Regulatory networks and signalling pathways

From Sachs et al Science 2005

Flow cytometry technology
possibly completely unknown

E.g.: Flow cytometry experiments
Here: Concentrations of (phosphorylated) proteins

Machine Learning statistical methods

Statistical Task
Extract a network from a data matrix

variables
\[ X_1, \ldots, X_n \]
genes/proteins

Either m independent (steady-state) observations of the system \( X_1, \ldots, X_N \)
Or time series of the system of length m: \( (X_1, \ldots, X_n)_{t=1, \ldots, m} \)
Elementary molecular biological processes

Transcription factors

a_2 Binding/unbinding

b_2

rC

Promoter

C

C_2

mRNA

Translation

Degradation

Transcription

Dimerization

Undimerization

Protein

Rates

Concentrations

Kinetic parameters q

Parameters q known: Numerically integrate the differential equations for different hypothetical networks

Given: Gene expression time series

Can we infer the correct gene regulatory network?

Description with differential equations

\[
\frac{d}{dt}[a_2 rC] = \lambda_{a_2, rC}[a_2] rC - \lambda_{a_2, rC}^{-1} a_2 rC
\]

\[
\frac{d}{dt}[C] = rC[C] + \lambda_{a_2, rC} a_2 rC - \lambda_{b_2, rC} b_2 rC - \lambda[C]
\]

\[
\frac{d}{dt}[c] = \lambda_{b}[c] - \lambda_{c}[c]
\]

\[
\frac{d}{dt}[c_2] = \lambda_{c_2}[c]^2 - \lambda_{c_2}[c_2]
\]
Model selection for known parameters $q$

Measured gene expression time series predicted with different models

Gene expression time series

Compare

Highest likelihood: best model

$P(D|q, M)$

Model selection for unknown parameters $q$

Gene expression time series predicted with different models

Gene expression time series

Highest likelihood: over-fitting

$P(D|q, M)$

Model selection: find the best pathway

Select the model $M$ with the highest posterior probability:

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

This requires an integration of the whole parameter space:

$$P(D|M) = \int P(D|q, M)P(q|M)dq$$

This integral is usually intractable especially for systems of non-linear differential equations

Static Bayesian networks

- Marriage between graph theory and probability theory.
- Directed acyclic graph (DAG) represents conditional independence relations.
- Markov assumption leads to a factorization of the joint probability distribution:

$$P(A, B, C, D, E, F) = P(A) \cdot P(B|A) \cdot P(C|A) \cdot P(D|B, C) \cdot P(E|D) \cdot P(F|C, D)$$
Bayesian networks versus causal networks

Bayesian networks represent conditional (in)dependency relations - not necessarily causal interactions.

Equivalence classes of BNs

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

\[
P(A, B | C) \neq P(A | C) \cdot P(B | C)
\]

Dynamic Bayesian networks for time series data

No need for the acyclicity constraint!!!

Interpretation:

Unfolding in time

(a) folded (cyclic) network

(b) unfolded (acyclic) dynamic network

Figure 1: State space graph and corresponding dynamic Bayesian network. The left panel shows a recurrent state space graph containing two nodes. Node X has a recurrent feedback loop and acts as a regulator of node Y. The right panel shows the same graph unfolded in time to obtain a valid DAG; this is the corresponding dynamic Bayesian network.
Bayesian networks

\[ P(\text{graph} | \text{data}) = \frac{P(\text{data} | \text{graph}) \cdot P(\text{graph})}{P(\text{data})} \propto P(\text{data} | \text{graph}) \cdot P(\text{graph}) \]

\[ = P(\text{graph}) \cdot \int P(\text{data}, \theta(\text{graph}) | \text{graph}) d\theta(\text{graph}) \]

**Parameterisation:** Gaussian BGe scoring metric:

- Data \(\sim N(\mu, \Sigma)\)
- \(\mu \sim N(\mu^*, (vW)^{-1})\) and \(W \sim \text{Wishart}(T_0)\)

\[ X_j \sim N(\mu_j + \sum_{i=1}^{n} b_{ij} (x_i - \mu_i), \sigma_j^2) \quad b_{ij} \neq 0 \iff X_j \rightarrow X_i \]

Learning the network structure

\[ \text{graph} \rightarrow \text{score}_{\text{BGe}}(\text{graph}) \]

<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>

**Idea:** Heuristically searching for the graph \(M^*\) that is most supported by the data \(P(M^*|\text{data}) > P(\text{graph}|\text{data})\), e.g.: greedy search algorithm
**MCMC sampling of Bayesian networks**

**Better idea:** Bayesian model averaging via Markov Chain Monte Carlo (MCMC) simulations

Construct and simulate a Markov Chain \( (M_t) \) in the space of DAGs whose distribution converges to the graph posterior distribution as stationary distribution, i.e.:

\[
P(M_t = \text{graph}|\text{data}) \to P(\text{graph}|\text{data})
\]

\[ t \to \infty \]

to generate a DAG sample: \( G_1, G_2, G_3, \ldots G_T \)

**Structure MCMC sampling scheme**

(based on single edge operations)

![Structure MCMC sampling scheme diagram]

**Equivalence classes of BNs**

Markov chain Monte Carlo (MCMC)

- Proposal
- Accept
- Rejct

Acceptance probability: \( \min \left\{ 1, \frac{P(D|M_\text{new})}{P(D|M_\text{old})} \times \frac{P(M_\text{new})}{P(M_\text{old})} \times \frac{Q(M_\text{new}|M_\text{old})}{Q(M_\text{old}|M_\text{new})} \right\} \)

- \( P(A) \cdot P(B) \cdot P(C | A) \)
- \( P(A) \cdot P(B | C) \cdot P(C | A) \cdot P(A)^{-1} \)

- \( P(A | C) \cdot P(B | C) \cdot P(C) \)
- \( P(A | C) \cdot P(B | C | B) \)

- \( P(A, B) = P(A) \cdot P(B) \)
- \( P(A, B | C) = P(A | C) \cdot P(B | C) \)

completed partially directed graphs (CPDAGs)
Utilise the DAG (CPDAG) sample for estimating the posterior probability of edge relation features:

\[
\hat{p}(A \rightarrow B) = \frac{1}{T} \sum_{t=1}^{T} I(G_t)
\]

where \(I(G_t)\) is 1 if the CPDAG of \(G_t\) contains the directed edge \(A \rightarrow B\), and 0 otherwise.

In practice: \(t\) is not infinite!!!
Probabilistic inference

data

true regulatory network

data

tresholding

high

TP: 1/2
FP: 0/4

low

TP: 2/2
FP: 1/4

edge posterior probabilities

concrete network predictions