Network propagation as a tool for deciphering disease mechanisms

Roded Sharan
Blavatnik School of Computer Science
Tel Aviv University

Guilt by association

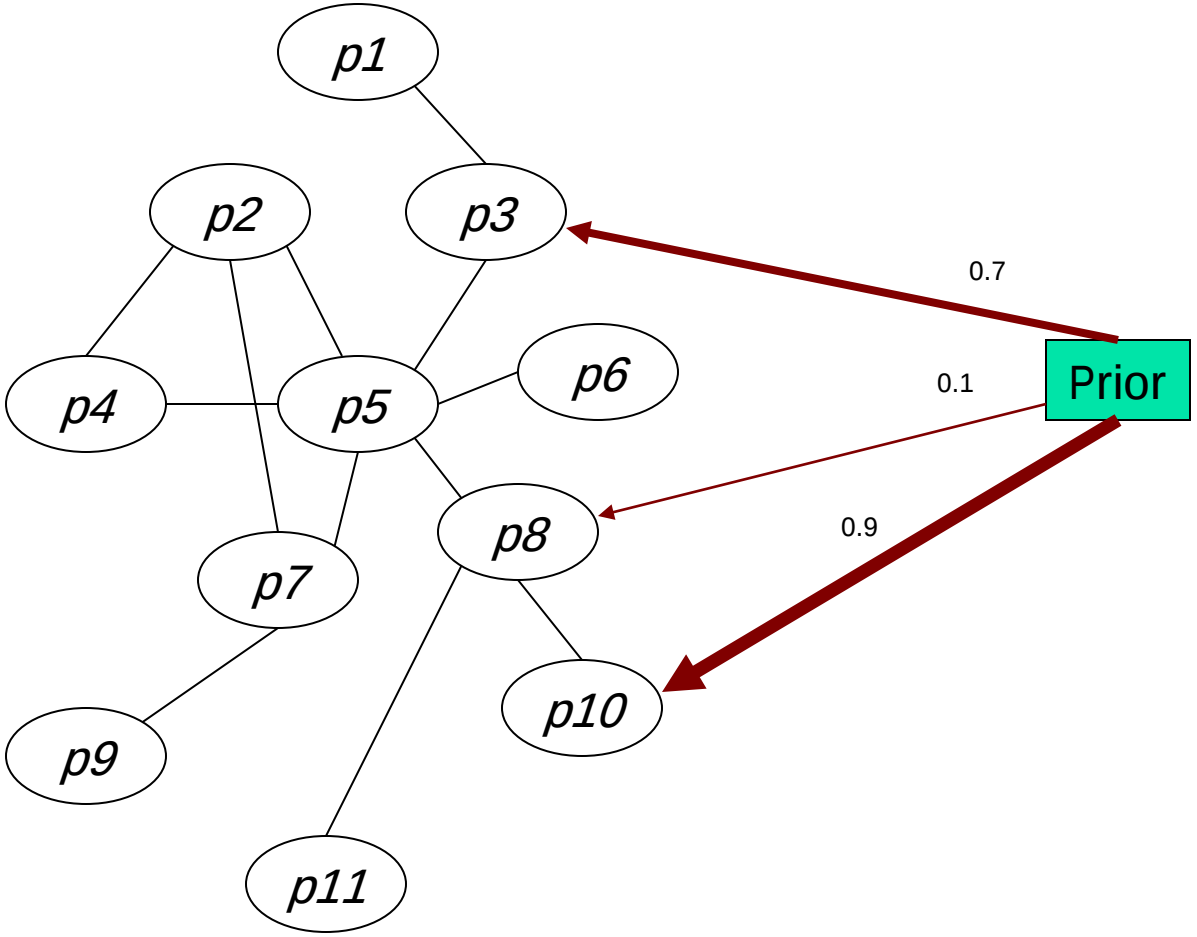
Input: Partial knowledge (genes) on a process/disease of interest

Goal: score genes for relation to the process/disease in the context of a network

Common methods:

- #interactions
- Average distance
- Hypergeometric p-value

Network propagation



The propagation score function

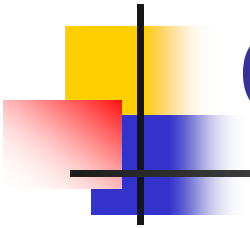
$$F(v) = \alpha \left[\sum_{(u,v) \in E} F(u) w(u,v) \right] + (1-\alpha) Y(v)$$

Two desirable properties/terms:

1. Smoothness over the network
2. Accounts for Prior knowledge

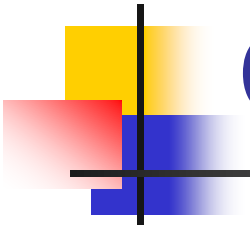
Propagation in network biology

- Nabieva et al.’07, Cao et al.’13 – function prediction
- Kohler et al.’08, Vanunu et al.’10, Shrestha et al.’14 – gene-disease association
- Vandin et al.’11; Leiserson et al.’15 – pathway-disease association
- Hofree et al.’13 – disease stratification



Outline

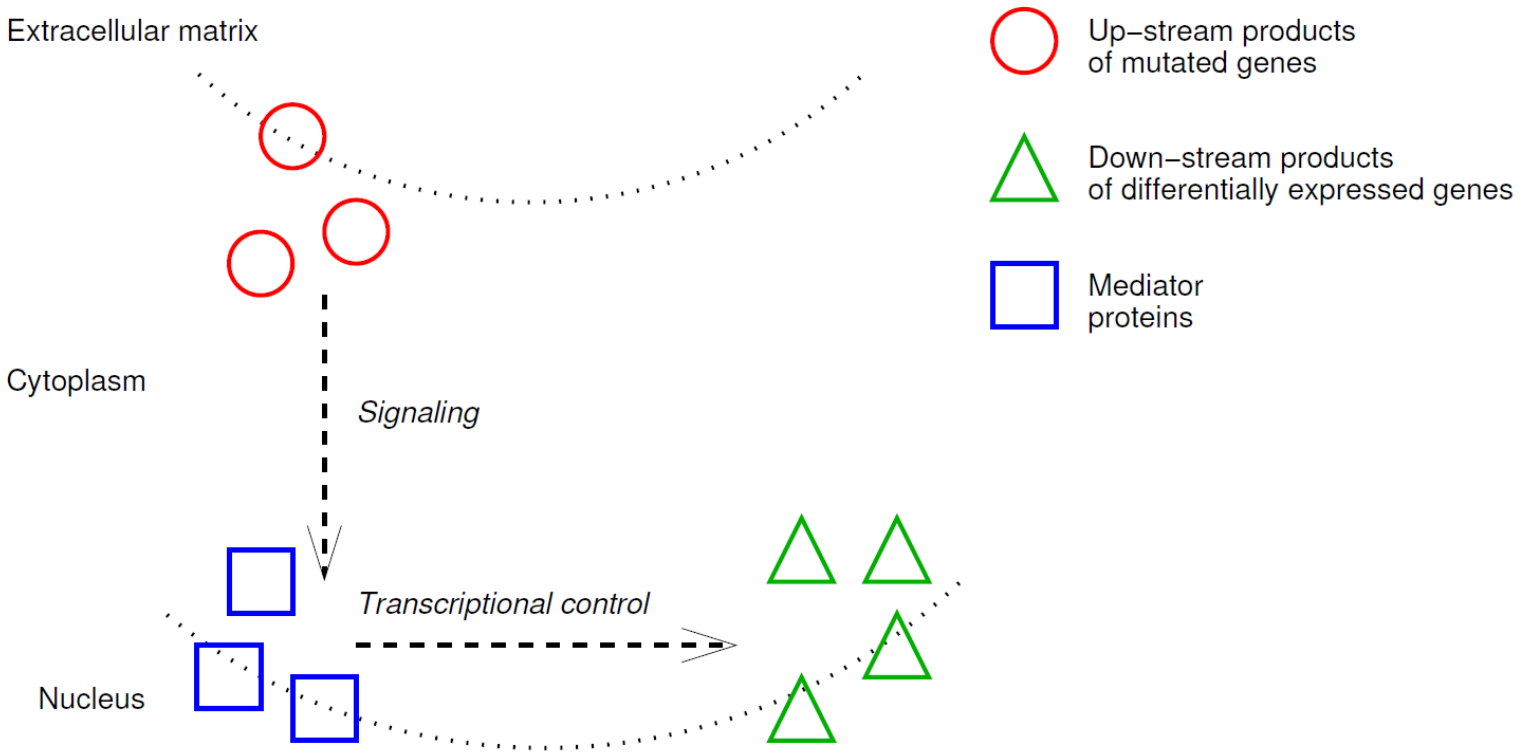
- Finding driver genes (Rufallo, Koyuturk, S.; PLoS Comp. Biol. ‘15)
- Finding disease modules (Mazza, Klockmeier, Wanker, S.; ISMB ’16)

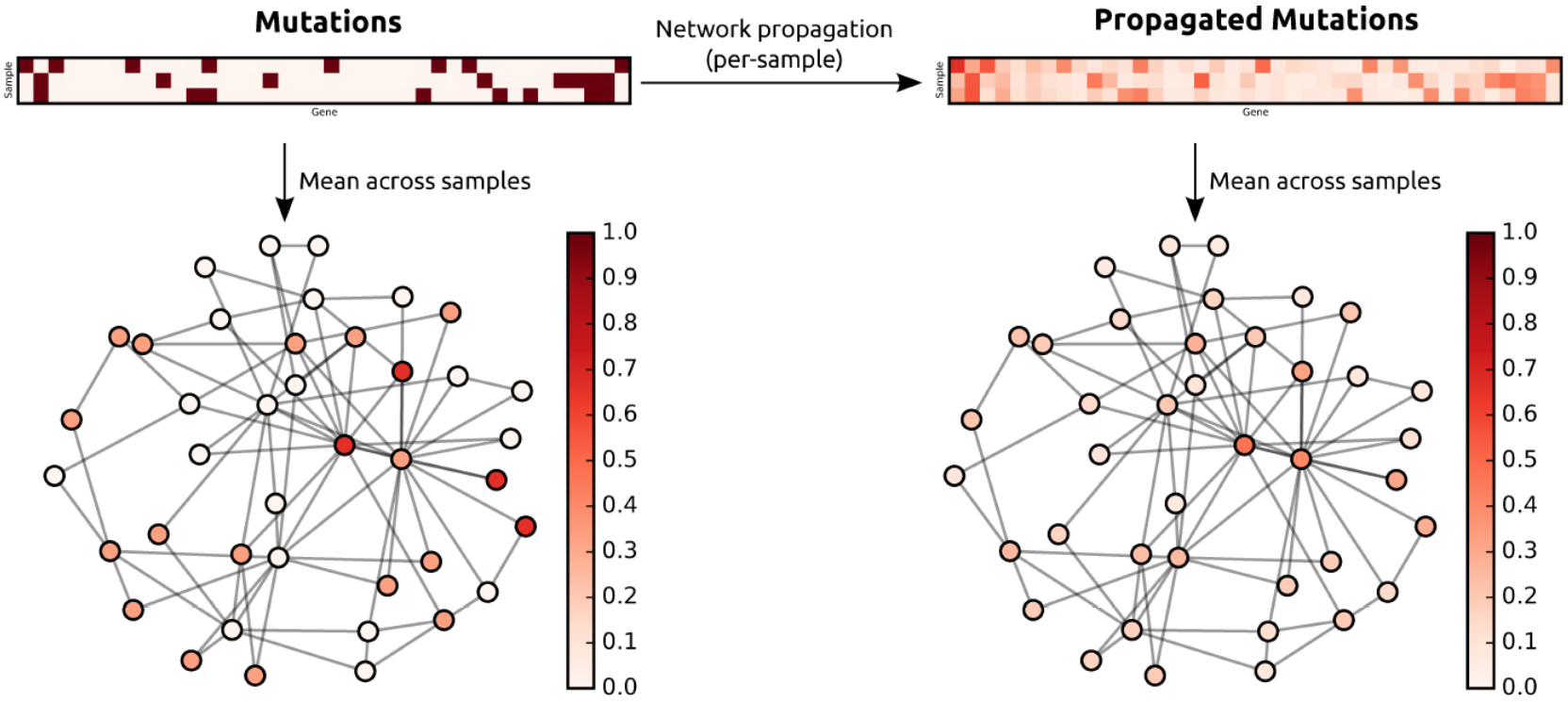


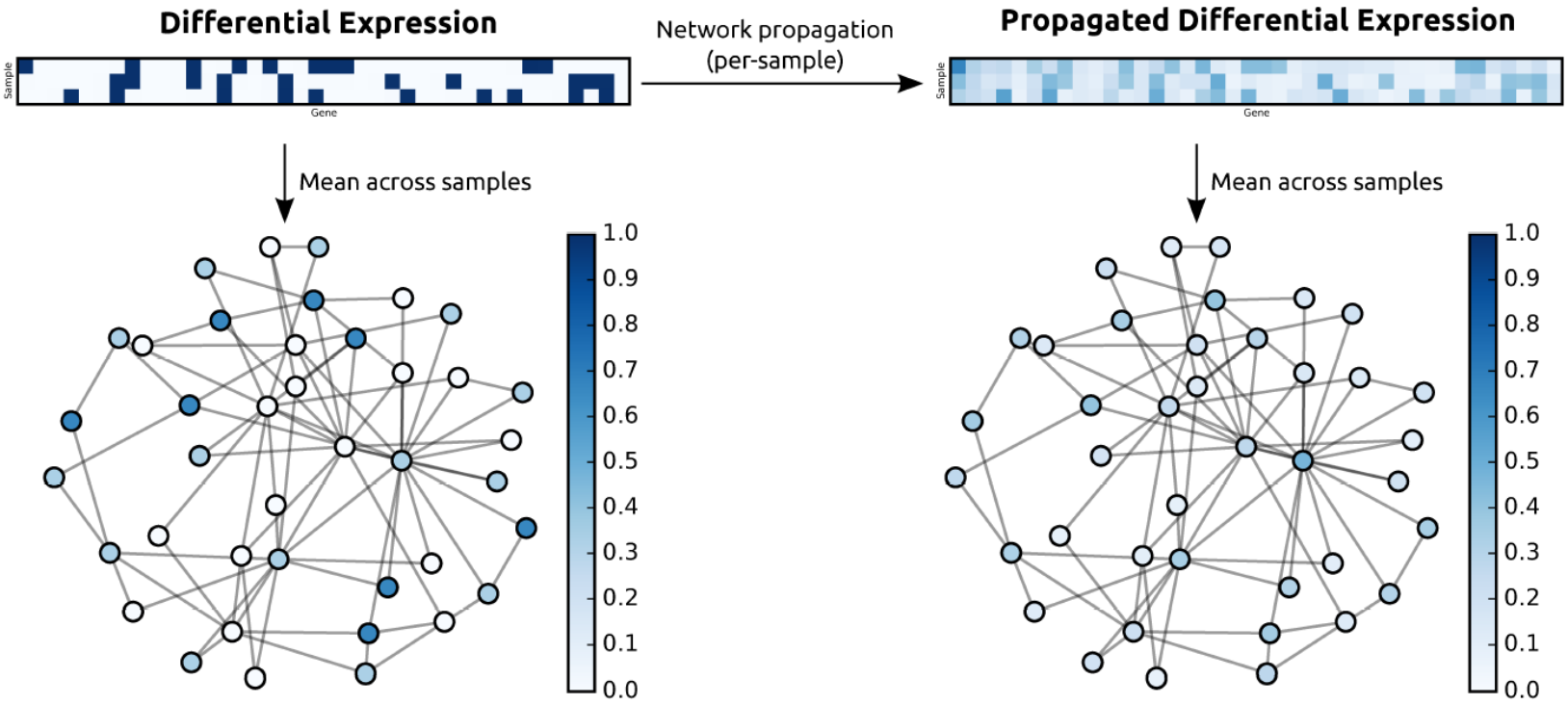
Outline

- Finding driver genes (Rufallo, Koyuturk, S.; PLoS Comp. Biol. ‘15)
- Finding disease modules (Mazza, Klockmeier, Wanker, S.; ISMB ‘16)

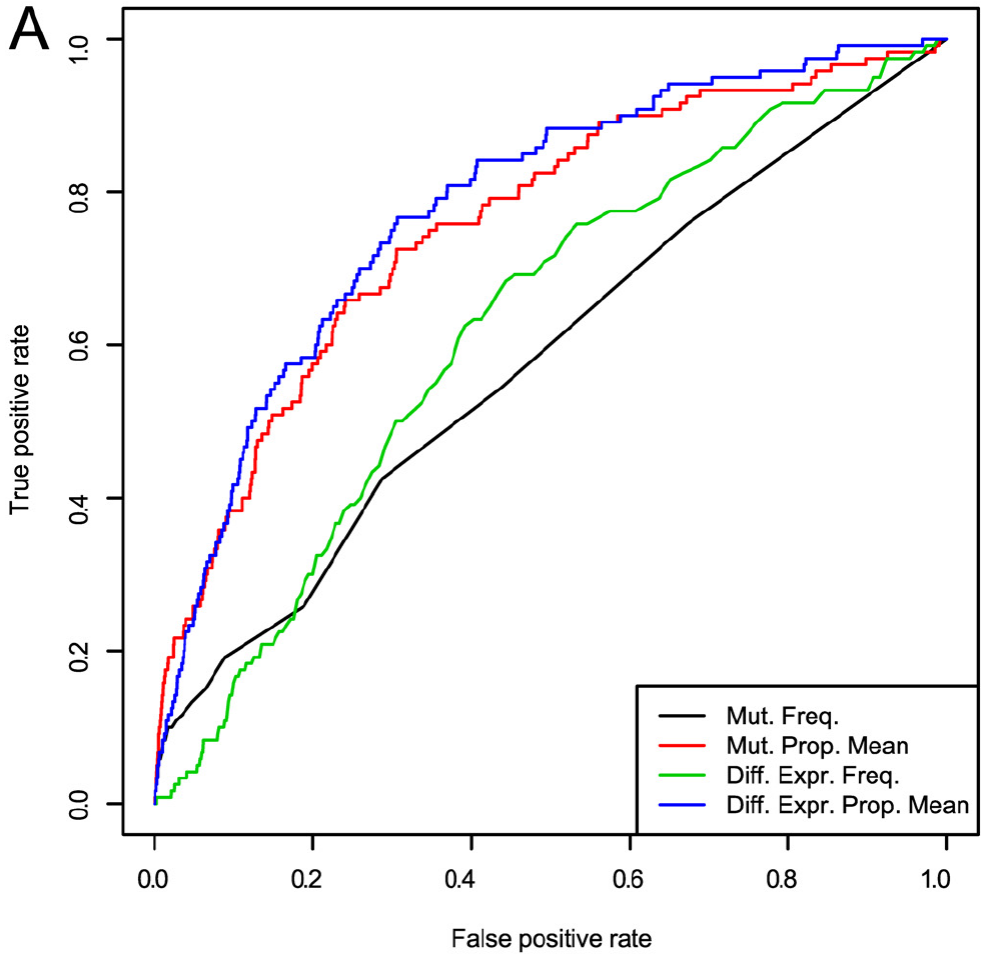
Motivation


Extracellular matrix
Cytoplasm
Signaling
Transcriptional control
Nucleus
Up–stream products of mutated genes
Down–stream products of differentially expressed genes
Mediator proteins


Mutations
Sample
Gene
Network propagation
(per-sample)
Propagated Mutations
Sample
Gene
Mean across samples
Mean across samples
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

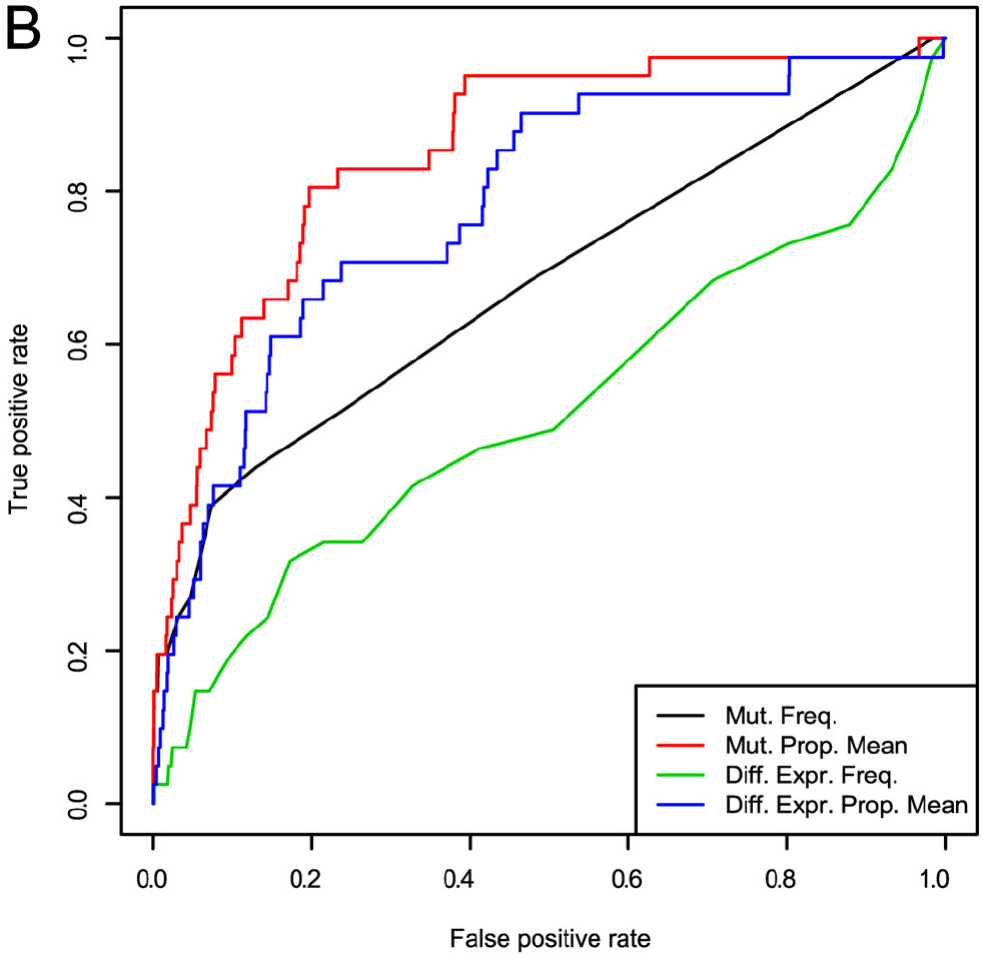

Differential Expression
Sample
Gene
Network propagation
(per-sample)
Propagated Differential Expression
Sample
Gene
Mean across samples
Mean across samples
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

The effect of propagation (BRCA)



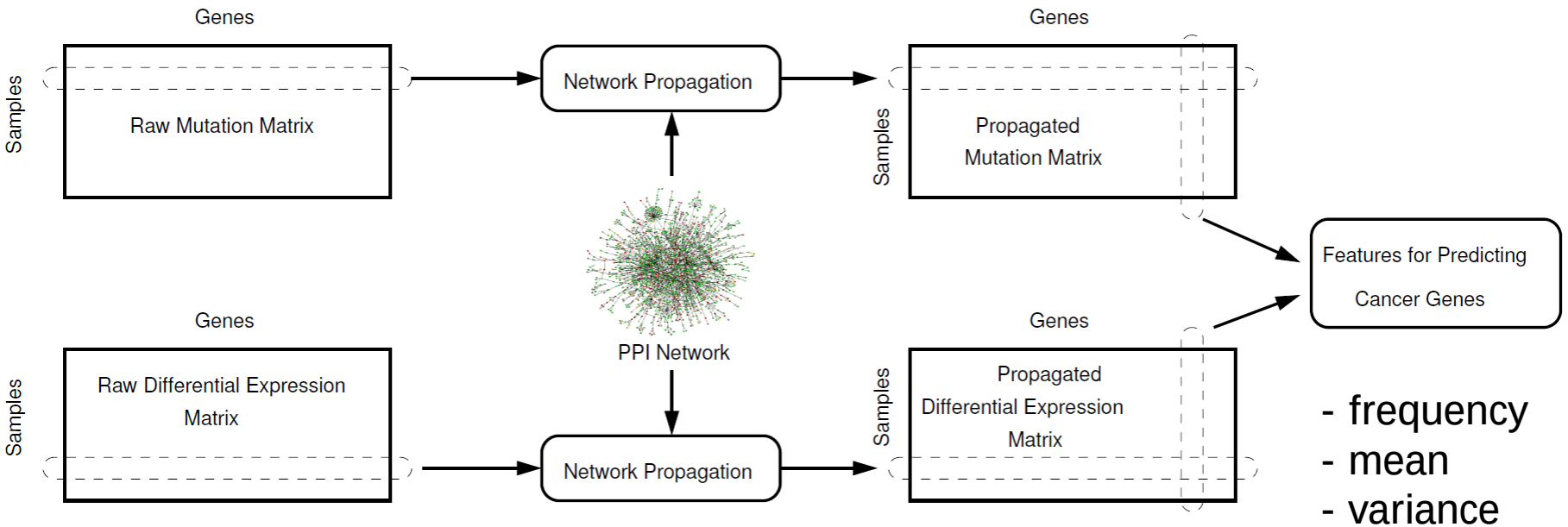
Score	AUC
Mut. Freq	0.581
Mut. Prop. Mean	0.757
Diff. Expr. Freq.	0.625
Diff. Expr. Prop. Mean	0.781

The effect of propagation (GBM)



Score	AUC
Mut. Freq	0.679
Mut. Prop. Mean	0.854
Diff. Expr. Freq.	0.511
Diff. Expr. Prop. Mean	0.782

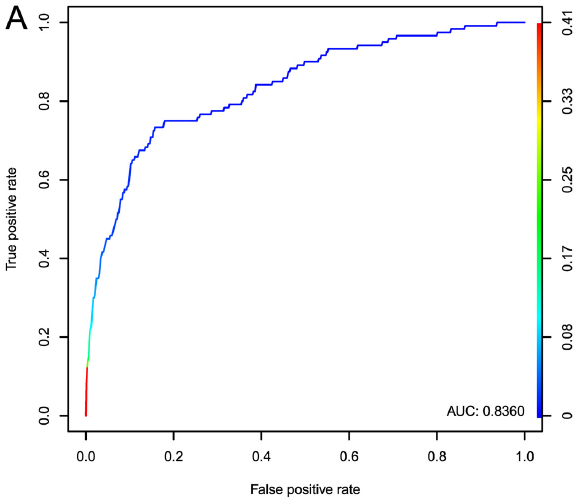
The computational workflow

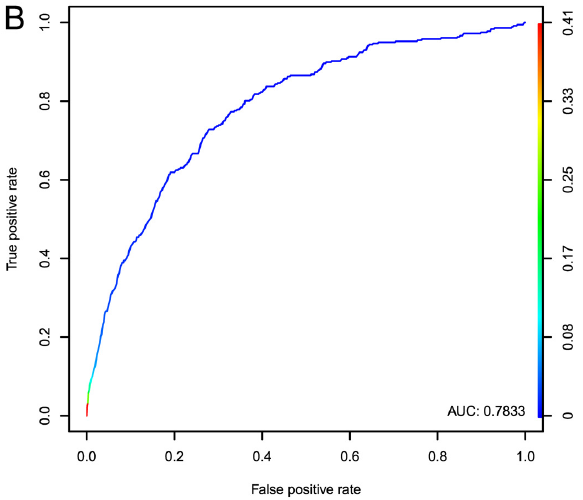


- frequency
- mean
- variance
- correlation
- min/max

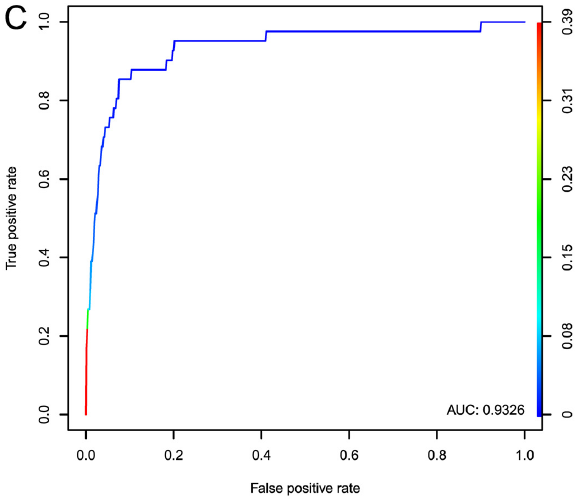
Performance evaluation

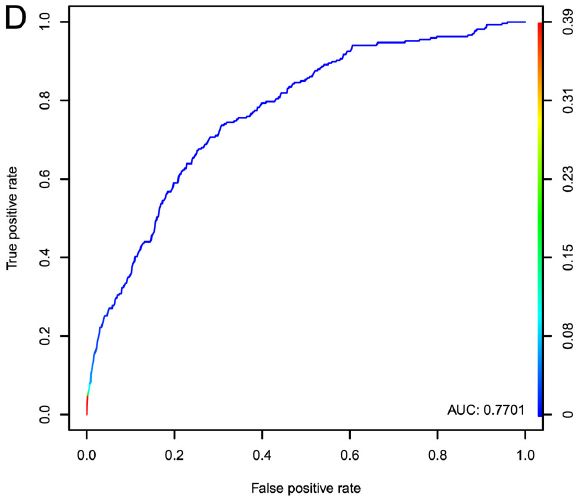
BRCA



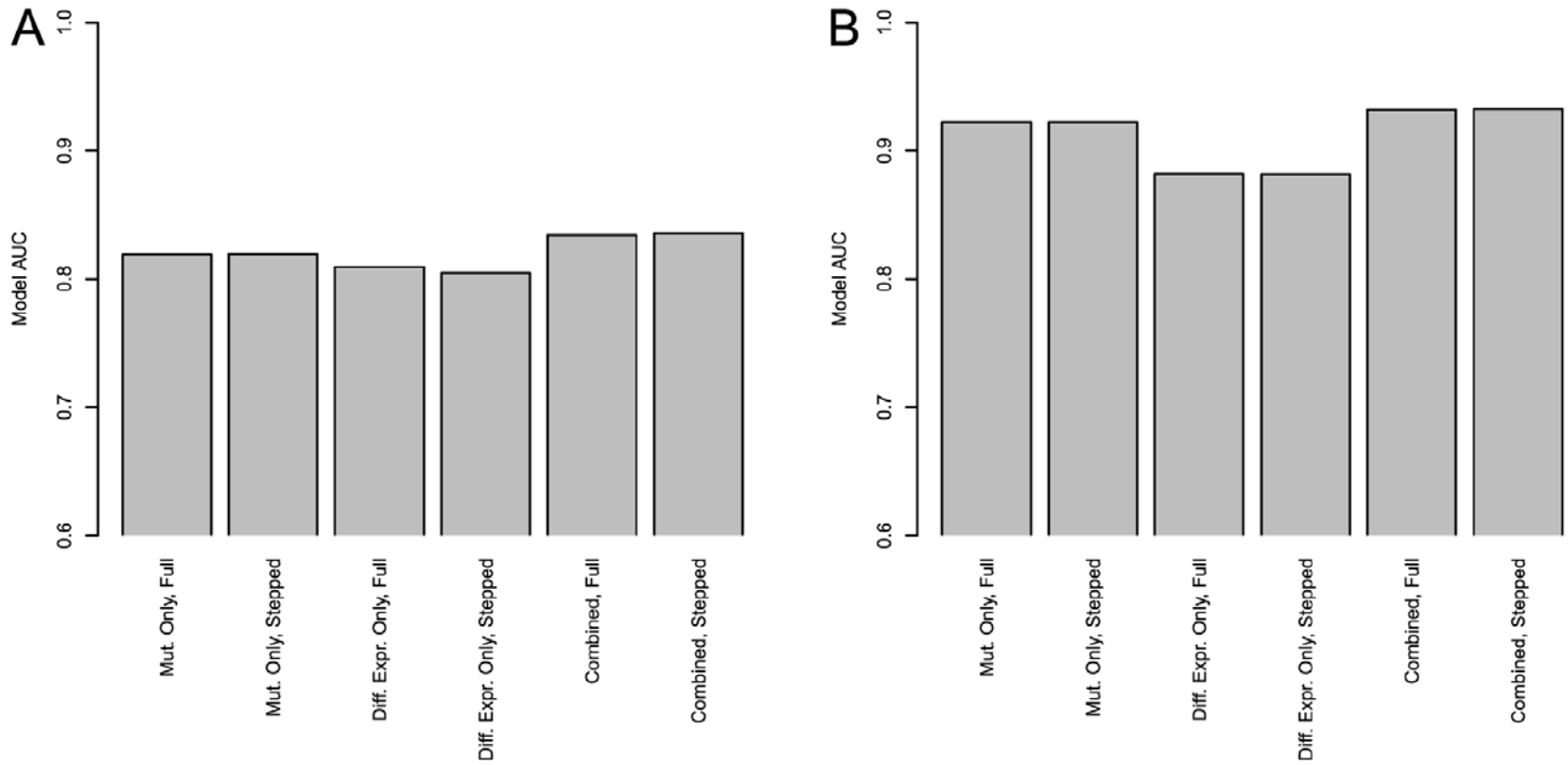


GBM



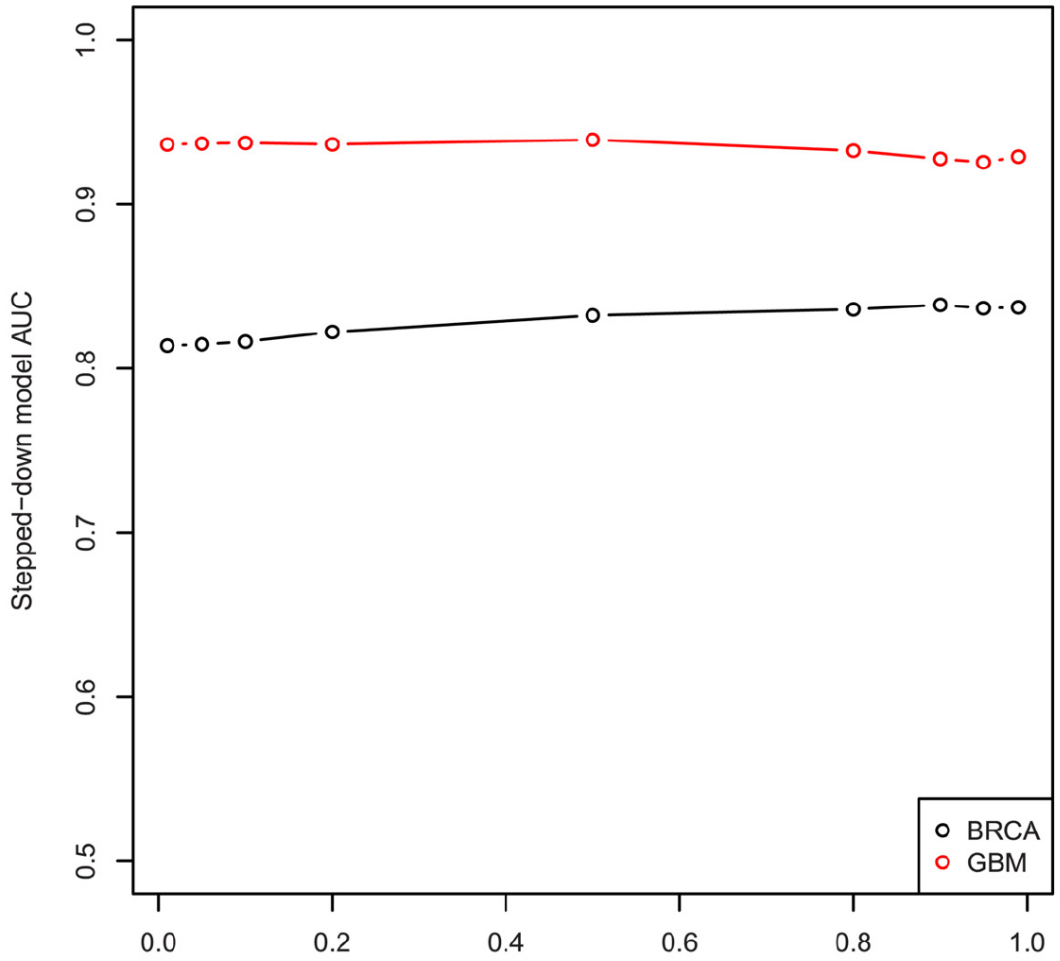


Mutations vs. expression



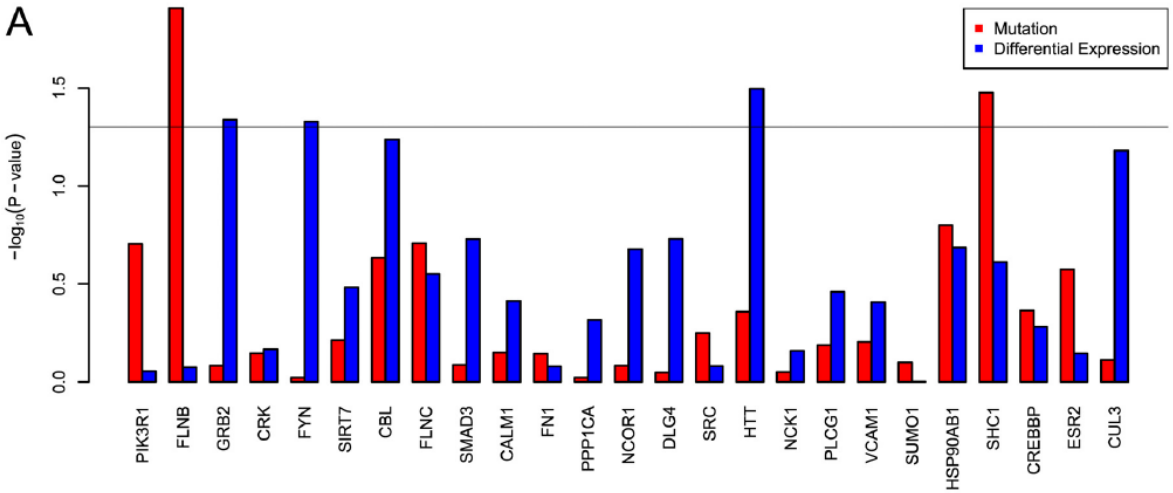
Top performing feature: min(mutation propagation, expression propagation)

Robustness to alpha

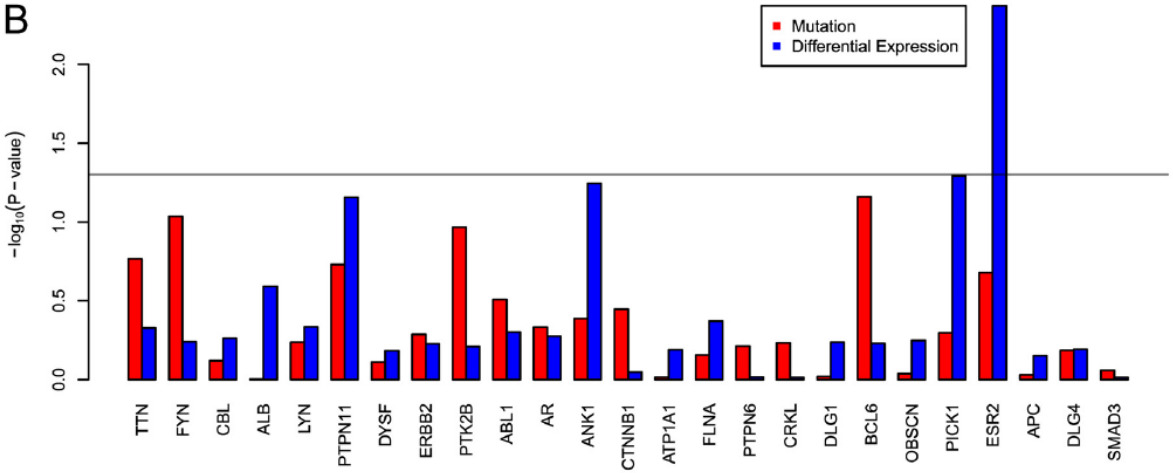


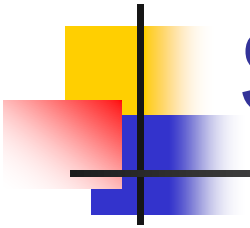
Association with patient outcome

BRCA



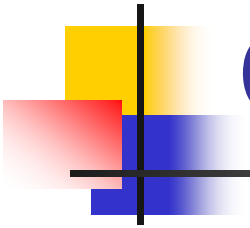
GBM





Summary (part I)

- Propagation is a tool for “extending” limited prior information to scoring the entire network.
- Integration helps: mutations and expression both inform the prediction



Outline

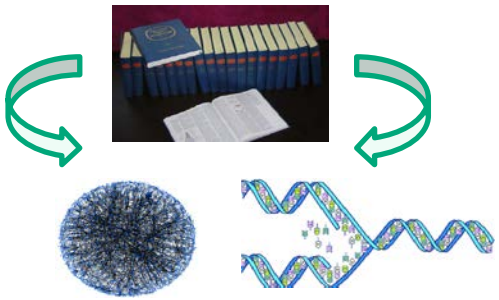
- Finding driver genes (Rufallo, Koyuturk, S.; PLoS Comp. Biol. ‘15)
- Finding disease modules (Mazza, Klockmeier, Wanker, S.; ISMB ’16)

Associating diseases with complexes

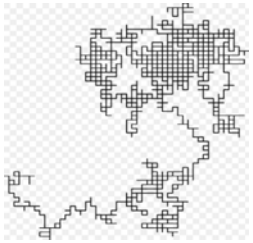
- Many studies link diseases to dysfunctions of protein assemblies working in concert.
  - Leigh syndrome – caused by disruption of mitochondrial complexes
- Previous methods:
  - PRINCE (Vanunu et al.'10)
  - HotNet2 (Leiserson et al.'15)

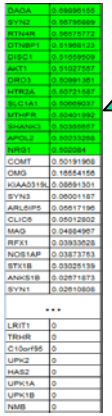
The general workflow

PPI network + disease causing proteins



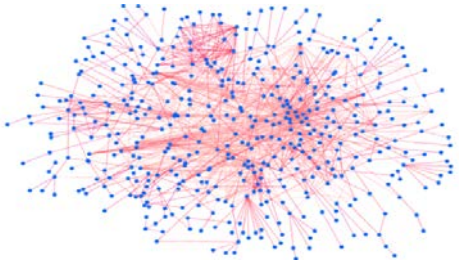
Network propagation and thresholding



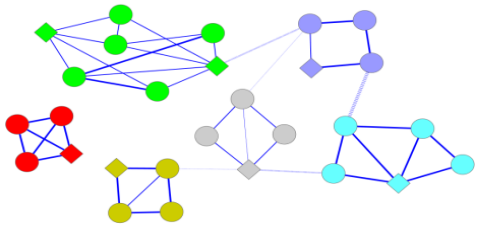


Statistical scoring

Input network for clustering

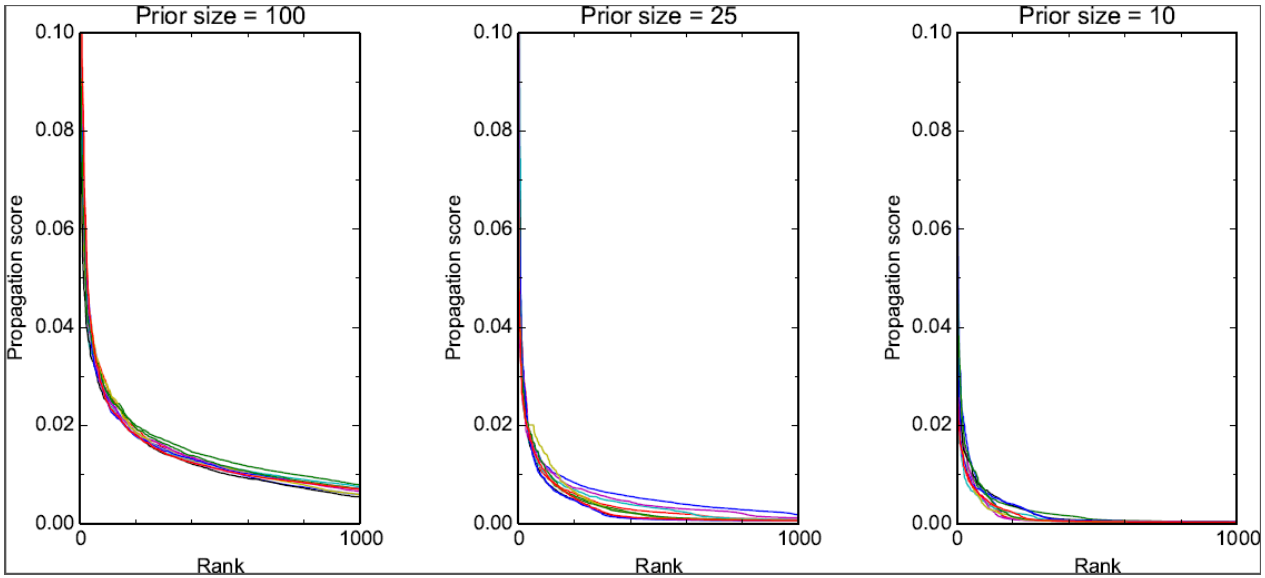


Detection of high scoring clusters



Integer program

Statistical scoring



- Propagation scores depend on prior size
- We normalize them by computing empirical p-values w.r.t. random priors of the same size

Finding dense clusters

- Clusters are scored via a likelihood ratio
- Protein complex model: edges occur indep. with high probability p.
- Random model: degree preserving. Probability of an edge depends on the degrees of its vertices

$$C = (V', E')$$

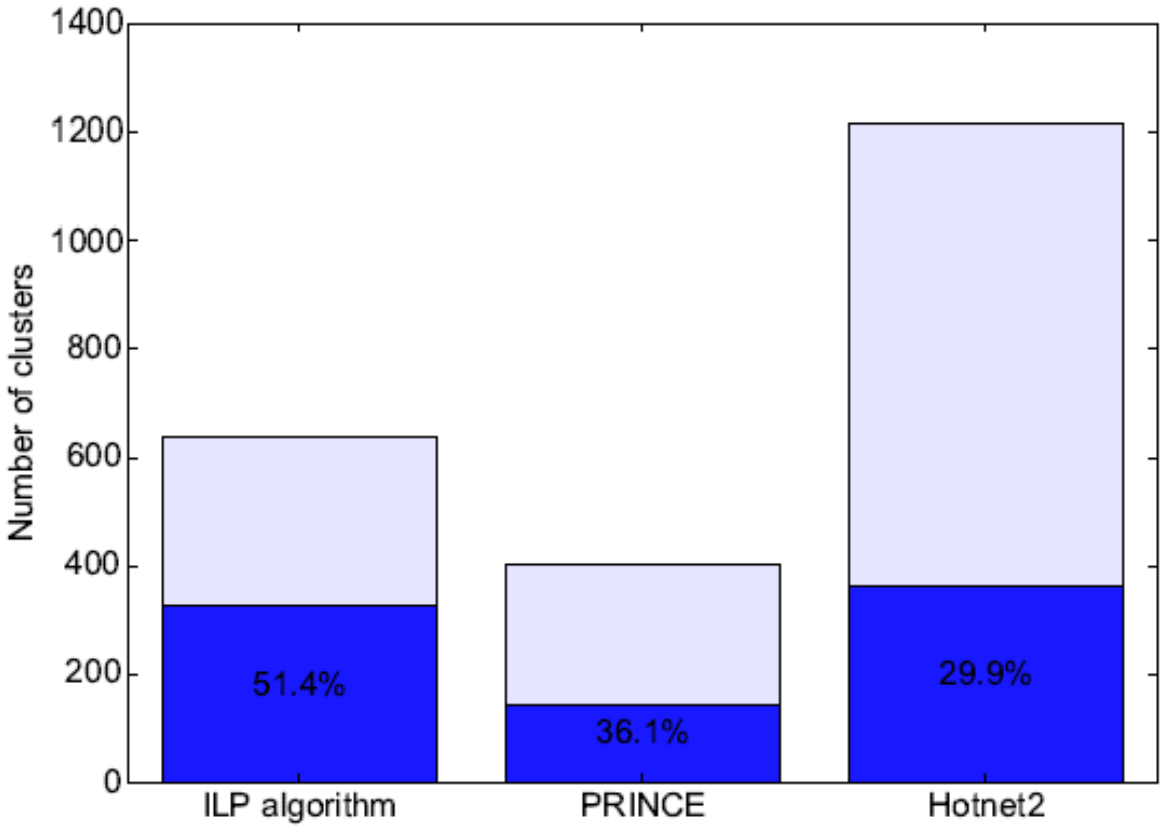
$$L(C) = \prod_{(u,v)\in E'} \frac{p}{p(u,v)} \prod_{(u,v)\notin E'} \frac{1-p}{1-p(u,v)}$$

- Linearity of scoring (under log) allows recasting the problem as an integer program

Data sets

- Disease associated genes were retrieved from 3 databases: OMIM, OrphaData and DISEASES (115 diseases, 8K associations)
- PPI data were taken from HIPPIE (150K interactions)

Performance evaluation: overlap with known complexes



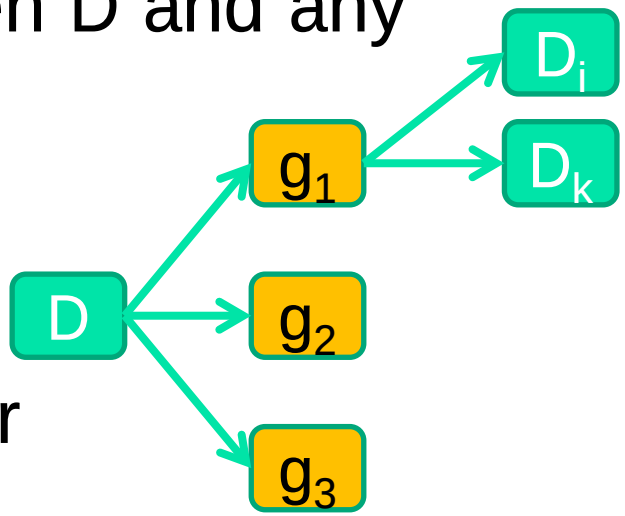
- % predicted clusters that significantly overlap one of 2276 GO/CORUM complexes

Performance evaluation: external validation

- Test enrichment of predictions per disease, using external validation sets from DISEASES:
  - Our ILP algorithm significantly captured 34 diseases (FDR corrected hypergoem. $p<0.05$)
  - PRINCE – 33
  - HotNet2 – 2 (of 23 diseases with significant modules)

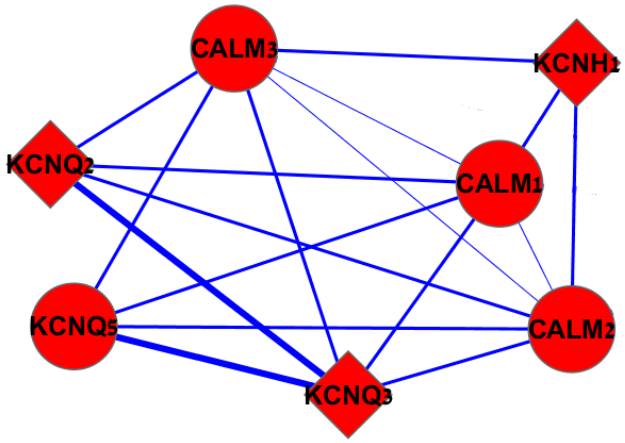
Performance evaluation: using information from similar diseases

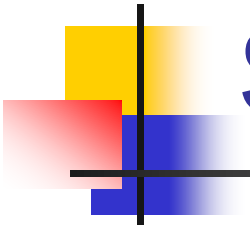
- Given disease D and prediction g_i, compute the max phenotypic similarity between D and any disease associated with g_i.
- Define score(D) = average over all g_i, the higher the better.
- Comparing score distributions, our algorithm's scores were significantly higher than HotNet2 *(*Wilcoxon rank sum $p < 3e\text{-}3$*)*



Case analysis – epilepsy syndrome

- 97 prior genes (diamonds).
- Top cluster yields two predictions: KCNQ5 and calmodulin proteins, both supported by the literature.
- Mice lacking functional KCNQ5 channels displayed increased excitability of neurons.
- Epilepsy-causing mutations led to alterations in calmodulin binding; calmodulin overexpression restored normal channel function.





Summary (part II)

- Propagation can be used to zoom in on disease regions in the network.
- The resulting module inference problem can be solved to optimality via ILP
- Global approaches that simultaneously consider all diseases can harness disease similarity measures to improve predictions (Silbeberg et al., submitted)

- Matthew Rufallo & Mehmet Koyuturk (Case Western)
- Konrad Klockmeier & Erich Wanker (MDC Berlin)
- Cytoscape plugin – *Propagate*