Reconstruction of Metabolic Networks from High Throughput Metabolic Data: In Silico Analysis of RBC metabolism

Ilya Nemenman¹, Michael Wall¹
Sean Escola², William Hlavacek¹

¹Los Alamos National Laboratory
²Columbia University Medical School
Metabolic Networks: (future) Inference Problem from MassSpec / isotopically labeled data

Reconstruction of B. anthracis (especially during the host colonization)

Robotic sampling:
- every 60 s
- ~500 metabolites
- ~100s of growth conditions
- 20-100 cells at a time
Steady states because…

- Destructive measurements
- Uncorrelated errors
- Smaller errors (const. sample sz.)
- Less samples, but repeatable
- Only need topologies
- Only relative concentrations

- Similar to mRNA arrays data
Statistical dependency model

Better model than for mRNA
- Direct coupling of nodes
- Simpler noise model
- Known modulators
- Interactions microscopically pairwise
- No directionality in steady state

\[
\frac{dA}{dt} = f(A, B, C, D, Enz)
\]

\[
f(A, B, C, D, Enz) = 0
\]

\[
P(A, B, C, D | Enz) = \delta(g)
\]

\[
P(A, B, C, D) = \langle \delta(g) \rangle = \exp[-\lambda_{ABCD}]
\]

\[
P(ABCD) \approx \exp[-\lambda_{AB} - \lambda_{AC} - \cdots - \lambda_{CD}]
\]
From activity to networks

Small data set  ✔  ✔  ×  ×
Robustness      ✔  ✔  ✔  ×
Complexity      ×  ✔  ×  ×
Conditional ints. ✔  ×  ✔  ✔
Reparam inv.    ✔  ✔  ×  ✔
Irreducibility  ✔  ×  ✔  ✔

Figure 4a
**ARACNE** (Califano & Co)

A → B

\[ I(A,C) = \int dA \, dC \, P(A,C) \log \frac{P(A,C)}{P(A)P(C)} \]

\[ I(A,C) \leq \min[I(A,B), I(B,C)] \]

Reparam. invariance; small sample; low complexity.

**Performance?**
Performance:
Few false positives

- No false positives for tree networks
- No false positives under very general conditions for networks with only a few strong loops
- No false negatives under stronger conditions (many otherwise, but it’s ok)
- Need to estimate MI reliably
MI estimation
B-cell dataset: cMYC network

~400 arrays (Dalla-Favera et al.)
No dynamics
~250 naturally occurring, ~150 perturbed
~25 phenotypes (normal, tumors, experimental perturbations)

- Protooncogene,
- 12% background binding,
- one of top 5% hubs
- significant MI with 2000 genes

Total interactions: 56
Pre-known: 22
New Ch-IP validated: 11/12
Does good microarray performance guarantee good results for metabolites?

- Different noises
- Different nonlinearities
- Very dense
- \( \sim 1 \times 10^7 \) ratios in kinetic rates/steady state concentrations
  - Interactions of low-abundance metabolites washed out
  - These are essential parts of environmental response pathways (intermediates)
- **Need benchmark metabolic data sets**
  (DREAM workshop, NYC, 09/06
  [http://dimacs.rutgers.edu/Workshops/ReverseEng/](http://dimacs.rutgers.edu/Workshops/ReverseEng/))
Synthetic model

- 39 metabolites
- 44 individual reactions
- 107 pairwise interactions between distinct metabolites

Jamshidi et al., 2001
Ni, Savageau, 1996
Data sets

• Jamshidi et al. Mathematica code: generate ~1000 steady states with different values for Donnan ratio, glucose, intracellular Pi, Mg, and extracellular Na:
  – chemostat (ranges consistent with survival of RBCs in culture)
  – natural (ranges consistent with normal human blood work)
  – natural correlated (same with correlated parameters)

• Also got ~100hrs of time-dependent data with naturalistic evolution of control parameters

• The chemostat dataset:
  – Smallest mean concentration 4.5e-5
  – Largest mean concentration 1.2e+2
  – Smallest ratio std/mean 1.3e-14
  – Largest ration std/mean 4.5e-1
Adding noise

- Experimental noise simulated by adding additive noise and multiplicative noise
  \[ X = X_0 + A \cdot \text{randn()} + B \cdot X_0 \cdot \text{randn()} \]

- Remove nodes with std<noise
- Couple all neighbors of removed nodes for validation
Example

$I > 0$

$I = 0$

$I > 0$

$I = 0$
Adjusting ARACNE parameters

- Best kernel: leave-one-out cross validation -- $h = 0.1173$ of the variable range (rank-order transform).
- To assure $<1$ falsely significant MI out of $39*38/2 = 741$, select threshold corresponding to $p$-value $= 1/741$; $I_0 = 0.019$. 
Performance on RBC data for different noise levels

PRC for changing noise, $I$ threshold, tolerance

\[
p = \frac{N_{TP}}{N_{TP} + N_{FP}} = \frac{N_{TP}}{N_{P, \text{found}}}
\]

\[
r = \frac{N_{TP}}{N_{TP} + N_{FN}} = \frac{N_{TP}}{N_{P, \text{tot}}}
\]

- Different DPI tolerances (0, 0.05, 0.1 for solid, dashed, dotted).

- Operation point for predefined $I$ threshold
Problems

- Low abundance metabolites
- Bootstrapped data sets to increase $r$
- Start with constrained networks (by mass transfer)
- Regulated interactions (metabolic/transcriptional data sets needed)

Enzyme (protein or mRNA)

Substrate $\rightarrow$ Product
Most interesting next thing
(detection of enzyme-coding genes)

\[
\frac{d[pE]}{dt} = \alpha \cdot [mE] - k_+ [pE][pS] + k_- [pES] - \beta [pE]
\]
\[
\frac{d[cS]}{dt} = V_0 - k_+ [pE][cS] - \beta [pS]
\]
\[
\frac{d[cES]}{dt} = k_+ [pE][cS] - k_- [cES] - k_{dis} [cES] - \beta [cES]
\]
\[
\frac{d[cP]}{dt} = k_{dis} [cES] - \beta [cP]
\]

\(I(cS,cP | mE)\)

(Wang et al, 2006)