Using Bayesian Networks to Analyze Expression Data

Xu Siwei, s0789023
Muhammad Ali Faisal, s0677834
Tejal Joshi, s0677858
Outline

- Introduction
- Bayesian Networks
- Equivalence Classes
- Applying to Expression data
- Experimental Analysis
- Discussion & Future work
Biological Background

- Recently developed technologies (e.g. DNA microarrays) allow parallel measurement of the expression level of thousands of genes/proteins.
- This allows biologists to view the cell as a complete system.
Challenge

- Extracting meaningful information from the expression data:
  - Infer regulatory mechanisms
  - Reveal function of proteins
- Experiment planning
Prior Work

- Clustering of expression data
  - Groups together genes with similar expression patterns
  - Disadvantage: does not reveal structural relations between genes

- Boolean Network
  - Deterministic models of the logical interactions between genes
  - Disadvantage: Deterministic, impractical for real data
Outline

- Introduction
- **Bayesian Networks**
- Equivalence Classes
- Applying to Expression data
- Experimental Analysis
- Discussion & Future work
Bayesian Networks

- A BN is a graphical representation of a joint probability distribution
- Advantages:
  - Useful for describing processes composed of locally interacting components
  - Statistical foundations for learning BN from observations, and computational algorithms to do so are well understood and successfully used
  - Provide models of causal influence
  - Capable of handling noise and estimating the confidence in the different feature of the networks.
Each variable is described as a stochastic function of its parents.

- The biological processes we want to model are stochastic.
- The measurements of the underlying biological system are noisy.

Figure 1: An example of a simple Bayesian network structure.

This network structure implies several conditional independence statements: $I(A; E)$, $I(B; D | A, E)$, $I(C; A, D, E | B)$, $I(D; B, C, E | A)$, and $I(E; A, D)$.

The network structure also implies that the joint distribution has the product form

$$P(A, B, C, D, E) = P(A)P(B | A, E)P(C | B)P(D | A)P(E)$$
Representing Distributions with BN

- **Representation**
  - $G$, a directed acyclic graph
  - Describes a conditional distribution on each variable, given $G$
- **Markov Assumption:**
- **Product form:** $\forall \ i, \ I(X_i; \ NonDescendents(X_i) \mid Pa(X_i))$
- **Generally:** $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$. BNs can accommodate many forms of conditional distribution, including various continuous models.
Learning BNs

- Common approach
- Efficient algorithms exist for learning a BN from data
- Learning a BN can:
  - Reveal underlying structure of domain
  - Direct relations between variable
  - Find causal influence
  - Discover hidden variables
Learning Causal Patterns

- Why?
- The difference between causal network and BN: Causal network interprets the parents of a variable as its immediate cause
- Causal Markov Assumption: given the values of a variable’s immediate causes, it is independent of its earlier causes
- $X \rightarrow Y \not\equiv X \leftarrow Y$
Outline

- Introduction
- Bayesian Networks
- **Equivalence Classes**
- Applying to Expression data
- Experimental Analysis
- Discussion & Future work
Equivalence Classes

Likelihood Score \((G:D)\) for 1 and 2 is equal

which one to choose ??

Joint distribution for 1 and 2:

\[ P(A,B) = P(B|A)P(A) = P(A|B)P(B) \]

Same underlying undirected graph
Equivalent Classes... cont

- For equivalent networks
- DAGs have the same underlying undirected graphs
- PDAGS are used to represent them
Outline

- Introduction
- Bayesian Networks
- Equivalence Classes
- Applying to Expression data
- Experimental Analysis
- Discussion & Future work
Applying to expression data

E.g.: Flow cytometry experiments

Images taken from: (Werhli, et al., 2006 Comparative Analysis of Graphical Modelling techniques: BioSS Scotland)
Applying to expression data

 Possibly completely unknown

 Machine Learning

 Statistical methods

 Images taken from: (Werhli, et al., 2006 Comparative Analysis of Graphical Modelling techniques: BioSS Scotland)
Applying to expression data

• Model expression levels of genes as random variable
• Random var. are the nodes of BN
• These include
  – Genes expression measurements
  – Experimental conditions
  – Clinical procedures
• Issues: Order of 10,000s of genes but few samples
Issues

<table>
<thead>
<tr>
<th>n</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>#DAGs</td>
<td>543</td>
<td>3,7 · 10^6</td>
<td>7,8 · 10^{11}</td>
<td>4,2 · 10^{18}</td>
</tr>
</tbody>
</table>

Genetic networks are sparse:
Set of plausible networks instead of a single best model

Images taken from: (Husmeier, et al., Probabilistic Modelling in Bioinformatics and Medical Informatics, 2005)
Solution: Feature Modelling

• Learn **Common features** in plausible networks instead of predicting the whole regulatory networks

• 2 different “Common Features”
  - Markov Relations (Biological interaction process)
    • Local Property
  - Order Relations (Causality)
    • Global Property
Feature 1: Markov Relations

• Is Gene Y in the Markov Blanket of X
  – X only depends upon set of genes in the blanket
• How do we identify them?
  – A direct edge between X and Y    OR
  – Both are parents of another variable/gene
Feature 2: Order Relations

- Is X an ancestor of Y
- Path from X to Y in which all edges are directed
- Does this mean X is a cause of Y
  - Yes provided Causal Markov assumption holds
    - Given X immediate causes it is independent of earlier causes
  - The assumption does not always hold
  - Thus causality is only an indication
Applying to Expression Data

• An optimum graph... is it possible?
  – Equivalent Classes
• Mapping to expression data
  – A broader picture and issues
• Is the result Significant
  – Feature modelling
• Quantifying the significance
  – Use bootstrap method
• Can we improve Complexity?
  – Heuristics: Sparse candidate algorithm
Revisiting... Significance

Genetic networks are sparse: Set of plausible networks instead of a single best model

<table>
<thead>
<tr>
<th>n</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>#DAGs</td>
<td>543</td>
<td>3,7 \times 10^6</td>
<td>7,8 \times 10^{11}</td>
<td>4,2 \times 10^{18}</td>
</tr>
</tbody>
</table>

Images taken from: (Husmeier, et al., Probabilistic Modelling in Bioinformatics and Medical Informatics, 2005)
Quantifying significance

• Ideally we should compute

\[ P(f \mid D) = \sum_{G} f(G)P(G \mid D) \]

\[ f(G) = 1 \text{ if } G \text{ satisfies the feature} \]

\[ 0 \text{ otherwise} \]

• But plausible G are v. high

• A Heuristic - Bootstrap method (Efron et al, 1993)

  for i = 1 to m (m=200)
  Re sample with replacement, N instances from D – D_i
  Apply learning procedure on D_i to produce G_i

  for each feature

  confidence \( C(f) = 1/m \sum_{i=1}^{m} f(G_i) \)
Improving Learning

- BN Optimization problem .... Huge search space...

- Sparse Candidate algorithm (Friedman, et al 1999)
  
  Identify candidate parents for each gene i: \( \text{Ci} = \{ Y_1 \ldots Y_k \} \)
  
  Restrict search to candidate parents only
  
  Problem....
  
  early choices lead to overly restricted search space
  
  Solution
  
  use iterative mechanism: \( n \) iterations
  
  Iteration 1: \( \text{Ci} \) leads to BN1 ...
  
  Iteration 2: check \( \text{Pa(Xi)} \) redefine \( \text{Ci} \)
  
  \( \text{Ci} = \text{previous parents} + \text{new parents} \)
  
  So that Score (BN2) increases
Outline

- Introduction
- Bayesian Networks
- Equivalence Classes
- Applying to Expression data
- Experimental Analysis
- Discussion & Future work
Experimental Analysis

- Dataset
- Robustness analysis
- Biological Analysis
Dataset

- 800 genes of S. Cerevisiae (Yeast) at various cell-cycle stages
- Each gene has 76 expression levels
- Discretized gene expression values according to threshold
  - Under expressed (-1), Normal (0), Over expressed (1)
- Bayesian Network having 800 + a root variable (representing cell-cycle phase)
- Sparse candidate algorithm with 200-fold bootstrap
- Computation of confidence in predictions
Robustness Analysis

• Testing credibility of confidence estimates
  – Perturbations in dataset
  – Addition of genes in dataset
  – Change of threshold
Robustness Analysis

- Order of experiments permuted independently for each gene
- Genes become independent of each other
- Order and Markov relations have significantly lower confidence than the original dataset
- High confidence near 0 => Sparser networks on original dataset

Friedman et al: Using Bayesian networks to analyze expression data. 1999
Robustness Analysis

Addition of more genes to the dataset

• Confidence levels of learnt features of 250 Vs 800 genes dataset – linear correlation

Change of threshold

• Linear relation in the confidence estimates of features between different discretization threshold

Friedman et al: Using Bayesian networks to analyze expression data. 1999
Biological Analysis

(Order relations)

• Discovers dominant genes
• Dominance score of \( X = \sum_{Y, C_o(X,Y) > t} C_o(X,Y)^k \)

where \( C_o(X,Y) \) is confidence in \( X \) being ancestor of \( Y \), constant \( k \) rewards high confidence features, \( t \) discards low confidence ones

• Potential causal sources in cell-cycle process.
• Dominated genes also important in replication and transcription regulation
Image from Friedman et al: Using Bayesian networks to analyze expression data. 1999
<table>
<thead>
<tr>
<th>Gene/ORF</th>
<th>Multinomial</th>
<th>Gaussian</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCD1</td>
<td>550</td>
<td>525</td>
<td>Mitotic Chromosome Determinant, null mutant is inviable</td>
</tr>
<tr>
<td>MSH6</td>
<td>292</td>
<td>508</td>
<td>Required for mismatch repair in mitosis and meiosis</td>
</tr>
<tr>
<td>CS12</td>
<td>444</td>
<td>497</td>
<td>Cell wall maintenance, chitin synthesis</td>
</tr>
<tr>
<td>CLN2</td>
<td>497</td>
<td>454</td>
<td>Role in cell cycle START, null mutant exhibits G1 arrest</td>
</tr>
<tr>
<td>YLR183C</td>
<td>551</td>
<td>448</td>
<td>Contains forkhead associated domain, thus possibly nuclear</td>
</tr>
<tr>
<td>RFA2</td>
<td>456</td>
<td>423</td>
<td>Involved in nucleotide excision repair, null mutant is inviable</td>
</tr>
<tr>
<td>RSR1</td>
<td>352</td>
<td>395</td>
<td>GTP-binding protein of the RAS family involved in bud site selection</td>
</tr>
<tr>
<td>CDC45</td>
<td>—</td>
<td>394</td>
<td>Required for initiation of chromosomal replication, null mutant lethal</td>
</tr>
<tr>
<td>RAD53</td>
<td>60</td>
<td>383</td>
<td>Cell cycle control, checkpoint function, null mutant lethal</td>
</tr>
<tr>
<td>CDC5</td>
<td>209</td>
<td>353</td>
<td>Cell cycle control, required for exit from mitosis, null mutant lethal</td>
</tr>
<tr>
<td>POL30</td>
<td>376</td>
<td>321</td>
<td>Required for DNA replication and repair, null mutant is inviable</td>
</tr>
<tr>
<td>YOX1</td>
<td>400</td>
<td>291</td>
<td>Homeodomain protein</td>
</tr>
<tr>
<td>SRO4</td>
<td>463</td>
<td>239</td>
<td>Involved in cellular polarization during budding</td>
</tr>
<tr>
<td>CLN1</td>
<td>324</td>
<td>—</td>
<td>Role in cell cycle START, null mutant exhibits G1 arrest</td>
</tr>
<tr>
<td>YBR089W</td>
<td>298</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>
## Dominant Genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN1</td>
<td>Role in cell cycle start</td>
</tr>
<tr>
<td>CLN2</td>
<td>Role in cell cycle start</td>
</tr>
<tr>
<td>CDC5</td>
<td>Cell cycle control, required for exit from mitosis</td>
</tr>
<tr>
<td>RAD53</td>
<td>Cell cycle control: checkpoint function</td>
</tr>
<tr>
<td>RFA2</td>
<td>Involved in nucleotide excision repair</td>
</tr>
<tr>
<td>PLO30</td>
<td>Required for DNA replication and repair</td>
</tr>
<tr>
<td>MSH6</td>
<td>Required for mismatch repair in mitosis and meiosis</td>
</tr>
</tbody>
</table>
Biological Analysis
(Markov relations)

Pairs of genes related in joint biological interaction

1) Pairs of functionally related known genes (50% of pairs)
2) Unknown-known pairs – Firm homologies between known and other unknown gene
3) Unknown gene pairs were physically adjacent to each other on the chromosome, and so regulated by same mechanism
## Markov Relations

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>YKL163W-PIR3</td>
<td>YKL164C-PIR1</td>
<td>Close locality on chromosome</td>
</tr>
<tr>
<td>0.985</td>
<td>PRY2</td>
<td>YKR012C</td>
<td>No homolog found</td>
</tr>
<tr>
<td>0.985</td>
<td>MCD1</td>
<td>MSH6</td>
<td>Both bind to DNA during mitosis</td>
</tr>
<tr>
<td>0.985</td>
<td>PHO11</td>
<td>PHO12</td>
<td>Both nearly identical acid phosphatases</td>
</tr>
<tr>
<td>0.975</td>
<td>HHT1</td>
<td>HTB1</td>
<td>Both are Histones</td>
</tr>
<tr>
<td>0.97</td>
<td>HTB2</td>
<td>HTA1</td>
<td>Both are Histones</td>
</tr>
<tr>
<td>0.94</td>
<td>YNL057W</td>
<td>YNL058C</td>
<td>Close locality on chromosome</td>
</tr>
<tr>
<td>0.94</td>
<td>YHR143W</td>
<td>CTS1</td>
<td>Homolog to EGT2 cell wall control, both do cytokinesis</td>
</tr>
<tr>
<td>0.92</td>
<td>YOR263C</td>
<td>YOR264W</td>
<td>Close locality on chromosome</td>
</tr>
<tr>
<td>0.91</td>
<td>YGR086</td>
<td>SIC1</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>FAS1</td>
<td>ASH1</td>
<td>Both part of a mating type switch, <strong>expression uncorrelated</strong></td>
</tr>
<tr>
<td>0.89</td>
<td>CLN2</td>
<td>SVS1</td>
<td>Function of SVS1 unknown, possible regulation mediated through SWI6</td>
</tr>
<tr>
<td>0.88</td>
<td>YDR033W</td>
<td>NCE2</td>
<td>Homolog to transmembrane proteins, suggesting both involved in protein secretion</td>
</tr>
<tr>
<td>0.86</td>
<td>STE2</td>
<td>MFA2</td>
<td>A mating factor and receptor</td>
</tr>
<tr>
<td>0.85</td>
<td>HHF1</td>
<td>HHF2</td>
<td>Both are Histones</td>
</tr>
<tr>
<td>0.85</td>
<td>MET10</td>
<td>ECM17</td>
<td>Both are sulfite reductases</td>
</tr>
<tr>
<td>0.85</td>
<td>CDC9</td>
<td>RAD27</td>
<td>Both participate in Okazaki fragment processing</td>
</tr>
</tbody>
</table>

Results from Friedman et al: Using Bayesian networks to analyze expression data. 1999
Image from Friedman et al: Using Bayesian networks to analyze expression data. 1999
Summary

• Successful and novel application of Bayesian networks for gene expression analysis
• Discretized gene expression data, based on multinomial distribution
• Discovers biologically significant dependencies, gene interactions and positive correlation
• Overcomes clustering by discovering relationship between genes having lower correlation
• Uses no prior biological knowledge
Future Work

• Incorporation of biological knowledge
• Time phase as a root of network – how about time-series data? (Murphy et. al)
• Can we identify hidden causes of regulations - such as protein activation/inhibition?
• Learning causality based on interventions rather than observations (knock-out experiments) (Cooper & Yoo)
• Improvement of search heuristics and confidence estimation
Thank You