Genetische Netzwerke  
Wintersemester 2012/2013  
Lecture 1: 10-Oct-12  
Marco Grzegorczyk

Cell Biology

(Gen-)Regulatory Network

Microarray Chips

(Gen-)Regulatory Network

(Gen-)Regulatory Network

(Gen-)Regulatory Network
Regulatory networks and signalling pathways
From Sachs et al Science 2005

Flow cytometry technology

Protein activation cascade

Protein activation cascade
possibly completely unknown

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

Statistical Task

Extract a network from a data matrix
cells →
variables
$X_1, \ldots, X_n$
gen/ proteins

either $m$ independent (steady-state) observations of the system $X_1, \ldots, X_n$

or time series of the system of length $m$: $(X_1^{(t)}, \ldots, X_n^{(t)})_{t=1, \ldots, m}$

Elementary molecular biological processes

Description with differential equations

Concentrations

Kinetic parameters $q$

Rates
Given: Gene expression time series

Can we infer the correct gene regulatory network?

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

Parameters $q$ unknown: Gene expression time series predicted with different models

Model selection for known parameters $q$

Measured gene expression time series

Gene expression time series predicted with different models

Highest likelihood: best model $P(D|q,M)$

Model selection for unknown parameters $q$

Measured gene expression time series

Gene expression time series predicted with different models

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

Bayesian networks

Dynamic Bayesian networks for time series data

No need for the acyclicity constraint!!!

Interpretation:

Unfolding in time

Bayesian networks

Parameterisation: Gaussian BGe scoring metric:

Data ~ N(μ, Σ)

with normal-Wishart distribution of the (unknown) parameters, i.e.: μ ~ N(μ*, (vW)^-1) and W ~ Wishart(T_0)

BGe metric: closed form solution

y ≥ 0 ⇔ X_j → X_i

15.10.2012
Learning the network structure

graph $\rightarrow$ score_{BGe}(graph)

<table>
<thead>
<tr>
<th>n</th>
<th>#DAGs</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/8</td>
<td>543</td>
<td>3.7 $\cdot$ 10^6</td>
<td>7.8 $\cdot$ 10^11</td>
<td>4.2 $\cdot$ 10^{18}</td>
<td></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 $(M_t)_t$ in the space of DAGs whose distribution converges to the graph posterior distribution as stationary distribution, i.e.: $P(M_t=\text{graph}|\text{data}) \rightarrow P(\text{graph}|\text{data})$ $\quad t \rightarrow \infty$

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

**Structure MCMC sampling scheme**

(based on single edge operations)

**Equivalence classes of BNs**

1. $P(A)P(B)P(C|A) = P(A)P(B)P(C|B) = P(A)P(B)P(C)$
2. $P(A|B)P(B|C) = P(A|C)P(C|B)$
3. Completed partially directed graphs (CPDAGs)
Utilise the DAG (CPDAG) sample for estimating the posterior probability of edge relation features:

\[ P(A \rightarrow B) = \frac{1}{T} \sum_{i=1}^{T} I(G_i) \]

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

**MCMC sampling of Bayesian networks**

The DAG sample \( G_1, G_2, G_3, \ldots, G_T \) is generated via Markov Chain Monte Carlo (MCMC) simulations. That is, via simulation of a Markov Chain (\( M_t \)), in the space of DAGs whose distribution converges to the graph posterior distribution:

\[ P(M_t = \text{graph}|\text{data}) \rightarrow P(\text{graph}|\text{data}) \quad t \rightarrow \infty \]

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