Reconstructing Signaling Pathways with Probabilistic Boolean Threshold Networks

Lars Kaderali
ViroQuant Research Group „Modeling“
University of Heidelberg
Viruses rely on many host factors for cell entry, replication within the host cell, and spread.

RNAi knock-downs of host genes can help identify these factors:
- RNAi knockdown of genes in infected cells
- Observe whether virus can still replicate

Human T-cell lymphotropic virus with host cell
Encyclopaedia Britannica Online, 2007
A Pipeline for the Analysis of RNAi Screens


RNAi knockdowns are well suited to **identify** genes, that are important for specific phenotypic traits of interest.

The *temporal* and *spatial* placement of these genes in signal transduction *pathways* remains a huge challenge.

*Network Inference* is the process of reconstructing such pathways from the experimental data.

<table>
<thead>
<tr>
<th>Gene Knockdown</th>
<th>Observed Phenotypic Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>Strong Effect</td>
</tr>
<tr>
<td>Gene 2</td>
<td>No Effect</td>
</tr>
<tr>
<td>Gene 3</td>
<td>Weak Effect</td>
</tr>
<tr>
<td>Gene 4</td>
<td>Strong Effect</td>
</tr>
</tbody>
</table>
Network Inference from RNAi Data

- Experimental data differ in available readouts
- *Want general method that will run with missing observations, but improves when more data are available!*

<table>
<thead>
<tr>
<th>Gene Knockdown</th>
<th>Observation Gene 1 at timepoint 1</th>
<th>Observation Gene 2 at timepoint 1</th>
<th>Observation Gene 3 at timepoint 1</th>
<th>Observation Gene 4 at timepoint 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>Active</td>
<td>Active</td>
<td>Inactive</td>
<td>Inactive</td>
</tr>
<tr>
<td>Gene 2</td>
<td>Inactive</td>
<td>Inactive</td>
<td>Inactive</td>
<td>Active</td>
</tr>
<tr>
<td>Gene 3</td>
<td>Inactive</td>
<td>Active</td>
<td>Active</td>
<td>Active</td>
</tr>
<tr>
<td>Gene 4</td>
<td>Active</td>
<td>Inactive</td>
<td>Active</td>
<td>Active</td>
</tr>
</tbody>
</table>
Network Inference from RNAi Data

- Experimental data differ in available readouts
- \textit{Want general method that will run with missing observations, but improves when more data are available!}

<table>
<thead>
<tr>
<th>Gene Knockdown</th>
<th>Observation</th>
<th>Observation</th>
<th>Observation</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1 at timepoint 2</td>
<td>Active</td>
<td>Inactive</td>
<td>Active</td>
<td>Inactive</td>
</tr>
<tr>
<td>Gene 2 at timepoint 2</td>
<td>Active</td>
<td>Inactive</td>
<td>Inactive</td>
<td>Inactive</td>
</tr>
<tr>
<td>Gene 3 at timepoint 2</td>
<td>Active</td>
<td>Active</td>
<td>Inactive</td>
<td>Active</td>
</tr>
<tr>
<td>Gene 4 at timepoint 2</td>
<td>Active</td>
<td>Inactive</td>
<td>Active</td>
<td>Inactive</td>
</tr>
</tbody>
</table>
Complexity of Network Inference

- For \( n \) genes, there are \( n^2 \) different possible edges between two genes.
- In a given network, each of these \( n^2 \) edges is present or absent.
- This yields a total of \( 2^{n^2} \) possible, different network topologies.
- How much data is required to decide which is the true topology?

<table>
<thead>
<tr>
<th>( n )</th>
<th># Topologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>512</td>
</tr>
<tr>
<td>4</td>
<td>65,536</td>
</tr>
<tr>
<td>5</td>
<td>33,554,432</td>
</tr>
<tr>
<td>10</td>
<td>1,267,650,600,228,229,401,496,703,205,376</td>
</tr>
</tbody>
</table>
Iterative Network Reconstruction

Experiment

Regularization!

1
2
3
4
R

p=0.6

p=0.1

1
2
3
4
R

p=0.3

Candidate Models

Experiment Planning
Bayesian Network Model

» Each node is either „active“ (1) or „inactive“ (0)

» State of node at time $t$ depends stochastically on states of „parents“ at time $t-1$

Mathematical Model

$$p \{ x_i(t) = 1 | x(t-1) \} = \frac{1}{1 + \exp \left( - \sum_{j=1}^{n} w_{j,i} x_j + w_i^0 \right)}$$
State Transition Matrix

- For a system with $n$ nodes, there are $2^n$ possible states.

- If in state $i$ at time $t$, we can compute the probability of being in state $j$ at time $t+1$.

- Hence, we can calculate the state transition matrix $M \in \mathbb{R}^{2^n \times 2^n}$ as

$$M_{i,j} = p \left\{ x(t) = \eta^{(i)} | x(t-1) = \eta^{(j)} \right\}$$

$$= \prod_{k=1}^{n} p \left\{ x_k(t) = \text{active}(k, \eta^{(i)}) | x(t-1) = \eta^{(j)} \right\}$$
If $p$ is a $2^n$ Row-Vector giving the probability distribution over the initial states, then

$$p \cdot M$$

is the column Vector giving the distribution after 1 timestep.

Similarly,

$$p \cdot M^\tau$$

gives the distribution after $\tau$ timesteps.
• Knockouts can be taken into account simply by „taking out“ the corresponding gene from the model.

• In terms of M, this amounts to removing rows where the knockout gene is active, and summing up the corresponding columns.
• Assume we have an initial state distribution $p_0$.
• Given model Parameters $\theta = (w, w_0, T)$, the likelihood of seeing a particular set of experimental outcomes $D$ after knockdown experiments is

$$p \{ D| w, w^0, T \} = \prod_{k=1}^{n} p \left( \eta^{(k)}(T)|M^{-k}, p_0 \right)$$

<table>
<thead>
<tr>
<th>Gene Knockdown</th>
<th>Observed Phenotypic Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>Strong Effect</td>
</tr>
<tr>
<td>Gene 2</td>
<td>No Effect</td>
</tr>
<tr>
<td>Gene 3</td>
<td>Weak Effect</td>
</tr>
<tr>
<td>Gene 4</td>
<td>Strong Effect</td>
</tr>
</tbody>
</table>
• We cannot compute an exact likelihood $p(D | \theta)$ for „larger“ networks, because $M$ is growing exponentially.
• BUT we can use the stochastic model to simulate data, and compare the simulated data with the measured data!
• We then approximate the likelihood by the percentage of trials where we are getting the observed data back:

$$p \{ D | \theta \} \approx \frac{1}{N} \sum_{m=1}^{N} I (S_m = D)$$

• This is of particular usefulness since it automatically takes into account the marginalization over unobserved nodes.
Parameters $w$ in model correspond to strength of interaction between two genes / proteins.

$$p \{ x_i(t) = 1 | x(t-1) \} = \frac{1}{1 + \exp \left( -\sum_{j=1}^{n} w_{j,i} x_j + w_{i}^0 \right)}$$

Expect network to be *sparse*, i.e. most pathway components should have NO interaction between them.

$$p(w) = N \exp \left[ -\frac{|w|^q}{qs^q} \right]$$

Ritter et al., submitted
Sampling from the Posterior

B1 Initialize \( \theta(0) = (w(0), w^0(0), T(0)) \), \( t = 0 \)

B2 Sample \( \bar{\theta} \) from a proposal distribution \( q(\cdot | \theta(t)) \)

B3 Simulate a dataset \( S \) using the stochastic model described by equations (1) and (2), with parameters \( \bar{\theta} \).

B4 If \( S \neq D' \) (based only on the observed nodes in \( D' \)),
   let \( \theta(t + 1) = \theta(t) \), increase \( t \) and go to B2.

B5 Compute \( \gamma = \min \left( 1, \frac{\pi(\bar{\theta})q(\theta(t)|\bar{\theta})}{\pi(\theta(t))q(\theta(t)|\theta(t))} \right) \)

B6 Accept \( \theta(t + 1) = \bar{\theta} \) with probability \( \gamma \),
   otherwise stay at the old point \( \theta(t + 1) = \theta(t) \).

B7 Increase \( t \) and go to B2, until enough points sampled.

Combines Metropolis Hasting algorithm with simulation approximation of the likelihood.


We furthermore integrated Mode Hopping steps

Senderowitz (1995)
Combining genetic algorithm and Markov chain Monte Carlo

- A population of $N = mk$ Markov chains are divided equally into $k$ subpopulations
- Genetic operators, mutation, cross over, migration are used to generate next generation in each chain in each subpopulation
• Experimental measurements at different time points, but „real time“ is continuous!
• Model requires discrete time steps
• How many „model time“ steps between two experimental measurements?

➢ Sample additional parameter Delta_T!
Application: Jak-Stat Signaling

- Experimental Data: Eva Dazert (Dept. of Virology)
- Huh-7 cell lines
- Knockdown of all genes in the pathway, stimulation with IFNα and IFNγ
- Signal: HCV Replication
Jak / Stat Signaling

Kaderali et al., Bioinformatics, 2009
Method to reconstruct signal transduction networks from RNAi phenotypes based on Bayesian networks

Approximation of likelihood using stochastic simulation

Regularization to Sparse Networks using Prior Distribution

Sampling from posterior allows computation of distributions over alternative topologies and parameters.
  – Important application in experiment design
  – Cost efficient method to reconstruct networks from data

Application to Jak/Stat data shows core topology can be reconstructed even from single downstream readouts.

Multiple readouts, time series data, ... easily integrated
Acknowledgements

**Molecular Virology:**
Eva Dazert
Ulf Zeuge
Michael Frese
Ilka Wörz
Anil Kumar
Andreas Merz
Marion Pönisch
Marco Binder
Alessia Rugieri
Wolfgang Fischl
Oliver Wicht
Ralf Bartenschlager

**Viroquant Modeling:**
Narsis Kiani
Bettina Knapp
Matthias Boeck
Nora Rieber
Johanna Mazur
Daniel Ritter
Nurgazy Sulaimanov
Gajendra Suryavanshi
Samta Malhotra
Sandep Amberkar
Cindy Nürnberg
Natalia Drost
Thorsten Stumpf

**Dept. of Virology:**
Johannes Hermle
Kathleen Boerner
Maik Lehmann
Oliver Keppler
Silvia Geuenich
Hans-Georg Kräusslich

**Viroquant NWG „Screening“:**
Vytaute Starkuviene

**TBI Bioinformatics, DKFZ:**
Petr Matula
Karl Rohr
Roland Eils

**Inst. f. Scientific Computing:**
Christoph Sommer
Fred Hamprecht
Julian Kunkel
Gerhard Reinelt

**Tokyo Medical University:**
Soichi Ogishima

**EMBL / Bioquant:**
Nigel Brown
Reinhard Schneider

**Forschungseinheiten FORSYS**
der Systembiologie

**Bundesministerium**
für Bildung und Forschung
Thank you for your attention!

Lars Kaderali
Viroquant Research Group Modeling
Bioquant, University of Heidelberg

lars.kaderali@bioquant.uni-heidelberg.de
Identifiability

• If only downstream readouts at steady state are available, some topological features cannot be reconstructed!
Identifiability

![Graph showing the average number of acyclic networks per dataset against network size (number of nodes). The graph exhibits a clear upward trend as the network size increases.]
Identifiability

• Situation improves considerably, when
  – Observations of several genes are available
  – Several time points are available
  – Double or multiple knock-downs are available
  – Different Stimulations / Conditions are available

• Method should be adaptable for these cases!
A Pipeline for the Analysis of RNAi Screens

• siRNA Spotting
• Experiment
• Microscopy
• Image Recognition
• Quality Control
• Statistical Analysis
• Bioinformatics
• Modeling

HCV infection
fixation and IF
36 h
seeding Huh7.5 cells

\[ dT_v / dt = k_1 R_{in} R_{c} - k_2 T_v - \mu T_v \]
\[ dP / dt = k_1 T_v - k_2 P \]
\[ dE^{cov} / dt = k_1 P - k_{in} E^{cov} - \mu E^{cov} \]