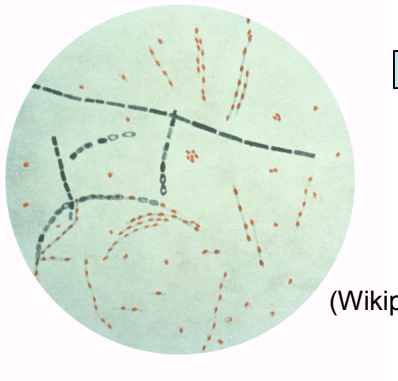
Reconstruction of Metabolic Networks from High Throughput Metabolic Data: In Silico Analysis of RBC metabolism

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

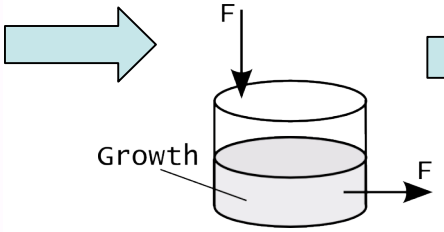
[1]Los Alamos National Laboratory
[2]Columbia University Medical School

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

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



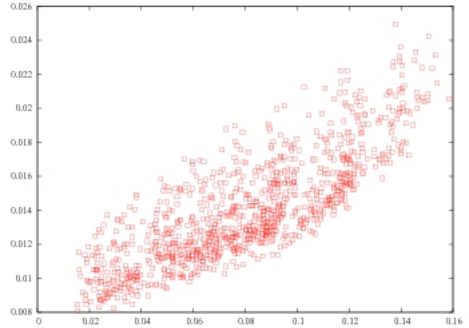
(Wikipedia)



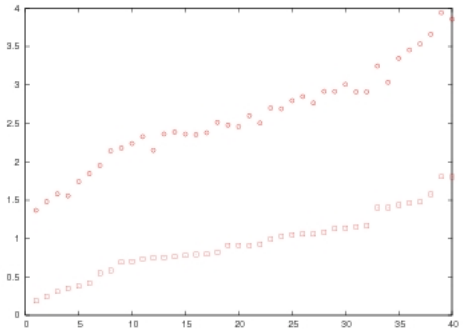
chemostat

Robotic sampling:

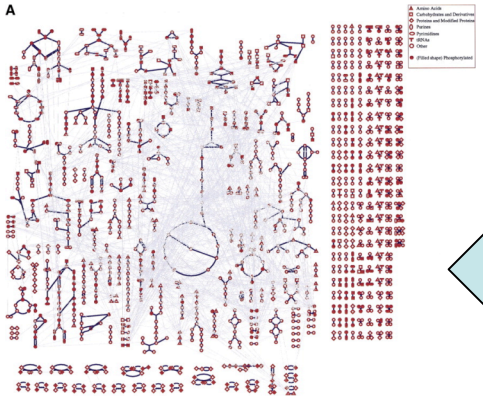
- every 60 s
- ~500 metabolites
- ~100s of growth conditions
- 20-100 cells at a time



Steady states



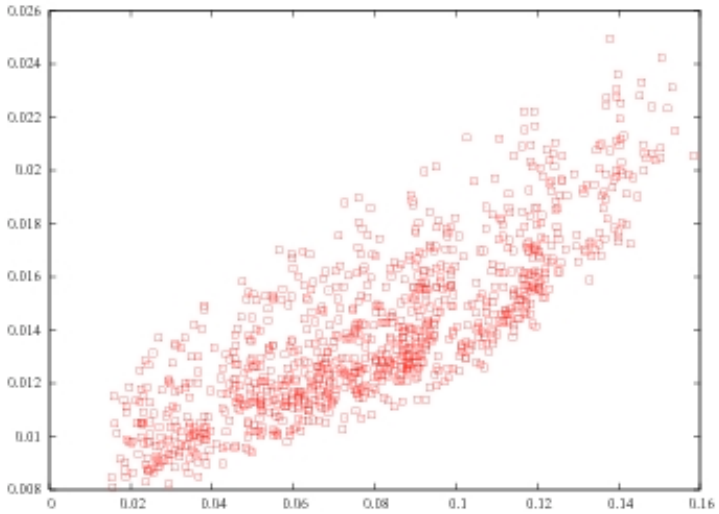
Time series



(Ouzounis, Karp 2000)

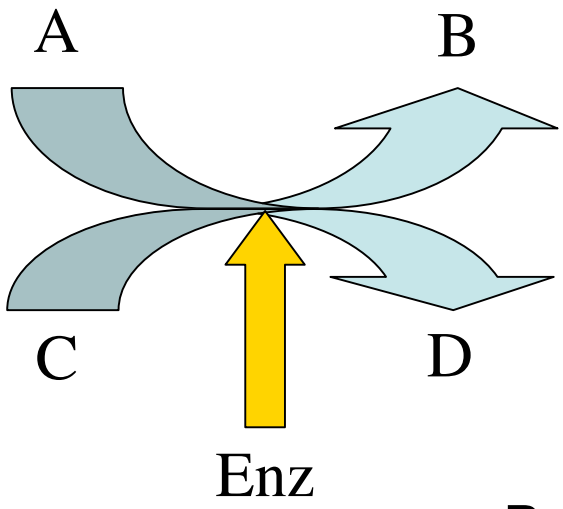
Steady states because…

- 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



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

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

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

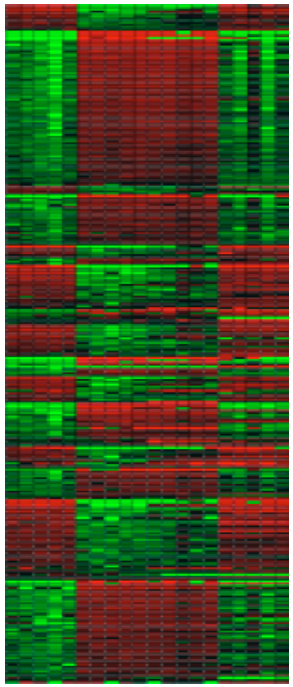
$$P(A,B,C,D) = \langle \delta(g) \rangle = \exp[-\lambda_{ABCD}]$$

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

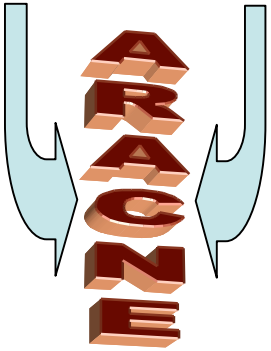
Better model than for mRNA

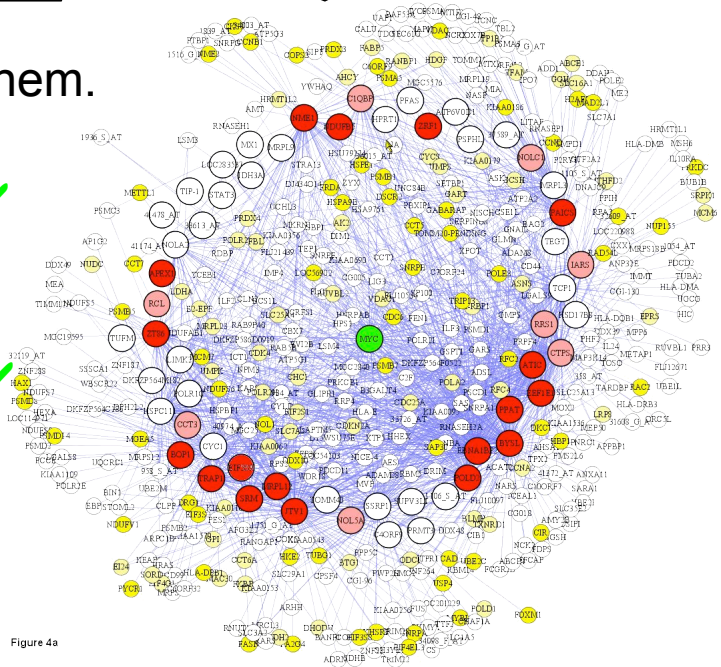
- Direct coupling of nodes
- Simpler noise model
- Known modulators
- Interactions microscopically pairwise
- No directionality in steady state

From activity to networks

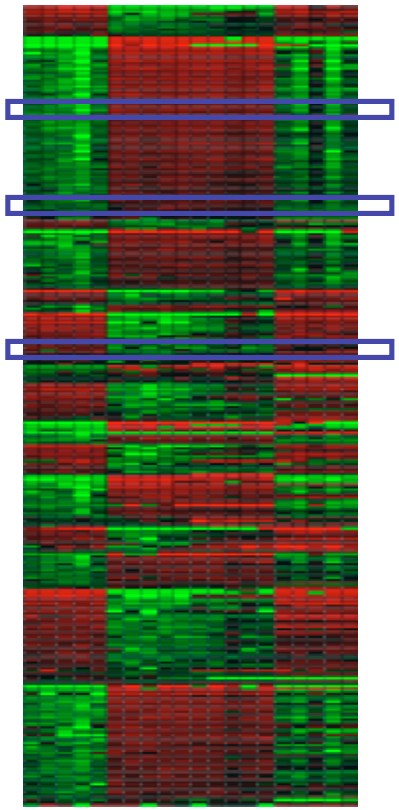


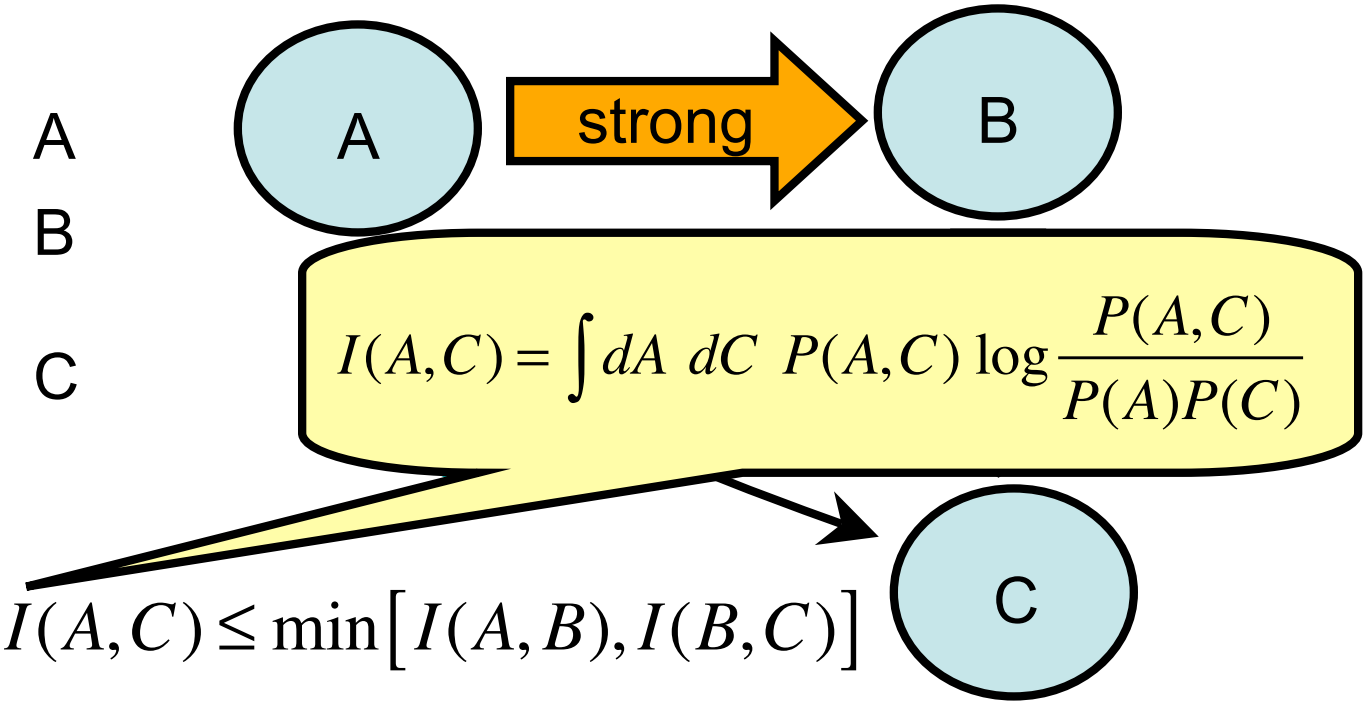
	Stat	Co/GM	Biochem.
Small data set	✖✔	✔	✖
Robustness	✔	✔	✖✔
Complexity	✖	✔	✖
Conditional ints.	✔	✖✔	✔
Reparam inv.	✖✔	✖✔	✖✔
Irreducibility	✔	✖	✔





ARACNE (Califano & Co)



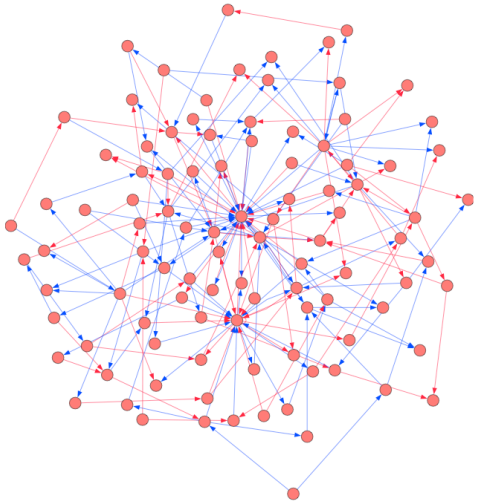


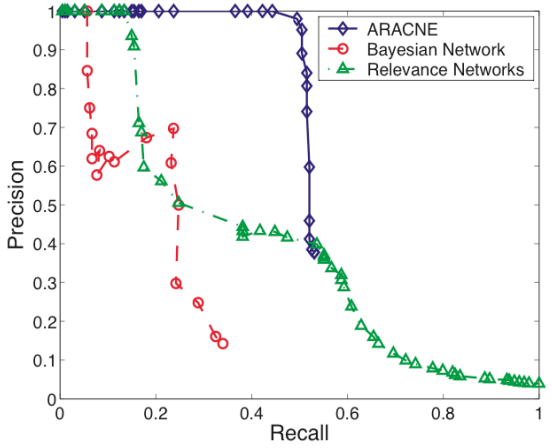
Reparm. invariance; small sample; low complexity.

Performance?

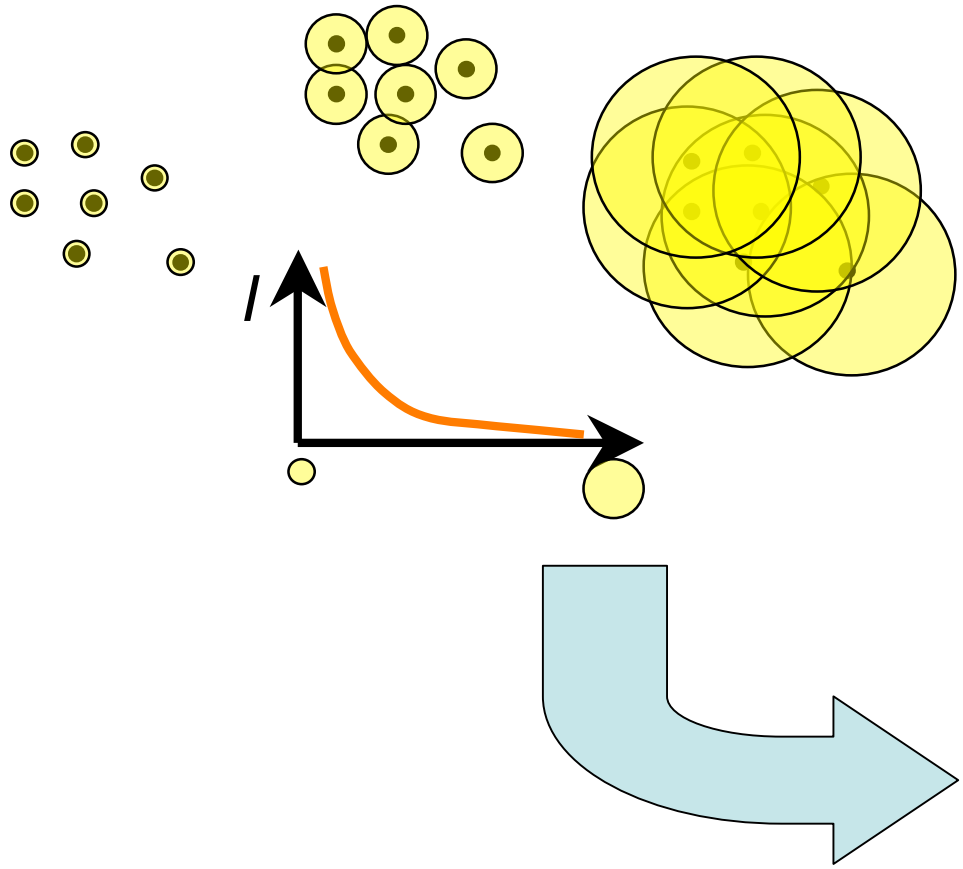
Performance: Few false positives

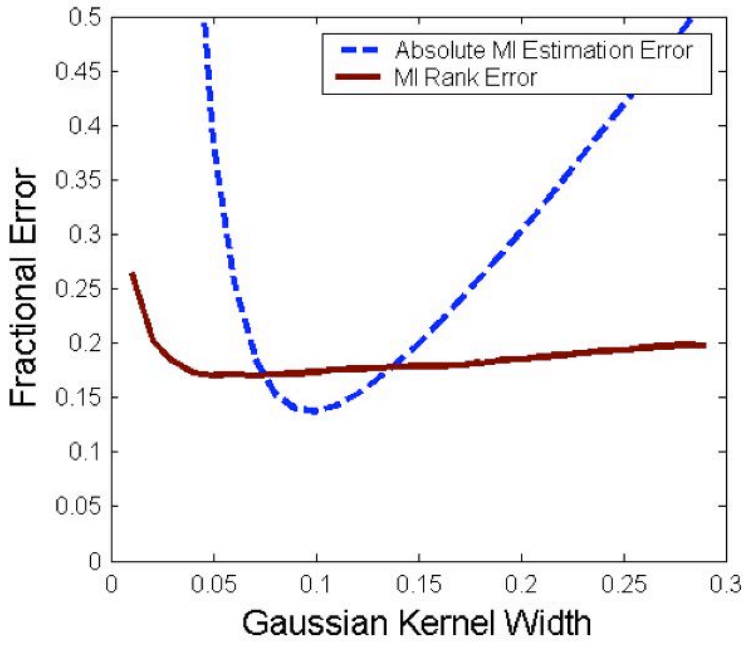
- No false positives for tree networks
- No false positives under very general conditions for networks with only a few strong loops
- No false negatives under stronger conditions (many otherwise, but it's ok)
- Need to estimate MI reliably





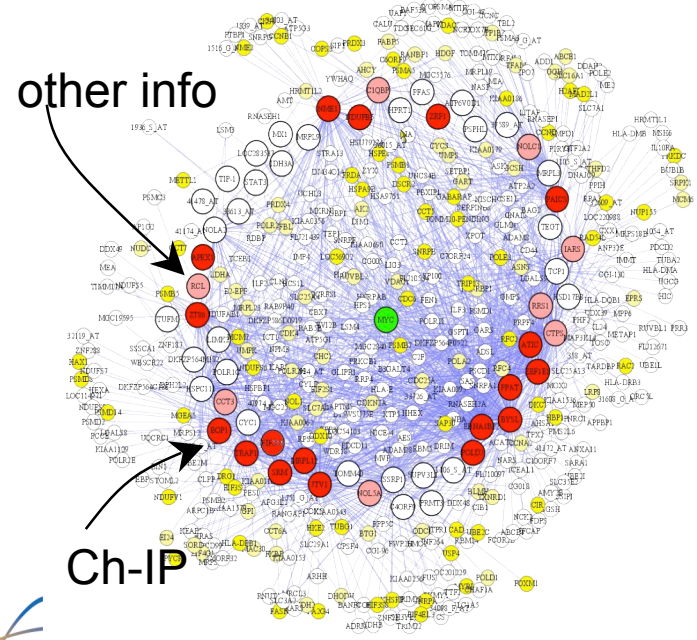
MI estimation





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

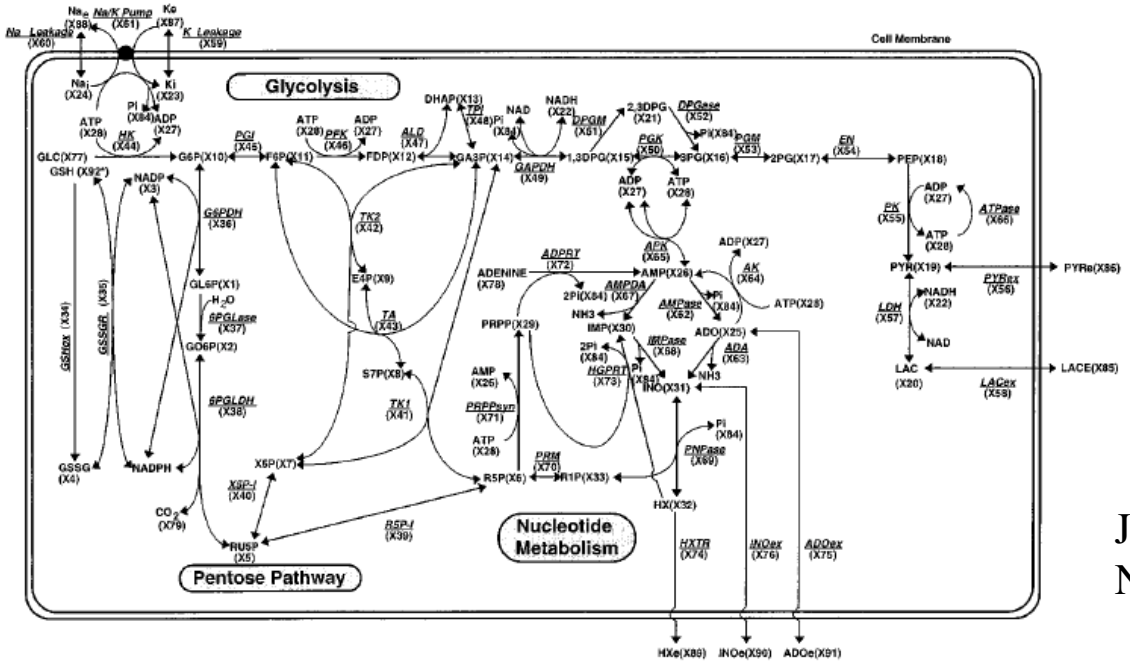
Total interactions: 56
Pre-known: 22
New Ch-IP validated: 11/12

Does good microarray performance guarantee good results for metabolites?

- Different noises
- Different nonlinearities
- Very dense
- ~1e7 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/)

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

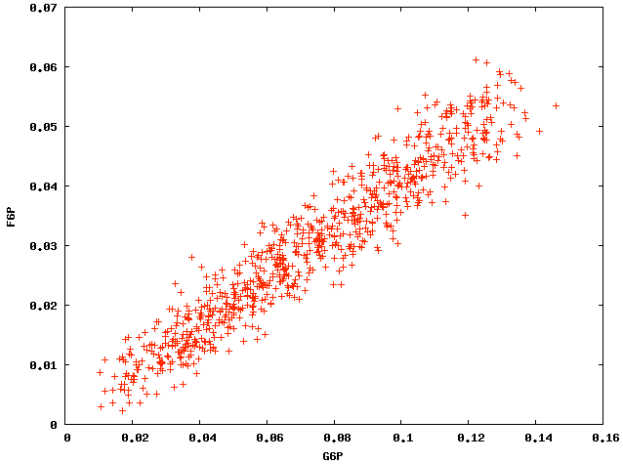
Adding noise

- Experimental noise simulated by adding additive noise and multiplicative noise

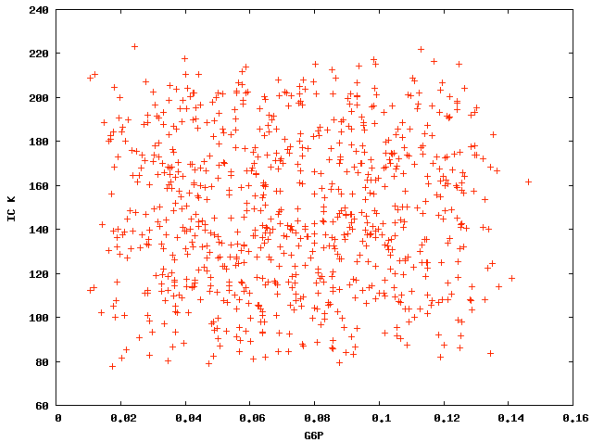
$$X = X_0 + A \cdot \text{randn}() + B \cdot X_0 \cdot \text{randn}()$$

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

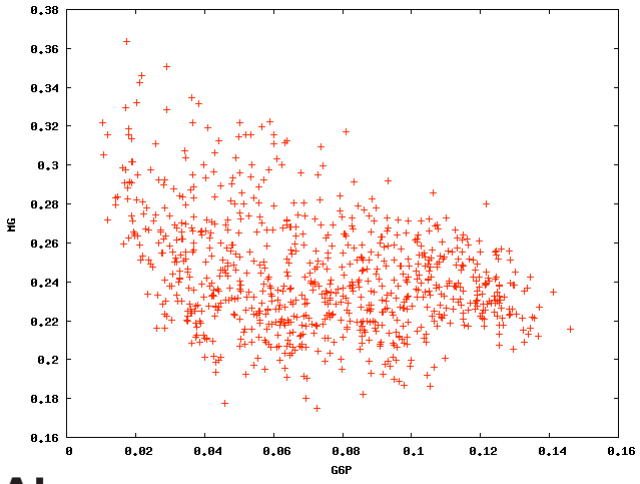
Example



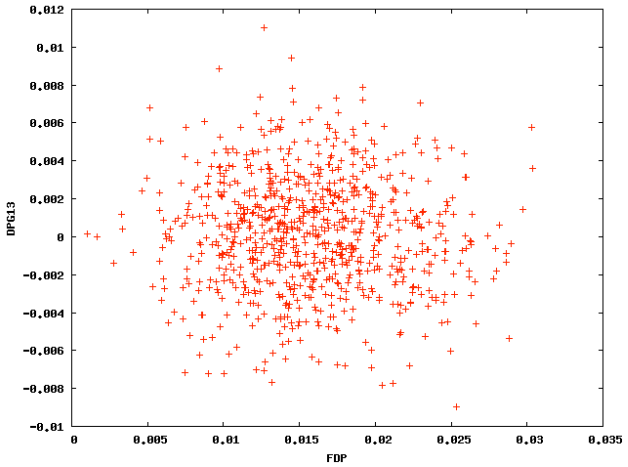
I>0



I=0



I>0



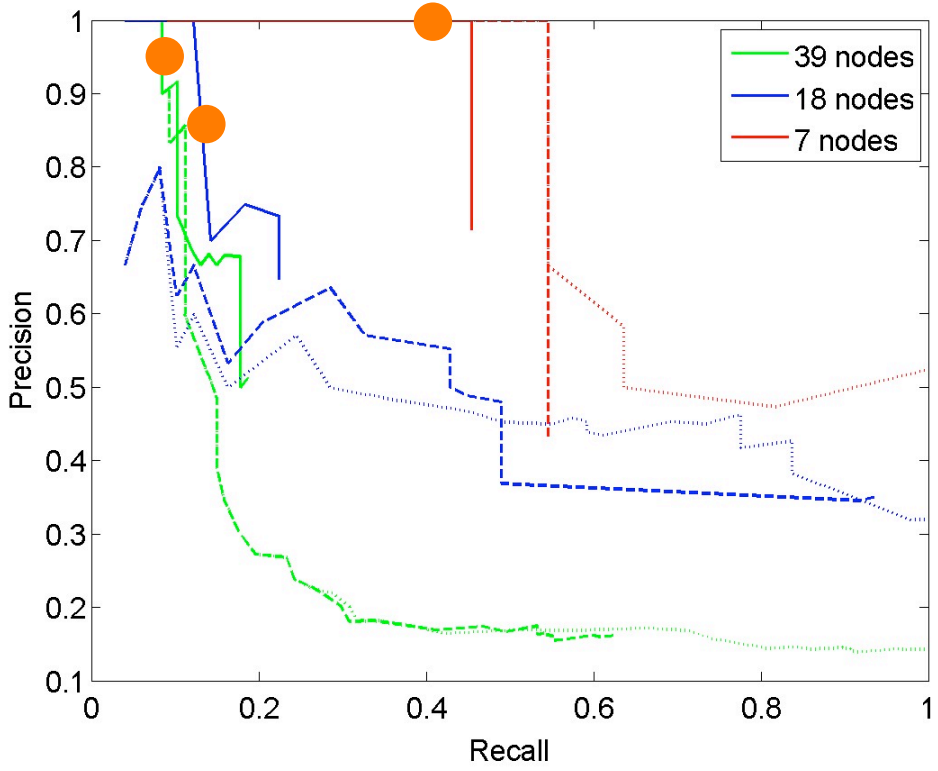
I=0

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



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

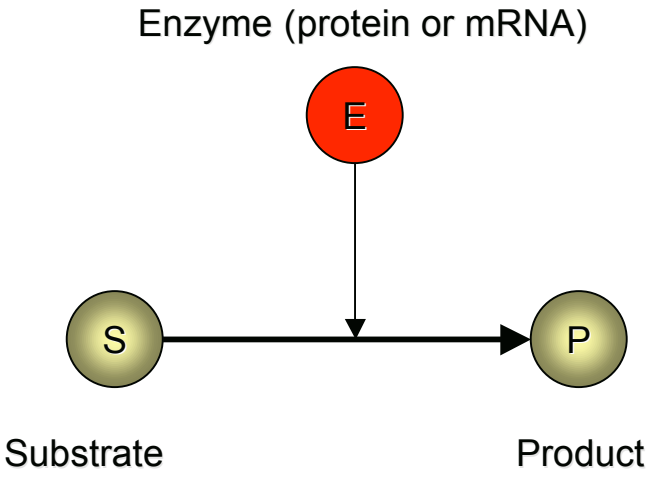
$$r = \frac{N_{TP}}{N_{TP} + N_{FN}} = \frac{N_{TP}}{N_{P,\ tot}}$$

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

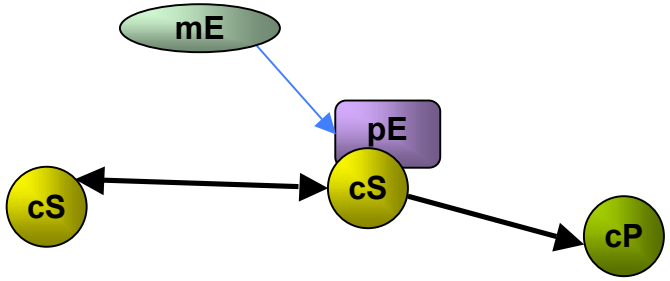
- Operation point for pre-determined *I* threshold

Problems

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



Most interesting next thing
(detection of enzyme-coding genes)

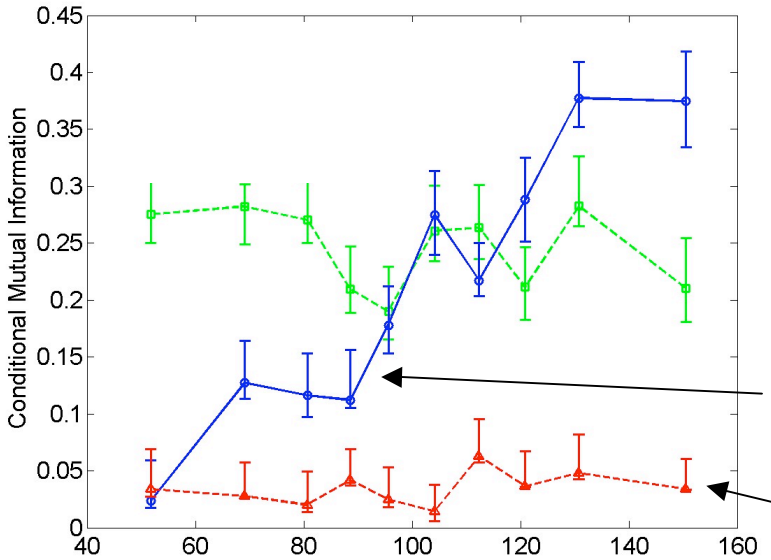


$$\frac{d[\mathrm{pE}]}{dt} = \alpha \cdot [\mathrm{mE}] - k_{+}[\mathrm{pE}][\mathrm{pS}] + k_{-}[\mathrm{pES}] - \beta[\mathrm{pE}]$$

$$\frac{d[\mathrm{cS}]}{dt} = V_0 - k_{+}[\mathrm{pE}][\mathrm{cS}] - \beta[\mathrm{pS}]$$

$$\frac{d[\mathrm{cES}]}{dt} = k_{+}[\mathrm{pE}][\mathrm{cS}] - k_{-}[\mathrm{cES}] - k_{dis}[\mathrm{cES}] - \beta[\mathrm{cES}]$$

$$\frac{d[\mathrm{cP}]}{dt} = k_{dis}[\mathrm{cES}] - \beta[\mathrm{cP}]$$



$I(cS, cP \mid mE)$

other mRNA

(Wang et al, 2006)