Network Reconstruction & Network Analysis

Ulf Leser
Content

- Network reconstruction
  - Boolean models
  - Correlation-Based Approaches: REVEAL / ARACNE
  - Example
- Quantitative network analysis
Networks

How do we know?
What does the network tell us?
Approaches to Network Reconstruction

• By many, many small-scale experiments
• By mathematical modeling from high-throughput data sets

• By evolutionary inference from model organisms
• By curation from the literature (see first bullet)
Reconstruction from Indirect High-Throughput Data

- Network reconstruction, re-engineering, inference, ...

- Idea: Derive network from indirect observations
  - **Network**: Links and their effect (strength, activation, …)
    - We usually assume the players (genes, metabolites, …) to be given
  - **Observation**: High-throughput measurements
    - Here: Transcriptome, microarrays, RNA-Seq
  - **Indirect**: We try to infer mechanistic causality by correlation

- Dynamic networks
  - Nodes get states (active / passive)
  - Current states determine future states of nodes
  - Leads to dynamic behavior

- Warning: All current methods are highly reductionist
Boolean Network Models

• Definition

A **Boolean Network** is a digraph $G=(V,E)$ where

- Every node has an associated Boolean state (on/off)
- Every node is labeled with a **Boolean function over the states** of all incoming nodes

• Usage

- Vertices = genes
- Edge $(v,w)$ models an effect of $v$ on $w$
- The state of a node $v$ is determined by its Boolean function over all “incoming” states
- Simplistic: No cofactors, no cellular context, no binding affinity, no time, no kinetics, …

![Diagram of a Boolean network with nodes A, B, C and functions $f_A$, $f_B$, $f_C$.]
Network Dynamics

• Definition
  A *Dynamic Boolean Network* (DBN) is a Boolean network where every node \( v \) is assigned a sequence of states \( v_0, v_1, v_2, \ldots \) such that the state of \( v_t \) is defined over the Boolean function of \( v \) applied to the states \( w_{t-1} \) of all incoming nodes \( w \)

• Remarks
  - Models the state of every gene (on / off) over time
  - States at time point \( t \) (only) depend on states at time point \( t-1 \)
    - No buffering, synchronized time, …
  - **Deterministic**: Given all states at any time point \( t \) and the Boolean functions, any state at any later time point can be uniquely determined
Example

Transition table

Source: Filkov, „Modeling Gene Regulation“, 2003
Example

\[ f_A(B) = B \]
\[ f_B(A, C) = A \text{ and } C \]
\[ f_C(A) = \text{not } A \]

Boolean Network

<table>
<thead>
<tr>
<th>genes time</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Network Analysis

- Many things can be analyzed using DBN
- For instance, an attractor is a (set of) states towards which the network state converges
  - Point attractor: State which cannot be left any more
  - Cyclic attractor: A series of states which will repeat forever
  - Probability of attractors depend largely on size of network and complexity of Boolean functions
- Skipped – we want to reconstruct networks
Network Reconstruction

- Assume we know all genes, but not their relationships.
- Assume we observe the states of $n$ genes over $m$ time points (a matrix $S$; the observations).
- Can we re-engineer the Boolean function of every gene given a sequence of states?

<table>
<thead>
<tr>
<th>genes time</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Network Reconstruction

- Assume we know all genes, but not their relationships
- Assume we observe the states of n genes over m time points (a matrix S; the observations)
- Can we re-engineer the Boolean function of every gene given a sequence of states?

\[ f(A) = \neg B \]
\[ f(B) = A \land \neg B \]
\[ f(C) = B \]

<table>
<thead>
<tr>
<th>genes time</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Formal Problem

• Definition
  Let $S_t$, $0 \leq t \leq m$, be the vector of all observed states of all genes $V$ at time point $t$. A DBN $G=(V,E)$ with functions $f_1, \ldots f_n$, $n=|V|$, is called
  - consistent with $S_t$ iff $S_t = [f_1(S_{t-1}), f_2(S_{t-1}), \ldots f_n(S_{t-1})]$
  - consistent with $S$ iff it is consistent for all $S_t$, $1 \leq t \leq m$

• The Boolean network reconstruction problem
  Given an observation $S$ over a set $V$, find a DBN $G=(V,E)$ that is consistent with $S$.

• Remark
  - Reconstruction means finding the functions $f_1, \ldots f_n$
  - This also determines network topology (nodes appearing in a $f_i$)
Solutions

• Clearly, there are many observations $S$ for which no consistent $G$ exists
  - Recall that DBN are deterministic
  - Imagine $S_t$, $S_{t+1}$ and $S_u$, $S_{u+1}$ with $S_t=S_u$ but $S_{t+1} \neq S_{u+1}$

• Also, there are many observation $S$ for which more than one consistent $G$ exists

• Every time point narrows the options for $G$ – the longer $S$, the less (or no) consistent $G$’s exist
Optimal Networks

- **Definition**
  - *For a DBN G, let size(G) be the total number of variables (edges) appearing in the Boolean functions of G*
  - *A DBN $G$ is minimal for observation $S$, if $G$ is consistent with $S$ and there is no $G'$ which is also consistent with $S$ and size($G'$) < size($G$)*

- **Remark**
  - **Parsimony assumption**: Small models are better
  - Thus, the smallest network is the best – functions are as simple as possible, nothing is inferred that is not enforced by the data
  - Not necessarily unique
Naïve Algorithm

N = V;
for k=1...n  # length of functions
    for every n in N  # all unexplained nodes
        test all functions f of size k for n on S;
        if f is consistent for n on S
            N := N \ n;  # n is explained
            Add f to network;
        end if;
    end for;
end for;

- Exhaustive algorithm for finding minimal networks
- **Very complex** (AND, OR, NOT, no paranthesis)
  - k=1: 2n functions
  - k=2: 2*2n*2n=O(n^2) functions
  - ...
  - General: O(2^{2k-1}*n^k) functions
Pros and Cons

- **Application (transcriptome data)**
  - Perform time-series gene expression experiments
  - Brutally discretize each measurement: Genes are on or off
  - Reconstruct DBN

- **Pros: Simple**

- **Cons**
  - Binary values are not capturing reality
  - Synchronized, clocked time is nonsense
  - No quantification (It needs 2*A and one B to regulate C)
  - Only small networks are computable
  - …
Content

- Network reconstruction
  - Boolean models
  - Correlation-Based Approaches: REVEAL / ARACNE
  - Example
- Quantitative network analysis
Towards Reality

• There are less complex & more robust algorithms

• REVEAL replaces Boolean functions by mutual information; correlations rather than deterministic switching

• ARACNE is even simpler: Only removal of some (presumably indirect) correlations
Foundations

- **Definition**
  
  *Let X, Y be two discrete random variables. The mutual information \( MI(X,Y) \) is defined as*

  \[
  MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \log \left( \frac{p(x,y)}{p(x)*p(y)} \right)
  \]

- **Remark**
  
  - Measure the variable’s mutual dependency
  - Dependency: Deviation of \( p(X,Y) \) from \( p(X)\cdot p(Y) \)
  - How much does the state of X determines the state of Y?
  - Many similar measures (information gain, conditional entropy, cross entropy, …)
Example

\[ MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \times \log \left( \frac{p(x, y)}{p(x) \times p(y)} \right) \]

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
p(x,y) & y=0 & y=1 \\
\hline
\hline
x=0; p(x=0)=0.2 & 0,12 & 0,08 \\
\hline
x=1; p(x=1)=0.8 & 0,48 & 0,32 \\
\hline
\end{tabular}
\end{table}

\[ MI(X, Y) = 0 \]

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
p(x,y) & y=0 & y=1 \\
\hline
\hline
x=0; p(x=0)=0.2 & 0,18 & 0,03 \\
\hline
x=1; p(x=1)=0.8 & 0,05 & 0,74 \\
\hline
\end{tabular}
\end{table}

\[ MI(X, Y) = 0,53 \]
Two more Facts

- With a little math, we find
  \[ \text{MI}(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \]
  - \( H(X) \): Entropy of \( X \)
  - \( H(X|Y) \): Conditional entropy of \( X \) given \( Y \)

- It follows that the maximal value of \( \text{MI}(X,Y) = H(X) \) (\( H(Y) \))
  - \( H(X|Y) = 0 \), which means that \( X \) (\( Y \)) completely determines \( Y \) (\( X \))

- MI can be extended to sets of three, four, ... variables
  - Like Boolean functions over three, four, ... variables
  - Multivariate mutual information
REVEAL

N = V;
for k=1...n  # number of nodes/variables
    for every X in N  # all unexplained nodes
        find subset T=(Y₁,...Yₖ) with MI(X,Y₁,...Yₖ) = H(X);
        if T exists
            N := N \ X;  # n is explained
        end for;
    end for;
end for;

• Again, we have observations of n genes at m time points
  - Or m different conditions, treatments, …
• Again, we discretize expression values to 0 or 1
  - More bins are possible
• MI(X,Y) means looking at pairs (x₁,y₀), (x₂,y₁), …
REVEAL in Practice

- In the formulation given, REVEAL would be as strict as Boolean functions
  - Dependencies must be perfect
- In the presence of noise, one must be satisfied with almost maximal MI
  - I.e., $|\text{MI}(X,Y)-H(X)| < \varepsilon$
- Can be extended to more than two possible states
  - Less strict discretization, more realistic model
- Most other restrictions of DBN remain
ARACNE

- **Fast** variation of REVEAL which (a) considers each pair in isolation and (b) gives up model minimality

- **Idea**
  - Compute mutual information between all pairs of genes
    - This gives a complete network
  - Remove edges where $|\text{MI}(X,Y)-\text{H}(X)| > \varepsilon$
    - $\varepsilon$ can be estimated from the distribution of MI – created at random?
  - Remove certain *indirect effects* (“data processing inequalities”)

- **Under certain assumptions, ARACNE provably converges to the true network**
  - Given unlimited input, no loops
  - “True”: Under all networks obeying our simplifying assumptions
Imagine with strong effects of A on B and B on C; will appear as

But if we find which edge most probably is an artifact?

- Assumption: If $\text{MI}(X,Z) \leq \min(\text{MI}(X,Y),\text{MI}(Y,Z))$, then the correlation between $X$-$Z$ is an indirect effect and removed
- Procedural: In every triangle, remove the smallest edge
  - But in which order should triangles be visited?
Content

- Network reconstruction
  - Boolean models
  - Correlation-Based Approaches: REVEAL/ ARACNE
    - Example
- Quantitative network analysis
Reconstructing the Mammalian Clock

- DA Sven Lund, 2015
- Data
  - ~630 rather unspecific arrays from GEO
  - Compared to two time-resolved clock-specific experiments
- Reconstruction quality of three algorithms
  - Aracne, Bayes Networks, Time-Delay Aracne
Results

- Filtering of ARACNE reduces recall a lot, while precision increases only marginally.
- Data set size outweighs specificity - reconstruction about as good using many untargeted arrays or using fewer targeted arrays.
Content

• Network reconstruction
• Quantitative network analysis
  - A model of transcriptional regulation
  - Metabolic network models
  - Kinetic modeling
Networks as Equations

- REVEAL / ARACNE infer relationships based on correlation
- Alternative: Describe states as sets of (linear) equations
  - No discretization
  - Extensibility: Incorporate different types of experiments (binding, epigenetic status, mRNA translation, …)
  - Still many limitations: Synchronized time, no kinetics, …
- We look at one simple approach in between reconstruction and analysis
  - We assume the network topology to be given
  - We infer the (probable) strengths of different effects
  - Schacht et al. (2014). "Estimating the activity of transcription factors by the effect on their target genes." Bioinformatics
Approach

• Assume a network $G=(V,E)$, where $V$ consists of a set of transcription factors $T$ and a set of genes $G$
  - Transcription factors regulate genes, but not vice versa
  - We also ignore that a TF may regulate TFs (other or even itself)
  - Each gene $g$ is regulated by all its incoming TFs (and no others)
• Measurements: $m$ observations for $n$ nodes (genes / TFs)
• We model the expression values of all genes as linear combinations of the values of its regulating TFs

$$g_{i,j} = \beta_0 + \sum_{t=1}^{\vert T \vert} \beta_t * \delta_{t,j} * e_{t,i}$$
Model

\[ g_{i,j} = \beta_0 + \sum_{t=1}^{\left| T \right|} \beta_t \cdot \delta_{t,j} \cdot e_{t,i} \]

- \( g_{i,j} \): Expression of gene \( i \) in observation \( S_j \)
- \( \beta_0 \): Fixed additive offset (sample differences)
- \( \beta_t \): Global activity parameter for transcription factor \( t \)
- \( \delta_{t,j} \): Observation specific (\( S_j \)) effect on \( t \)
- \( e_{t,i} \): Effect strength of TF \( t \) on gene \( i \)
  - Set to 0 if there is no edge between \( g \) and \( t \)
Optimization

- Typically, these (large) systems cannot be solved exactly
- Instead, minimize the error

\[
g_{i,j} - \left( \beta_0 + \sum_{t=1}^{T} \beta_t \cdot \delta_{t,j} \cdot e_{t,i} \right) \] = \min

- ... under a set of constraints
- Several solvers available
Many Other Models

• **Stoichiometric networks**
  - Model the turnover of molecules
    - Especially metabolism
  - Needs to consider enzymatic effects
  - What will a network produce given a certain input?
  - Is a network in flux balance?

• **Kinetic networks**
  - Takes into account reaction rates: How many in what time
    - No linear relationship
  - Leads to systems of differential equations
  - Can predict system behavior in time under realistic assumptions

\[
\text{CH}_4 + 2\text{O}_2 \rightarrow \text{CO}_2 + 2\text{H}_2\text{O}
\]