Learning genetic networks from gene expression data

Dr. Dirk Husmeier
Biomathematics & Statistics Scotland (BioSS)
JCMB, The King’s Buildings, Edinburgh EH9 3JZ
United Kingdom
http://www.bioss.ac.uk/~dirk
Paradigm Shift in Molecular Biology
Paradigm Shift in Molecular Biology

Pre-Genomic

- Reductionist (DNA or RNA or protein)
- Generally qualitative, non-numeric
- Hypothesis driven
Paradigm Shift in Molecular Biology

Pre-Genomic
- Reductionist (DNA or RNA or protein)
- Generally qualitative, non-numeric
- Hypothesis driven

Post-Genomic
- Holistic, systems approach: DNA and RNA and protein
- Quantitative, highly numeric
- Data driven
Paradigm Shift in Molecular Biology

**Pre-Genomic**
- Reductionist (DNA or RNA or protein)
- Generally qualitative, non-numeric
- Hypothesis driven

**Post-Genomic**
- Holistic, systems approach: DNA and RNA and protein
- Quantitative, highly numeric
- Data driven

⇒ Need for machine learning and statistics
Inferring genetic networks from microarray gene expression data
DNA → Transcription → mRNA → Translation → Protein

Microarrays
<table>
<thead>
<tr>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
</tr>
</tbody>
</table>

[Image of a grid with colors and annotations]
Advantage of clustering

Fast, computationally cheap
Advantage of clustering

Fast, computationally cheap

Shortcoming of clustering

It is NOT reverse engineering
oscillator

hysteretic

external ligand

ligand binding

switch

cascades
Reverse engineering

Learn the network structure from gene expression data.

Problem: Noise, sparse data
Bayesian networks

Probabilistic framework for robust inference of interactions in the presence of noise

Nir Friedman et al. (2000)
Journal of Computational Biology 7: 601-620
Outline of the talk

• Recapitulation: Bayesian networks
• Reverse engineering: Learning networks from data
• Application to the yeast cell cycle
• Estimating the accuracy of inference
Outline of the talk

- Recapitulation: Bayesian networks
- Reverse engineering: Learning networks from data
- Application to the yeast cell cycle
- Estimating the accuracy of inference
Nodes
Edges
Edges = directed
No directed cycles!
\[ P(A, B, C, D, E) = \prod_i P(\text{node}_i | \text{parents}_i) \]
\[ P(A, B, C, D, E) = P(A) \]
\[ P(A, B, C, D, E) = P(A)P(B|A) \]
\[ P(A, B, C, D, E) = P(A)P(B|A)P(C|A) \]
\[ P(A, B, C, D, E) = P(A)P(B|A)P(C|A)P(D|B, C) \]
\[ P(A, B, C, D, E) = \]
\[ P(A)P(B|A)P(C|A)P(D|B, C)P(E|D) \]
Biological interpretation
Initiation of cell (sub-)cycle
Mediation
Conditional independence relations
\[ P(A, B, C) = P(A|C') P(B|C) P(C) \]
\[ P(A, B, C) = P(A|C)P(B|C)P(C) \]

\[ P(A, B|C) = \frac{P(A, B, C)}{P(C)} = P(A|C)P(B|C) \]
\[ P(A, B, C) = P(A|C)P(B|C)P(C) \]

\[ P(A, B|C) = \frac{P(A, B, C)}{P(C)} = P(A|C)P(B|C) \]

But: \( P(A, B) \neq P(A)P(B) \)
Biological example

Yeast cell cycle

Nir Friedman et al. (2000)

Journal of Computational Biology 7: 601-620
Low osmolarity response genes

SLT2

- MAP kinase
- Rlm1p
- Swi4/6

Transcription factors

| SLT2          | Low osmolarity response genes |
Low osmolarity response genes

SLT2

MAP kinase

Rlm1p

Transcription factors

Swi4/6

---

SLT2

Low osmolarity response genes
Low osmolarity response genes

SLT2

MAP kinase

Rlm1p

Swi4/6

Transcription factors

Low osmolarity response genes
Low osmolarity response genes

SLT2

MAP kinase

Rlm1p

Swi4/6

Transcription factors

---

<table>
<thead>
<tr>
<th>SLT2</th>
<th>Low osmolarity response genes</th>
</tr>
</thead>
</table>
Low osmolarity response genes

SLT2

MAP kinase

Rlm1p

Swi4/6

Transcription factors

Low osmolarity response genes
Outline of the talk

• Recapitulation: Bayesian networks
• **Reverse engineering:** Learning networks from data
• Application to the yeast cell cycle
• Estimating the accuracy of inference
Classical learning paradigm
Classical learning paradigm

Find the best network structure $M$:

$$M^* = \text{argmax}\{P(M|D)\}$$
Classical learning paradigm

Find the best network structure $M$:

$$M^* = \arg\max \{P(M|D)\}$$

Find the best parameters $\theta^*$

$$\theta^* = \arg\max \{P(\theta|D, M^*)\}$$
Find the best model $M$, that is, the best network

$$P(M|D) \propto P(D|M)P(M)$$
Find the best model $M$, that is, the best network

$$P(M|D) \propto P(D|M)P(M)$$

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

When is the integral \textit{analytically} tractable?
Find the best model $M$, that is, the best network

$$P(M|D) \propto P(D|M)P(M)$$

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

When is the integral analytically tractable?

• Complete observation: No missing values.

• $P(D|\theta, M)$ and $P(\theta|M)$ must satisfy certain regularity conditions.

• Examples: Multimodal with a Dirchlet prior, linear Gaussian with a normal-gamma prior.
Naive approach

- Compute $P(M|D)$ for all possible network structures $M$.
- Select network structure $M^*$ that maximizes $P(M|D)$.
Naive approach

- Compute $P(M|D)$ for all possible network structures $M$.
- Select network structure $M^*$ that maximizes $P(M|D)$.

**Problem 1:**

Number of different network structures increases super-exponentially with the number of nodes.

<table>
<thead>
<tr>
<th>N of nodes</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of structures</td>
<td>3</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>

$\rightarrow$ Optimization problem intractable for large N of nodes.
Naive approach

- Compute $P(M|D)$ for all possible network structures $M$.
- Select network structure $M^*$ that maximizes $P(M|D)$.

Problem 2:
Data are sparse $\rightarrow$ Intrinsic uncertainty of inference

Large data set $D$: Best network structure $M^*$ well defined

Small data set $D$: Intrinsic uncertainty about $M^*$
Objective: Sample from the posterior distribution

\[ P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_i P(D|M_i)P(M_i)} \]

Direct approach intractable due to \( \sum_i P(D|M_i)P(M_i) \)
Objective: Sample from the posterior distribution

\[ P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_i P(D|M_i)P(M_i)} \]

Direct approach intractable due to \( \sum_i P(D|M_i)P(M_i) \)

Markov chain Monte Carlo (MCMC):

- **Proposal move**: Given network \( M_{\text{old}} \), propose a new network \( M_{\text{new}} \) with probability \( Q(M_{\text{new}}|M_{\text{old}}) \).

- **Acceptance/Rejection**: Accept this new network with probability

\[
\min \left\{ 1, \frac{P(D|M_{\text{new}})P(M_{\text{new}})}{P(D|M_{\text{old}})P(M_{\text{old}})} \times \frac{Q(M_{\text{old}}|M_{\text{new}})}{Q(M_{\text{new}}|M_{\text{old}})} \right\}
\]
Markov chain Monte Carlo (MCMC)

Accept move with probability: \[
\min \left\{ 1, \frac{P(M_{\text{new}}|D)}{P(M_{\text{old}}|D)} \times \frac{Q(M_{\text{old}}|M_{\text{new}})}{Q(M_{\text{new}}|M_{\text{old}})} \right\}
\]
MCMC moves

- Delete edge
- Reverse edge
- Create edge
Convergence of MCMC simulation

**Burn-in:** $T$ MCMC steps

**Sampling:** $T$ MCMC steps

How large do we have to choose $T$?
Convergence of MCMC simulation

Burn-in: \( T \) MCMC steps

Sampling: \( T \) MCMC steps

How large do we have to choose \( T \)?

Repeat MCMC simulations from different initializations

Scatter plots of posterior probabilities of edges
MCMC simulation 1

MCMC simulation 2

MCMC 2
MCMC 1
MCMC 2
MCMC 1
MCMC 2
MCMC 1
MCMC simulation 1 MCMC simulation 2

T infinite
T too short
T long enough
MCMC steps: 10,000 (left) versus 100,000 (right)
The structure prior $P(\mathcal{M})$
Fan-out unrestricted

Fan-in restricted

not permissible
$P(\mathcal{M})$ uniform over structures

or

$P(\mathcal{M})$ uniform over cardinalities of parent sets
\[ P(\mathcal{M}) \text{ uniform over structures} \]

or

\[ P(\mathcal{M}) \text{ uniform over cardinalities of parent sets} \]

Study with Marco Grzegorczyk and Wolfgang Urfer

- 100 genes
- 60 kidney cancer patients
- MCMC: 100,000 Metropolis-Hastings steps
$P(\mathcal{M})$ uniform over structures
$P(M)$ uniform over cardinalities
Problem: Statistical significance of the networks

- **Complex models**: Transcript levels of hundreds of genes.
- **Sparse data**: Typically a few dozen samples.

- Posterior probability $P(M|D)$ diffuse: Global network inference is meaningless.
Solution: Focus on features and subnetworks

Feature: Indicator variable for a property of interest, e.g.: Are X and Y close neighbours in the network?

\[ f(M) = \begin{cases} 
1 & \text{if } M \text{ satisfies the feature} \\ 
0 & \text{otherwise} 
\end{cases} \]
Solution: Focus on features and subnetworks

Feature: Indicator variable for a property of interest, e.g.: Are X and Y close neighbours in the network?

\[ f(M) = \begin{cases} 
1 & \text{if } M \text{ satisfies the feature} \\
0 & \text{otherwise} 
\end{cases} \]

Posterior probability of features: \( P(f|D) = \sum_M f(M)P(M|D) \)

assumed to be sufficiently informative.
Model network, data set size: $N = 50$
Model network, data set size: $N = 50$
Outline of the talk

- Recapitulation: Bayesian networks
- Reverse engineering: Learning networks from data
- Application to the yeast cell cycle
- Estimating the accuracy of inference
Experimental Results

- **Friedman, Linial, Nachman, Pe'er (2000)**
  Journal of Computational Biology 7 (3/4): 601-620

- **Pe'er, Regev, Elidan, Friedman (2001)**
  Bioinformatics S1: 215-224

  Molecular Biology of the Cell 9 (12) :3273-97
• Yeast cell cycle (*S. cerevisiae*).

• Six time series under different experimental conditions, altogether 76 gene expression measurements.

• 800 genes.

• No biological prior knowledge.

• Do not take into account the temporal aspect of the measurements. Introduce an additional root node representing the cell cycle phase.

• Discretization: Underexpressed (-1), normal (0), overexpressed (1).
Order relations

- Is A an ancestor of B in all the networks of a given equivalence class?
- Does the network contain a directed path from A to B?

Indication that A might be a causal ancestor of B.
Order relations

Confidence in $X$ being an ancestor of $Y$:
$$P(X \rightarrow Y|D)$$

Dominance score of $X$: $\sum_Y P(X \rightarrow Y|D)$

Genes with high dominance scores are indicative of potential causal sources of the cell cycle process.
Order relations

Confidence in $X$ being an ancestor of $Y$:

$$P(X \rightarrow Y|D)$$

Dominance score of $X$:

$$\sum_Y P(X \rightarrow Y|D)$$

Genes with high dominance scores are indicative of potential causal sources of the cell cycle process.

Finding: Only a few genes dominate the order.
Dominant genes in the ordering relations

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

DNA repair is associated with transcription initiation: DNA areas which are more active in transcription are also repaired more frequently.
Markov neighbours

- Variables that are not separated by any other measured variable in the domain.

- Indication that two genes are related in some *joint biological interaction or process*.

- **Parent-child**: One gene regulating another.

- **Spouse relations**: Two genes co-regulating another.
Markov relations

$P(X \leftrightarrow Y|D)$: Indication that genes are functionally related.

- Most Markov pairs: *Intracluster pairings* with high correlation in their expression.

- **But**: Genes where $P(X \leftrightarrow Y|D)$ is high and correlation is low.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAR1</td>
<td>Role in a mating type switch</td>
</tr>
<tr>
<td>ASH1</td>
<td>Role in a mating type switch</td>
</tr>
<tr>
<td>LAC1</td>
<td>GPI transport protein</td>
</tr>
<tr>
<td>YNL300W</td>
<td>Modified by GPI</td>
</tr>
<tr>
<td>SAG1</td>
<td>Induces the mating process</td>
</tr>
<tr>
<td>MF-ALPHA-1</td>
<td>Participates in the mating process</td>
</tr>
</tbody>
</table>
Separator relations

and

subnetworks
Low osmolarity response genes

SLT2

MAP kinase
Rlm1p
Swi4/6
Transcription factors

| SLT2 | Low osmolarity response genes |
Low osmolarity response genes

SLT2

MAP kinase

Rlm1p

Swi4/6

Transcription factors

Low osmolarity response genes
Outline of the talk

- Recapitulation: Bayesian networks
- Reverse engineering: Learning networks from data
- Application to the yeast cell cycle
- Estimating the accuracy of inference
Dirk Husmeier (2003)
Bioinformatics 19, 2271-2282
Gene A → Gene C → Gene B → Data

generate
Gene A → Gene C → Gene B

Gene A → Gene C

Gene A → Gene C

Data: learning → Data

Data: generating → Data
Deterministic inference
Probabilistic inference
Thresholding
Gene A
Gene B
Gene C
Gene A
Gene B
Gene C

Data
compare
learn
generate

Thresholding

True positives
False positives
Data: binary

Model:

Parameters:

Noisy boolean: $P \in \{0.1, 0.9\}$
ROC curve: Sample size = 3
ROC curve: Sample size = 6
ROC curve: Sample size = 12
Disadvantage:
Unrealistic, no mismatch between the model used for data generation and the model used for inference.
Transcription factors

Promoter

mRNA

Protein
Transcription factors

Promoter

Transcription

mRNA

Protein
Transcription factors

Promoter

Transcription

mRNA

Translation

Protein
Transcription factors

Promoter

Transcription

mRNA

Translation

Protein

Dimerization

Undimerization
\[
\frac{d}{dt}[a2.rC] = \lambda^+_{a2.rC}[a_2][rC] - \lambda^-_{a2.rC}[a_2.rC]
\]
\[
\frac{d}{dt}[C] = \lambda_{rC}[rC] + \lambda_{a2.rC}[a_2.rC] + \lambda_{b2.rC}[b_2.rC] - \lambda_{C}[C]
\]
\[
\frac{d}{dt}[c] = \lambda_{Cc}[C] - \lambda_{c}[c]
\]
\[
\frac{d}{dt}[c_2] = \lambda^+_{cc}[c]^2 - \lambda^-_{cc}[c_2]
\]
Zak et al., Int. Conf. Sys. Biol., 2001
Sampling Discretization

12 time points

Recover the true genetic network with reverse engineering.
Simulation Experiments

Ligand injection for 10 minutes.

Equilibrium

12 data points collected over 4000 min in equi-distant intervals.

Disequilibrium

12 data points collected over 500 min in equi-distant intervals.
ROC curve: Equilibrium

True positives (vertical axis) ←→ False positives (horizontal axis)
ROC curve: Disequilibrium

True positives (vertical axis) $\leftrightarrow$ False positives (horizontal axis)
Structure Prior

Max fan-in = 2, 3, 4
ROC curves

True positives (vertical axis) ←→ False positives (horizontal axis)

Left: max fan-in = 2  
Right: max fan-in = 3
ROC curves

True positives (vertical axis) ←→ False positives (horizontal axis)

Left: \( \text{max fan-in} = 2 \)  
Right: \( \text{max fan-in} = 4 \)
Sequence information

\[
P(\frac{y \rightarrow rX | r \in B[y])}{P(y \rightarrow rX | r \notin B[y])} = 2
\]

\(y \rightarrow rX\) denotes the event that transcription factor \(y\) binds to the promoter \(r\) upstream of gene \(X\), and \(B[y]\) is the set of (known) binding motifs for \(y\).

In words: The equation expresses that on identifying a binding motif for transcription factor \(y\) in the upstream region of gene \(X\), this transcription factor is twice as likely to bind to \(X\) than in the absence of such a motif.
ROC curves

True positives (vertical axis) ↔ False positives (horizontal axis)
Left: max fan-in = 2
Right: max fan-in = 3
Conclusions
Conclusions

• Learning the global network $\rightarrow$ impossible

• Intrinsic uncertainty due to lack of data
Conclusions

- Learning the **global network** $\rightarrow$ impossible
- Intrinsic **uncertainty** due to lack of data
- Inference **local substructures** possible, but obscured by noise.
Conclusions

- Learning the global network → impossible
- Intrinsic uncertainty due to lack of data
- Inference local substructures possible, but obscured by noise.
- Biologically realistic priors important
- Integrating post-genomic data.
DNA → Transcription → mRNA → Translation → Protein → Microarrays
Acknowledgments

Marco Grzegorczyk
Isabelle Grimmenstein
Wolfgang Urfer

Dortmund University