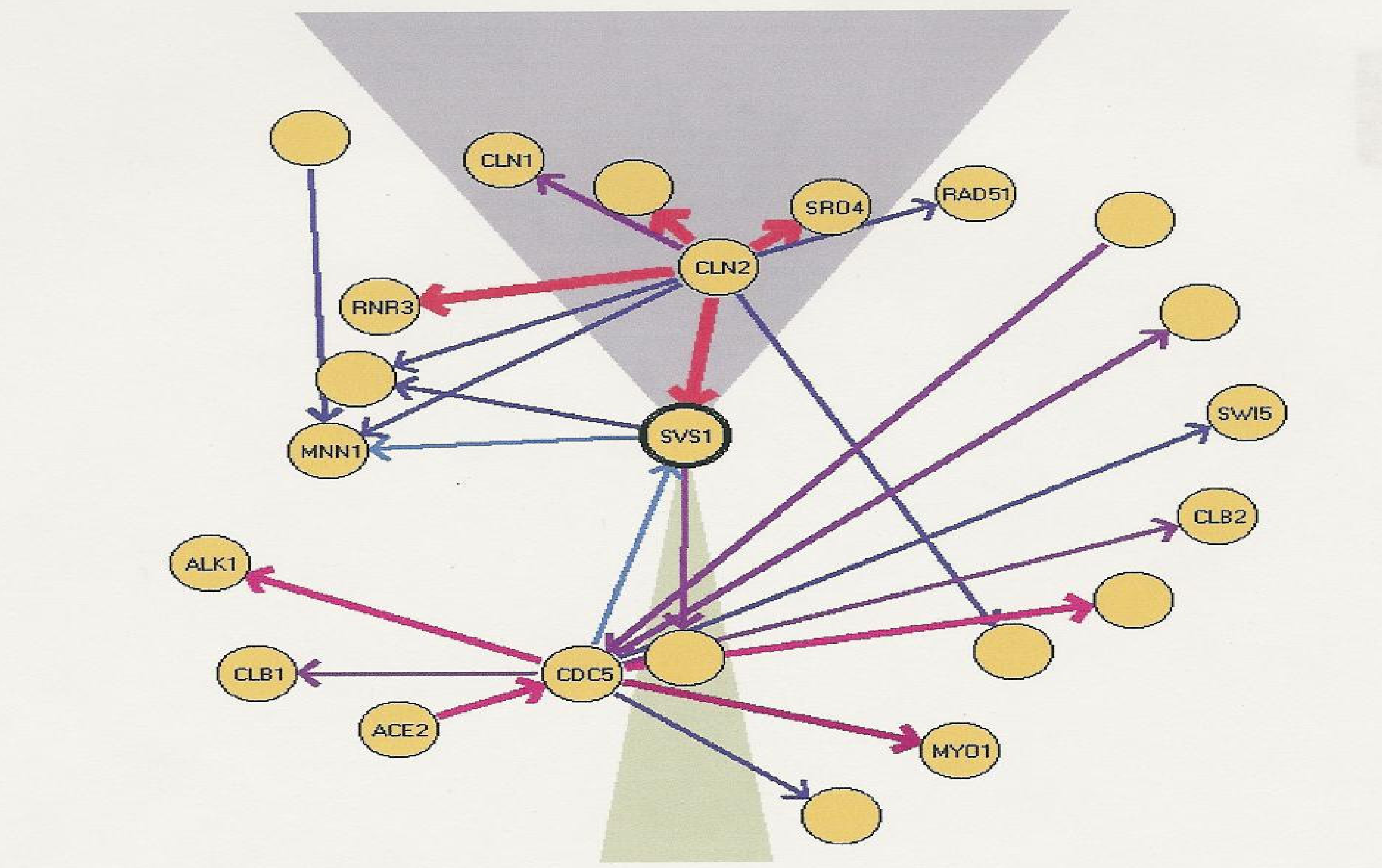


Figure 2: An example of the graphical display of Markov features. This graph shows a "local map" for the gene SVS1. The width (and color) of edges corresponds to the computed confidence level. An edge is directed if there is a sufficiently high confidence in the order between the genes connected by the edge. This local map shows that CLN2 separates SVS1 from several other genes. Although there is a strong connection between CLN2 to all these genes, there are no other edges connecting them. This indicates that, with high confidence, these genes are conditionally independent given the expression level of CLN2.

# Robustness Analysis: Thresholds

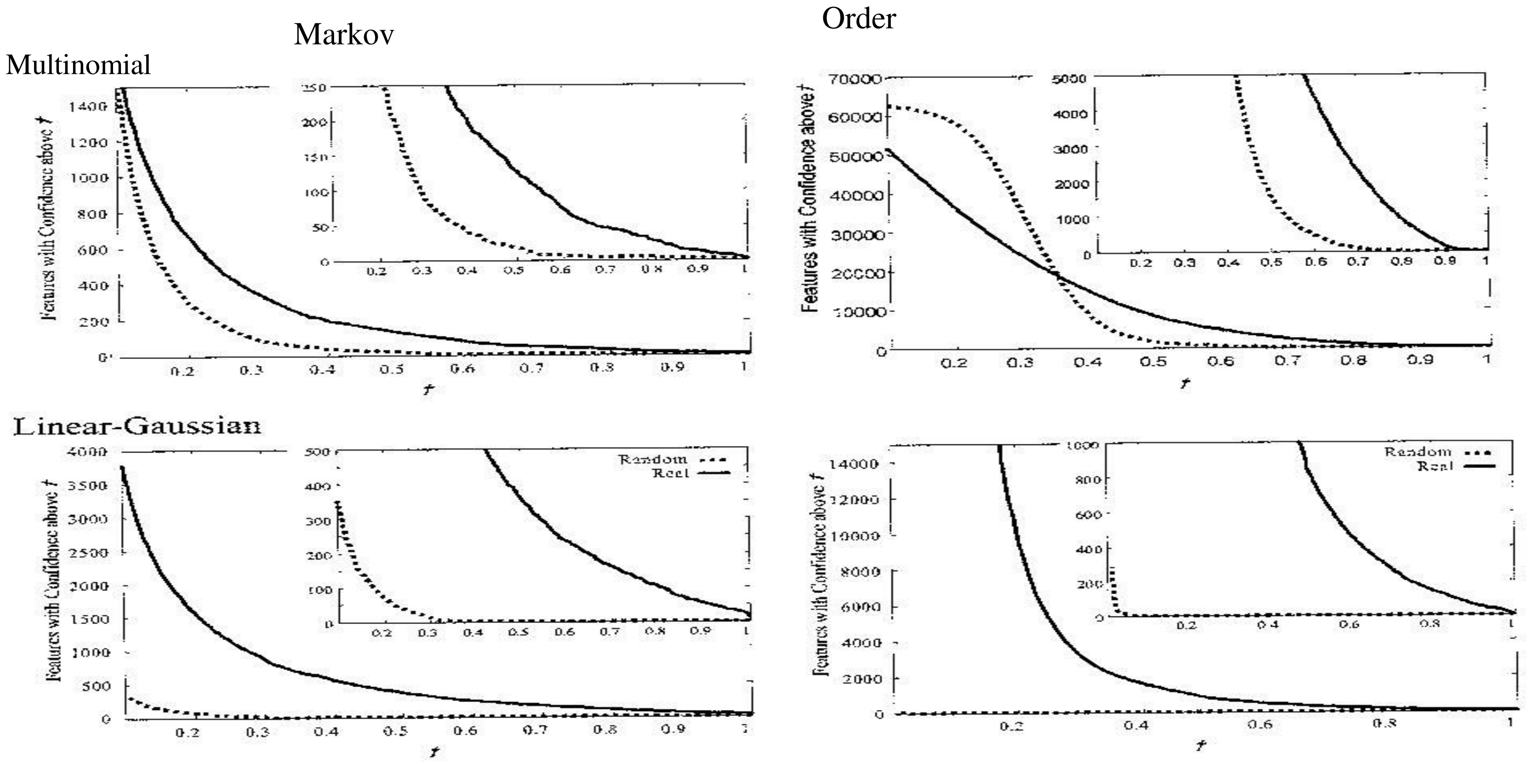
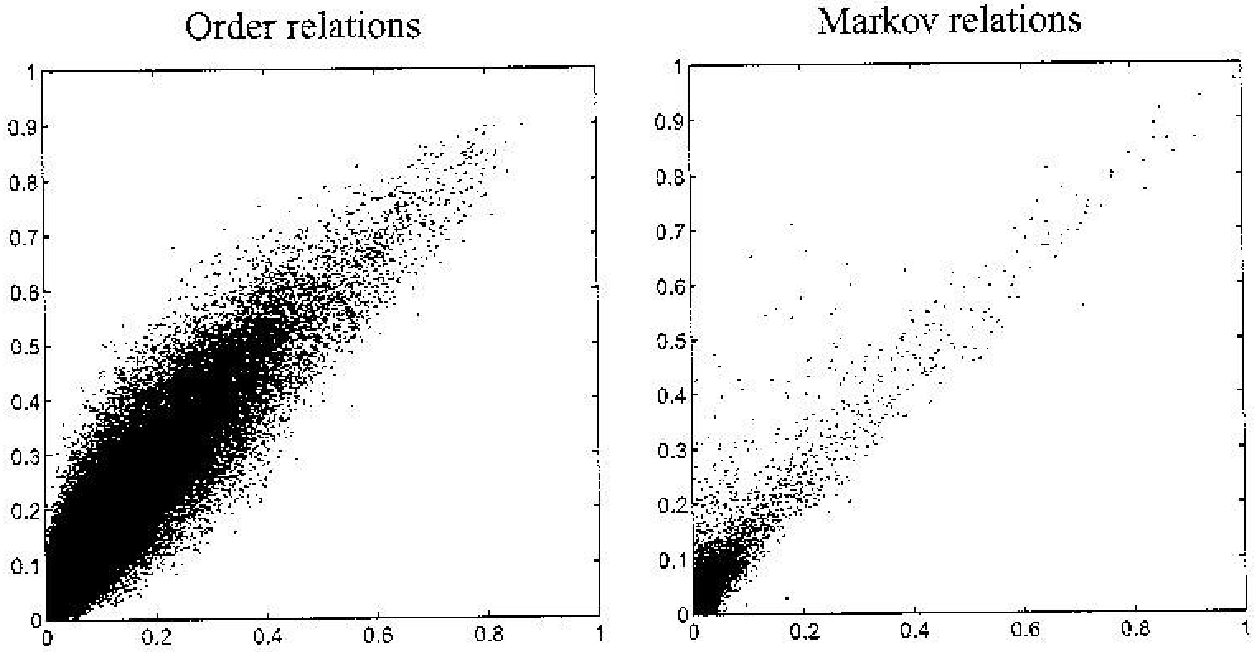


Figure 3: Plots of abundance of features with different confidence levels for the cell cycle data set (solid line), and the randomized data set (dotted line). The $x$-axis denotes the confidence threshold, and the $y$-axis denotes the number of features with confidence equal or higher than the corresponding $x$-value. The graphs on the left column show Markov features, and the ones on the right column show Order features. The top row describes features found using the multinomial model, and the bottom row describes features found by the linear-Gaussian model. Inset in each graph is plot of the tail of the distribution.

Comparing results on real vs. random data

# Robustness: Scaling



Order relations

Markov relations

Figure 4: Comparison of confidence levels obtained in two datasets differing in the number of genes, on the multinomial experiment. Each relation is shown as a point, with the $x$-coordinate being its confidence in the the 250 genes data set and the $y$-coordinate the confidence in the 800 genes data set. The left figure shows order relation features, and the right figure shows Markov relation features.

# Robustness: Discrete vs. Linear

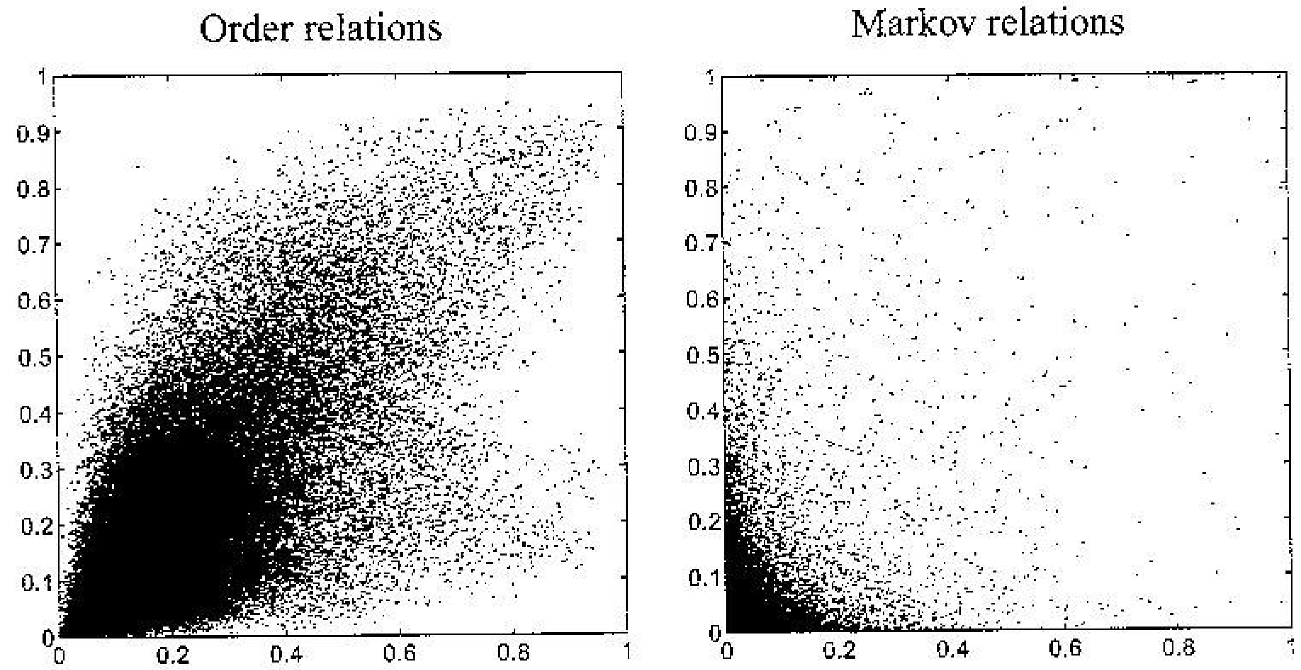


Figure 5: Comparison of of confidence levels between the multinomial experiment and the linear-Gaussian experiment. Each relation is shown as a point, with the $x$-coordinate being its confidence in the multinomial experiment, and the $y$-coordinate its confidence in the linear-Gaussian experiment. The left figure shows order relation features, and the right figure shows Markov relation features.

# Biological Analysis

- Order relations and Markov relations yield different significant pairs of genes
- Order relations: strikingly pronounced dominant genes, with many interesting known, or even key properties for cell functions
- Markov relations: all top pairs of known importance, some found beyond the reach of clustering (see CLN2 fig. for example)

Table 1: List of dominant genes in the ordering relations. Included are the top 10 dominant genes for each experiments.

| | Score in Experiment | | |
|---|---|---|---|
| Gene/ORF | Multinomial | Gaussian | Notes |
| MCD1 | 550 | 525 | Mitotic Chromosome Determinant,null mutant is inviable |
| MSH6 | 292 | 508 | Required for mismatch repair in mitosis and meiosis |
| CSI2 | 444 | 497 | cell wall maintenance, chitin synthesis |
| CLN2 | 497 | 454 | Role in cell cycle START, null mutant exhibits G1 arrest |
| YLR183C | 551 | 448 | Contains forkheaded associated domain, thus possibly nuclear |
| RFA2 | 456 | 423 | Involved in nucleotide excision repair, null mutant is inviable |
| RSR1 | 352 | 395 | GTP-binding protein of the RAS family involved in bud site selection |
| CDC45 | - | 394 | Required for initiation of chromosomal replication, null mutant lethal |
| RAD53 | 60 | 383 | Cell cycle control, checkpoint function, null mutant lethal |
| CDC5 | 209 | 353 | Cell cycle control, required for exit from mitosis, null mutant lethal |
| POL30 | 376 | 321 | Required for DNA replication and repair, null mutant is inviable |
| YOX1 | 400 | 291 | Homeodomain protein |
| SRO4 | 463 | 239 | Involved in cellular polarization during budding |
| CLN1 | 324 | - | Role in cell cycle START, null mutant exhibits G1 arrest |
| YBR089W | 298 | - | |

Table 2: List of top Markov relations, multinomial experiment.

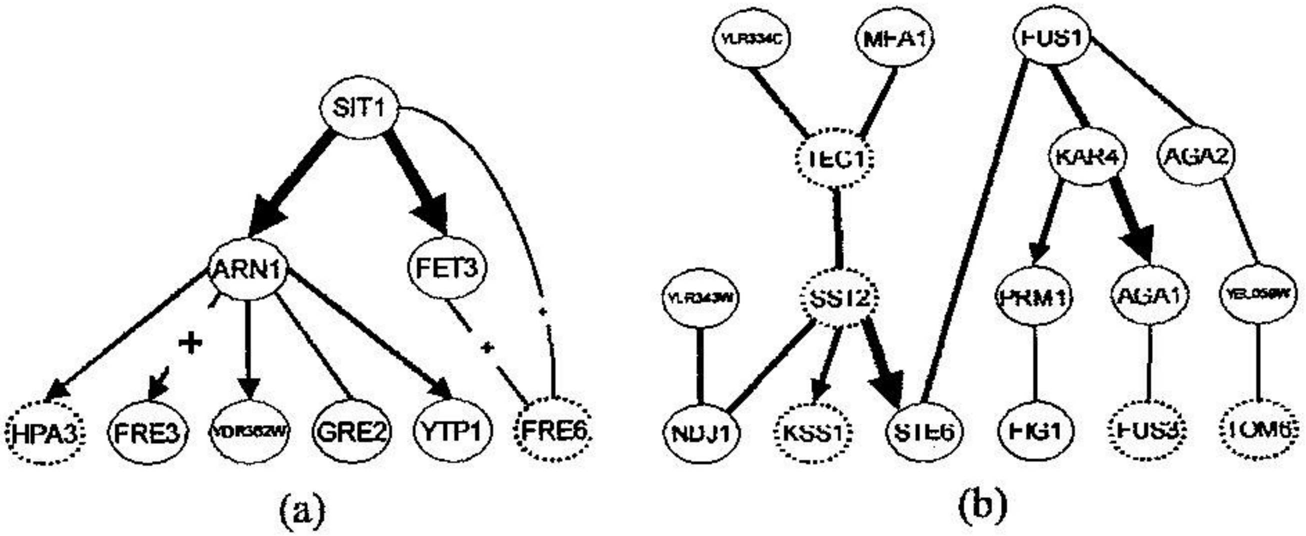
| Confidence | Gene 1 | Gene 2 | Notes |
|---|---|---|---|
| 1.0 | YKL163W-PIR3 | YKL164C-PIR1 | Close locality on chromosome |
| 0.985 | PRY2 | YKR012C | Close locality on chromosome |
| 0.985 | MCD1 | MSH6 | Both bind to DNA during mitosis |
| 0.98 | PHO11 | PHO12 | Both nearly identical acid phosphatases |
| 0.975 | HHT1 | HTB1 | Both are Histones |
| 0.97 | HTB2 | HTA1 | Both are Histones |
| 0.94 | YNL057W | YNL058C | Close locality on chromosome |
| 0.94 | YHR143W | CTS1 | Homolog to EGT2 cell wall control, both involved in Cytokinesis |
| 0.92 | YOR263C | YOR264W | Close locality on chromosome |
| 0.91 | YGR086 | SIC1 | Homolog to mammalian nuclear ran protein, both involved in nuclear function |
| 0.9 | FAR1 | ASH1 | Both part of a mating type switch, **expression uncorrelated** |
| 0.89 | CLN2 | SVS1 | Function of SVS1 unknown |
| 0.88 | YDR033W | NCE2 | Homolog to transmembrane proteins suggest both involved in protein secretion |
| 0.86 | STE2 | MFA2 | A mating factor and receptor |
| 0.85 | HHF1 | HHF2 | Both are Histones |
| 0.85 | MET10 | ECM17 | Both are sulfite reductases |
| 0.85 | CDC9 | RAD27 | Both participate in Okazaki fragment processing |

# Bayesian Networks and Perturbation Data (Pe'er et al. 2001)

- Similar study as above, but on a different, and bigger data set.
- Hughes et al. 2000
  - 6000+ genes in yeast
  - 300 full-genome perturbation experiments
    - 276 deletion mutants
    - 11 tetracycline regulatable alleles of essential genes
    - 13 chemically treated yeast cultures
- Pe'er et al. chose 565 significantly differentially expressed genes in at least 4 profiles

# Results

- Biologically meaningful pathways learned from the data!



Iron homeostasis

Mating response

Read the paper.....

# Limitations

- Bayesian Networks:
  - Causal vs. Bayesian Networks
  - What are the edges really telling us?
  - Dependent on choice of priors
  - Simplifications at every stage of the pipeline: analysis impossible
- Friedman et al. approach:
  - They did what they knew how to do: priors and other things chosen for convenience
  - Very little meaningful biology
  - Do we need all that machinery if what they discovered are only the very strong signals?

# References:

- Friedman et al., Using Bayesian Networks to Analyze Expression Data, RECOMB 2000, 127-135.
- Pe’er et al., Inferring Subnetworks from Perturbed Expression Profiles, Bioinformatics, v.1, 2001, 1-9.
- Ron Shamir's course, Analysis of Gene Expression Data, DNA Chips and Gene Networks, at Tel Aviv University, lecture 10 http://www.math.tau.ac.il/~rshamir/ge/02/ge02.html
- Spellman et al., Mol. Bio. Cell, v. 9, 3273-3297, 1998
- Hughes et al., Cell, v. 102, 109-26, 2000