Network Reconstruction

Johannes Starlinger

Content

- Network reconstruction
  - Boolean models
  - Correlation-Based Approaches: REVEAL / ARACNE
  - Example

Networks

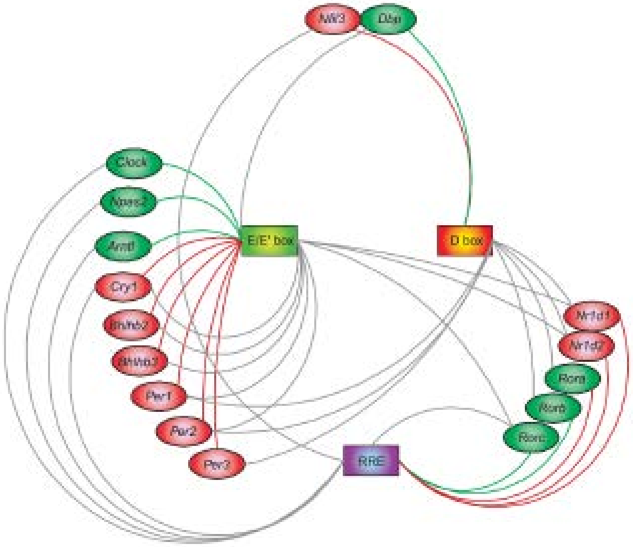
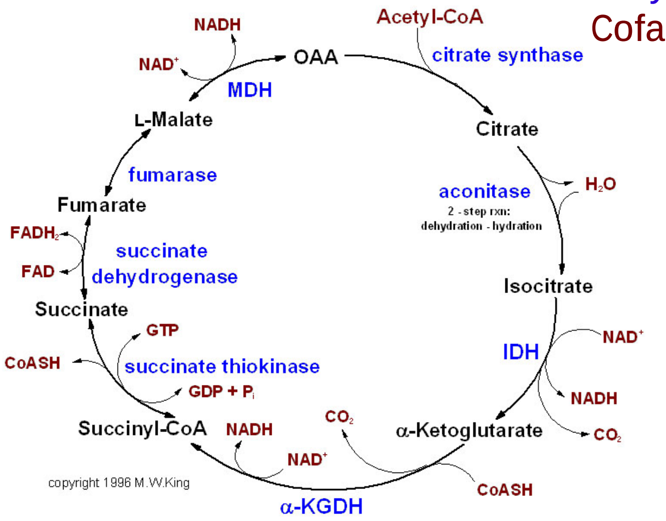


Abbildung 2: Zentrale Gene der zirkadianen Uhr und deren wechselseitiger Einfluss. [UHC+05] (Kästen: Cis-Elemente/Grüne Ovale: Positiv regulierende Gene/Rote Ovale: Negativ regulierende Gene/Regulationsrichtung 1: Von Gen über farbige Kante zu Cis-Element/Regulationsrichtung 2: Von Cis-Element über graue Kante zu Gen)



copyright 1996 M.W.King

Chemicals
Enzymes
Cofactors

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

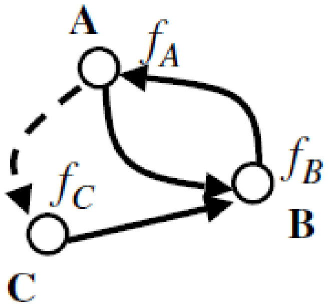
Boolean Network Models

- 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, ...



$f_A(B) = B$

$f_B(A, C) = A \text{ and } C$

$f_C(A) = \text{not } A$

Boolean Network

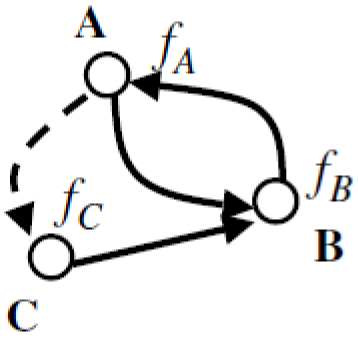
Network Dynamics

- 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

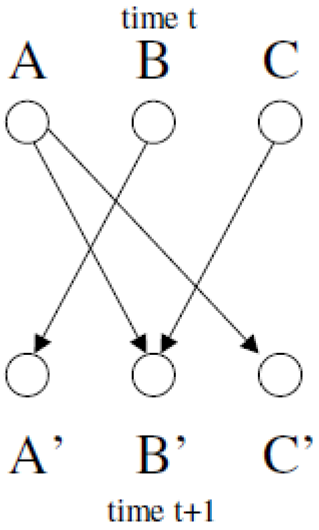


$f_A(B) = B$

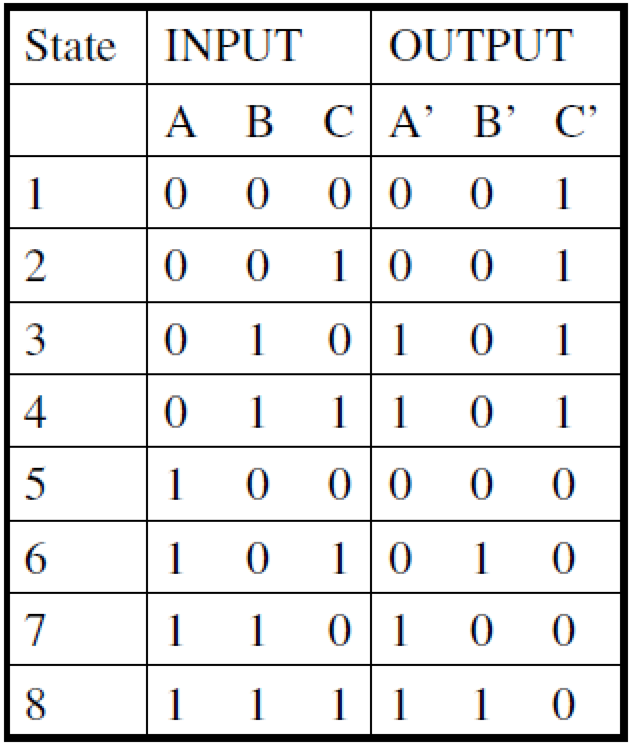
$f_B(A,C) = A \text{ and } C$

$f_C(A) = \text{not } A$

Boolean Network



Wiring Diagram

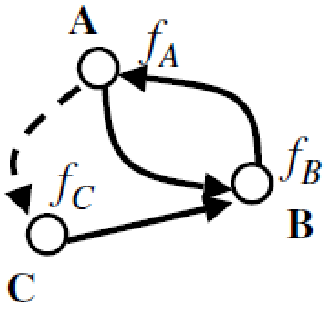


State	INPUT			OUTPUT		
	A	B	C	A'	B'	C'
1	0	0	0	0	0	1
2	0	0	1	0	0	1
3	0	1	0	1	0	1
4	0	1	1	1	0	1
5	1	0	0	0	0	0
6	1	0	1	0	1	0
7	1	1	0	1	0	0
8	1	1	1	1	1	0

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

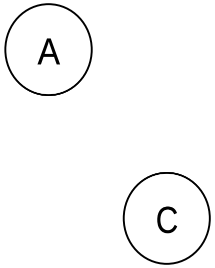
genes time	A	B	C
0	1	1	0
1	1	0	0
2	0	0	0
3	0	0	1
4	0	0	1
5	…	…	…

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?

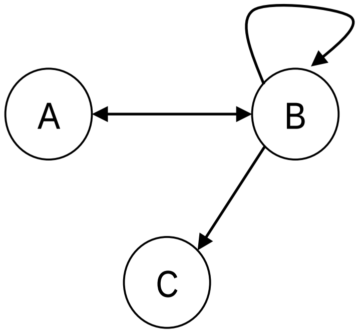


S

genes time	A	B	C
0	1	1	0
1	0	0	1
2	1	0	0
3	1	1	0
4	0	0	1
5	...	...	...

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)=not B

f(B) = A and not B

f(C)=B

genes time	A	B	C
0	1	1	0
1	0	0	1
2	1	0	0
3	1	1	0
4	0	0	1
5	…	…	…

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

Towards Reality

- There are less complex & more robust algorithms

- REVEAL replaces Boolean functions by mutual information; correlations rather than deterministic switching
  - Liang, S., S. Fuhrman and R. Somogyi (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing., Hawaii, US.
- ARACNE is even simpler: Only removal of some (presumably indirect) correlations
  - Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera and A. Califano (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." BMC Bioinformatics 7((Suppl 1), S7).

Foundations

- Definition

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

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

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

Example

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

p(x,y)	y=0 p(y=0)=0.6	y=1 p(y=1)=0.4
x=0; p(x=0)=0.2	0,12	0,08
x=1; p(x=1)=0.8	0,48	0,32

MI(X,Y)=0

p(x,y)	y=0 p(y=0)=0.6	y=1 p(y=1)=0.4
x=0; p(x=0)=0.2	0,18	0,03
x=1; p(x=1)=0.8	0,05	0,74

MI(X,Y)=0,53

Two more Facts

- With a little math, we find

$$MI(X,Y) = H(X) - H(X|Y) = H(Y)-H(Y|X)$$

  - H(X): Entropy of X
  - H(X|Y): Conditional entropy of X given Y

- It follows that the maximal value of 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 $T=(Y_1,…Y_k)$ with $MI(X,Y_1,…Y_k) = H(X)$;
    if T exists
      N := N \ X;                     # n is explained
  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 (x_1,y_0), (x_2,y_1), ...

REVEAL in Practice

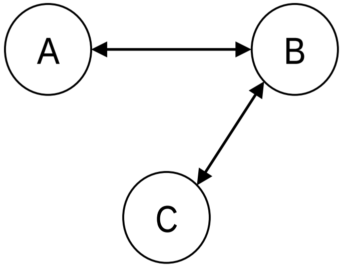
- In the formulation given, REVEAL would be as strict as Boolean functions
  - Dependencies must be perfect
- In the presence of noise, one must be satisfied with almost maximal MI
  - I.e., $|MI(X,Y)-H(X)| < \varepsilon$
- Can be extended to more than two possible states
  - Less strict discretization, 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 |MI(X,Y)-H(X)| > ε
    - ε 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

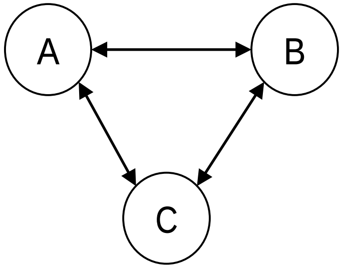
Data Processing Inequalities

Imagine

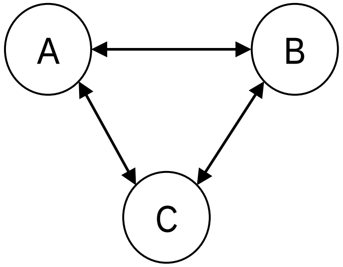


with strong effects of A on B and B on C;

will appears as



But if we find



which edge most probably is an artifact?

- Assumption: If $\mathrm{MI}(X,Z) \leq \min(\mathrm{MI}(X,Y),\mathrm{MI}(Y,Z))$, 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

Reconstructing the Mammalian Clock

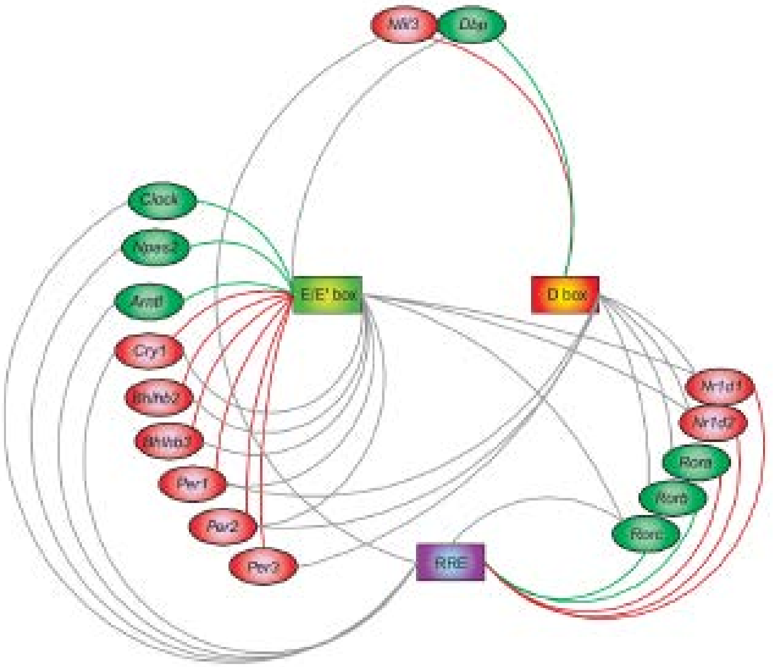
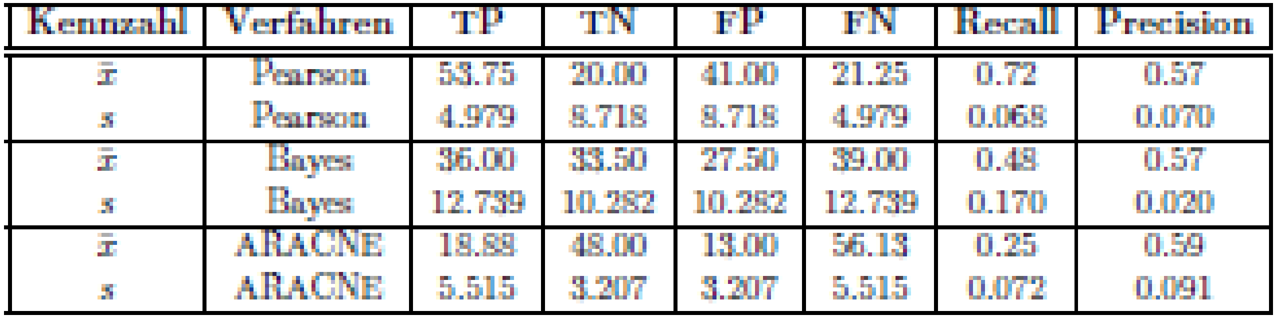


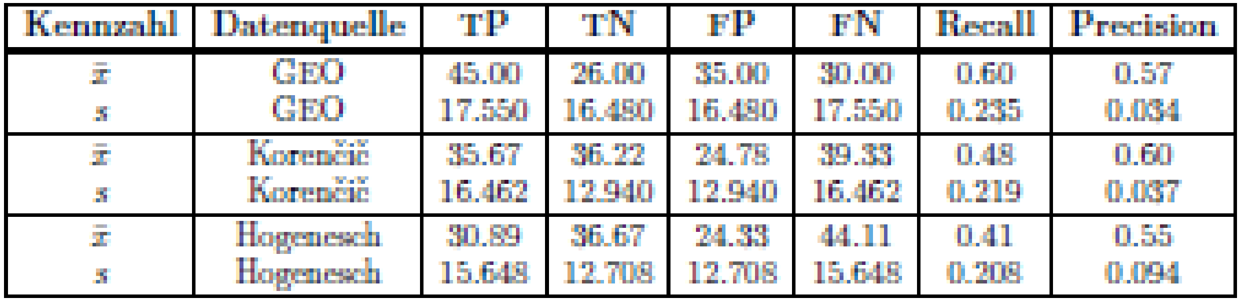
Abbildung 2: Zentrale Gene der zirkadianen Uhr und deren wechselseitiger Einfluss. [UHC$^+$05] (Kästen: Cis-Elemente/Grüne Ovale: Positiv regulierende Gene/Rote Ovale: Negativ regulierende Gene/Regulationsrichtung 1: Von Gen über farbige Kante zu Cis-Element/Regulationsrichtung 2: Von Cis-Element über graue Kante zu Gen)

- 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



Kennzahl	Verfahren	TP	TN	FP	FN	Recall	Precision
$\bar{x}$	Pearson	53.75	20.00	41.00	21.25	0.72	0.57
s	Pearson	4.979	8.718	8.718	4.979	0.068	0.070
$\bar{x}$	Bayes	36.00	33.50	27.50	39.00	0.48	0.57
s	Bayes	12.739	10.282	10.282	12.739	0.170	0.020
$\bar{x}$	ARACNE	18.88	48.00	13.00	56.13	0.25	0.59
s	ARACNE	5.515	3.207	3.207	5.515	0.072	0.091



Kennzahl	Datenquelle	TP	TN	FP	FN	Recall	Precision
$\bar{x}$	GEO	45.00	26.00	35.00	30.00	0.60	0.57
s	GEO	17.550	16.480	16.480	17.550	0.235	0.034
$\bar{x}$	Korenčič	35.67	36.22	24.78	39.33	0.48	0.60
s	Korenčič	16.462	12.940	12.940	16.462	0.219	0.037
$\bar{x}$	Hogenesch	30.89	36.67	24.33	44.11	0.41	0.55
s	Hogenesch	15.648	12.708	12.708	15.648	0.208	0.094

- 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