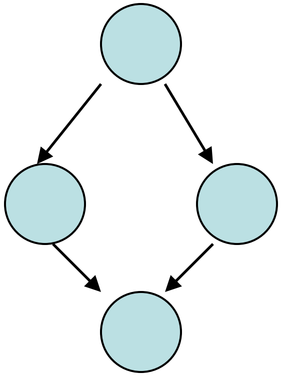
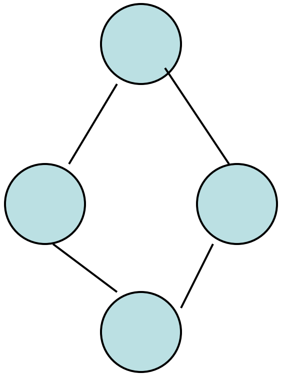
02-710
Computational Genomics

Physical networks & active learning

Graphical models

- Efficient way to represent and reason about *joint distributions*
- Graphs in which nodes represent random variables and edges correspond to dependency relationships
- Two major types: Directed and undirected





$$\prod_i p[x_i \mid Pa(x_i)]$$

$$\prod_{i,j} \psi_{i,j}(x_i, x_j)$$

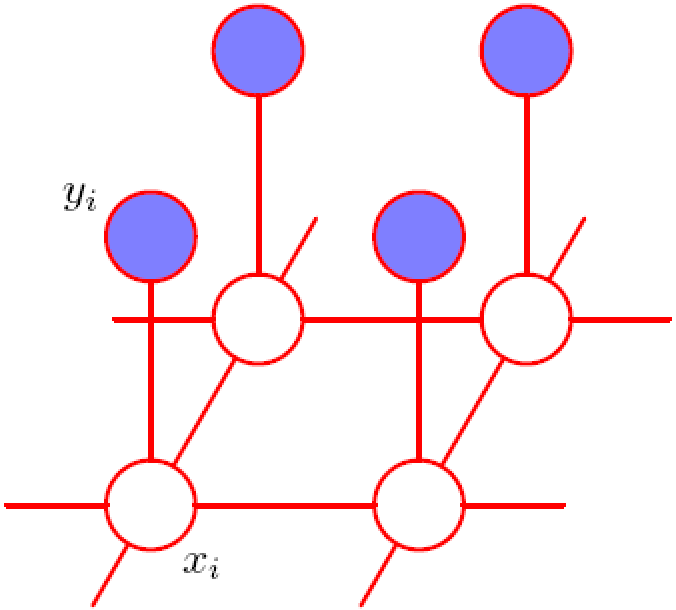
- Bayesian networks
- Hidden Markov models

- Markov random fields

Undirected – Markov Random Fields

- Popular in statistical physics, computer vision, sensor networks, social networks, protein-protein interaction networks etc.
- Example – Image Denoising x_i – value at pixel i

  y_i – observed noisy value



Factorization

- Joint distribution factorizes according to the graph

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

$\mathcal{C}$ is the set of maximal cliques in the graph

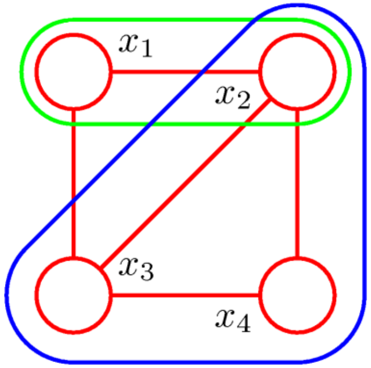
$\psi_C(x_C)$ is a potential function on the clique x_C

→ Arbitrary positive function

normalization factor

$$Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

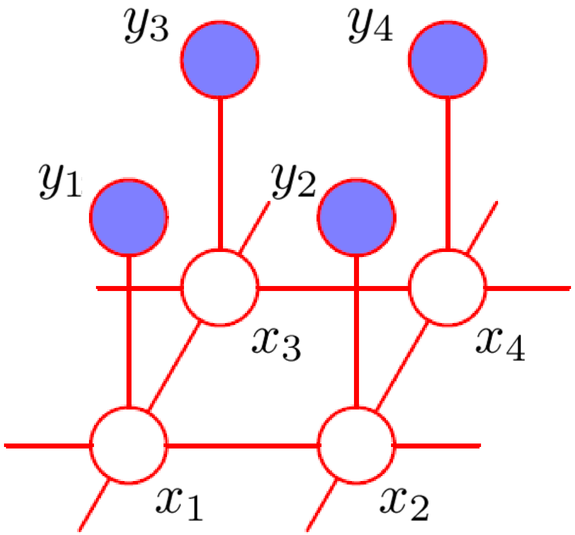
typically NP-hard to compute



Clique, $x_C = \{x_1, x_2\}$

Maximal clique
$x_C = \{x_2, x_3, x_4\}$

MRF Example

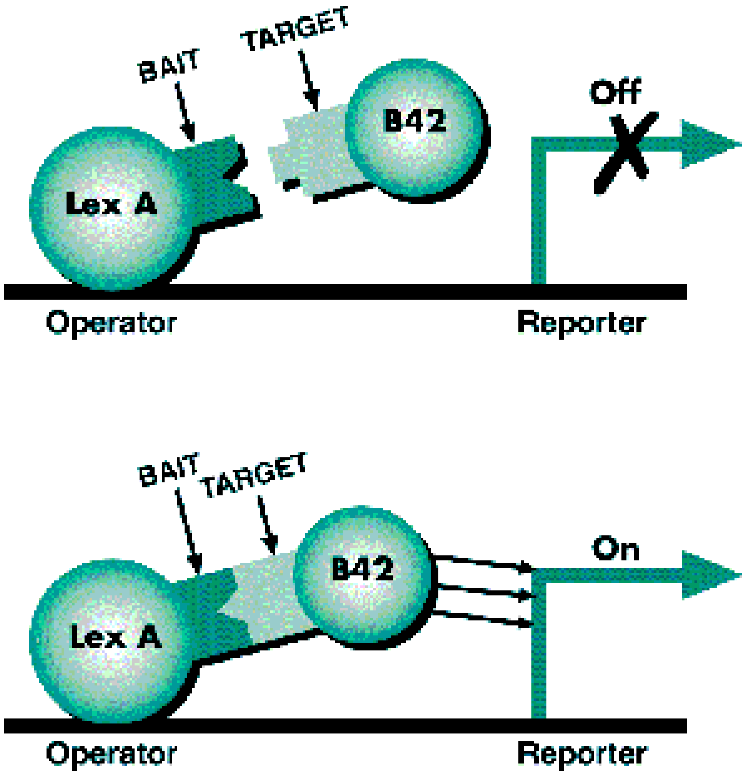


$$P(x,y) \propto \Psi(x_1,x_2)\Psi(x_1,x_3)\Psi(x_2,x_4)\Psi(x_3,x_4)\prod_{i=1}^{4}\Psi(x_i,y_i)$$

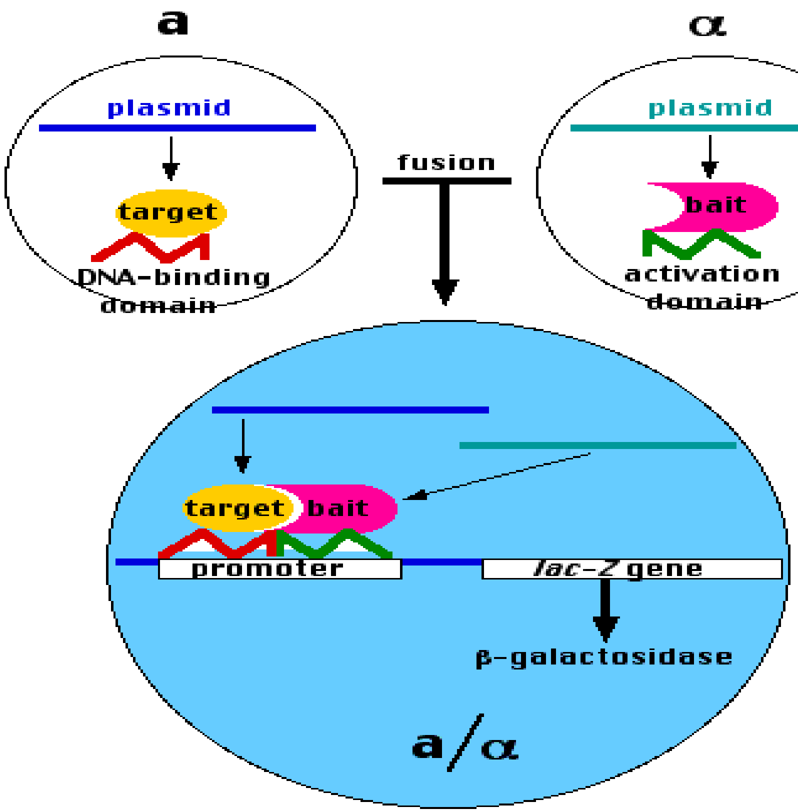
Protein interaction datasets

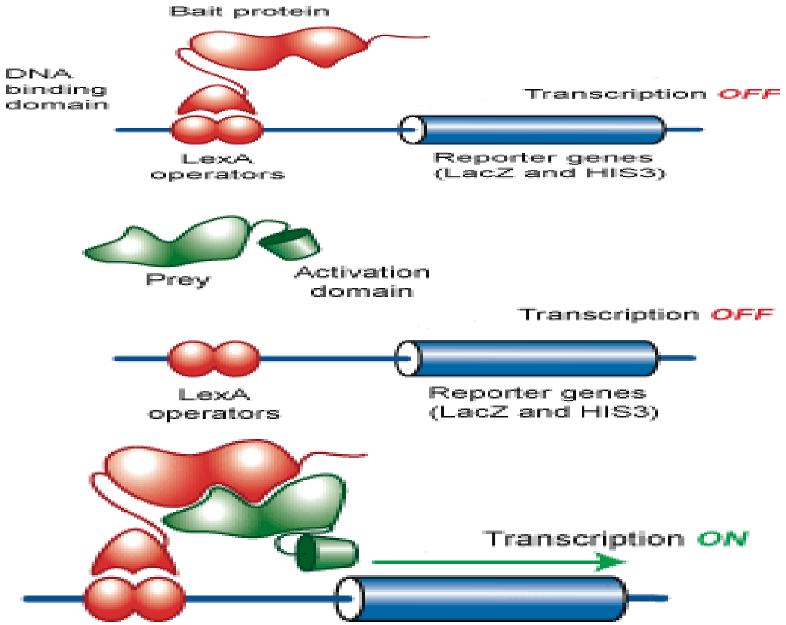
Yeast two-hybrid assay

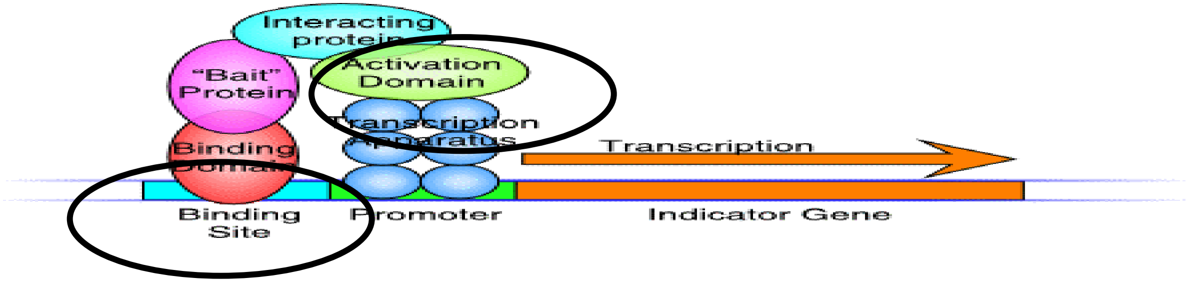
- Pairs of proteins to be tested for interaction are expressed as fusion proteins ('hybrids') in yeast:
- One protein is fused to a DNA-binding domain, the other to a transcriptional activator domain.
- Any interaction between them is detected by the formation of a functional transcription factor.



Yeast 2 Hybrid Technique







Mass spectrometry of purified complexes

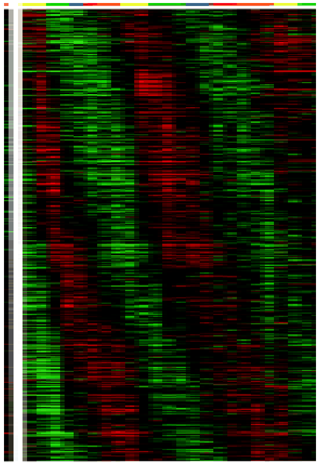
- Individual proteins are tagged and used as 'hooks' to biochemically purify whole protein complexes. These are then separated and their components identified by mass spectrometry.
- Can also be used to identify virus-host interactions

Interaction databases

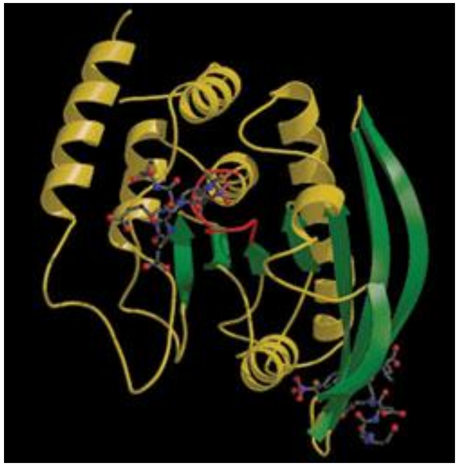
- STRING: string-db.org/
- BioGrid: thebiogrid.org/
- HPRD: www.hprd.org/
- KEGG: www.genome.jp/kegg/

Data integration

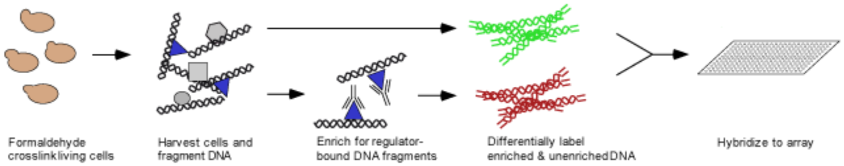
Gene expression



Protein interactions

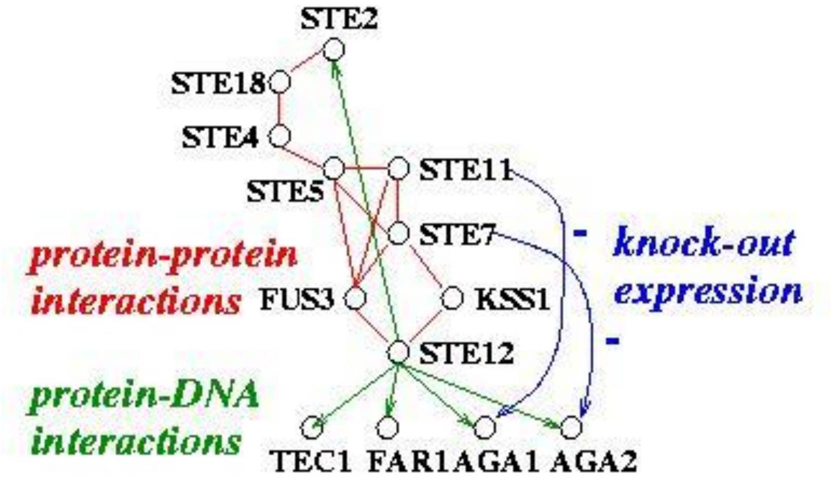


Protein-DNA binding


Formaldehyde crosslink living cells
Harvest cells and fragment DNA
Enrich for regulator-bound DNA fragments
Differentially label enriched & unenriched DNA
Hybridize to array

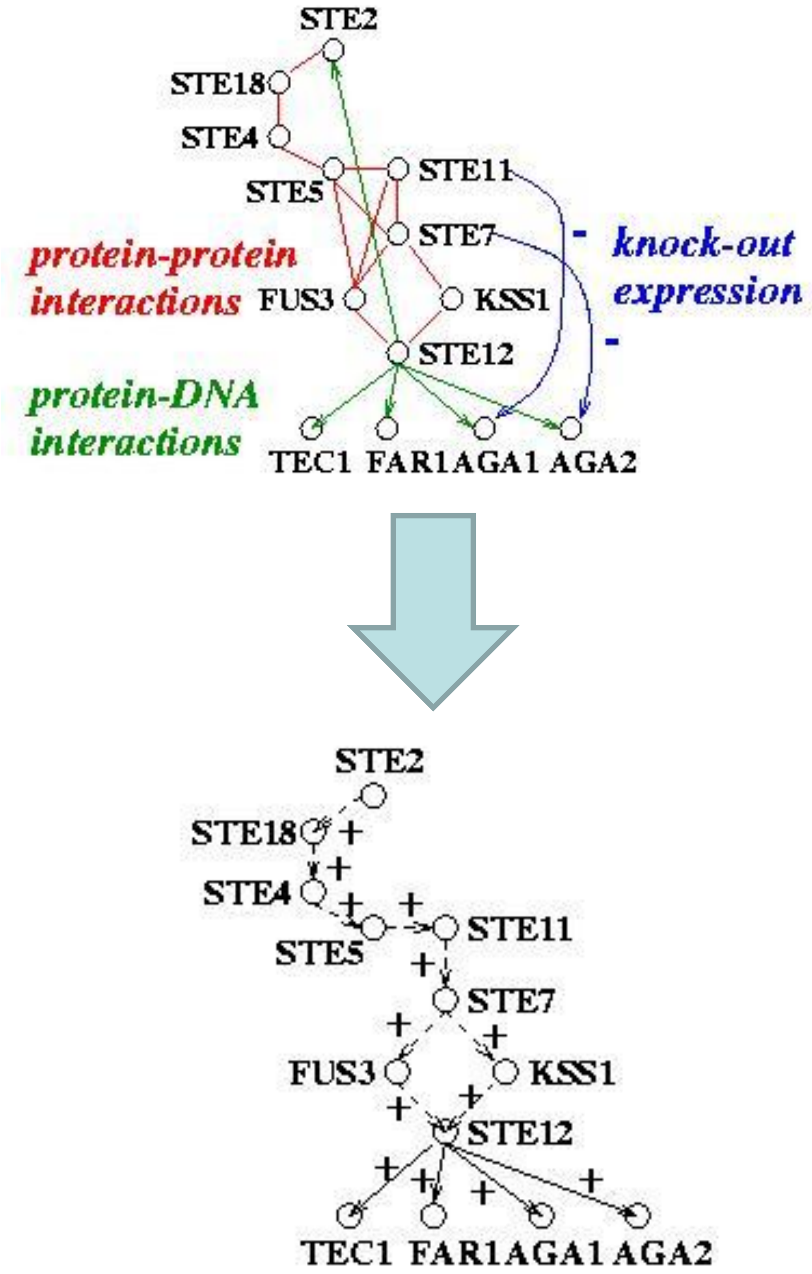
Yeast mating pathway

- Physical data:
  - Yeast binding data
  - DIP database (PPI)
- Functional data:
  - Rosetta compendium knockout data



A mechanistic model of gene regulation

- A graph depicting physical interactions and functional annotations.
- Nodes: Proteins
- Edges: PPI or Protein-DNA
- Signs on the edges: Activation or repression

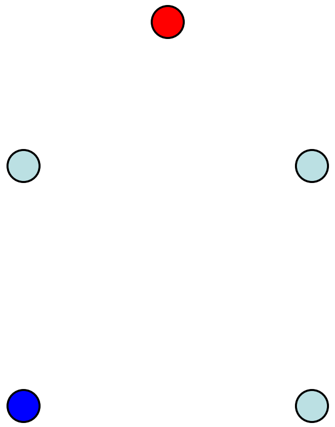


Inferring the mechanistic model from observed data

Key question: How do we construct the model from known mechanisms and constraints from observed data?

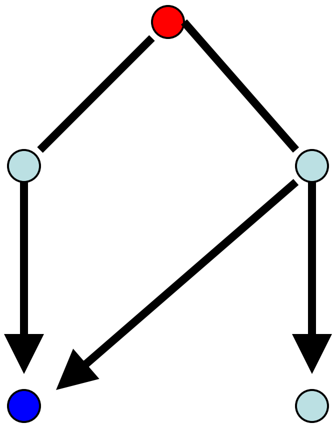
- Decompose data into pairwise items.
- Construct potential functions specifying constraints of each item.
- Combine potential functions by multiplication.

Requirements to explain knock-out data



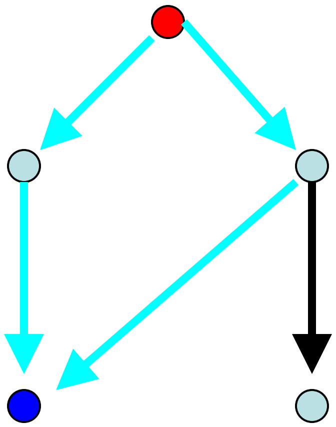
Requirements to explain knock-out data

- There is at least one connecting path.



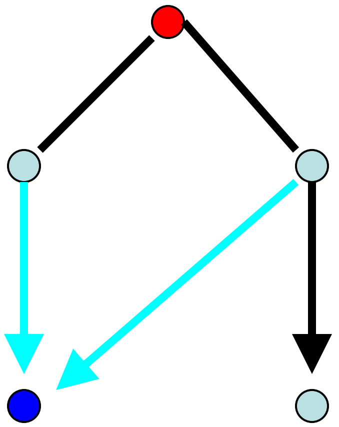
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.



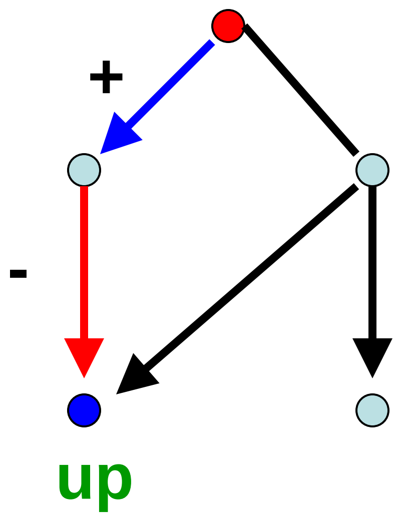
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.



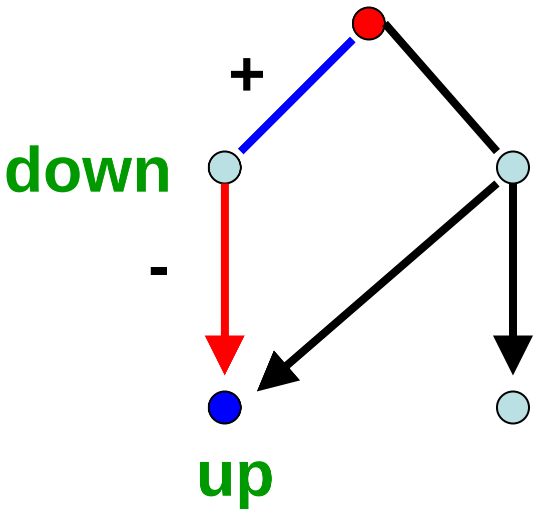
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.
- The aggregate sign along the path is consistent with the knock-out effect.



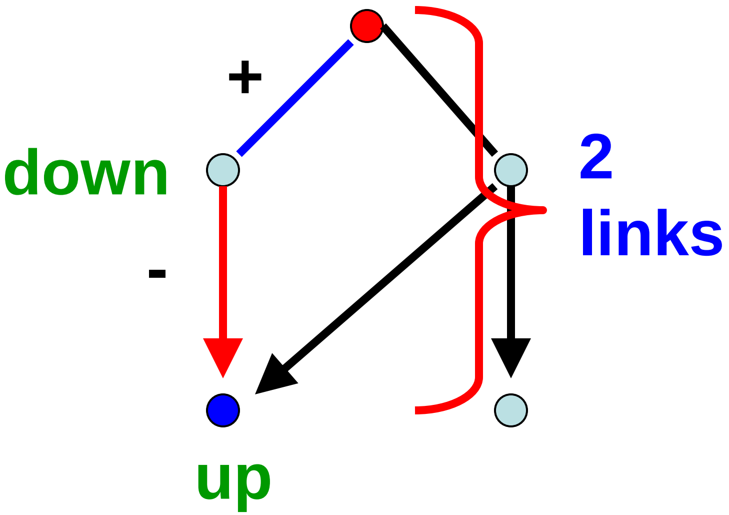
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.
- The aggregate sign along the path is consistent with the knock-out effect.
- Intermediate genes along the path either have knock-out effects on or were not tested.



Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.
- The aggregate sign along the path is consistent with the knock-out effect.
- Intermediate genes along the path either have knock-out effects or were not tested.
- The path length is upper bounded.

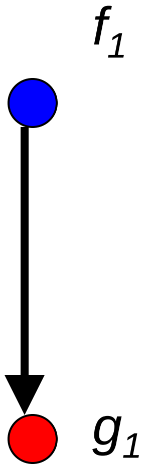


Factor graph formalism

- Factor graph is an undirected bipartite graph where edges represent dependency
- The joint likelihood is written using a set of potential functions, one for each edge in the graph and others for paths in the graph
- The key challenge is to determine the set of potential functions and how to encode them

Associations with binding data

- Assume we have p-value *y* for the event *x* (binding of f_1 to g_1).
- How can we use this value in a probabilistic setting?
- Possible solution: use likelihood ratio:



$$\frac{p(y \mid x)}{p(y \mid \sim x)}$$

x – the event of f binding to g

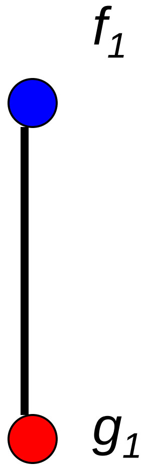
y – observed p-value

Associations with binding data

- Given a possible protein-DNA interaction e_i, the potential function $\phi_{ei}(x_{ei};y_{ei})$ is related to the direct evidence about this interaction:

$$\phi_{ei}(x_{ei};y_{ei}) = [\frac{p(y_{ei} \mid x_{ei} = 1)}{p(y_{ei} \mid x_{ei} = 0)}]^{x_{ei}}$$

- And similarly for protein interaction.



Determining the confidence in the observed data

- In order to determine the probabilistic term in the potential function we use an appropriate error model.
- As a crude approximation, $p(y_{ei} \mid x_{ei})$ can be obtained from the binding p-value
- First, set p(measurement | interaction does not exist) = p-value
- The other side p(measurement | interaction exists) is set to a fixed value.
- The potential term for the protein interaction case is defined analogously.

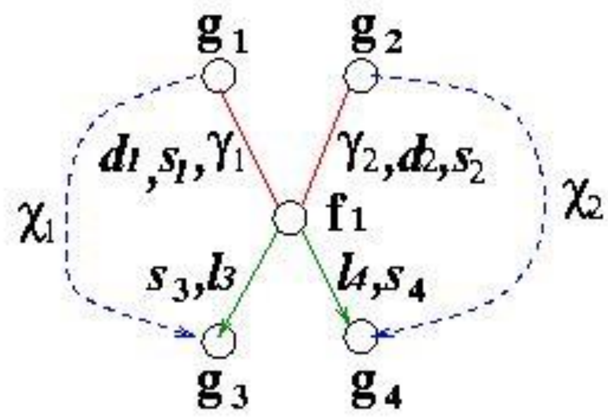
$$\phi_{ei}(x_{ei};y_{ei}) = [\frac{p(y_{ei} \mid x_{ei}=1)}{p(y_{ei} \mid x_{ei}=0)}]^{x_{ei}}$$

Associations with knock-out expression data

- Given knockout expression data, we need to determine whether or not the knockout of gene *i* influenced gene *j*
- The interaction effect is associated with the observed data o by:

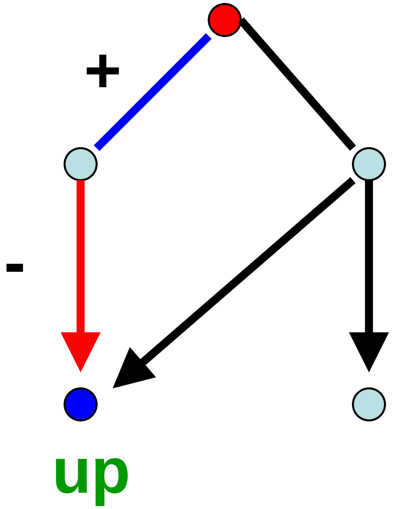
$$\phi_{i,j}(k_{i,j}, o_{k,i,j}) = [\frac{p(o_{k,i,j} \mid k_{i,j} = 1)}{p(o_{k,i,j} \mid k_{i,j} = 0)}]$$

- *k* can be explained by cascades of molecular interactions, i.e., paths in the physical model.



Knockout (cont.)

- Explanation of a KO can be expressed as a logic clause of variables along the paths connecting a knock-out pair:
  - the knock-out effect (χ_k)
  - edge presences (E_k),
  - edge directions (D_k), and sign (S_k),
  - and path selections (Σ_k).



- The potential term can also incorporate the situations of multiple paths and uncertainties of explanation.

Inference

- Potential functions are combined by multiplication.
- Goal: find the optimal configuration of the variables.
- This is done using a maximum likelihood approach using a variant of belief propagation.
- Using a graph known as a factor graph, the max-product algorithm is applied to obtain a MAP configuration.
- If the network is small, we can apply the max-product recursively to obtain all MAP configurations.

Datasets

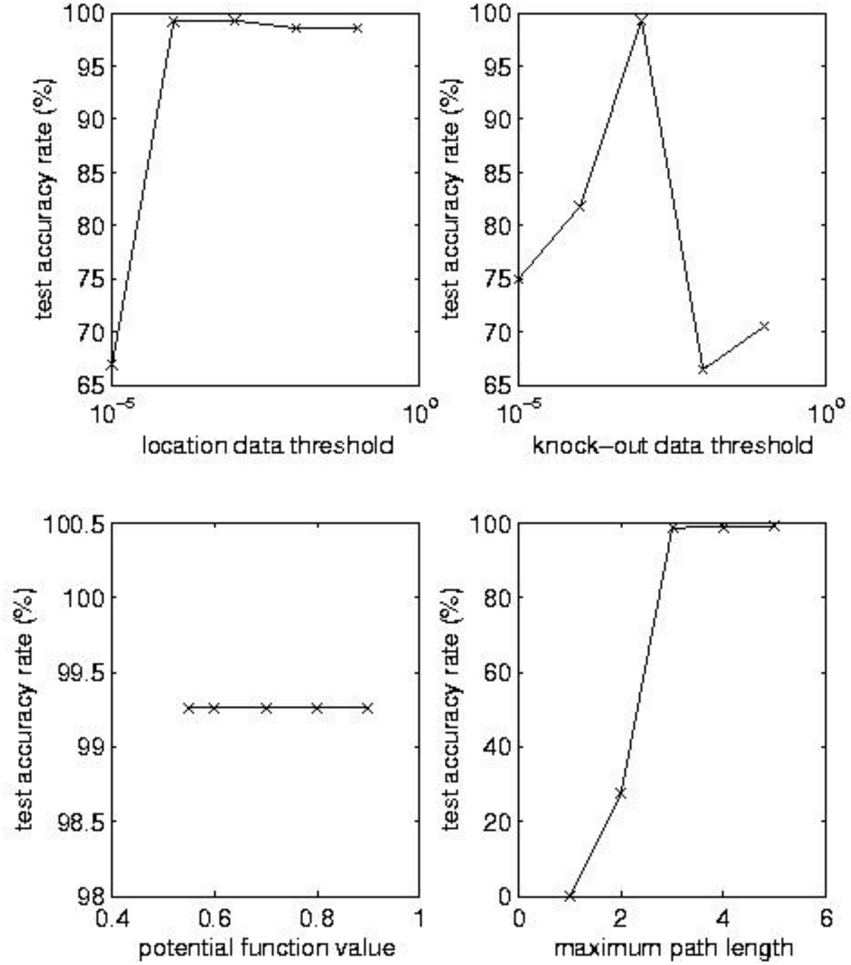
- 46 genes including 2 transcription factors (STE12 and MCM1).
- Binding p-value threshold 0.001 result in 34 protein-DNA edges (Lee et al., 2003).
- 30 protein-protein edges (DIP).
- 164 knock-out pairs from 10 experiments (Hughes et al., 2000).
- Maximal path length set to 5.

Results: yeast mating pathway

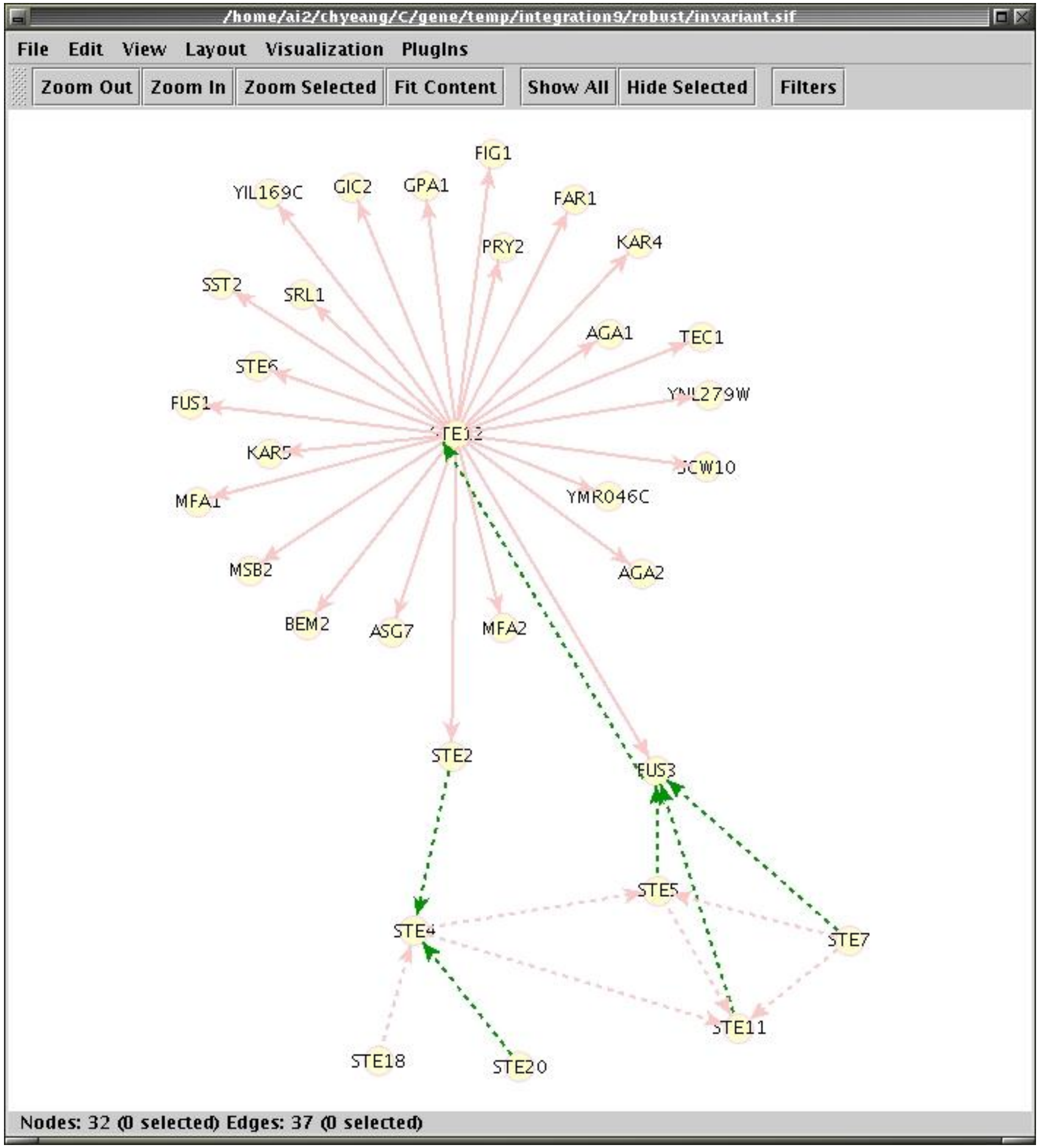
- 129 knock-out pairs are connected via valid paths.
- 8 MAP configurations.
- 129 knock-out pairs are explained by all MAP models.
- 106 knock-out pairs are explained by non-trivial inference.
- 2 knock-out pairs whose explanatory paths are not constrained by other knock-out pairs

Robustness of the model

- Are prediction outcomes sensitive to parameter settings?
- Robustness tests on location and knock-out p-value cutoffs, potential values and path length

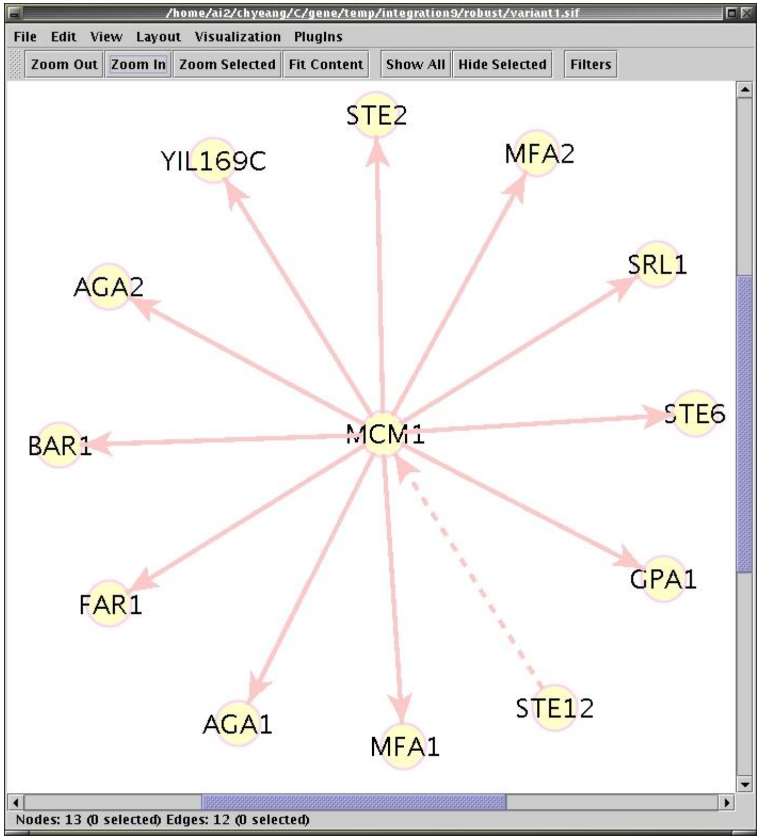


Common features for all MAP models

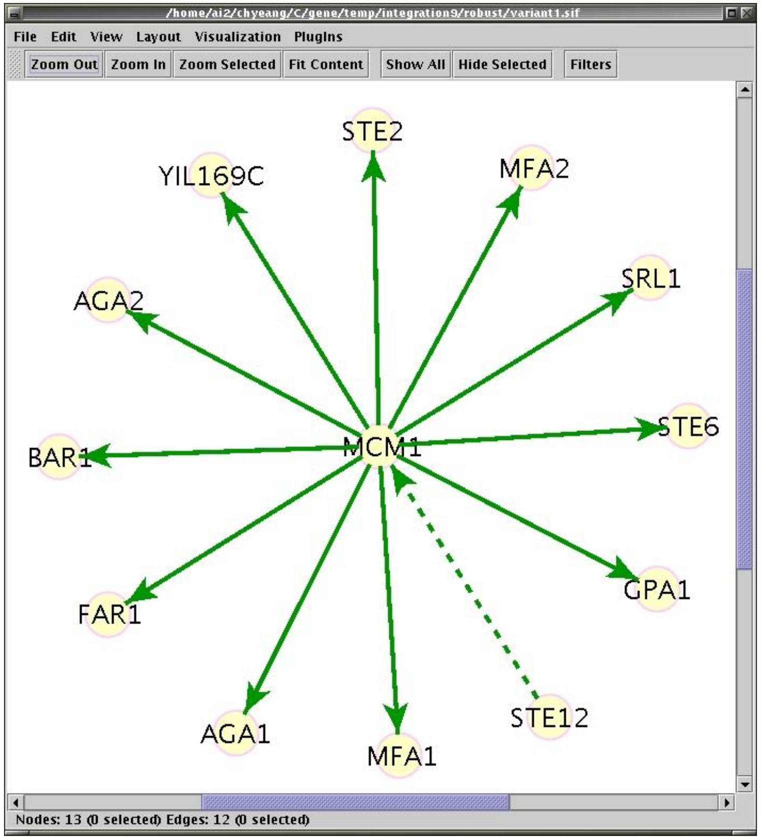


Nodes: 32 (0 selected) Edges: 37 (0 selected)

Variant features

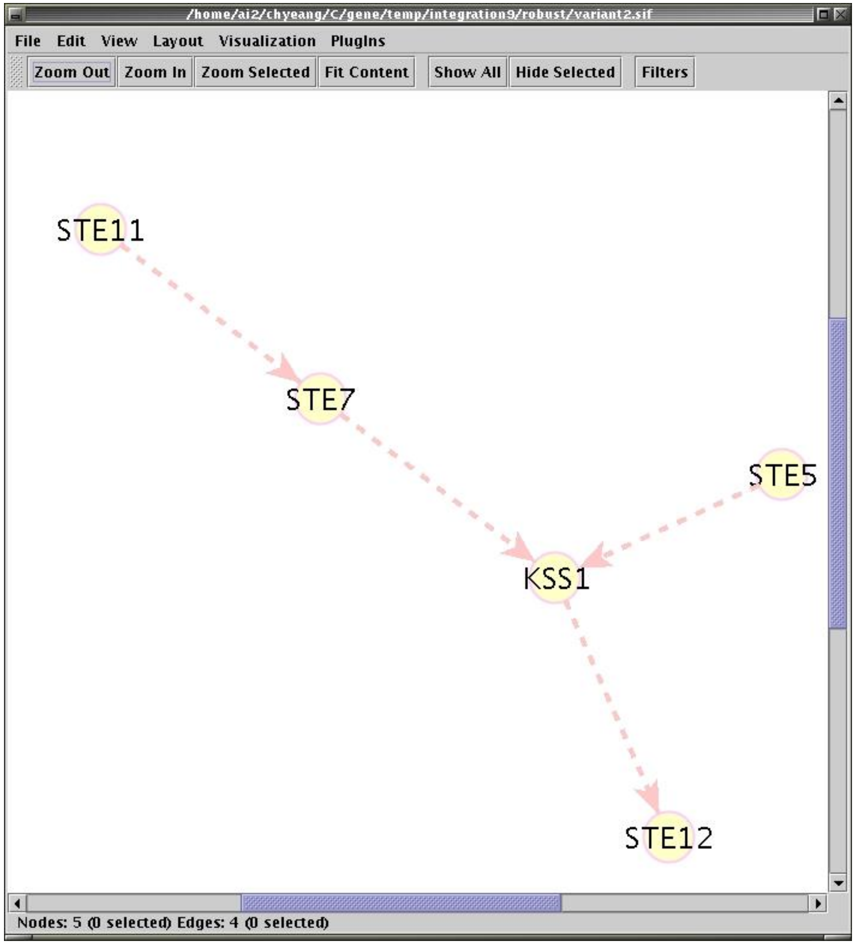


Nodes: 13 (0 selected) Edges: 12 (0 selected)

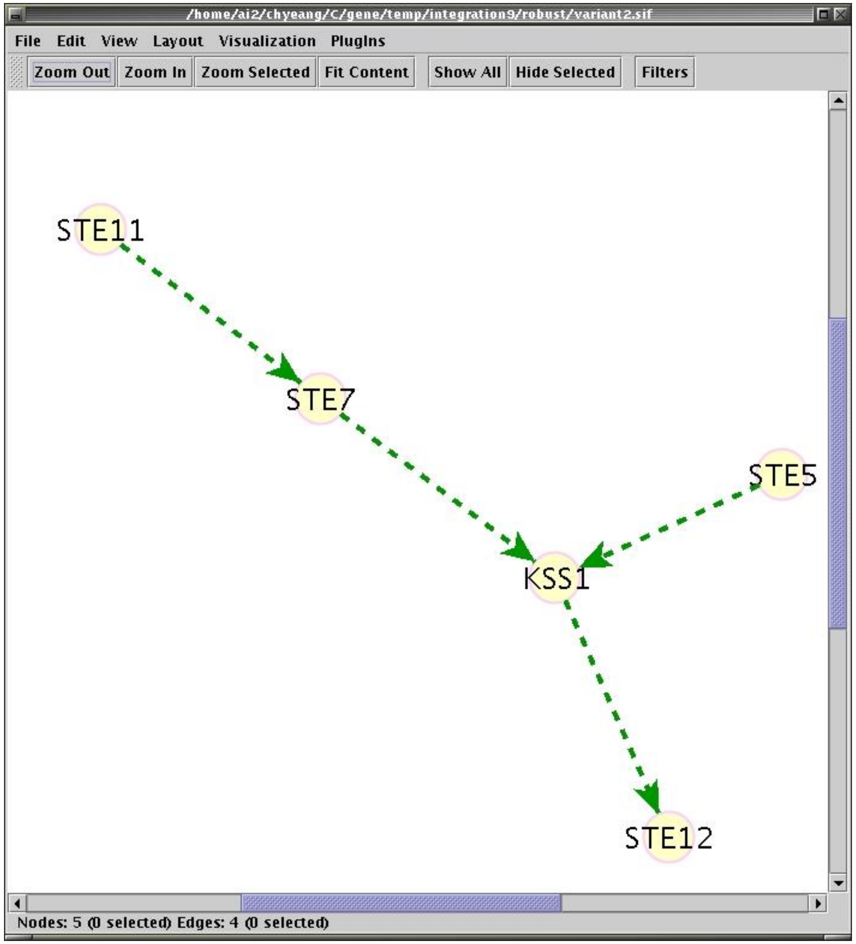


Nodes: 13 (0 selected) Edges: 12 (0 selected)

Variant features



Nodes: 5 (0 selected) Edges: 4 (0 selected)



Nodes: 5 (0 selected) Edges: 4 (0 selected)

Resolving ambiguities in the model

- Resolving ambiguities in the model requires new experiments
- There are many possible experiments (knockout of every gene)
- How can we chose which one to perform?

Active learning

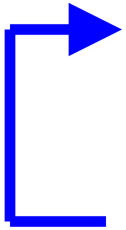
- Assume we want to teach a computer to distinguish between cats and dogs …

Can you give me some outdoor dog and indoor cat pictures?

Sure!

Active Learning for designing experiments

- On the basis of current model, *M*, the learner



  - predicts the answers O_x to various possible queries q_x
  - computes which query's answer will be most beneficial in improving model quality (or minimizing the loss)
  - Perform the experiment, updates model with the answer

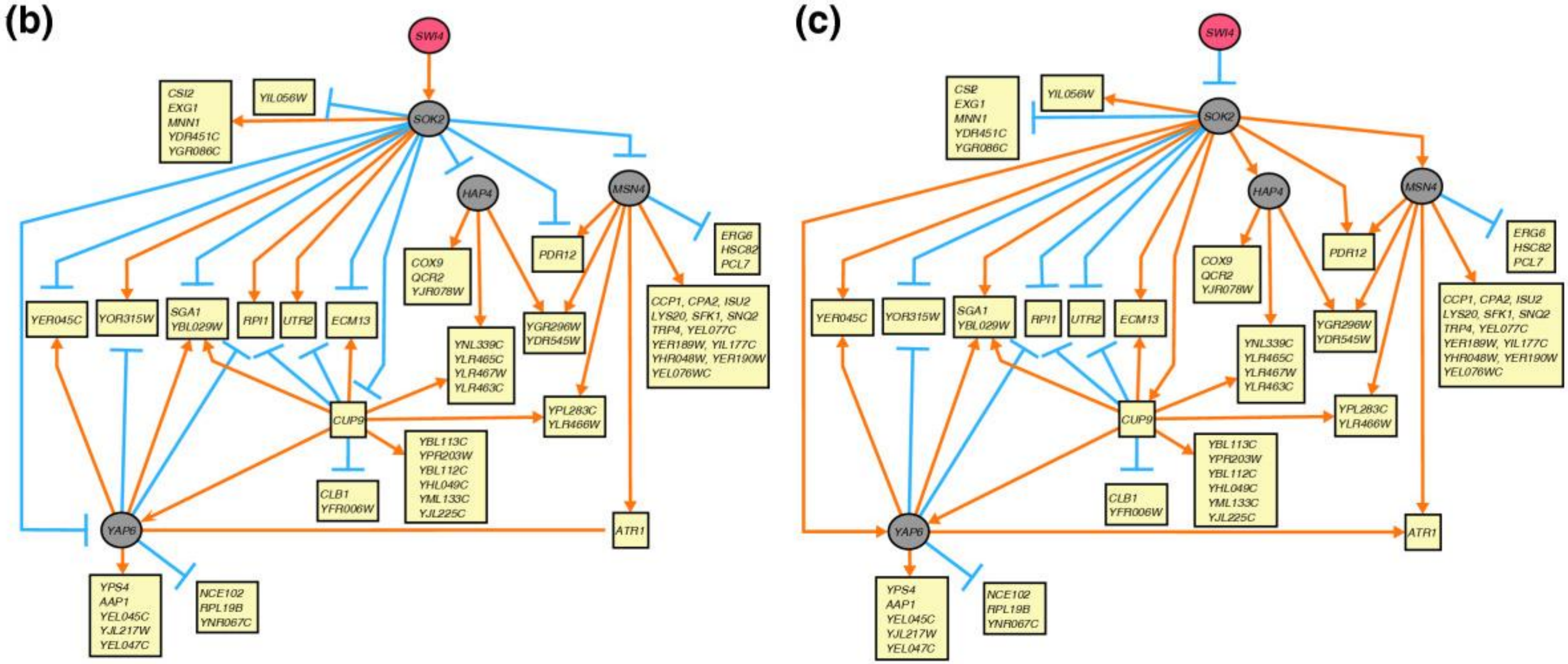
$$\min\langle Loss(q_x)\rangle = \min E[Loss(M^{O_x})]$$

Table 2

Top-ranking knock-out experiments proposed for model discrimination

Gene	Function	Score	Downstream genes	Rank	Model
HHF1	Histone	52.1429	74	1	2
*SOK2**	Regulator for meiosis and PKA pathway	45.0279	64	2	1
CKA1	Protein kinase of cell cycle	45.0075	64	3	5
A2	Mating response	40.9023	58	4	4
*YAP6**	Stress response regulator	35.1652	50	5	1, 3
NRG1	Regulator of glucose dependent genes	31.6501	45	6	3
FKH1	Regulator of cell cycle	29.1194	41	7	2
FKH2	Regulator of cell cycle	26.7131	38	8	7
SLT2	Protein kinase of cell wall integrity pathway	23.4727	31	9	8
*MSN4**	Regulator of stress response	21.8224	31	10	1
*HAP4**	Regulator of cellular respiration	6.3310	9	34	1

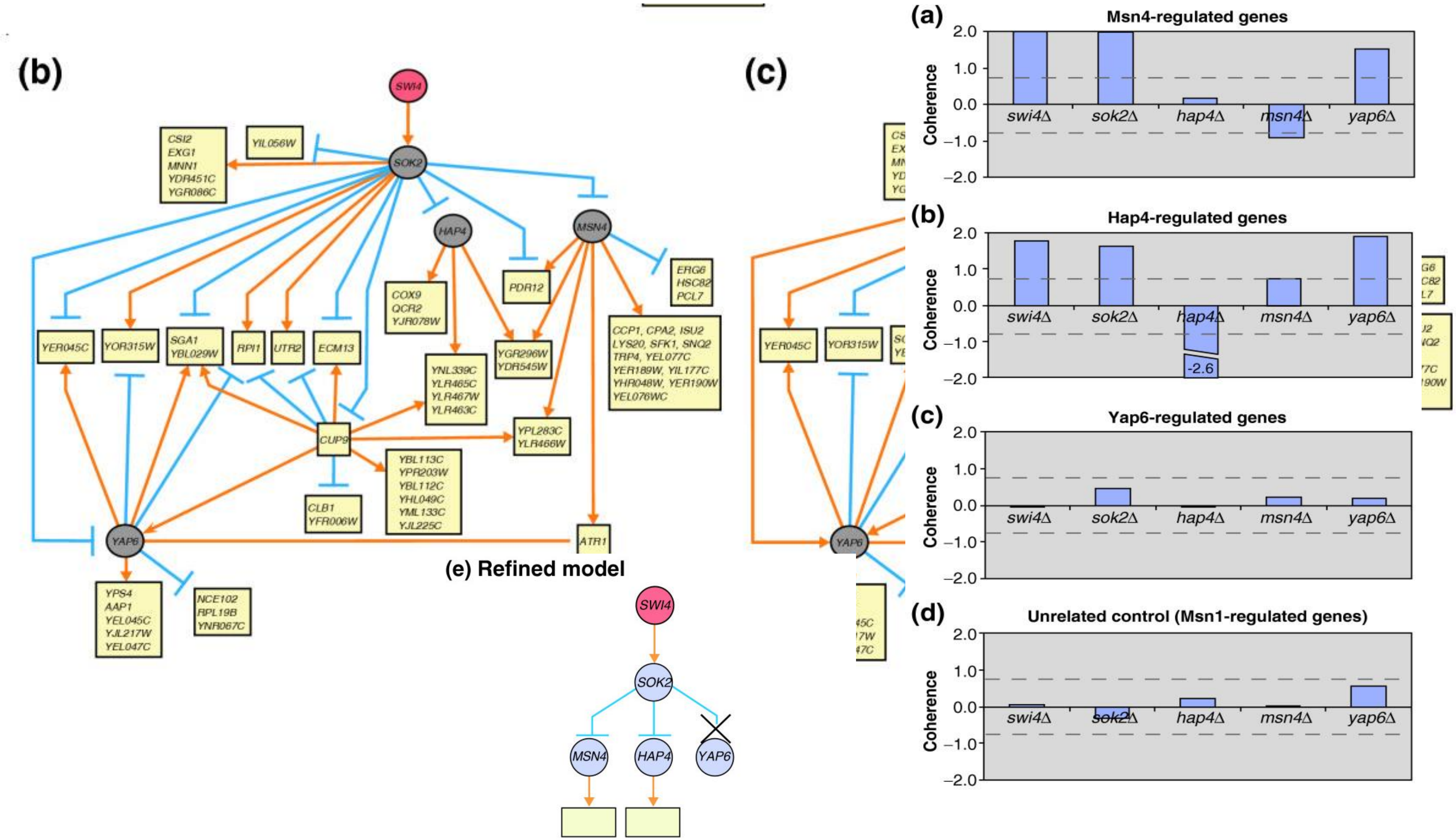
Targeting specific network



Using the ranked list

- How should we use the list in the previous table?
- Performing all the experiments at once ignores the dependency between these experiments
- Its much better to carry them one at a time
- However, that may cause other problems that are less desirable.

Experiments carried out


(b)
SWI4
SOK2
CSI2
EXG1
MNN1
YDR451C
YGR086C
YIL056W
HAP4
MSN4
COX9
QCR2
YJR078W
PDR12
ERG6
HSC82
PCL7
CCP1, CPA2, ISU2
LYS20, SFK1, SNQ2
TRP4, YEL077C
YER189W, YIL177C
YHR048W, YER190W
YEL076WC
YGR296W
YDR545W
YNL339C
YLR465C
YLR467W
YLR463C
YER045C
YOR315W
SGA1
YBL029W
RPI1
UTR2
ECM13
CUP9
YPL283C
YLR466W
YBL113C
YPR203W
YBL112C
YHL049C
YML133C
YJL225C
CLB1
YFR006W
ATR1
YAP6
YPS4
AAP1
YEL045C
YJL217W
YEL047C
NCE102
RPL19B
YNR067C
(c)
YER045C
YOR315W
YAP6
(e) Refined model
SWI4
SOK2
MSN4
HAP4
YAP6
(a)
Msn4-regulated genes
(b)
Hap4-regulated genes
-2.6
(c)
Yap6-regulated genes
(d)
Unrelated control (Msn1-regulated genes)
Coherence
2.0
1.0
0.0
−1.0
−2.0
swi4Δ
sok2Δ
hap4Δ
msn4Δ
yap6Δ