Reconstruction of Interaction Networks (With Applications to Transcriptional Regulation)

Ilya Nemenman

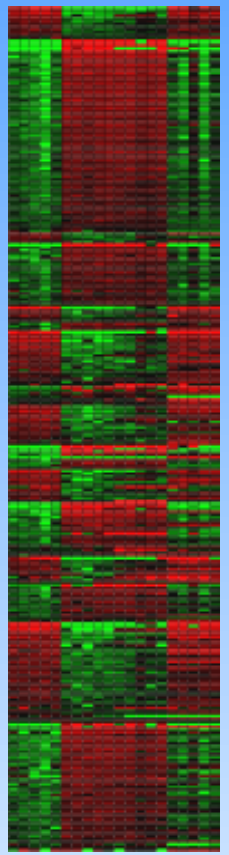
JCSB, Columbia

Andrea Califano, Adam Margolin, Chris Wiggins, Gustavo Stolovitzky, Kai Wang, Riccardo Dalla Favera, Katia Basso, Ulf Klein, Nila Banerjee

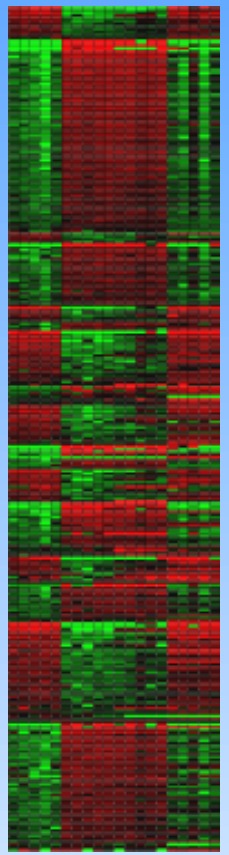
q-bio.MN/0411003, q-bio.MN/0410037

q-bio.MN/0410036, q-bio.QM/0406015

Gene expression analysis

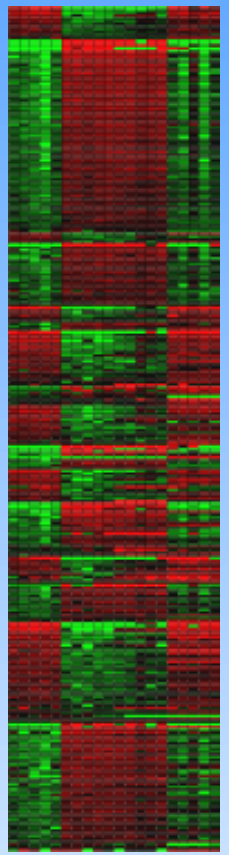


Gene expression analysis



- clustering – too coarse
- reconstructing networks – Holy Grail!

Gene expression analysis

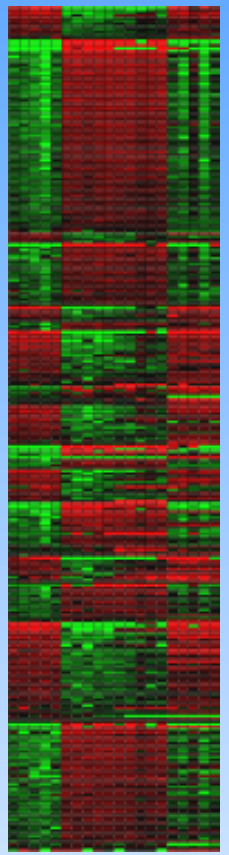


- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:

- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

Gene expression analysis



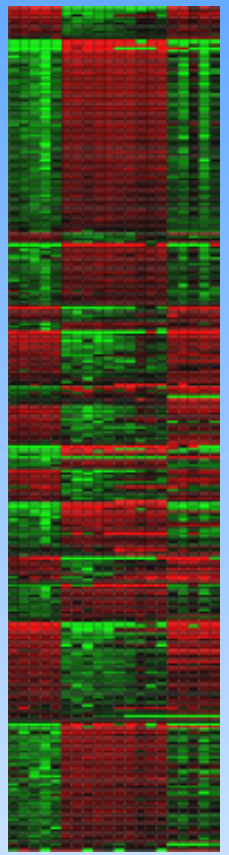
- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:

- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

Different conditions – different steady states.

Gene expression analysis



- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:

- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

Different conditions – different steady states.

Extremely many false positives (e. g. joint co-regulation).

Model of dependence

No time series $\longrightarrow$ steady state statistical dependencies only.

Model of dependence

No time series $\longrightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \ldots$$

Model of dependence

No time series $\longrightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \ldots$$

- use MaxEnt to define ϕ
- connections with spin glasses, MNs, belief propagation

Model of dependence

No time series $\longrightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \ldots$$

- use MaxEnt to define ϕ
- connections with spin glasses, MNs, belief propagation
- enough data to evaluate 2-way marginals only;

Model of dependence

No time series $\longrightarrow$ steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \cancel{\sum_{ijk} \phi_{ijk}(g_i, g_j, g_k)} + \ldots$$

- use MaxEnt to define ϕ
- connections with spin glasses, MNs, belief propagation
- enough data to evaluate 2-way marginals only;
- truncate at 2nd order potential (cannot reconstruct `XOR`), Bethe approximation (but inverse problem)
- Mutual information $I(g_i, g_j) = I_{ij}$ is enough to establish dependencies.

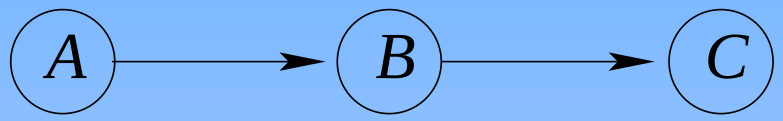
Notes

- changing $\phi(g_i)$ describes response to perturbations (but: directionality)

Notes

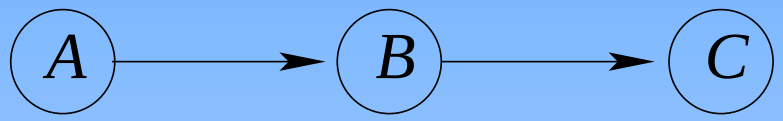
- changing $\phi(g_i)$ describes response to perturbations (but: directionality)
- biochemical dependencies persist as steady state statistical dependencies, but orders of interactions may change

Removing false positives – Data Processing inequality



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

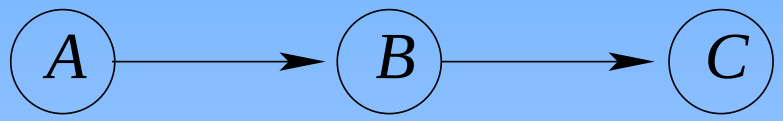
Removing false positives – Data Processing inequality



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

<u>ARACNE:</u> Look at every triplet and remove the weakest link.

Removing false positives – Data Processing inequality



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

ARACNE: Look at every triplet and remove the weakest link.

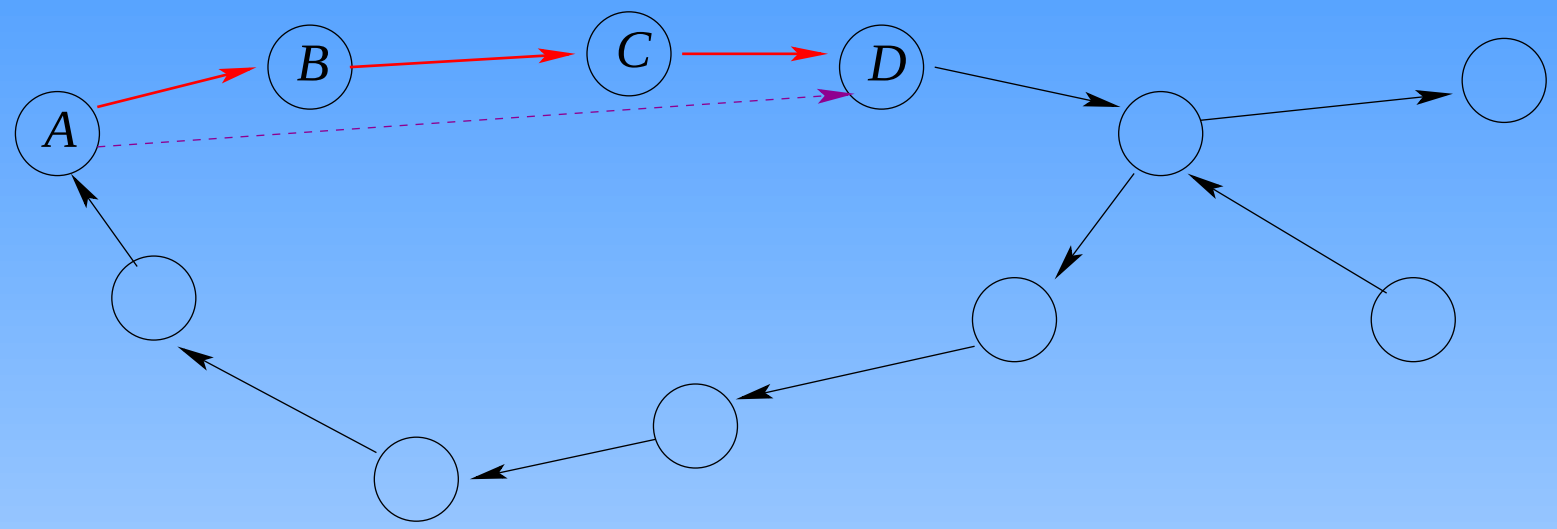
Every 3-gene loop is opened!

Guarantees

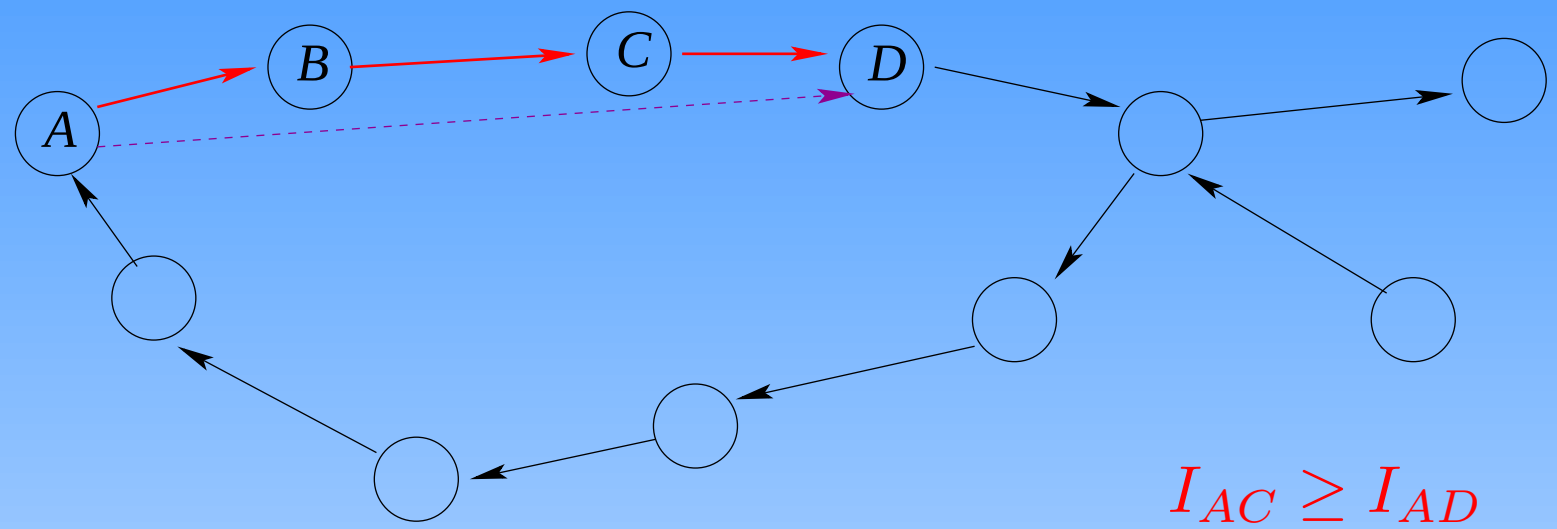
Theorem. If MIs can be estimated with no errors, then ARACNE reconstructs the underlying interaction network exactly, provided this network is a tree and has only pairwise interactions.

.

Theorem. The maximum Mutual Information spanning tree (Chow-Liu) is a subnetwork of the network reconstructed by ARACNE.



<u>Theorem</u>. Let π_{ik} be the shortest path between i and k. Then, if MIs are known, ARACNE reconstructs an interaction network without false positives edges, provided: (a) the network consists only of pairwise interactions, (b) for each $j \in \pi_{ik}$, $I_{ij} \geq I_{ik}$.



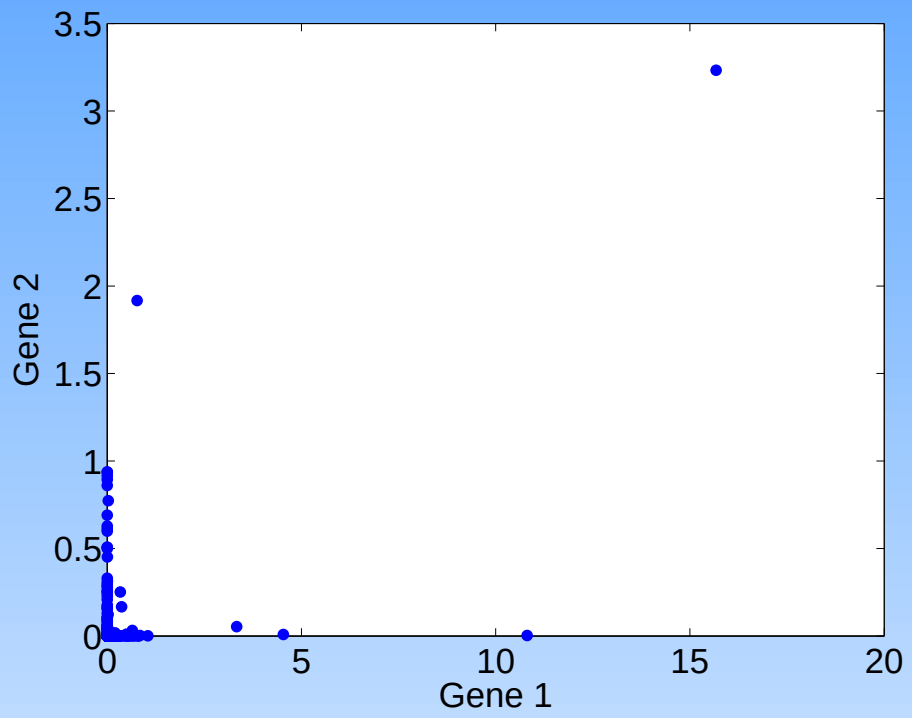
$$I_{AC} \geq I_{AD}$$
$$I_{AB} \geq I_{AD}$$

<u>Theorem</u>. Let π_{ik} be the shortest path between i and k. Then, if MIs are known, ARACNE reconstructs an interaction network without false positives edges, provided: (a) the network consists only of pairwise interactions, (b) for each $j \in \pi_{ik}$, $I_{ij} \geq I_{ik}$.

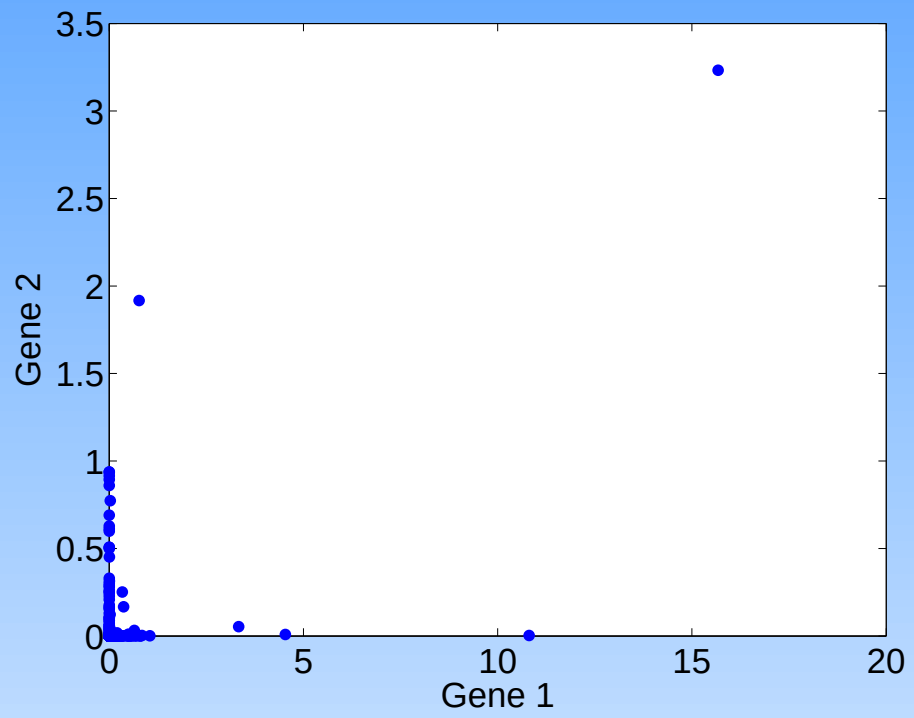
Why should it work?

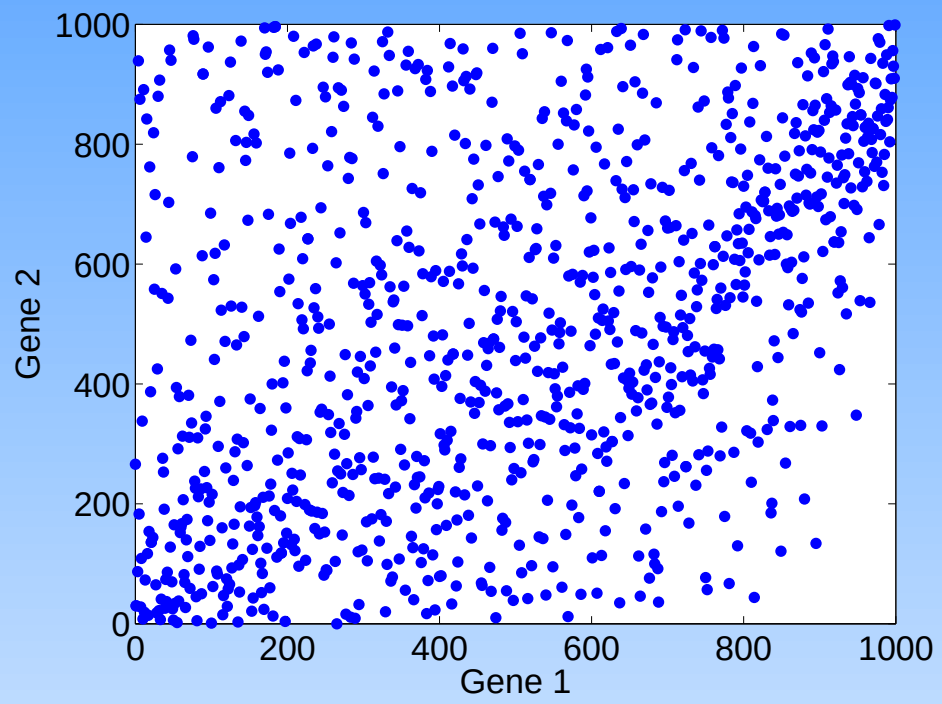
- higher order interactions project into lower order ones
- large loops are locally trees (biological signals decorrelate very fast: $I(\mathrm{cMYK}, \mathrm{cMYK}) \approx 8$ bits, $I(\mathrm{cMYK},$ second best$) \approx 1$ bit.
- small loops (e. g., feed forward) are often transient

Kernel MI estimation: Copula transform

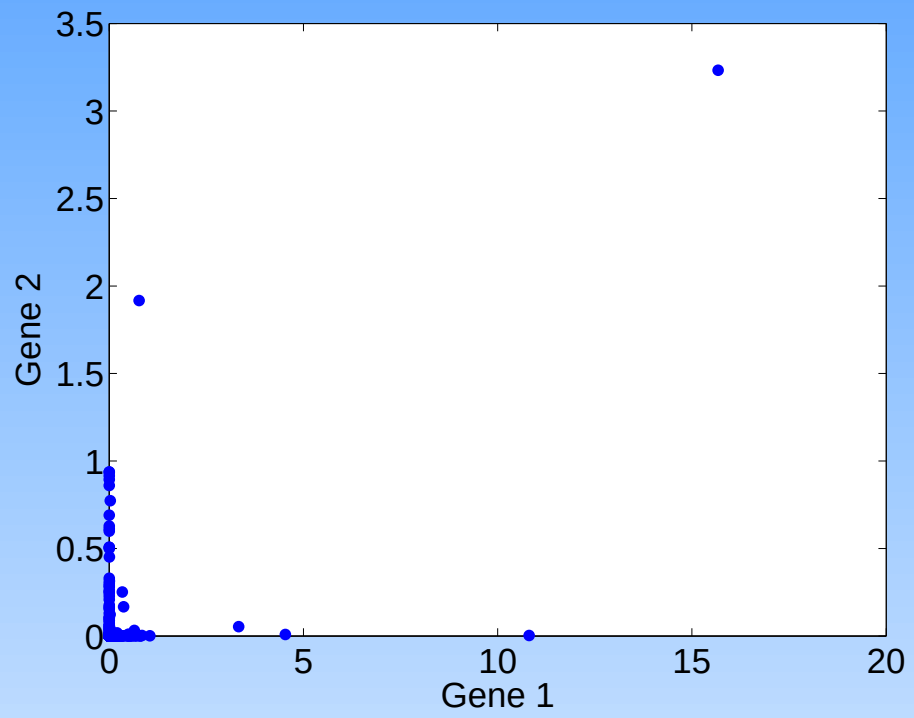


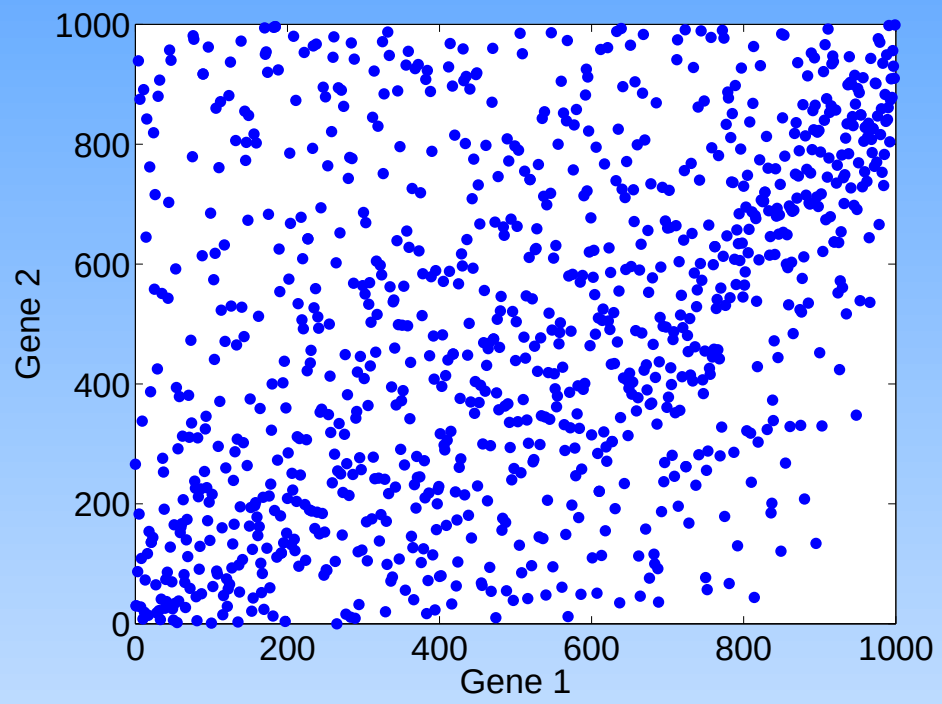
Kernel MI estimation: Copula transform





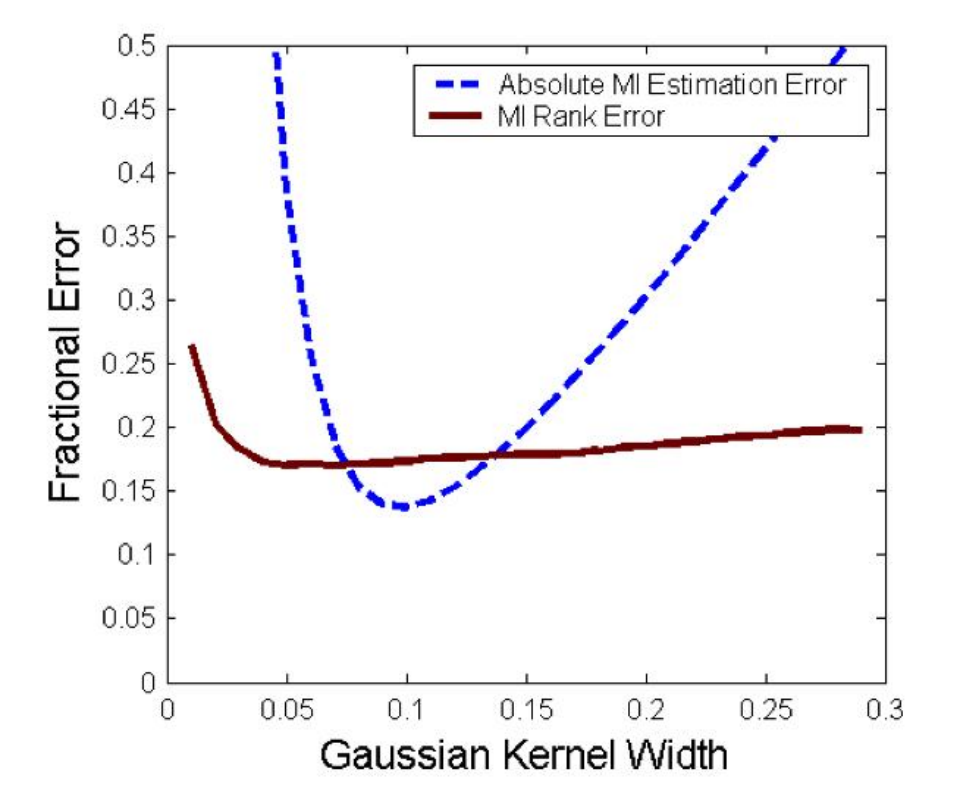
Kernel MI estimation: Copula transform



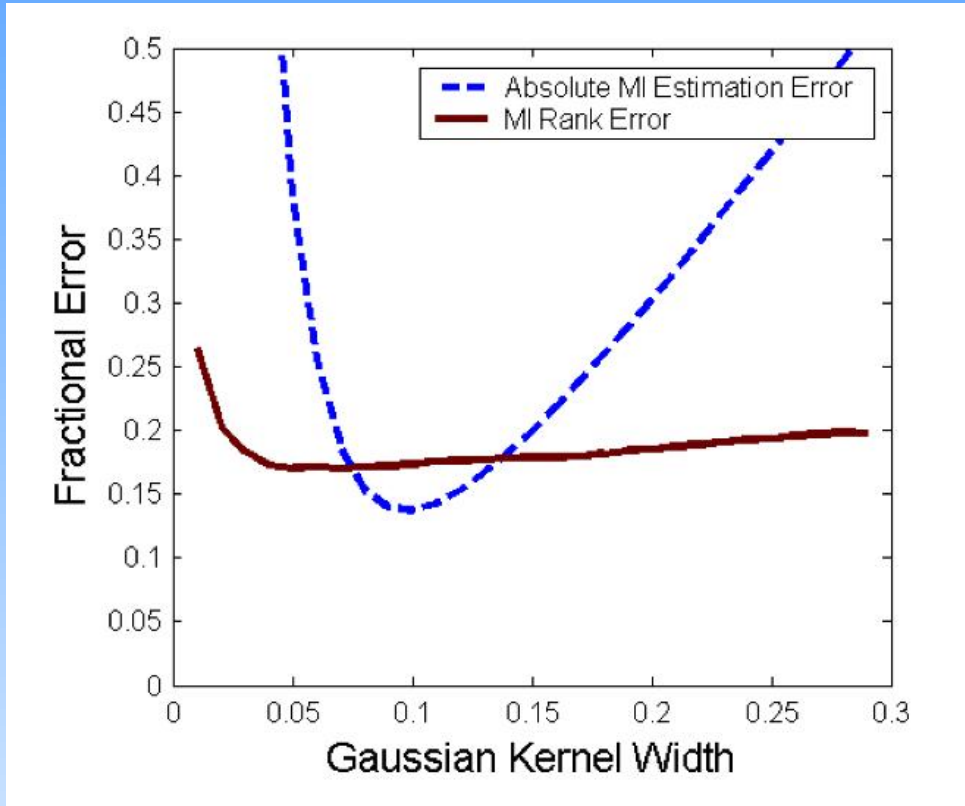


Do before estimating MI. No need for spatial inhomogeneity.

Mutual information error vs. ranking error

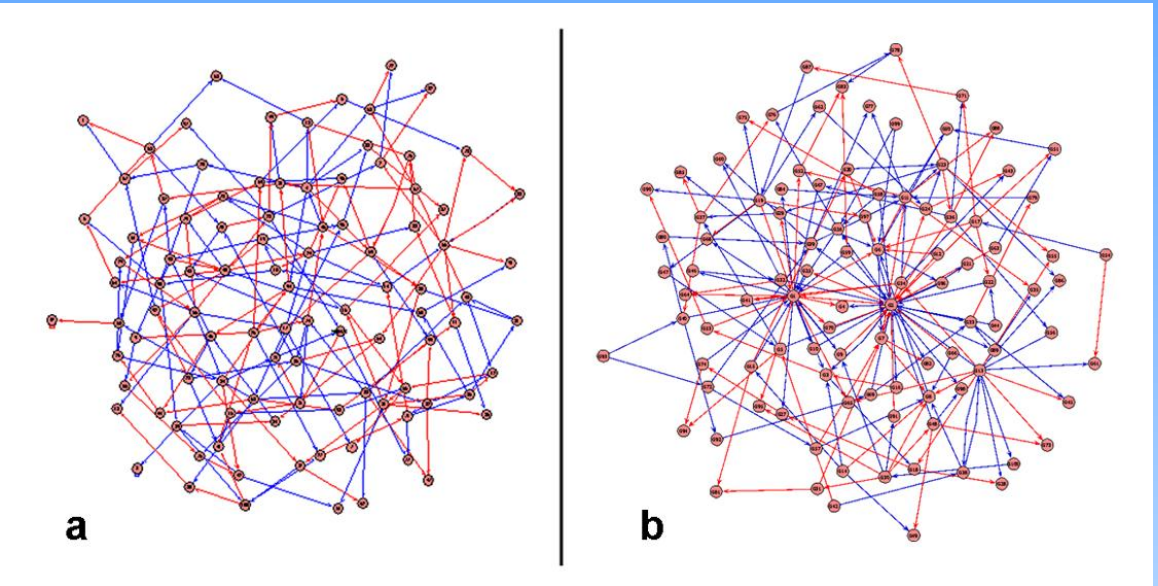


Mutual information error vs. ranking error

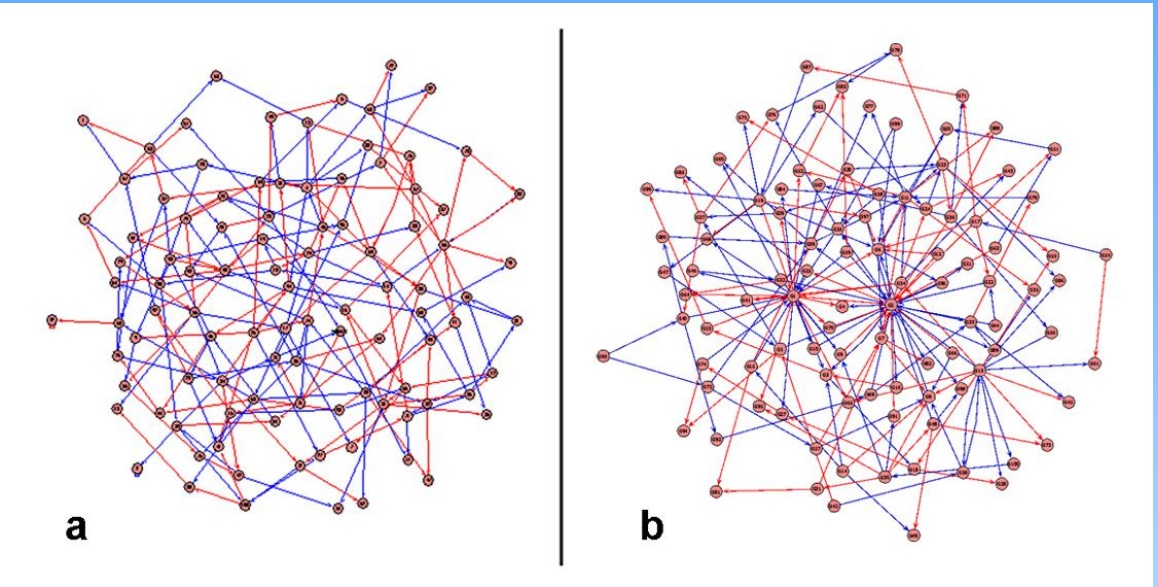


Can use universal best h.

Synthetic networks

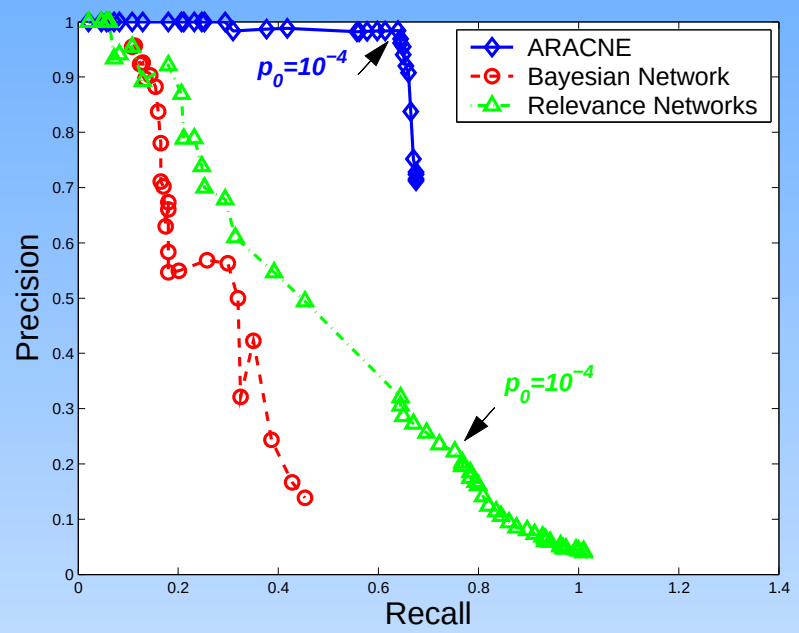


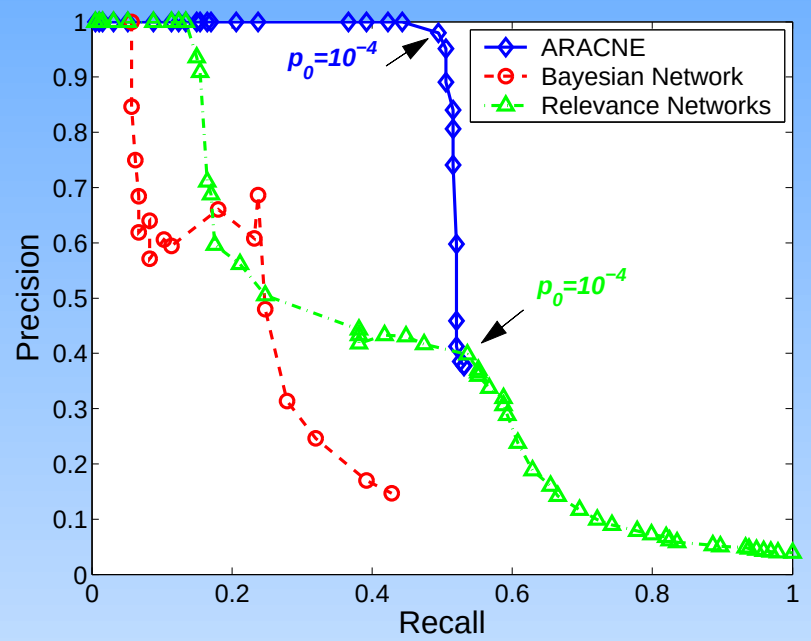
Synthetic networks



$$\frac{dx_i}{dt} = a_i \prod_j \frac{I_{0,j}^{\nu_j}}{I_j^{\nu_j} + I_{j,0}^{\nu_j}} \prod_j \left(1 + \frac{A_j^{\nu_j}}{A_j^{\nu_j} + A_{j,0}^{\nu_j}}\right) - b_i x_i$$

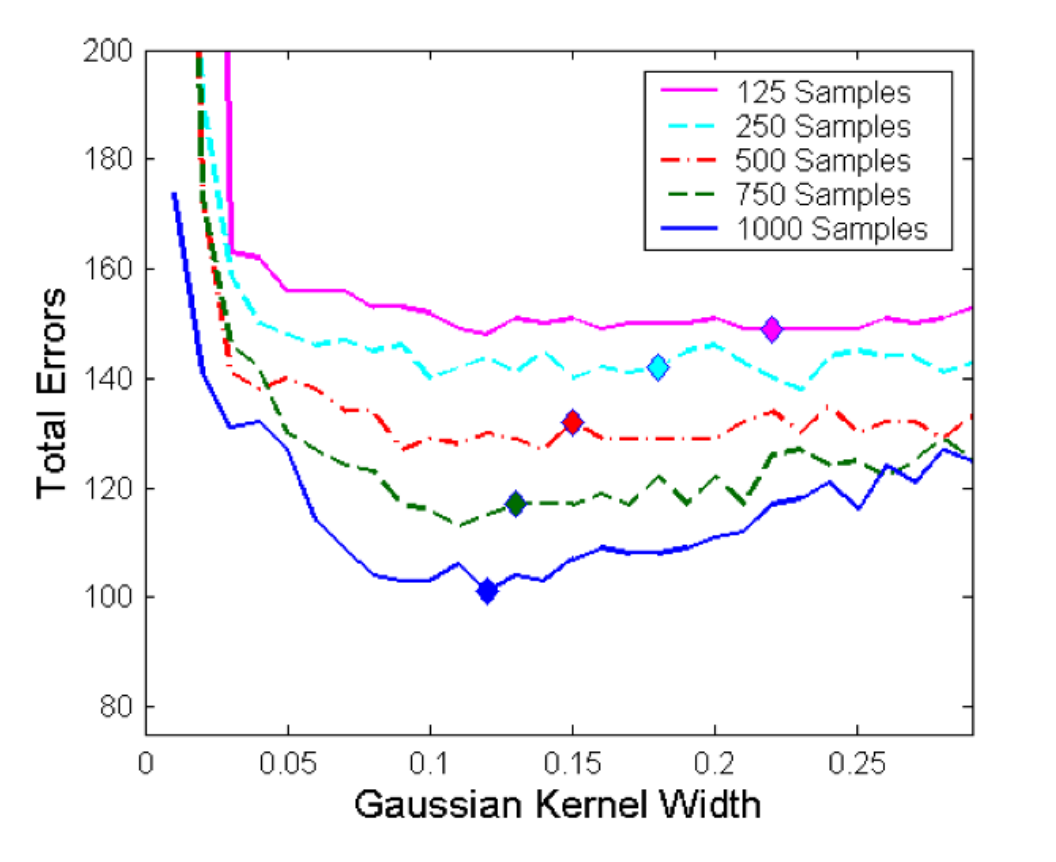
Benchmarks





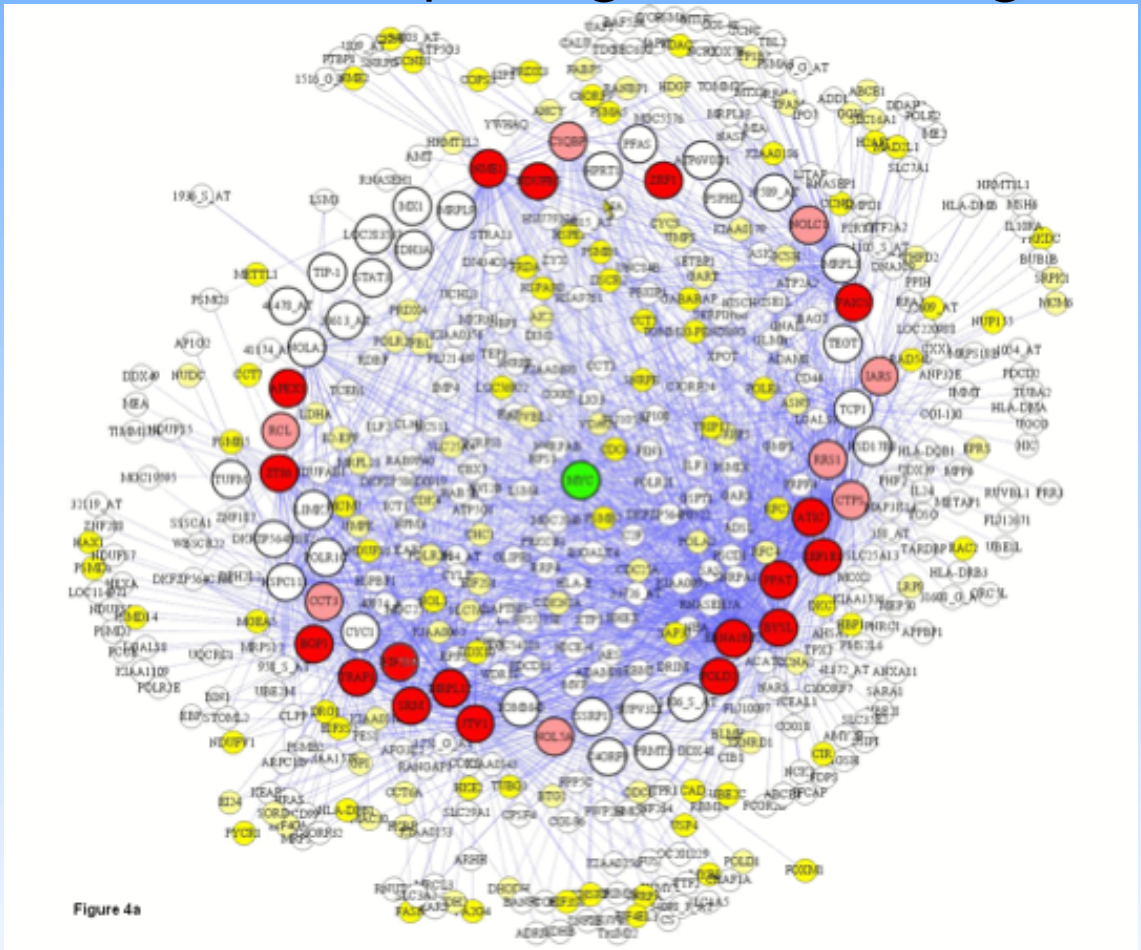
$$N_{TP} - N_{FP} = \max \text{ at } p = 10^{-4}.$$

No sampling catastrophe!



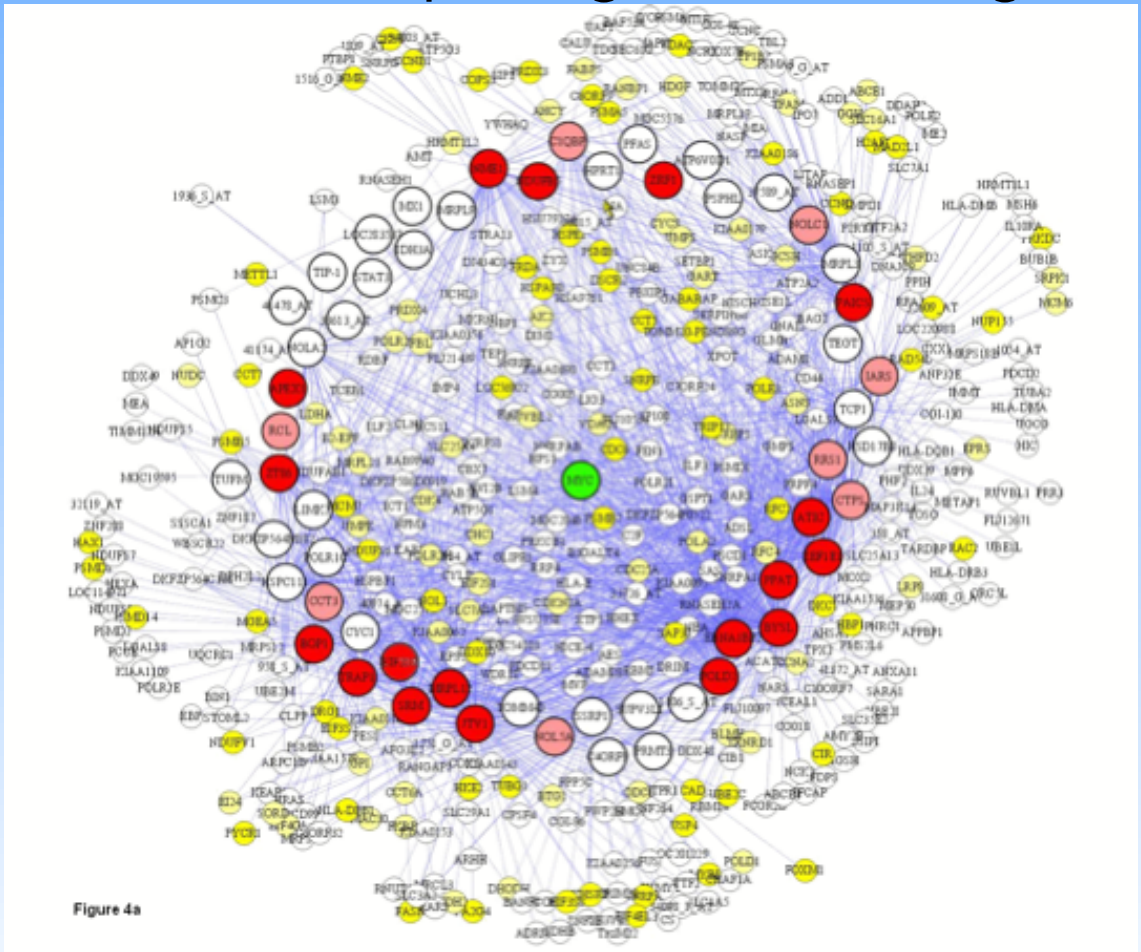
c-MYC TF centered network

Protooncogene, involved in many cellular processes, 12% background interactions, top 5% genetic hub, significant MI with ~ 2000 genes.



c-MYC TF centered network

Protooncogene, involved in many cellular processes, 12% background interactions, top 5% genetic hub, significant MI with ~ 2000 genes.



- 56 1st neighbors
- pre-known targets – 22
- ChIp-proven targets – 11/12
- 2nd neighbors – weaker enrichments
- Most 1st – major hubs

Hub 3-way interactions (conditional analysis)

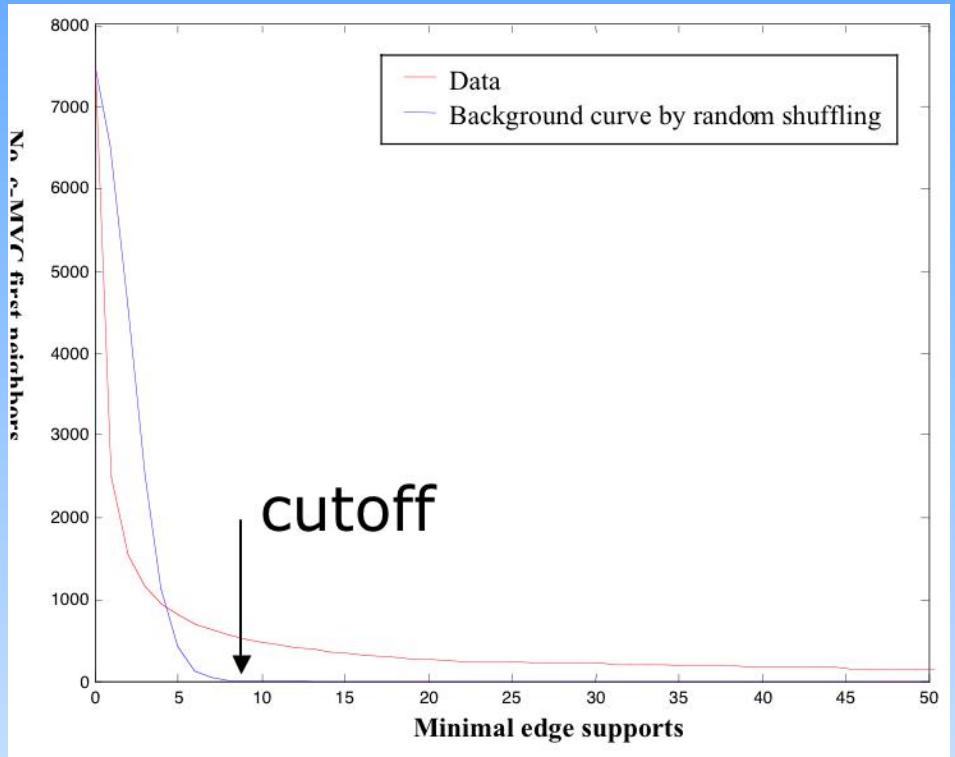
G^*_μ – coarse conditions (+/-) of correlated gene clusters

$$I(g_i, g_j | G^*_\mu)$$

- Independent of the hub (true 3-way interactions)
- Large dynamic range

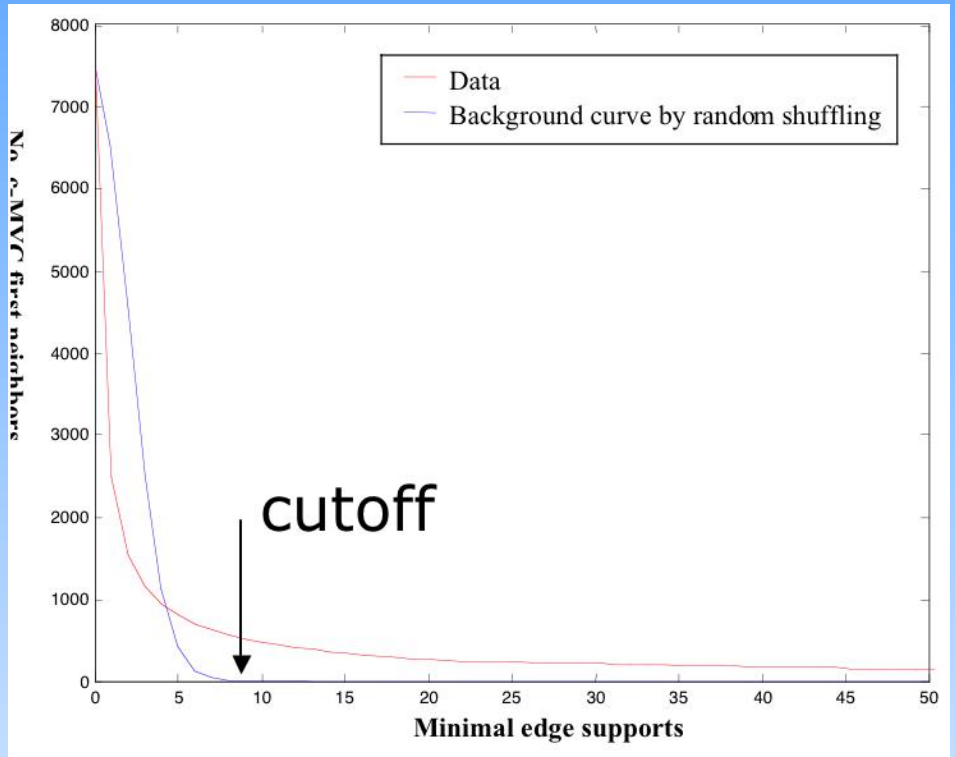
	G_1^+	G_1^-	G_2^+	G_2^-	...	...	G_M^+	G_M^-
Edge 1	1	0	1	0	0	0	1	0
Edge 2	0	1	0	0	0	1	0	0
:	0	0	1	0	1	0	0	0
Edge N	1	0	0	1	0	0	0	1

Edge support conditions set size



#	N_P	N_V	E.	N_{FP}	P
1	2422	437	0.18	6520.1	1
2	1458	278	0.19	4541.5	1
3	1066	224	0.21	2514.1	1
4	847	182	0.21	1131.6	1
5	710	157	0.22	423.13	0.60
6	591	132	0.22	136.04	0.23
7	511	119	0.23	37.03	0.072
8	459	110	0.24	9.18	0.02
9	406	104	0.26	1.9	< 0.01
10	367	96	0.26	0.37	< 0.01

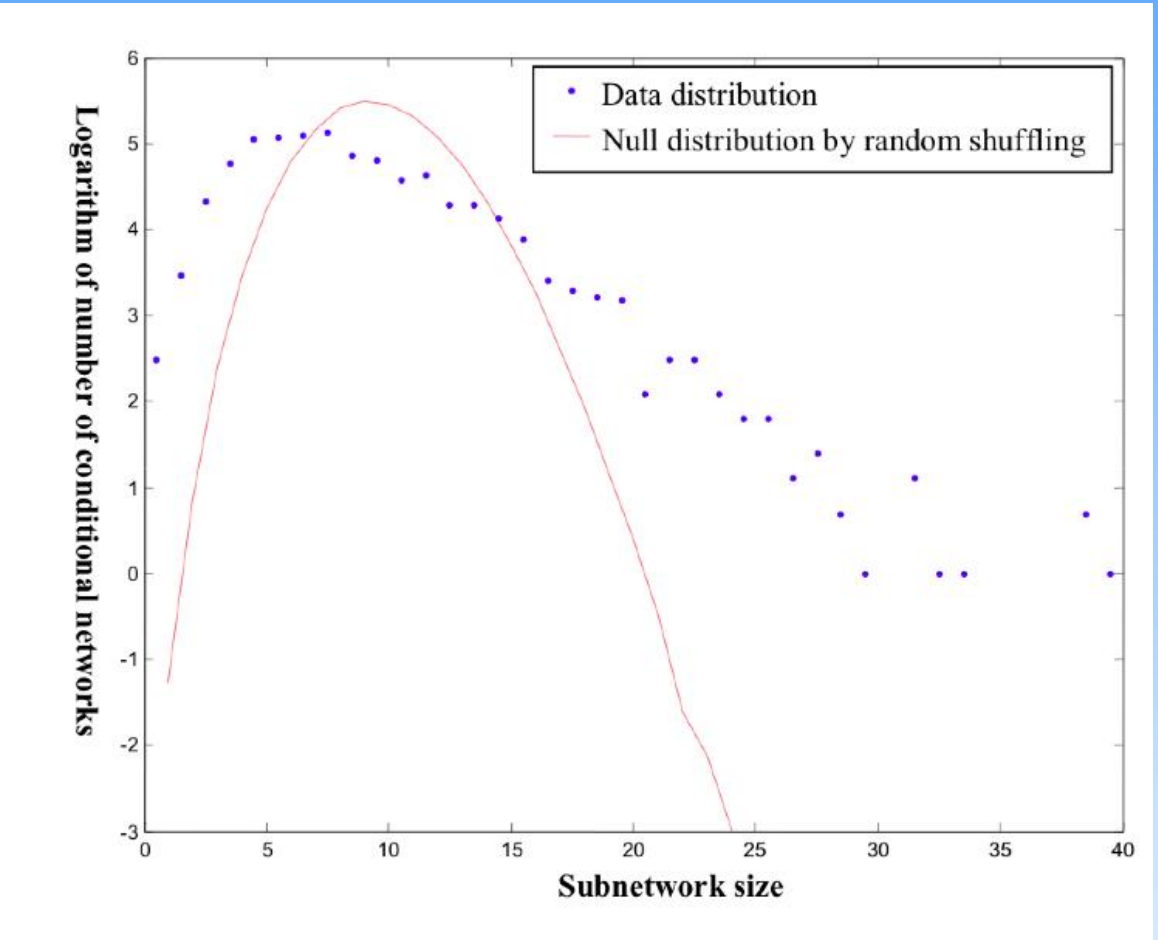
Edge support conditions set size



#	N_P	N_V	E.	N_{FP}	P
1	2422	437	0.18	6520.1	1
2	1458	278	0.19	4541.5	1
3	1066	224	0.21	2514.1	1
4	847	182	0.21	1131.6	1
5	710	157	0.22	423.13	0.60
6	591	132	0.22	136.04	0.23
7	511	119	0.23	37.03	0.072
8	459	110	0.24	9.18	0.02
9	406	104	0.26	1.9	< 0.01
10	367	96	0.26	0.37	< 0.01

Probably better than original algorithm.

Conditional network sizes



Regulators, indeed

Of 168 c-MYC regulators:

GO Category	N_C	N_{Tot}	GO N_C	GO N_{Tot}	P
Transcription Regulator Activity (MF)	21	116	2089	21014	0.0049
Protein Kinase (MF)	13	116	1004	21014	0.003
IKBK/NFKB cascade (BP)	5	117	222	24373	0.004
Immune Response (BP)	19	117	1664	24373	0.0003
Humoral IR (BP)	9	117	378	24373	0.0001
Reg. of Transcription, DNA-dep. (BP)	26	117	1697	24373	1.2×10^{-7}