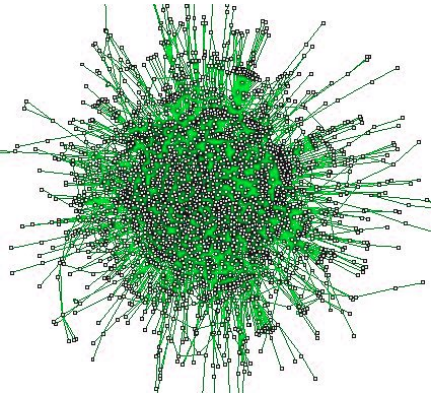
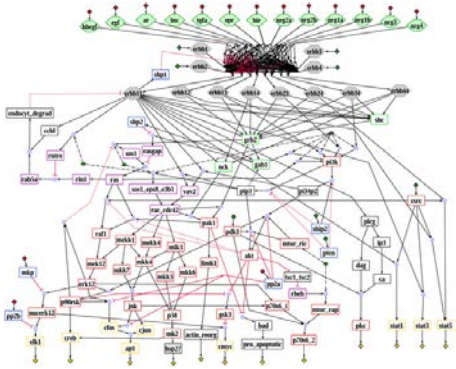
Protein networks: from topology to logic





Roded Sharan

School of Computer Science, Tel Aviv University &

International Computer Science Institute at Berkeley

Motivation

- Goal: an executable model of a process of interest
- Current experimental techniques yield only the global wiring of proteins
- What is missing:
  - Directionality information
  - Process specific subnetwork
  - The underlying logic

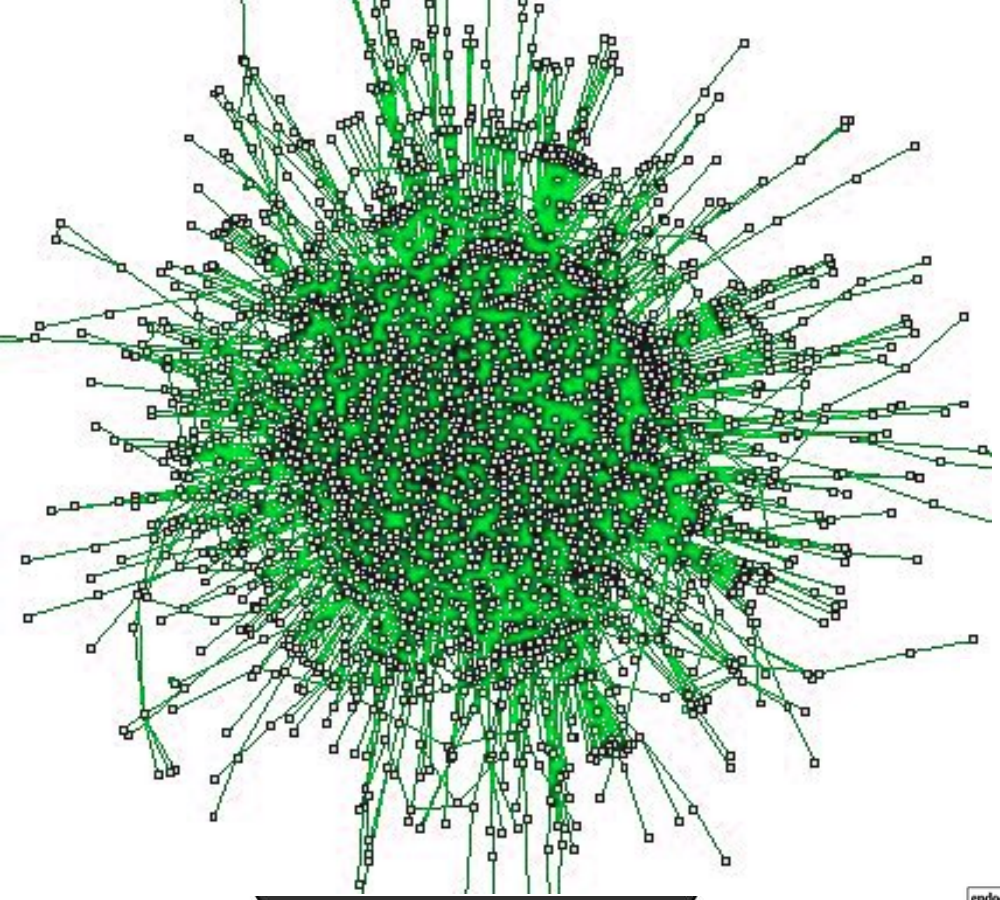
Sharan, EMBO Reports'13

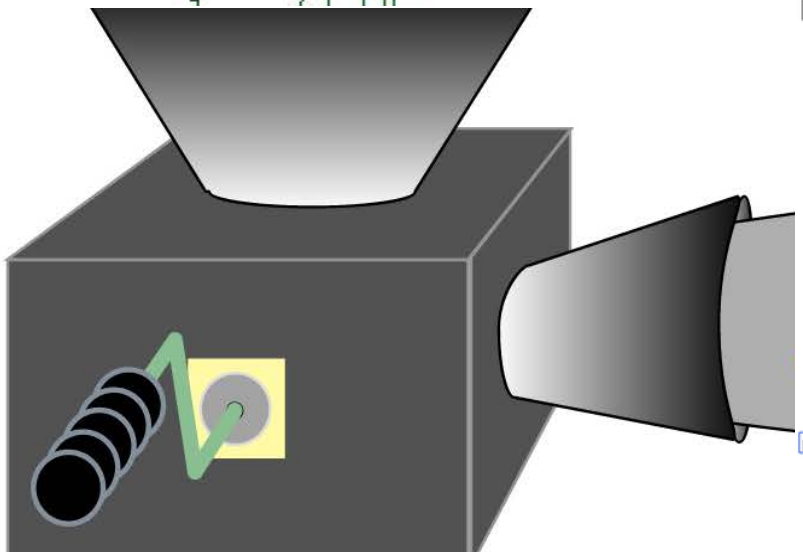
Our vision

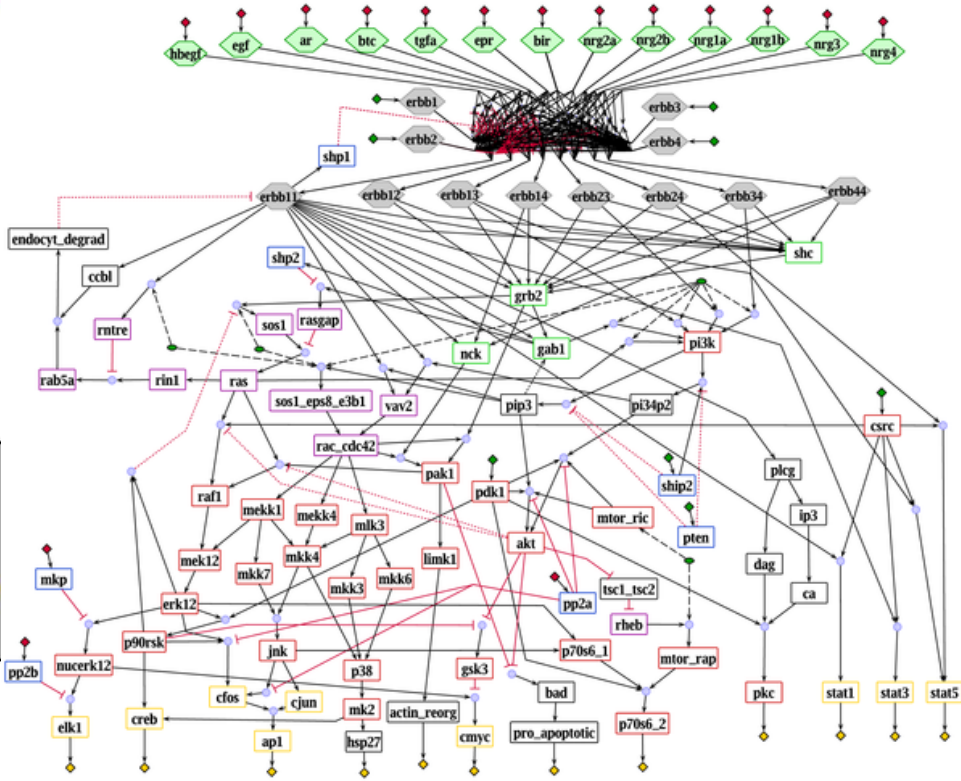
Network Orientation

Subnetwork inference

Logical model learning

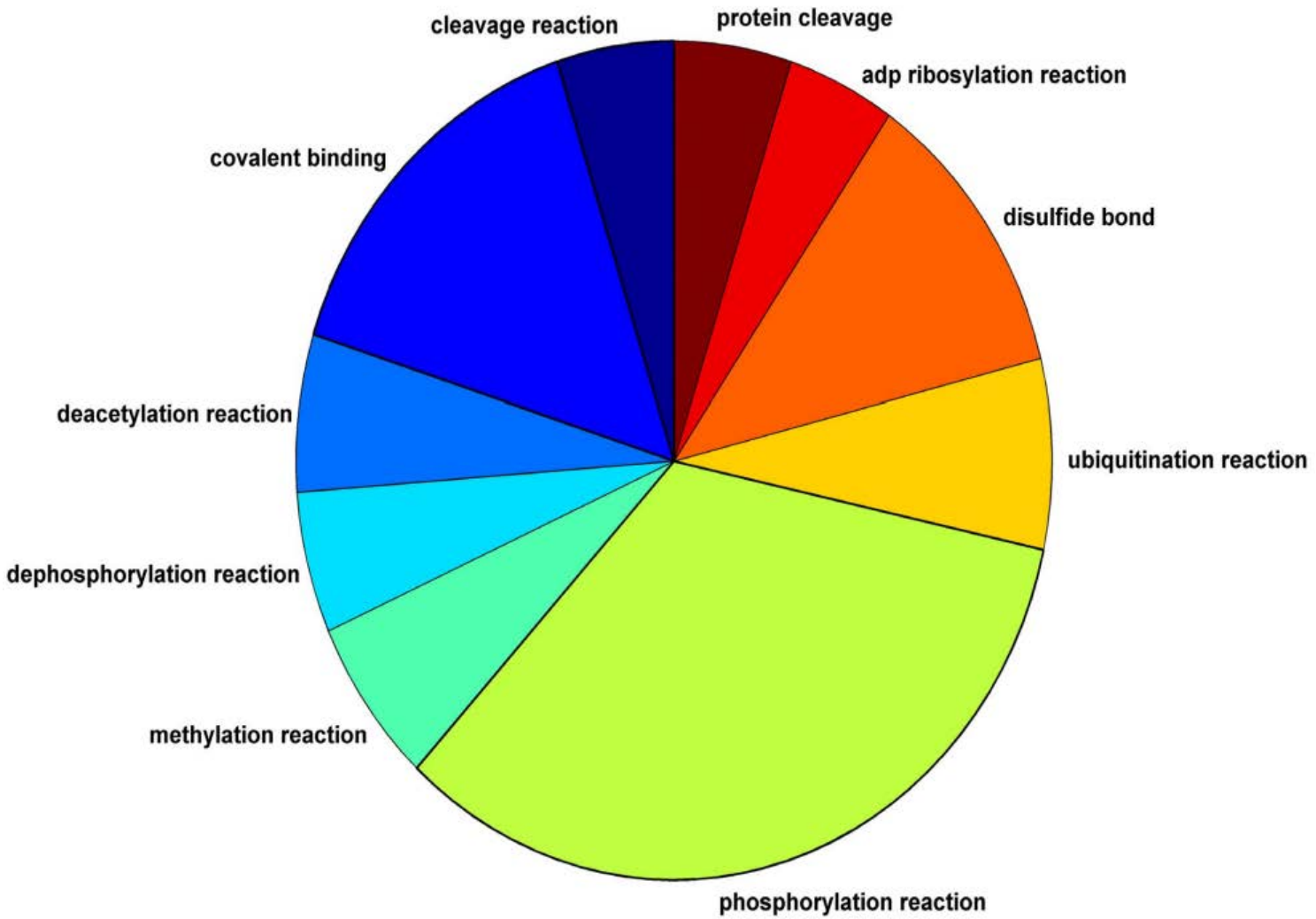






Network orientation

Are protein interactions directed?



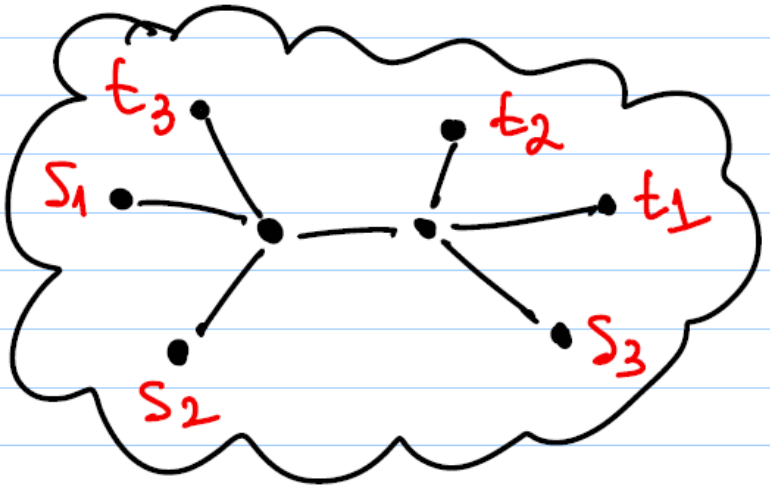
Silberberg et al., PLoS One'14

The computational problem

- Directionality is not revealed by the experiments
- Indirect information is obtained from knockout experiments:
  - Observe: knockout of protein *s* affects *t*
  - Assume: there is a directed *(s,t)* path
- Goal: predict directions to maximize #KO-pairs that can be "explained"

MAXIMUM GRAPH ORIENTATION

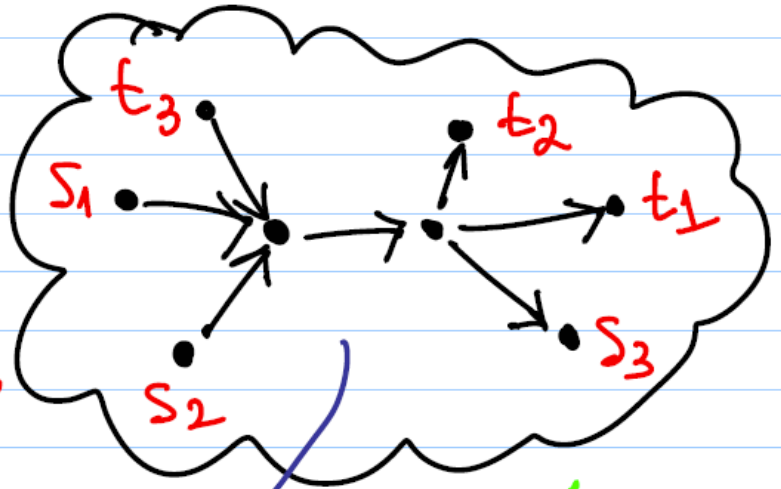
- Input: undirected graph $G=(V,E)$ with n vertices
  source-target pairs $(s_1,t_1),\dots,(s_k,t_k)$



MAXIMUM GRAPH ORIENTATION

- Input: undirected graph $G=(V,E)$ with n vertices
  source-target pairs $(s_1,t_1),\dots,(s_k,t_k)$

- goal: Compute an orientation in which the number of connected pairs is maximized



(s_1,t_1) ✓

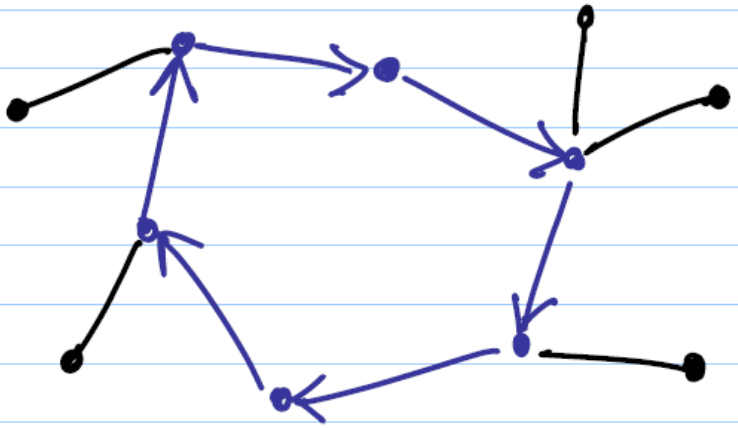
(s_2,t_2) ✓

(s_3,t_3) ✗

MAXIMUM GRAPH ORIENTATION

- Input: undirected graph $G=(V,E)$ with n vertices
  source-target pairs $(s_1,t_1),\dots,(s_k,t_k)$

- goal: compute an orientation in which the number of connected pairs is maximized



- remark: we may assume that the underlying graph is a tree

Complexity of Max. Tree Orientation

- NP-hard (reduction from MAX DI-CUT)
- Hard to approximate to within 12/13
- $\Omega(\log\log n / \log n)$ approximation
- Can we do better?

Medvedovsky et al., WABI 2008
Gamzu et al., WABI 2010
Elberfeld et al., Internet Math. 2011

An Integer Programming Formulation

- Assign a single direction for each edge

  $O(v,w) + O(w,v) = 1$

- Describe reachability relations

  $c(s,t) \leq O(x,y)$ for all edges in the path from s to t

- Objective: $\max \sum c(s,t)$

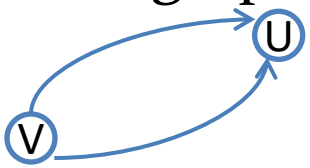
Medvedovsky et al., WABI 2008

A biological complication

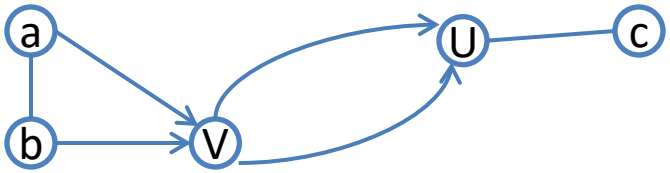
- In reality, some of the edges are pre-directed, e.g. kinase-substrate interactions.
- Can we deal with mixed graphs?
- On the theoretical side, large gap between upper (7/8) and lower ($\tilde{\Omega}(1/n^{1/\sqrt{2}})$) approximation bounds.

Mixed vs. undirected

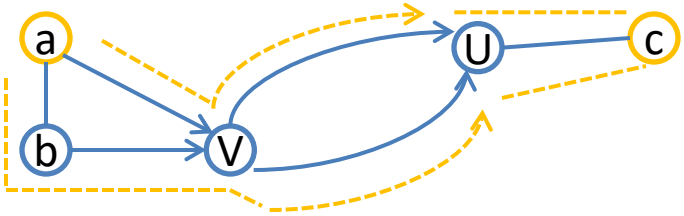
In the mixed graph there are cycles which cannot be contracted



The graph cannot be reduced to a tree

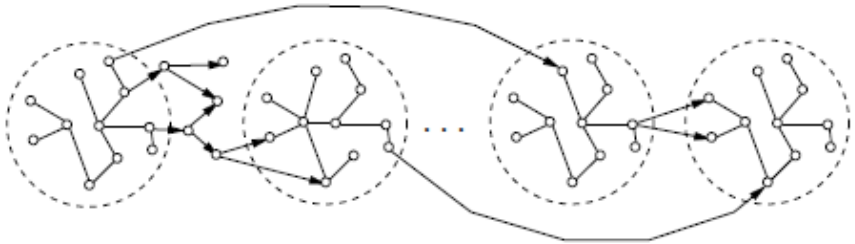


There may be multiple paths between a pair of vertices



An ILP for mixed graphs

- Contract all cycles, obtaining an acyclic graph
- Use topological sorting to create a graph of trees connected by left-to-right directed edges:



- Work recursively on pairs crossing from $G_i = T_1 \cup ... \cup T_i$ to T_{i+1}
- A path between trees decomposes to subpaths within trees and a single directed edge between the trees.

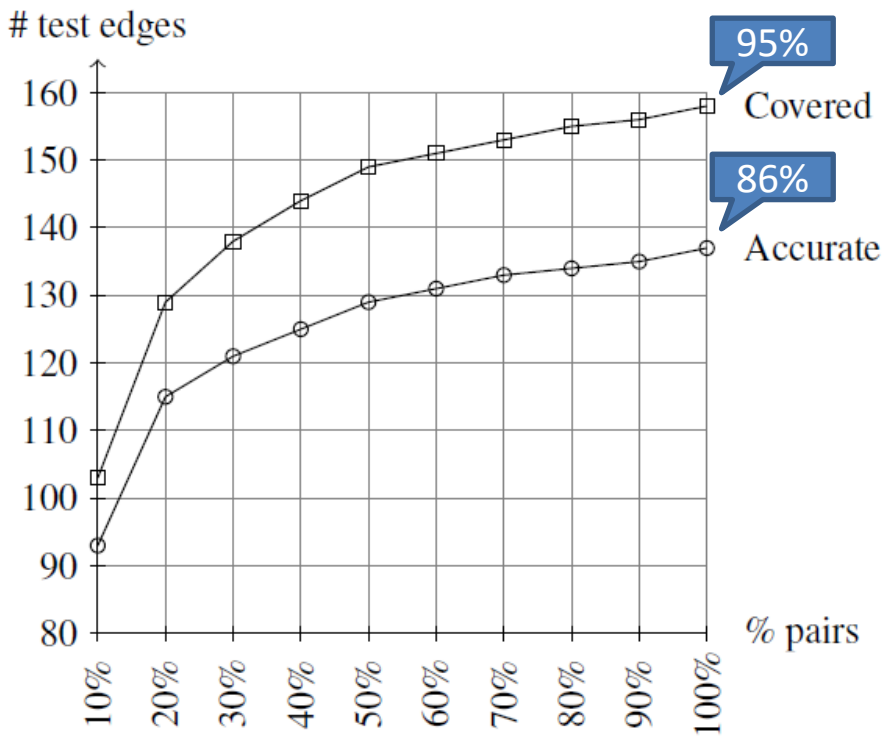
Silverbush et al., JCB 2011

A taste of the results

- Applied to yeast data: ~50K pairs, ~8,000 interactions (mixed) and 1361 test edges (KPIs) whose directions are hidden from the algorithm.

- After cycle contraction:
  - ~2,000 edges
  - 166 test edges

- Coverage: % oriented (with confidence)
- Accuracy: % correct (confident) orientations

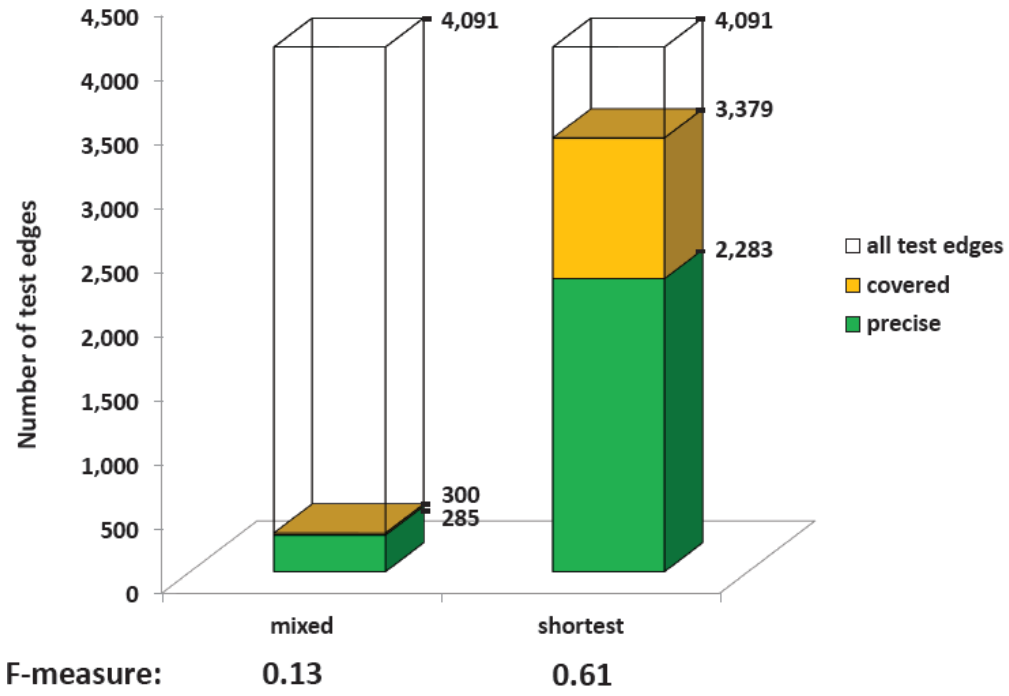


Increasing coverage

- Most edges are eliminated by the cycle contraction phase, hence their directions remain ambiguous.
- One “biologically-meaningful” attack is to require the connecting path to be SHORTEST
- Can be efficiently tackled via ILP by:
  - For any given pair (s,t) build a graph of all shortest paths
  - Perform flow computations in this graph to determine if the pair is connected under a given orientation.

Silverbush et al., Bioinformatics’14

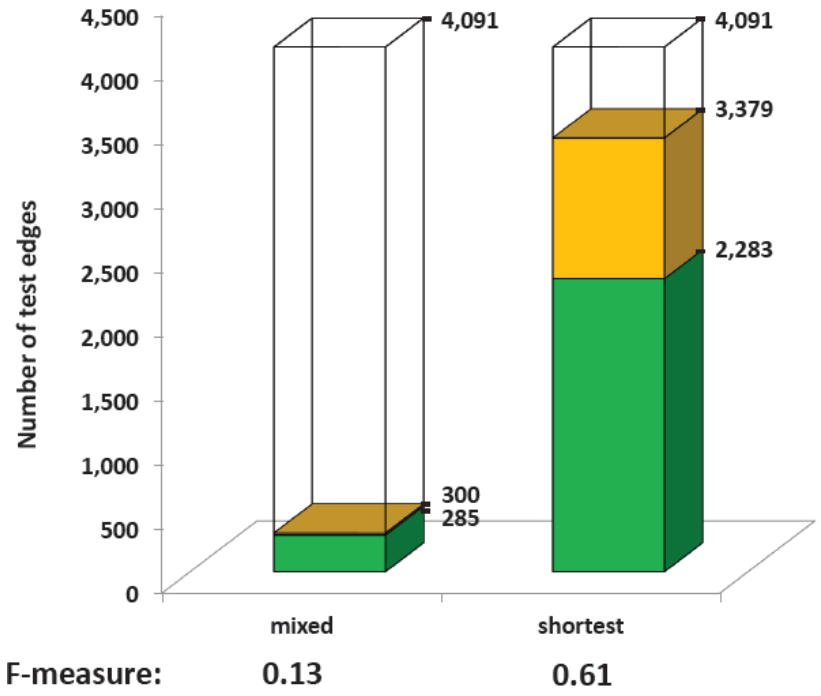
The SHORTEST approach (application)



- Yeast: similar accuracy, 8-fold more coverage!

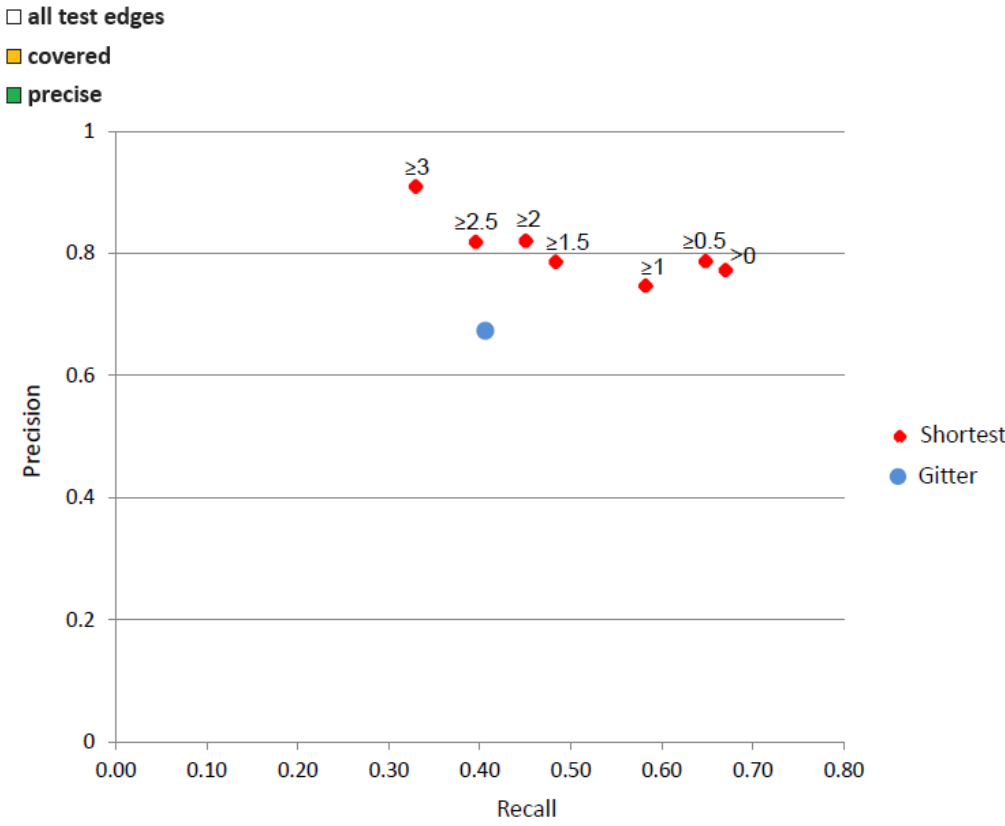
Silverbush et al., Bioinformatics'14

The SHORTEST approach (application)



- Yeast: similar accuracy, 8-fold more coverage!

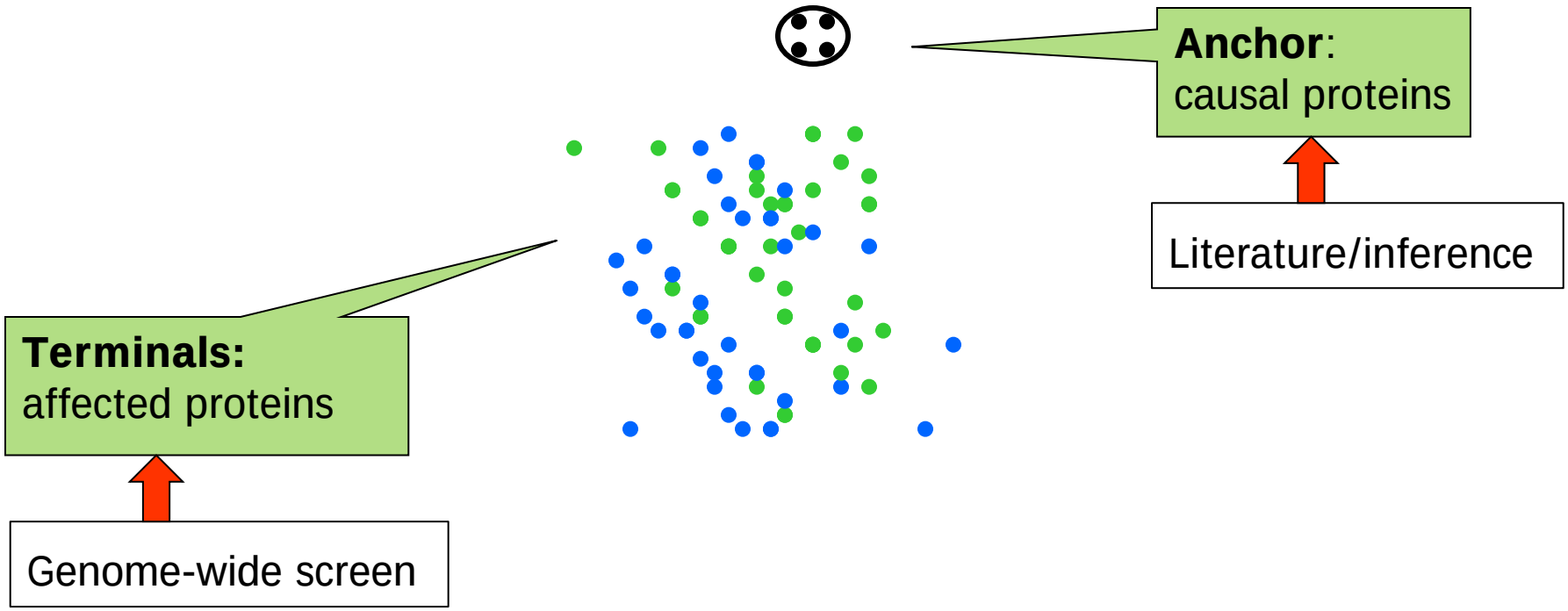
- Human: outperforms a previous method by Gitter et al.



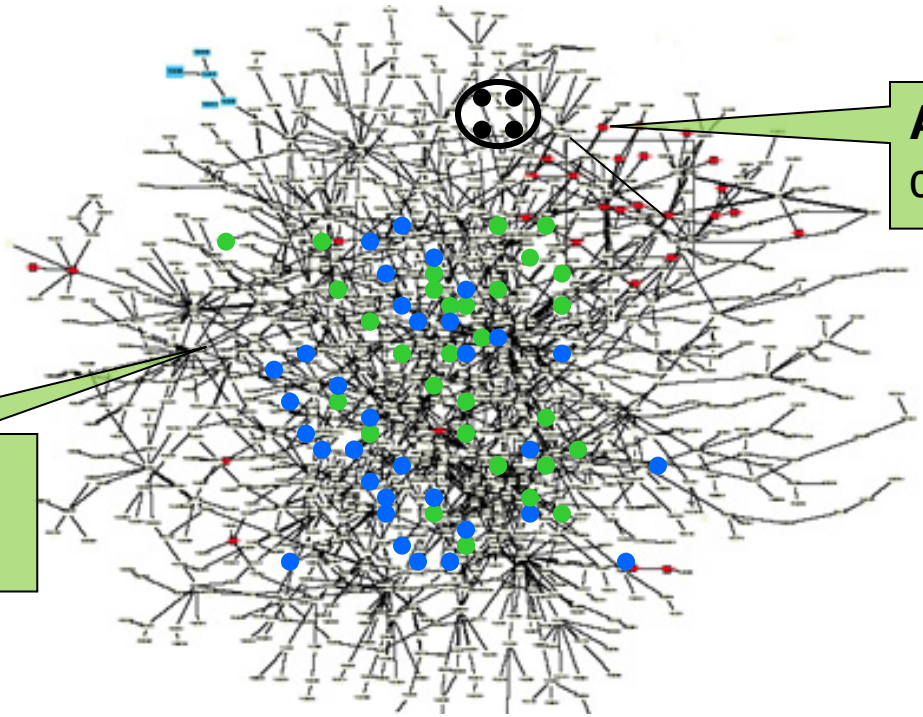
Silverbush et al., Bioinformatics'14

Subnetwork inference

Identifying process-specific proteins



From components to a map



Anchor:
causal proteins

Terminals:
affected proteins

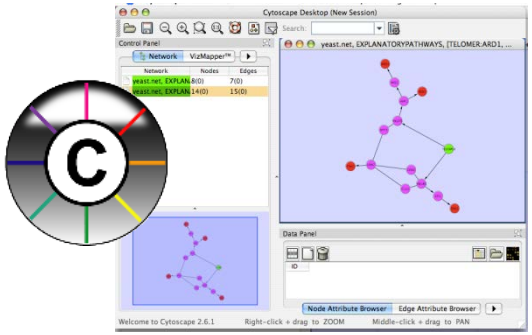
Goal: Infer the underlying subnetwork

Shachar et al., MSB 2008
Yosef et al., MSB 2009
Atias et al., MBS 2013

From components to a map (cont.)

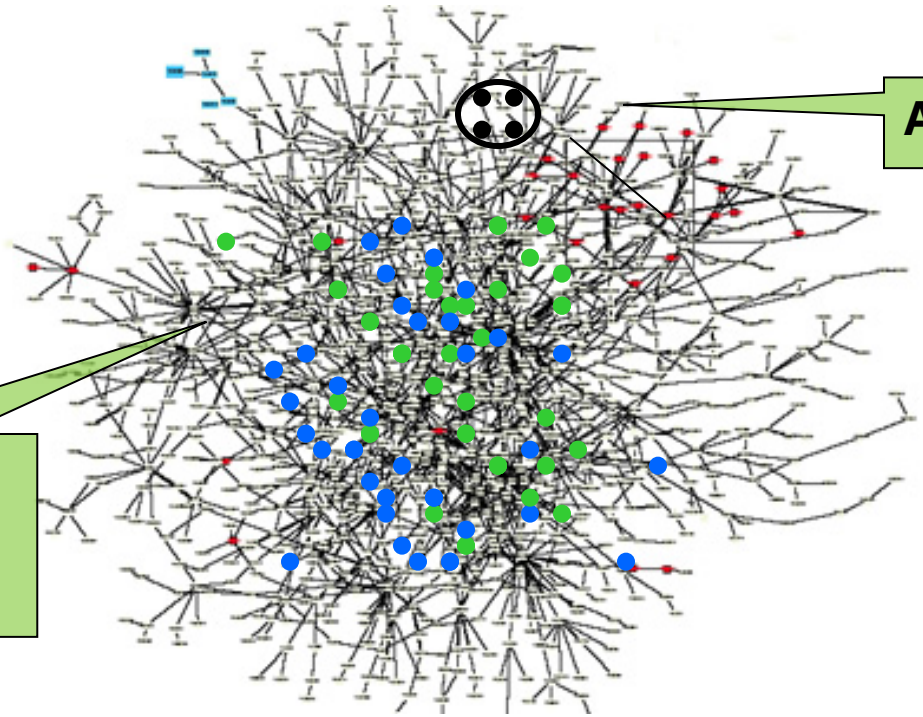
- Unique approach to simultaneously optimize subnetwork size and length of anchor-terminal paths.
- Shown to outperform existing tools on yeast and human data
- Implemented as a cytoscape plugin called **ANAT**

(www.cs.tau.ac.il/~bnet/ANAT)



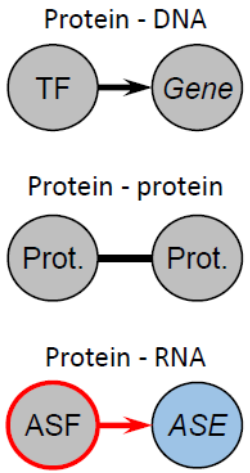
Yosef et al., Science Signaling'11
Atias et al., MBS'13

Application to alternative splicing events in cancer



Anchor: TF

Terminals:
Differentially spliced events

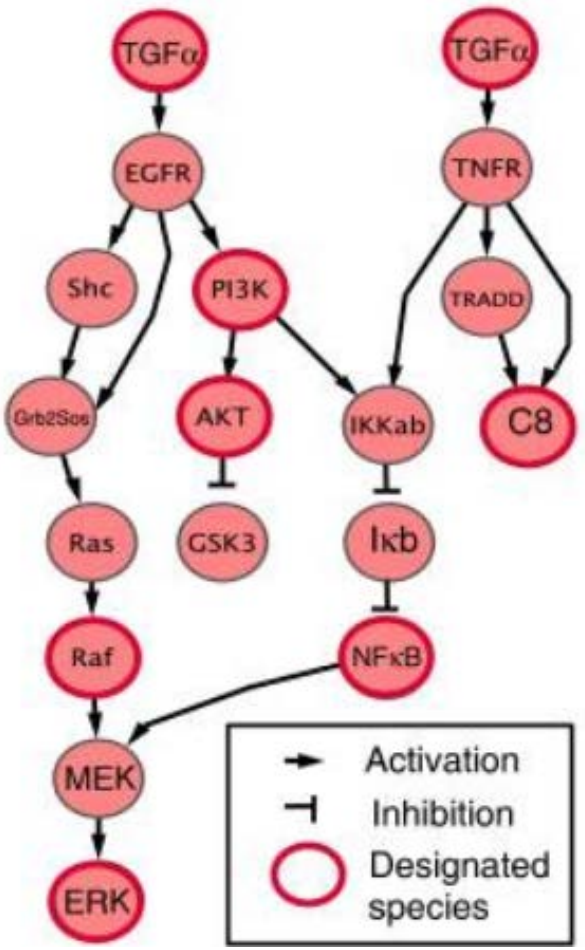


Dror Hollander, Gil Ast

Logical model learning

The Boolean model

- Each node=protein/ligand can be active (1) or inactive (0).
- The activity of a node is a *Boolean function* of the activities of its predecessors in the network.



The computational problem

Input: (i) Directed network

(ii) Protein activity readouts following different perturbations

Goal: learn the Boolean functions so as to minimize disagreements with experimental data

Stimuli							
TGFα	+	−	+	+	+	+	Design
TNF	−	+	+	−	+	−	
Inhibitors							
PI3K	−	−	−	+	+	−	
Raf	−	−	−	−	−	+	
Readouts							
NFκB	0	0	1	0	0	0	Measured
ERK	1	0	1	1	1	0	
C8	0	1	1	0	1	0	
AKT	1	0	1	0	0	1	

NFκB	0	0	0	0	0	0
ERK	1	0	1	1	1	0
C8	0	1	1	0	1	0
AKT	1	0	1	0	0	1

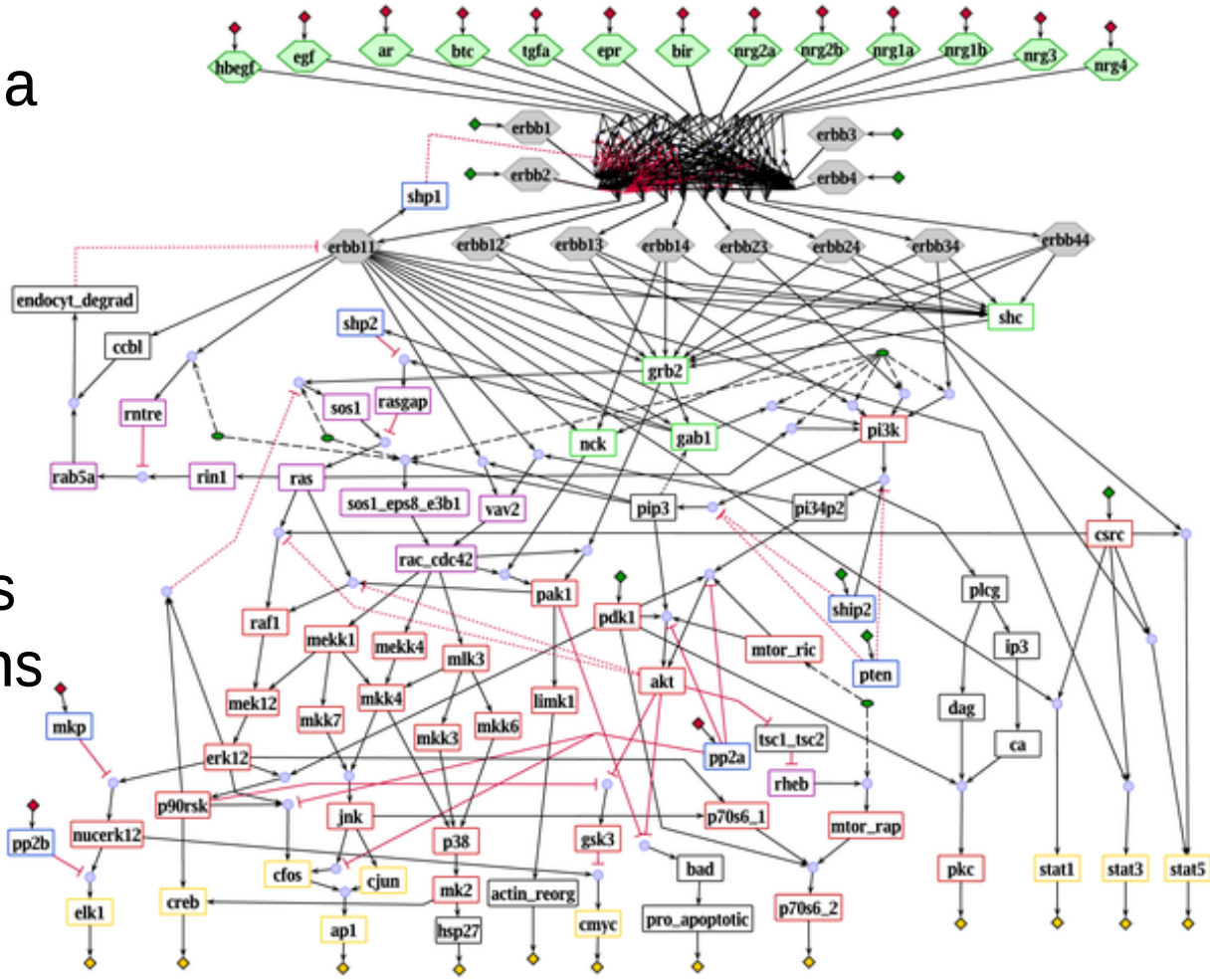
Algorithmic results

- *ILP* formulation, solved to *optimality*
- *Activation/repression effects* are automatically learned as part of the logic
- Particularly efficient solution for *threshold* functions (generalize AND & OR)

Sharan & Karp, JCB 2013

Application to EGFR signaling

- Detailed model by Oda et al. and Samaga et al. contains:
  - 112 nodes
  - 157 non-I/O reactions
- Readouts: 11 proteins under 34 perturbations
- **76%** fit to data



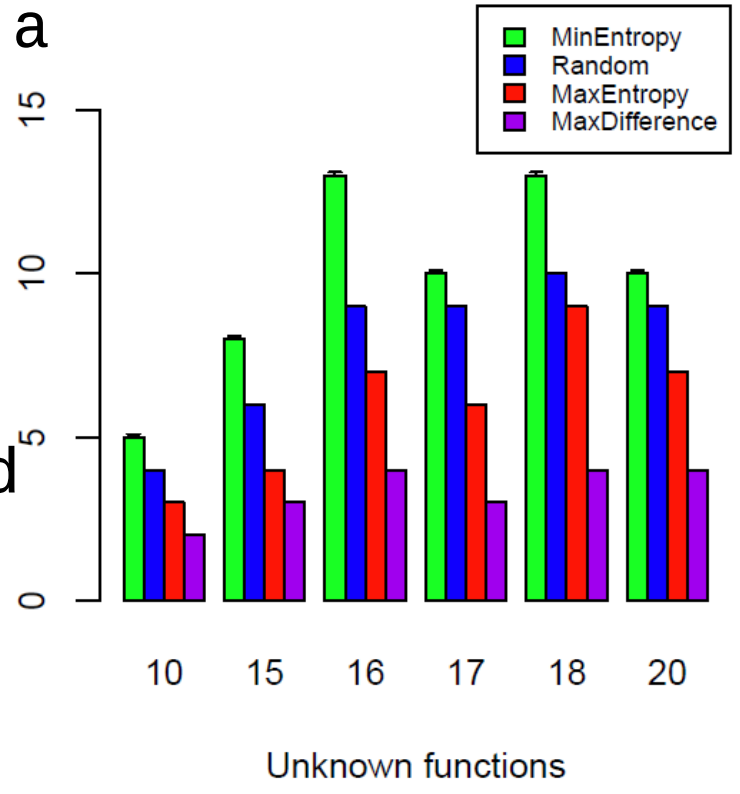
Improving the fit

- Focus on 16 uncertain gates (2^33 possible models), for 4 of which modifications were manually proposed
- 11 of 12 reconstructed functions matched the curated description
- 3 of 4 proposed changes were predicted correctly, the fourth rejected.
- The learned model achieved the same **90%** fit as the manual model!

Original function	Proposed modification	Reconstructed function
erb11 AND (pip3 OR pi34p2) → vav2	erb11 → vav2	**erb11 → vav2**
sos1eps8e3b1 → raccdc42	REMOVE	sos1eps8e3b1 → raccdc42
erb11 AND csrc → stat3	REMOVE	**REMOVE**
mk2 → hsp27	REMOVE	**REMOVE**

Challenges ahead

- Integrate the three phases (orientation, inference, logic) into a coherent pipeline
- Deal with multiple solutions:
  - Confidence computation
  - Experimental design
  - Rank via biologically-motivated secondary criteria
- Advance from static (acyclic) to dynamic models



Atias et al., Bioinformatics'14 (ECCB)

Acknowledgments

Orientation
Dana Silverbush
Michael Elberfeld
Danny Segev…

Logic
Richard Karp
Nir Atias…

Inference
Nir Yosef
Nir Atias
Assaf Gottlieb
Gil Ast
Dror Hollander
Martin Kupiec
Eytan Ruppin…