LA-UR #06-4270

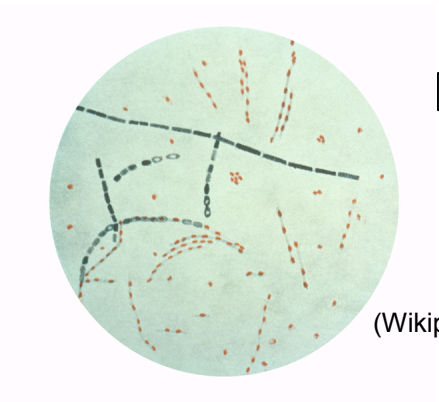
Reconstruction of Metabolic Networks from High Throughput Metabolic Data: In Silico Analysis of RBC metabolism

Ilya Nemenman[1], Michael Wall[1]
Sean Escola[1,2], William Hlavacek[1]

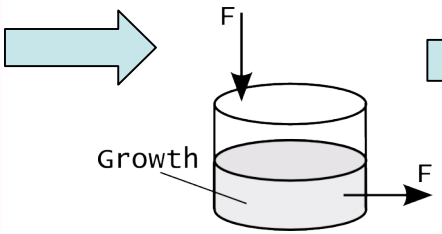
[1]Los Alamos National Laboratory
[2]Columbia University Medical School

Metabolic Networks: *(future)* Inference Problem from MassSpec / isotopically labeled data

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



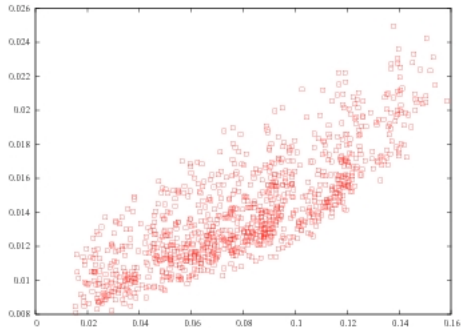
(Wikipedia)



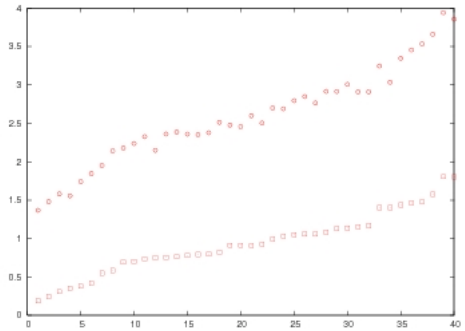
chemostat

Robotic sampling:

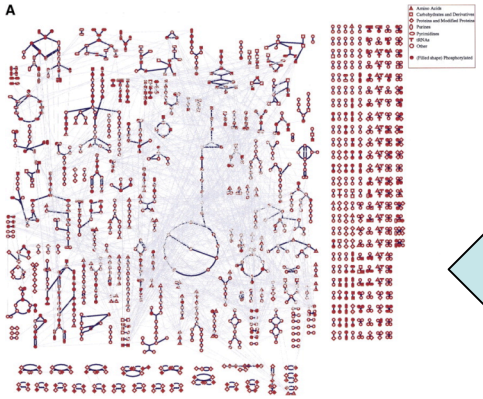
- every 60 s
- ~500 metabolites
- ~100s of growth conditions
- 20-100 cells at a time (uM)



Steady states



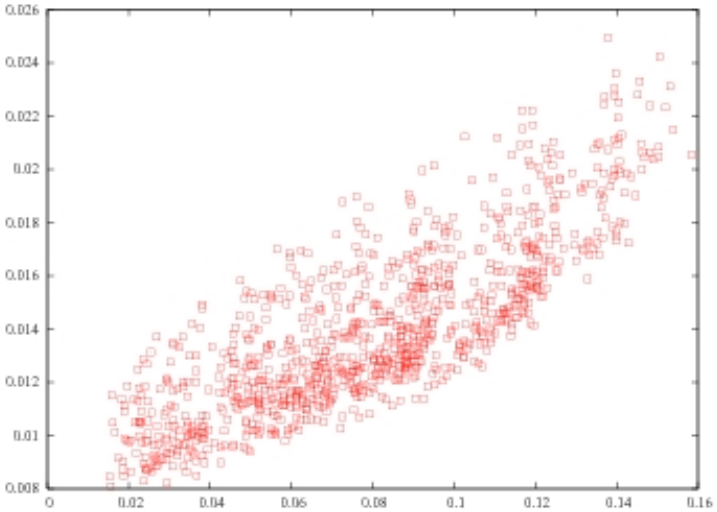
Time series



(Ouzounis, Karp 2000)

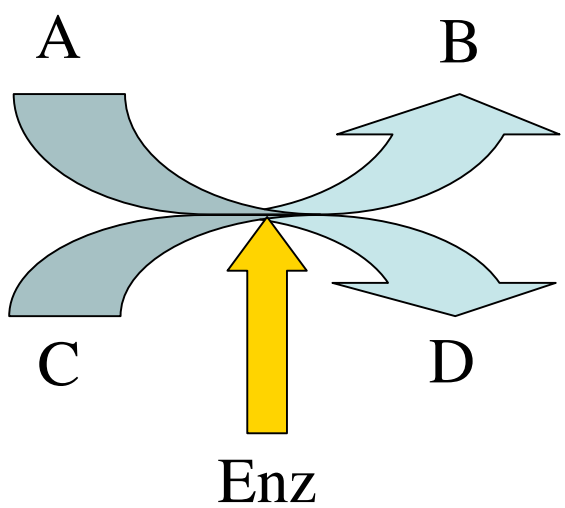
Steady states because…

- Destructive measurements
- Uncorrelated errors/measurements
- Smaller errors (const. sample sz.)
- Less samples, but repeatable
  (steady states more stable than kinetics)
- Only want topologies (not rates)
- Only relative concentrations
- Unknown species function



- Similar to mRNA arrays data

Statistical dependency model



$$f(A,B,C,D,Enz)=0$$

$$P(A,B,C,D \mid Enz)=\delta(\ldots)$$

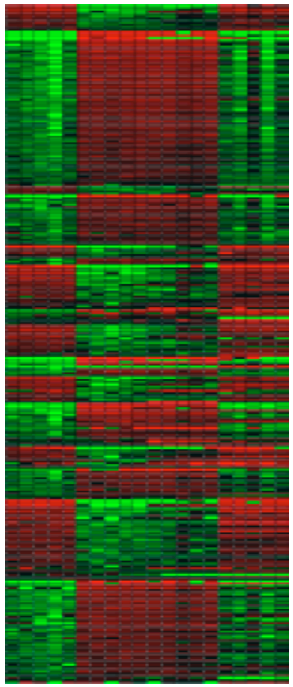
$$P(A,B,C,D)=\langle\delta(\ldots)\rangle=\exp[-\lambda_{ABCD}]$$

$$P(ABCD)\approx\exp[-\lambda_{AB}-\lambda_{AC}-\cdots-\lambda_{CD}]$$

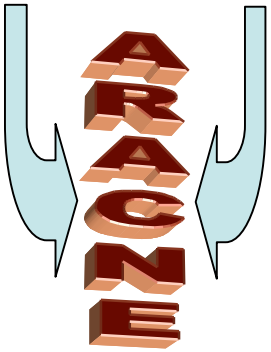
Better model than for mRNA

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

From activity to networks



	Stat	Co/GM	Biochem.
Small data set	✖✔	✔	✖
Robustness	✔	✔	✖✔
Complexity	✖	✔	✖
Conditional ints.	✔	✖✔	✔
Reparam inv.	✖✔	✖✔	✖✔
Irreducibility	✔	✖	✔



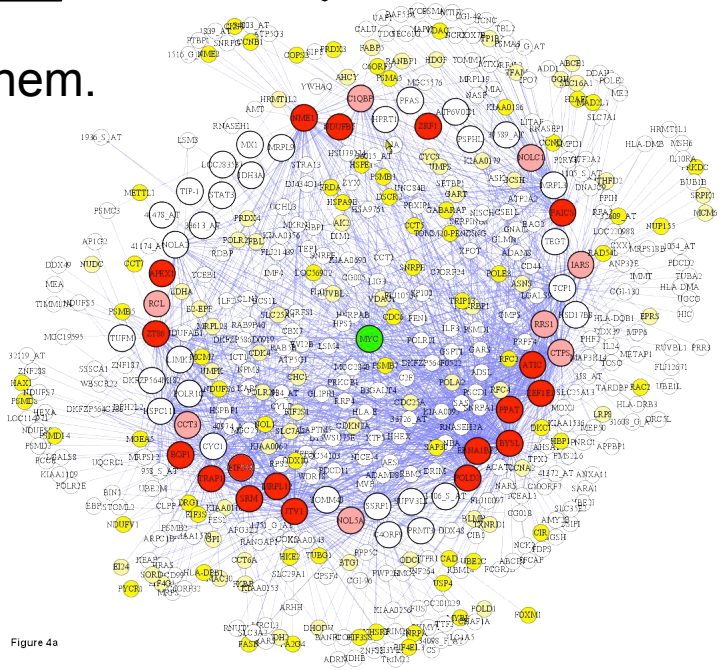
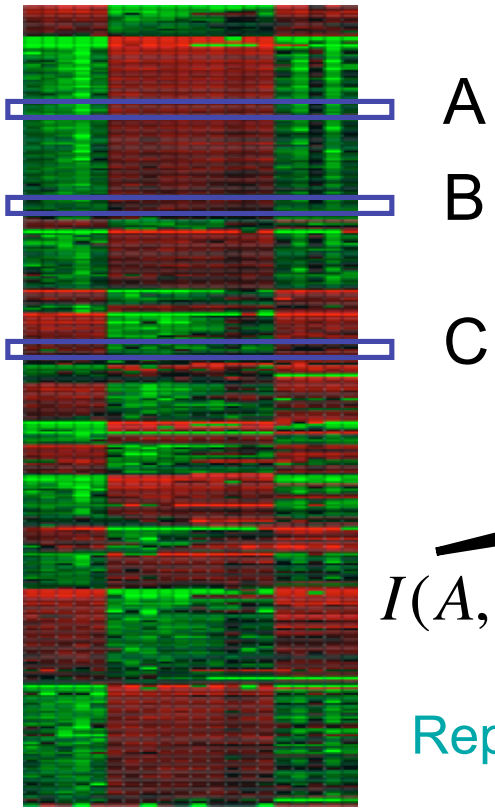
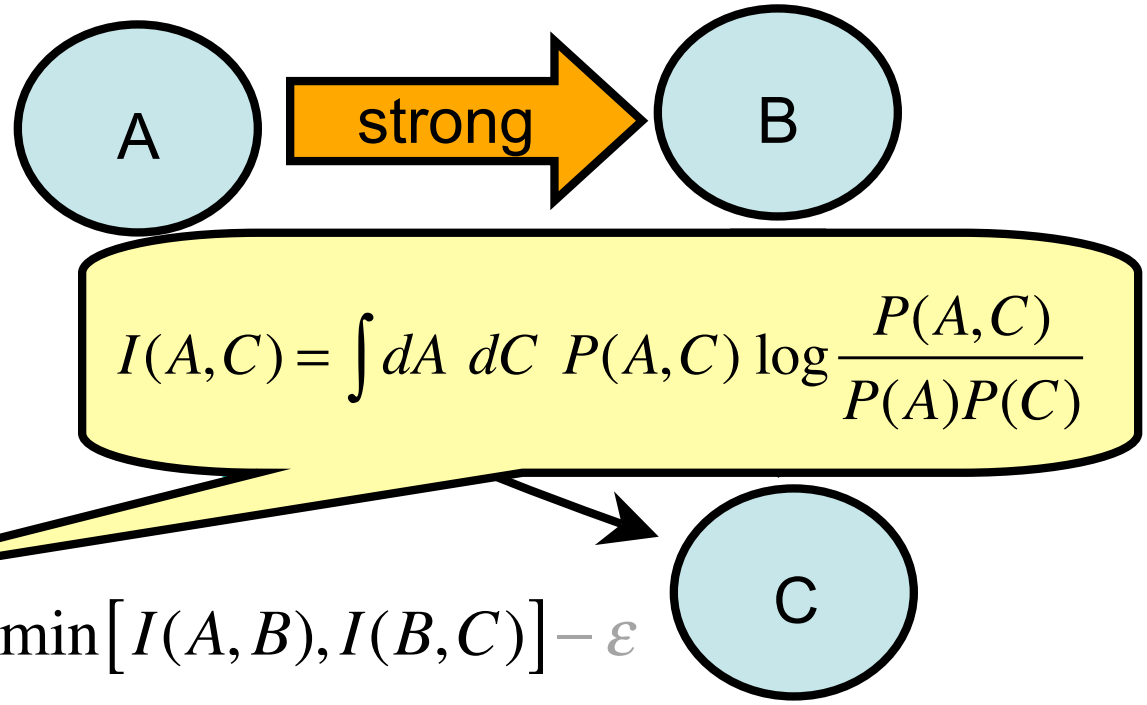


Figure 4a

ARACNE (Califano & Co)





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

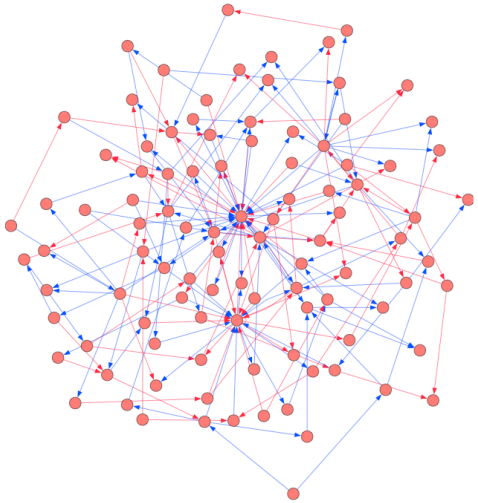
$$I(A,C) \le \min\left[I(A,B), I(B,C)\right] - \varepsilon$$

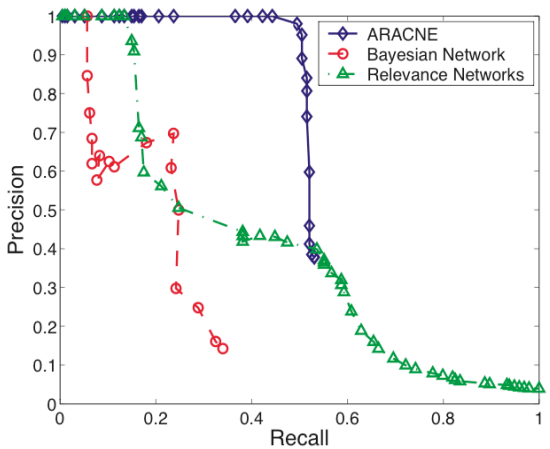
Reparm. invariance; small sample; low complexity.

Performance?

Performance: Few false positives

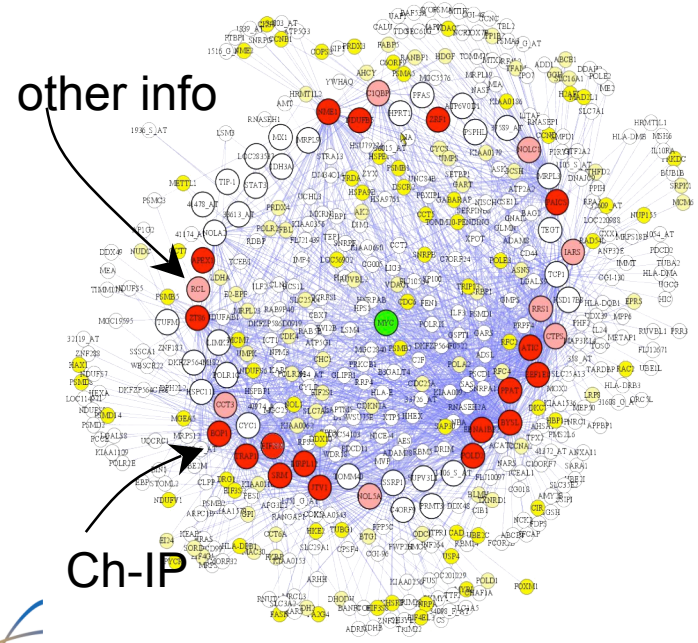
- No false positives for tree networks
- No false positives under very general conditions for networks with only a few strong loops
- No false negatives under stronger conditions (many otherwise, but it's ok)
- Need to estimate MI reliably





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

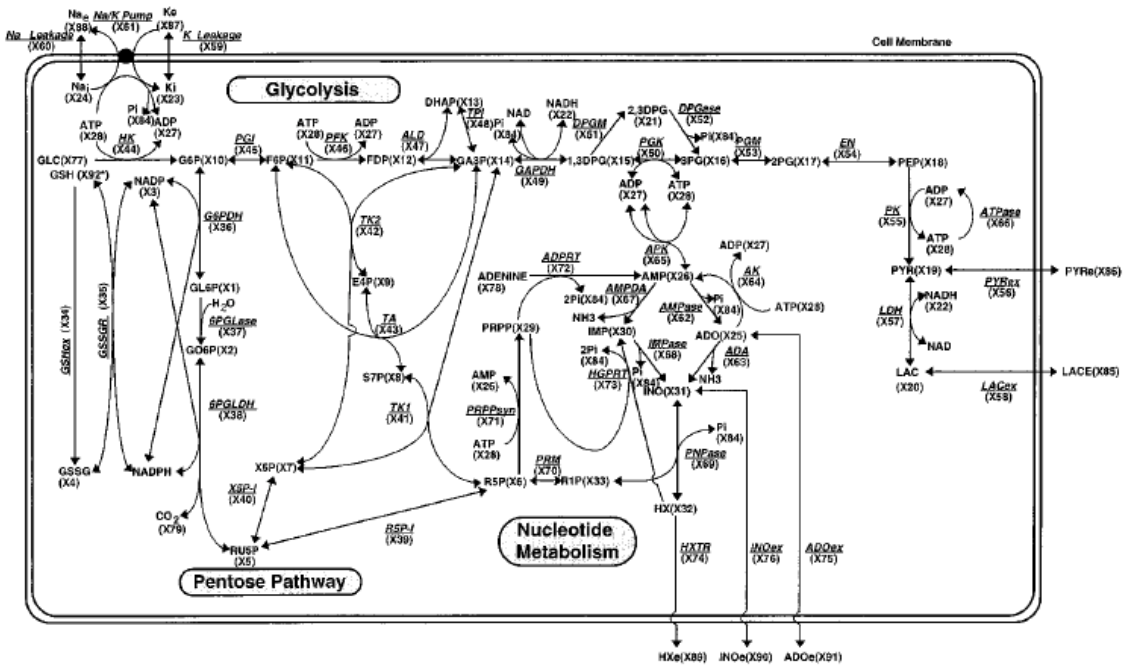
Total interactions: 56
Pre-known: 22
New Ch-IP validated: 11/12

Does good microarray performance guarantee good results for metabolites?

- Different noises
- Different nonlinearities
- Transformation instead of regulation
- Very dense (many loops)
- ~1e7 ratios in kinetic rates/steady state concentrations
  - Interactions of low-abundance metabolites washed out
  - These are essential intermediates of environmental response pathways
  - Steady states?
- **Need benchmark metabolic data sets**

Synthetic model

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



No questions on:
Is this biologically relevant?

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:
  1. chemostat (ranges consistent with survival of RBCs in culture)
  2. natural (ranges consistent with normal human blood work)
  3. natural correlated (same with human-like correlated parameters)
- Also time-dependent data with naturalistic evolution of control parameters
  4. evolution from natural-correlated params. (25 evolutions, 1000 samples)
  5. time-dependent evolving params. (100 hours)
- E.g.: chemostat (natural) dataset
  - Smallest mean concentration 5e-5 (5e-5)
  - Largest mean concentration 1.1e2 (1e2)
  - Smallest ratio std/mean 0 (0)
  - Largest ration std/mean 1.1 (3e-1)

Adding noise

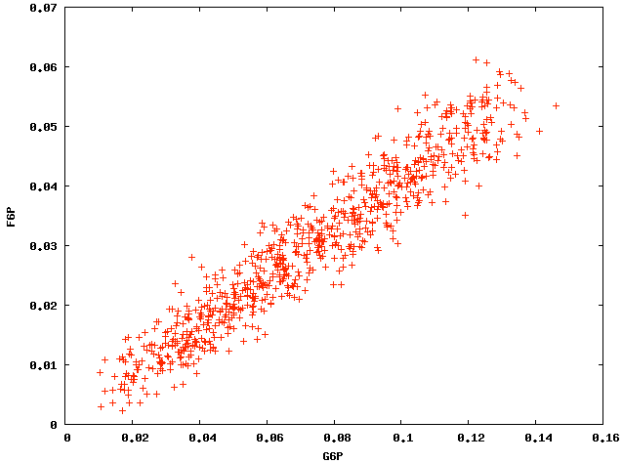
- Experimental noise simulated by adding additive noise and multiplicative noise

$$X = X_0 + A \cdot \mathrm{randn}() + B \cdot X_0 \cdot \mathrm{randn}()$$

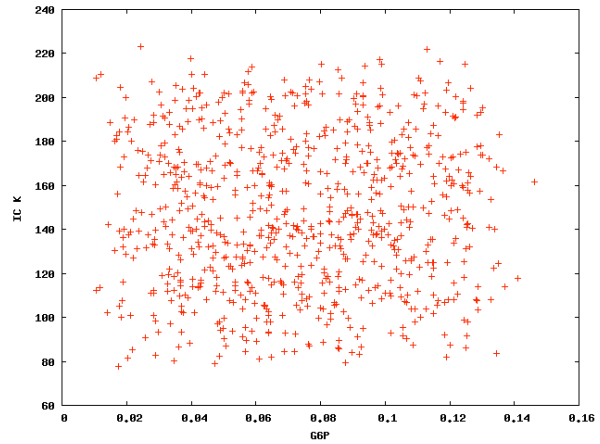
  for many different *A* and *B*
- Remove nodes with std<noise
- Connect all neighbors of removed nodes for validation purposes

Will be available at www.menem.com/~ilya

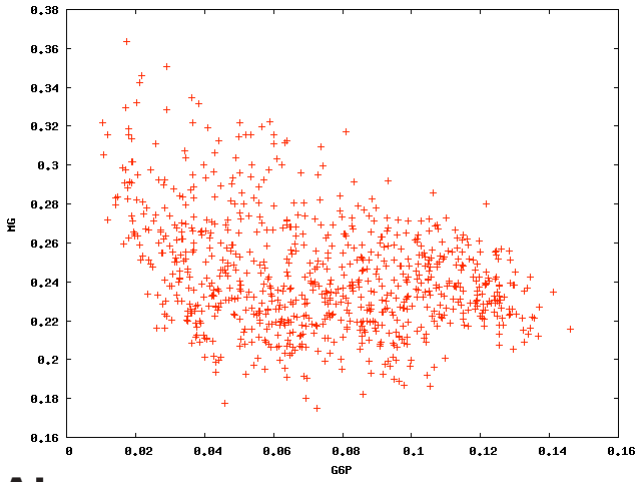
Example



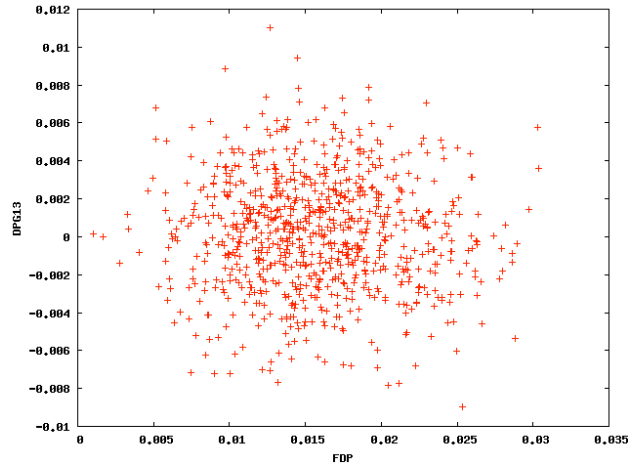
I>0



I=0



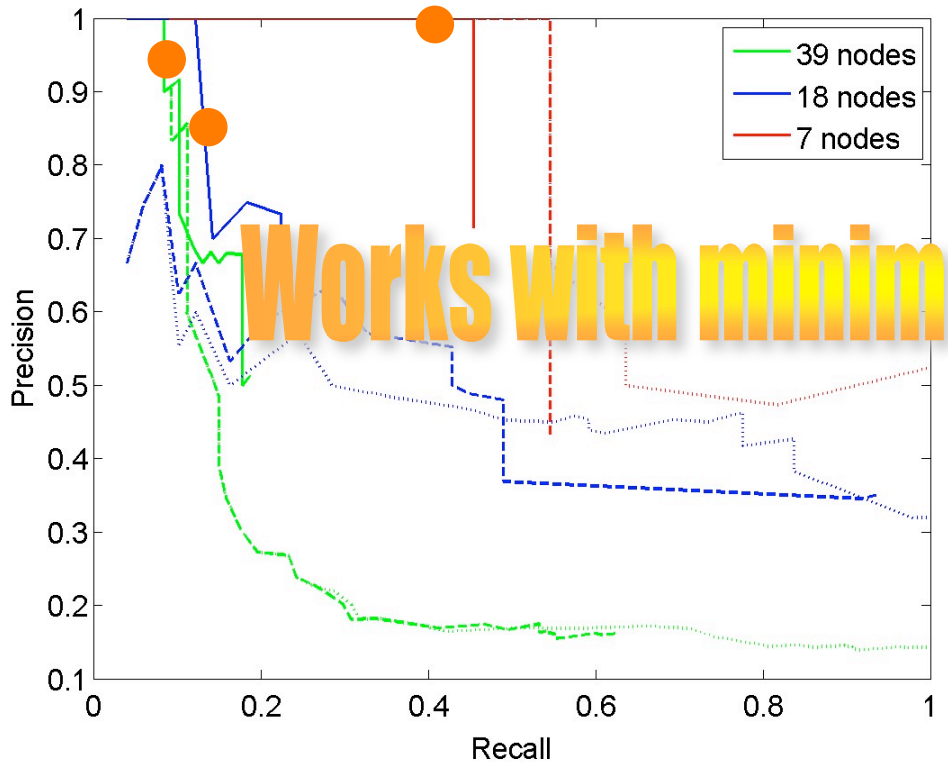
I>0



I=0

Performance on RBC data for different noise levels

PRC for changing noise, *I* threshold, tolerance



$$p = \frac{N_{TP}}{N_{TP} + N_{FP}} = \frac{N_{TP}}{N_{P,\ found}}$$

$$r = \frac{N_{TP}}{N_{P} - N_{FN}} = \frac{N_{TP}}{N_{P}}$$

Works with minimal modifications

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

● Operation point for pre-determined *I* threshold

Why low recall?

- Low abundance metabolites (bootstrapped data sets to increase *r*; correct for downwards bias in MI for this low metabs.)
- High connectivity (use mass differences)

Edge A--B

$$A + B \leftrightarrow X + Y \quad (\text{or} \quad Y = \varnothing); \qquad m_X + m_Y = m_A + m_B + (\text{small})$$

$$A + X \leftrightarrow B + Y \quad (\text{or} \quad X = \varnothing \quad \text{or} \quad Y = \varnothing); \qquad m_A - m_B = m_Y - m_X - (\text{small})$$

Masses known to 1e-2% -- can reconstruct pruned links.
Stay tuned for results.

1st International Conference on

Information Processing in Cellular Signaling and Gene Regulation

Building predictive physics-based mathematical models of biochemical interaction networks

August 2007
Santa Fe, NM

http://cnls.lanl.gov/Conference/q-bio/2007