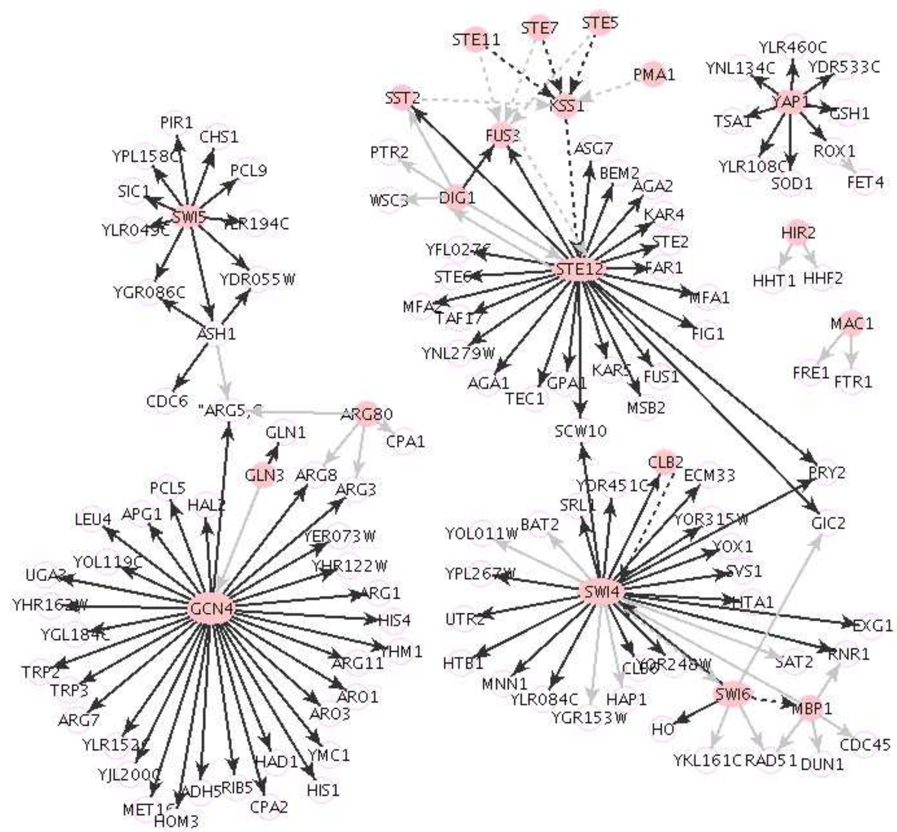
Inferring molecular interaction networks

Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

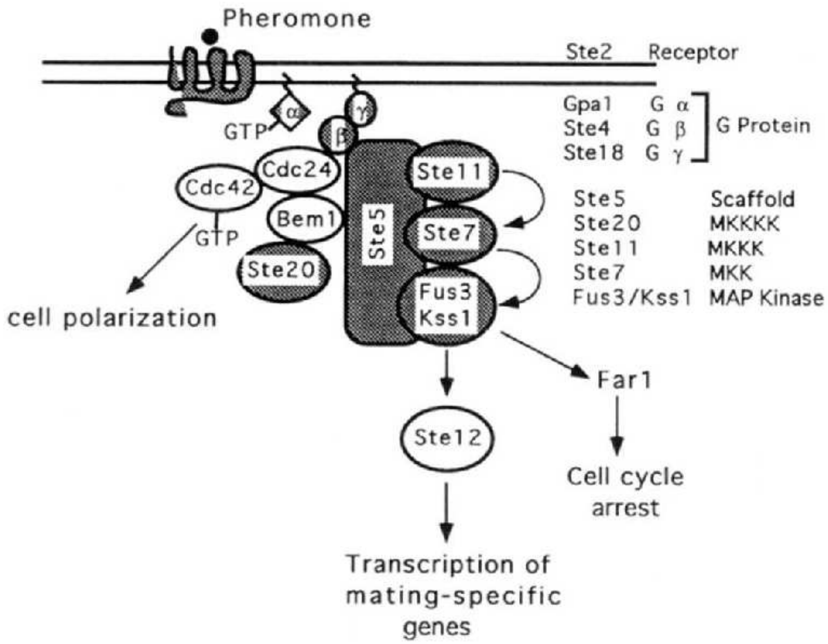
In collaboration with
Chen-Hsiang Yeang, MIT
Trey Ideker, UCSD

Models of transcriptional regulation

- Mechanistic models, annotated networks of molecular interactions
  - interpretable, verifiable
  - statistical models, emphasis on causation



Outline

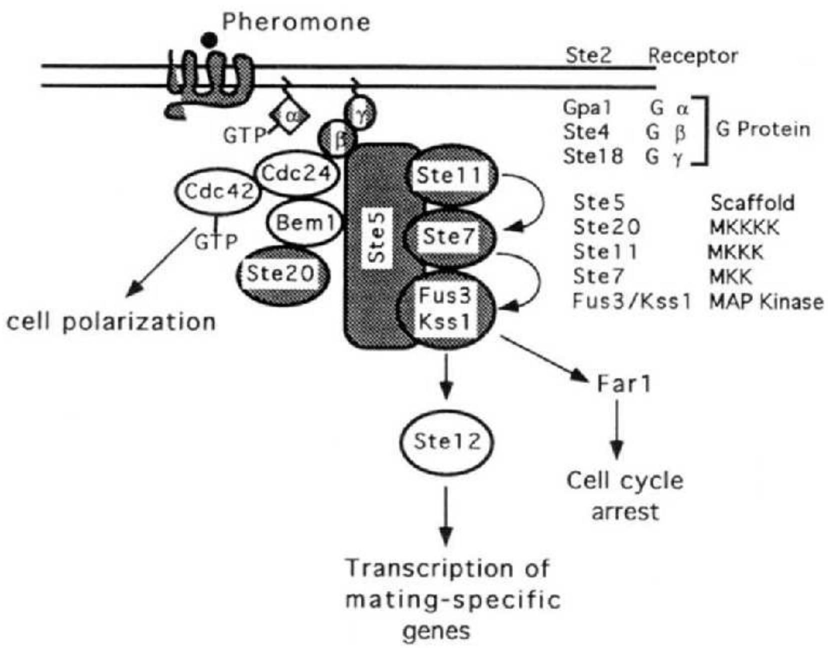


(Schrick et al)

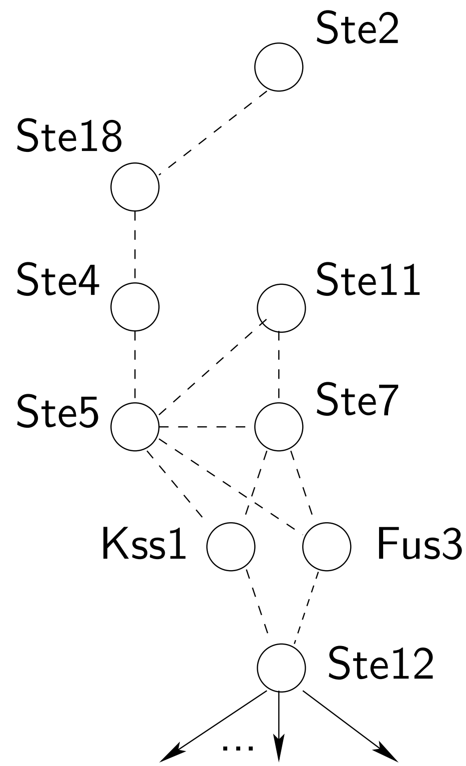
- Abstraction
- Data association and mechanisms
- Model inference
- Experiment design
- Model revision
- Model verification

Abstraction

- Mechanistic model: representation of molecular interactions

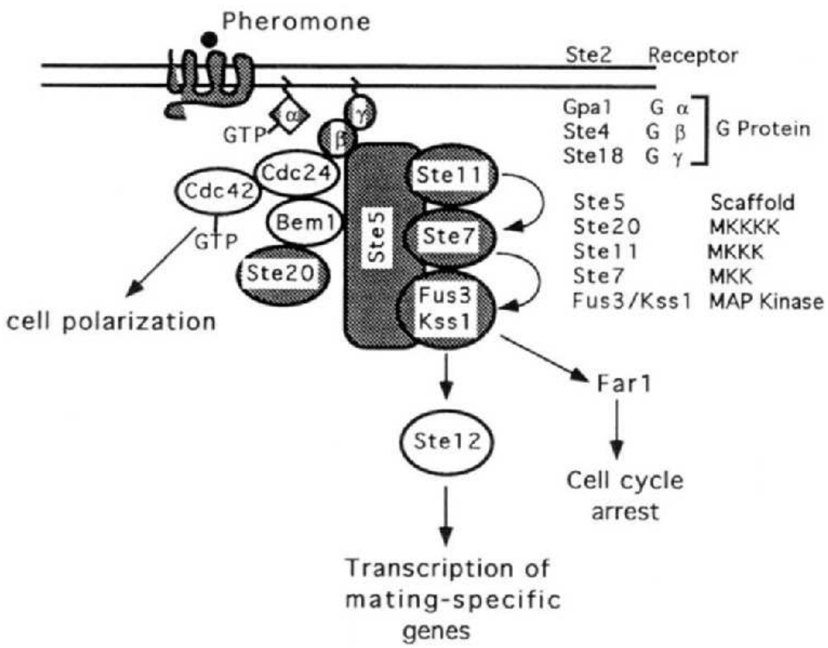


(Schrick et al)

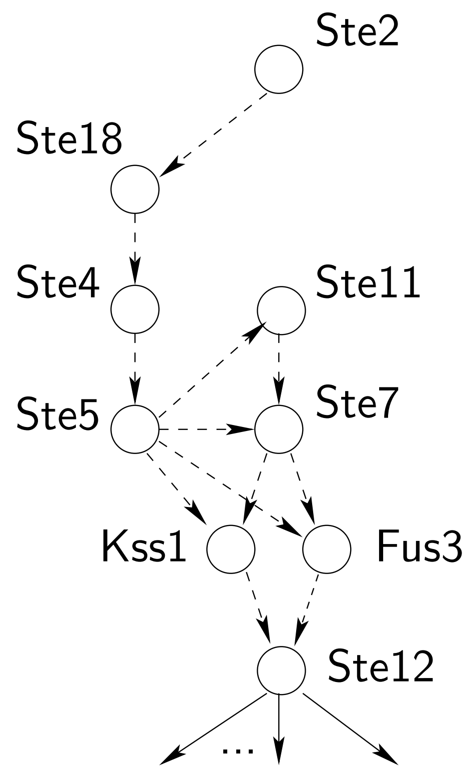


Abstraction

- Mechanistic model: representation of molecular interactions



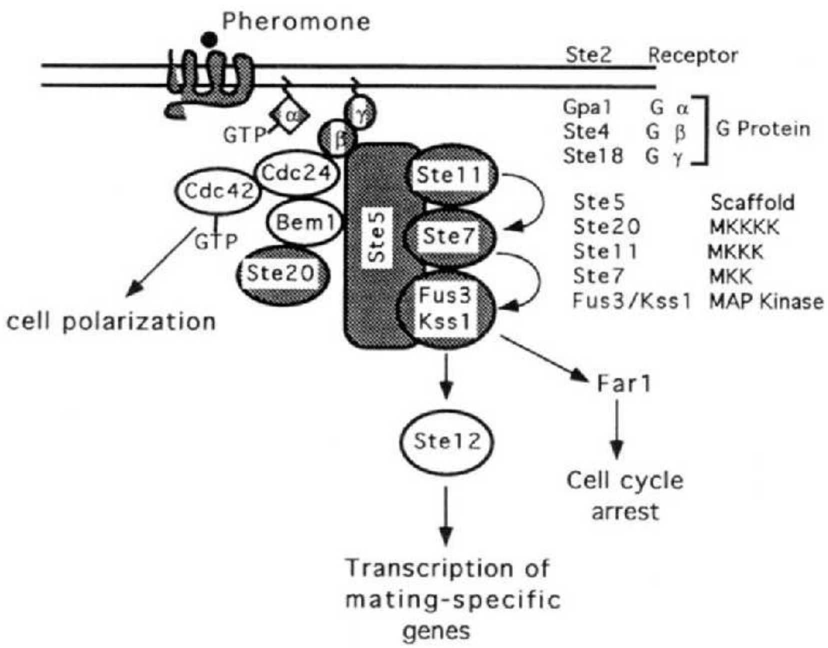
(Schrick et al)



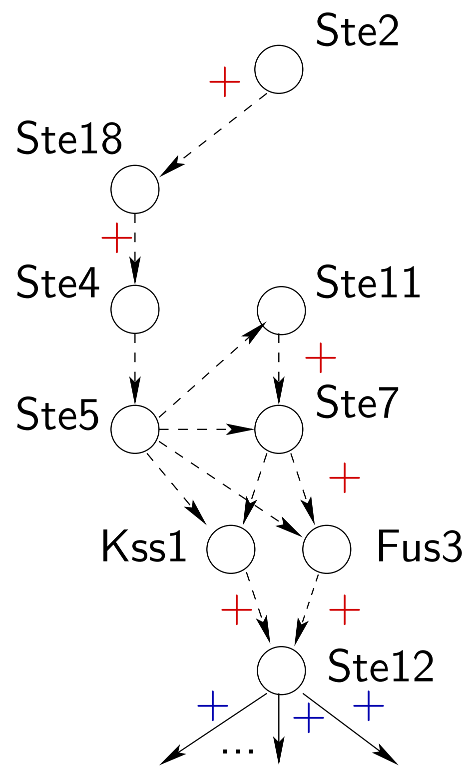
– pathways (information flow)

Abstraction

- Mechanistic model: representation of molecular interactions



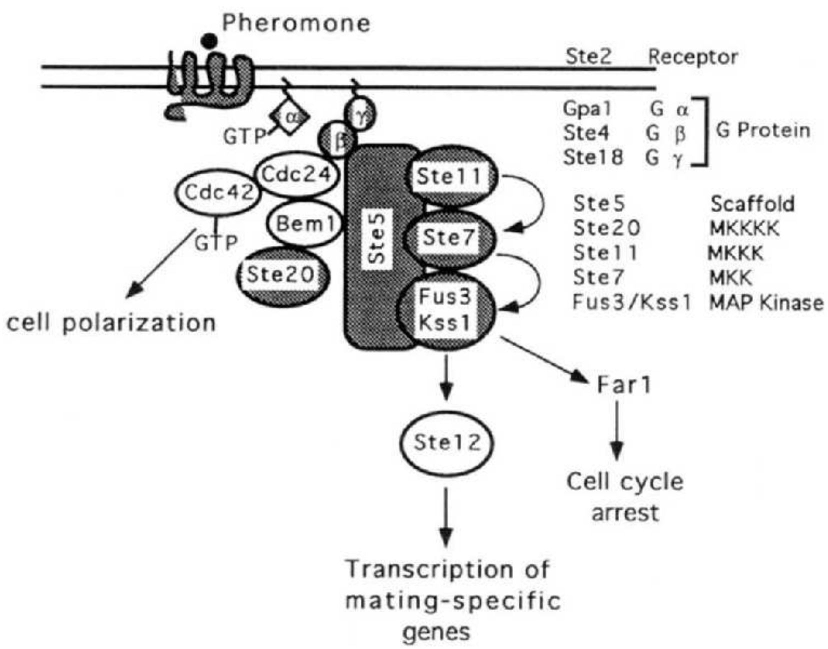
(Schrick et al)



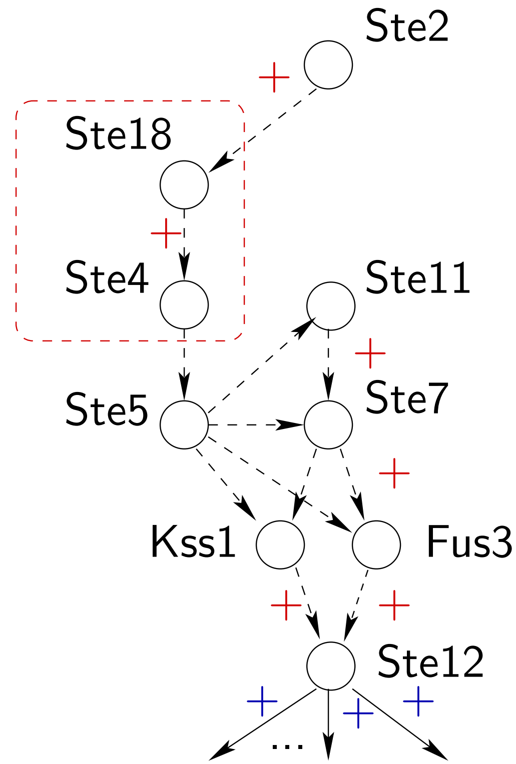
- pathways (information flow)
- effects of interactions

Abstraction

- Mechanistic model: representation of molecular interactions



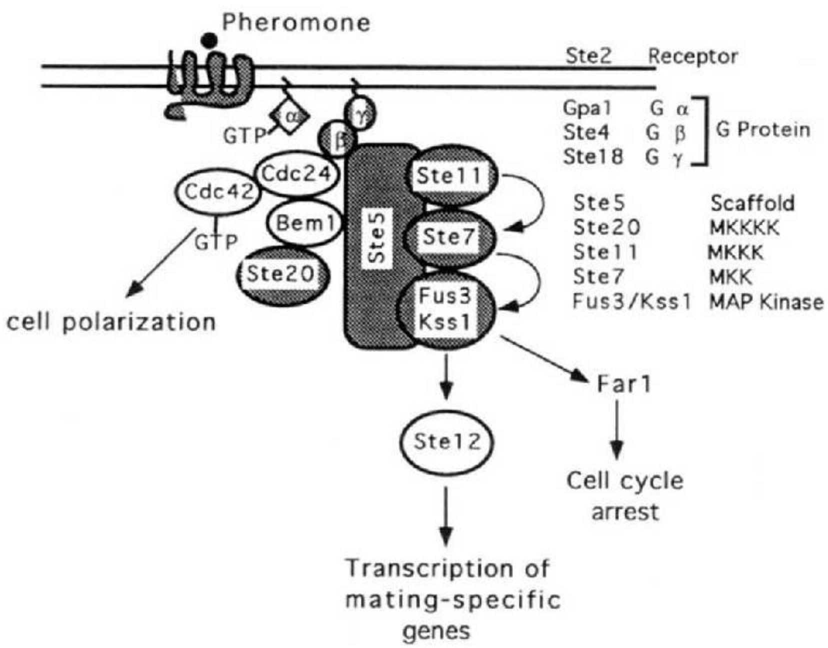
(Schrick et al)



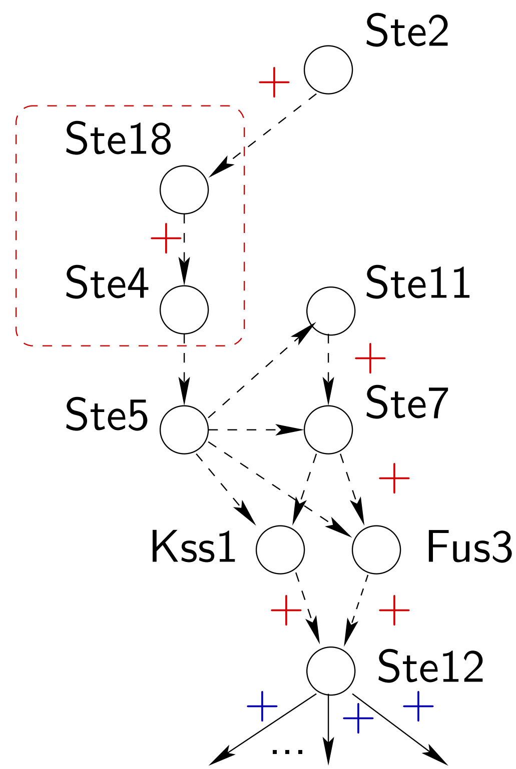
– pathways (information flow)
– effects of interactions
– complexes

Abstraction

- Mechanistic model: representation of molecular interactions



(Schrick et al)

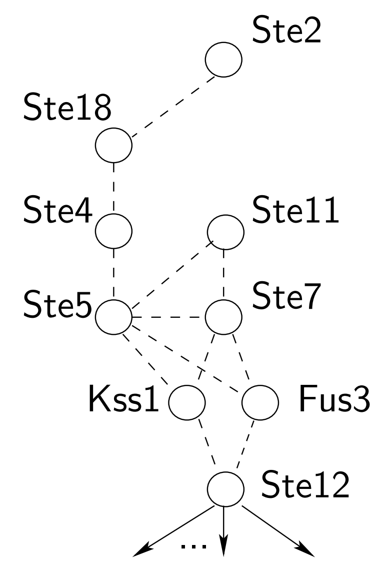


- pathways (information flow)
- effects of interactions
- complexes
- protein modification, etc.

Variables and empirical constraints

- The model (annotated graph) is specified by three types of variables
  - presence/absence of edges
  - directionality of protein-protein interactions
  - effects (signs) of interactions

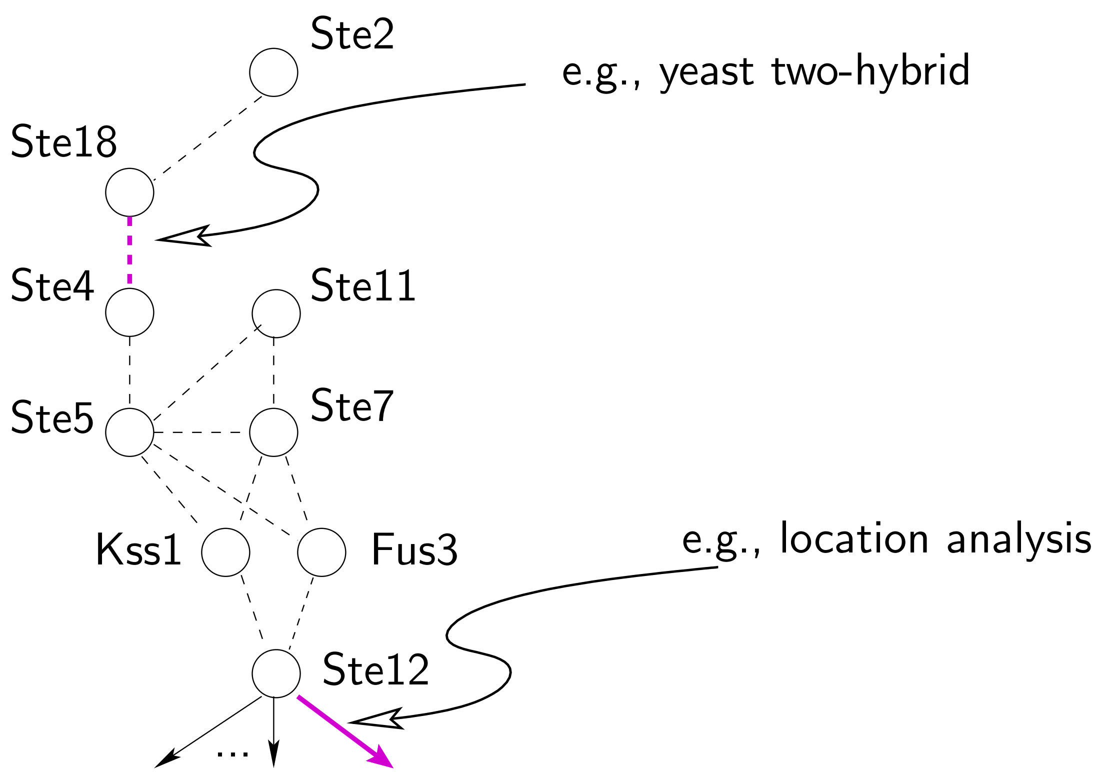
  (and the initial set of possible interactions)



- The empirical constraints on the settings of these variables come from both direct physical measurements and indirect functional measurements

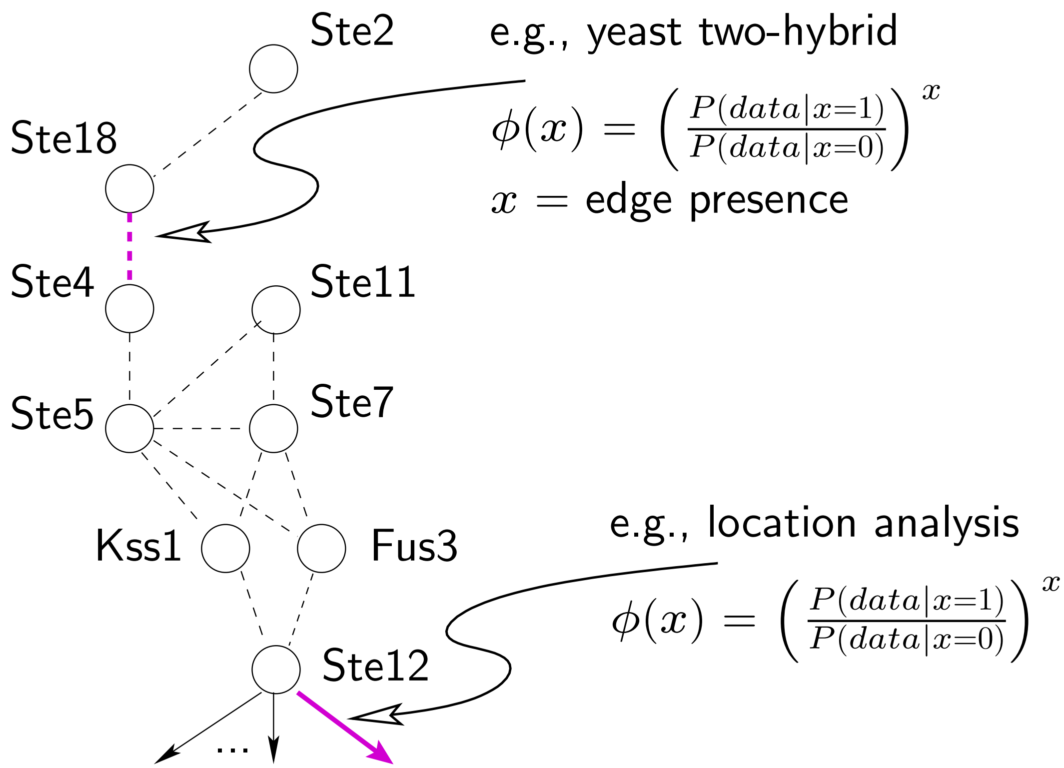
Direct physical measurements

- protein-protein
- protein-DNA



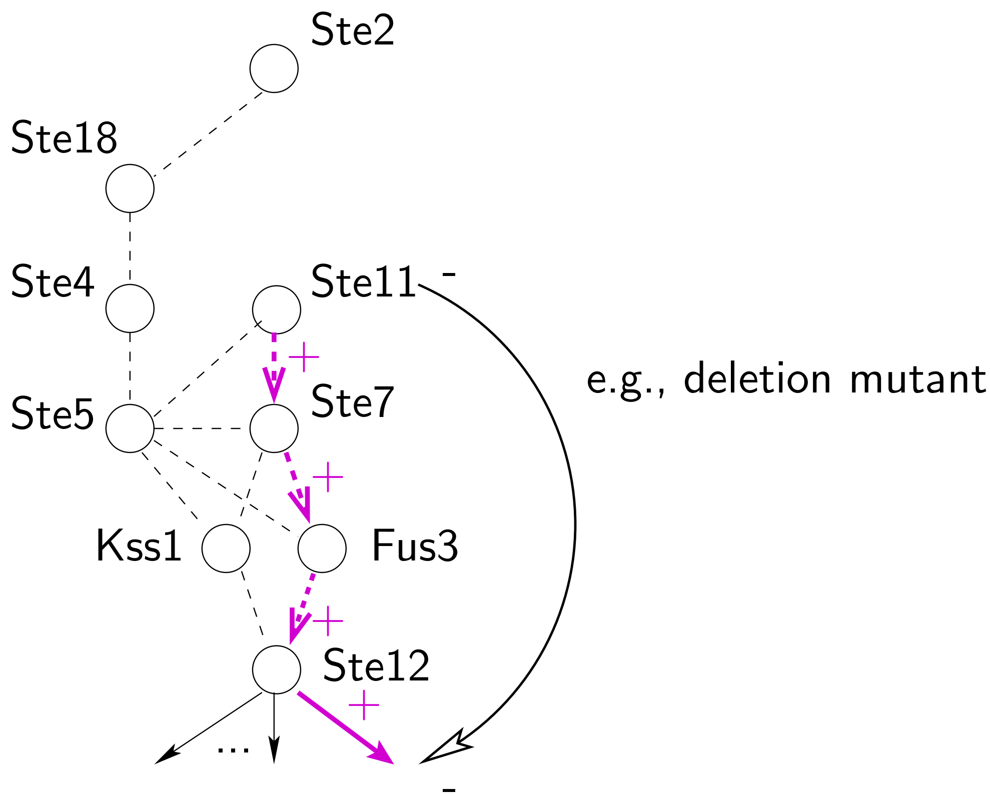
Direct physical measurements

- protein-protein
- protein-DNA



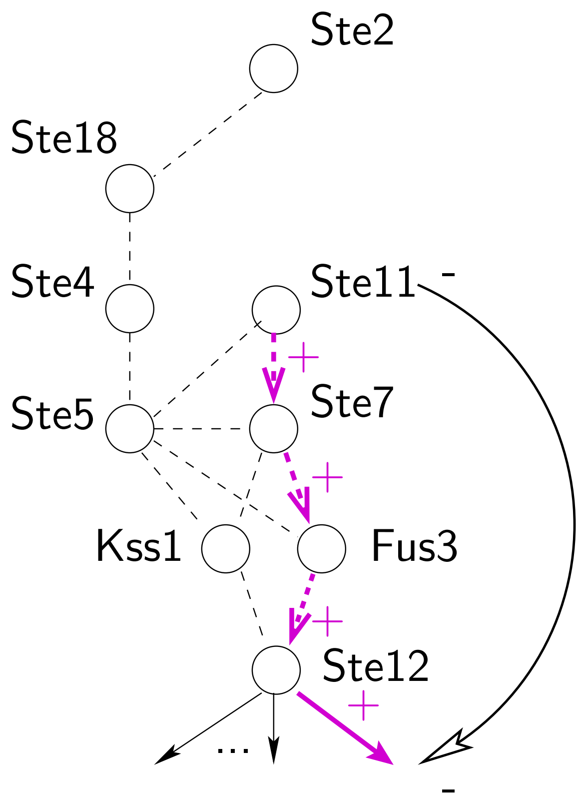
Indirect measurements: data association

- We need to introduce molecular mechanisms (e.g., cascades) to explain indirect functional measurements such as gene knock-outs



Indirect measurements: data association

- We need to introduce molecular mechanisms (e.g., cascades) to explain indirect functional measurements such as gene knock-outs



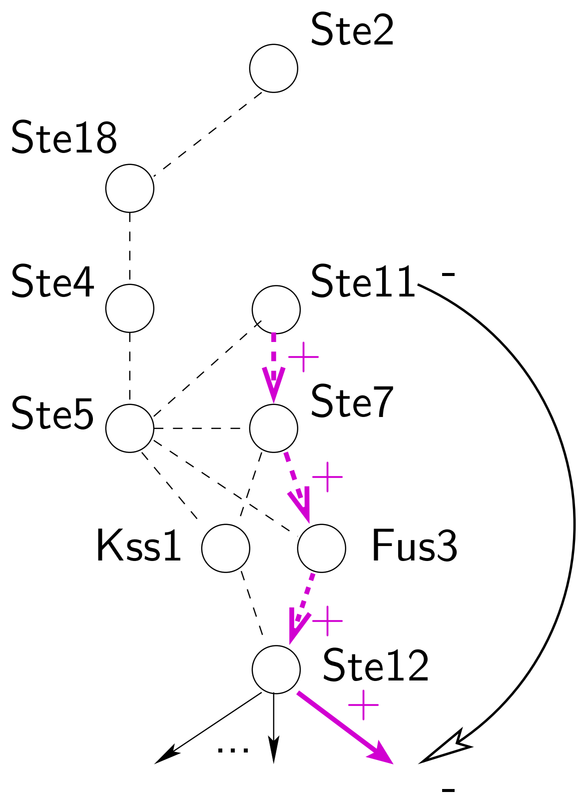
e.g., deletion mutant

$$\psi(k) = \frac{P(data|k)}{P(data|k=0)}$$

$k \in \{-1, 0, 1\}$, knock-out effect

Indirect measurements: data association

- We need to introduce molecular mechanisms (e.g., cascades) to explain indirect functional measurements such as gene knock-outs



e.g., deletion mutant

$\psi(k) = \frac{P(data|k)}{P(data|k=0)}$

$k \in \{-1, 0, 1\}$, knock-out effect

$\psi(k, \vec{x}_p, \vec{d}_p, \vec{s}_p)$: path constraints

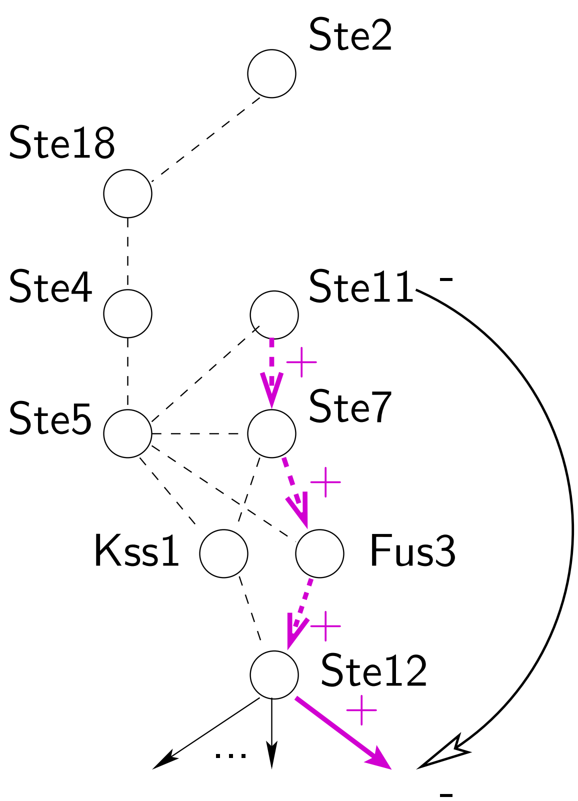
$\vec{x}$ edges along path p

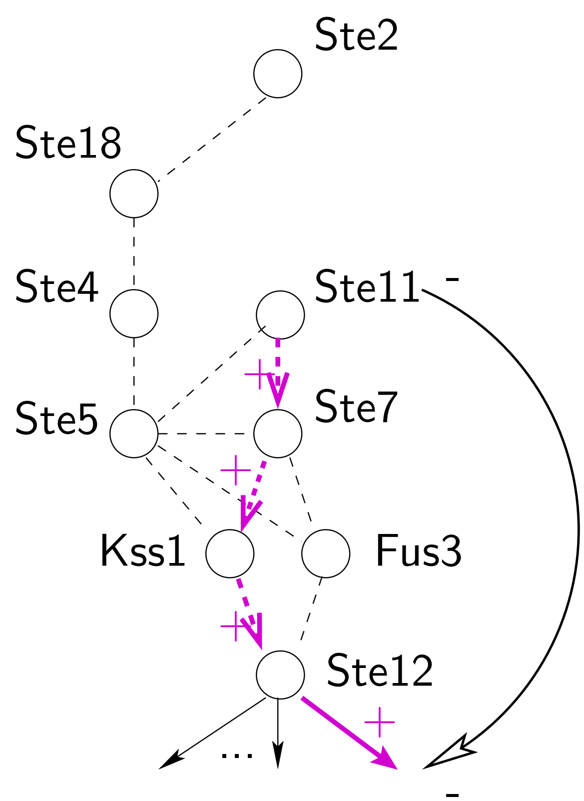
$\vec{d}$ protein-protein edge directions

$\vec{s}$ edge signs

Indirect measurements: data association

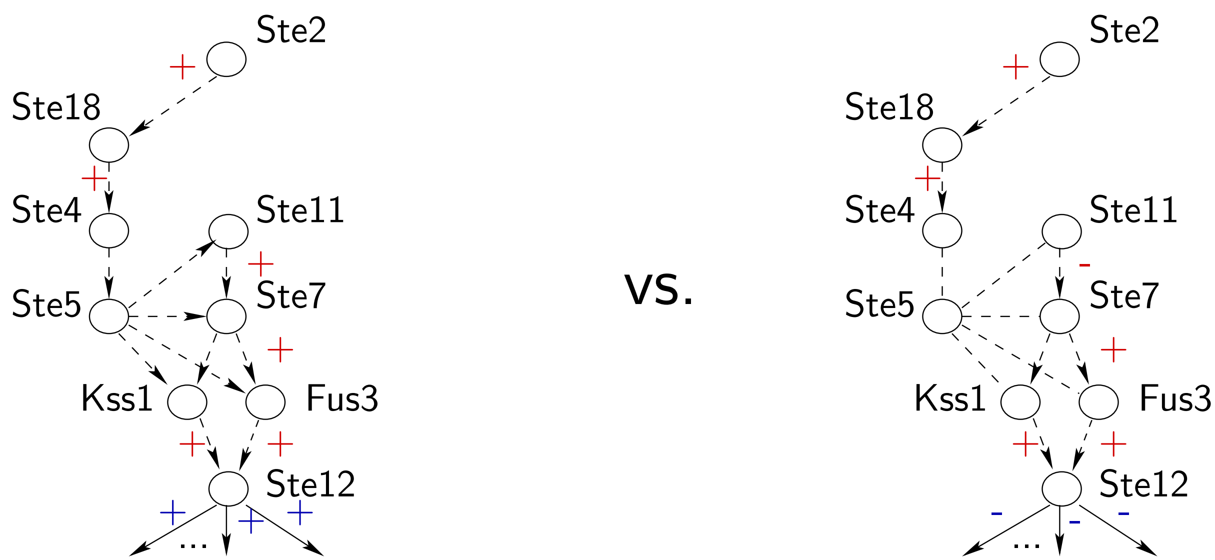
- Only some of the available paths (cascades) are “valid”
  - transcriptional control, length, consistency





Reconstruction

- Finding the most likely annotated graph can be cast as an inference problem



$$\text{Prob(annoted graph)} \;=\; \overbrace{(\;\cdots\;)}^{\text{protein-protein}} \;\times\; \overbrace{(\;\cdots\;)}^{\text{protein-DNA}} \;\times$$

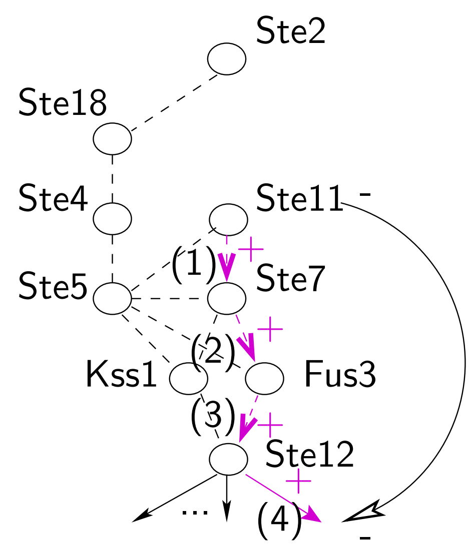
$$\times\; \underbrace{(\;\cdots\;)}_{\text{knock-out}}$$

Inference algorithms

- A number of approximation algorithms such as max-product, generalized max-product, or tree re-weighted max-product are available to solve the inference problem
- The nature of the aggregate constraints permits efficient application of such algorithms.

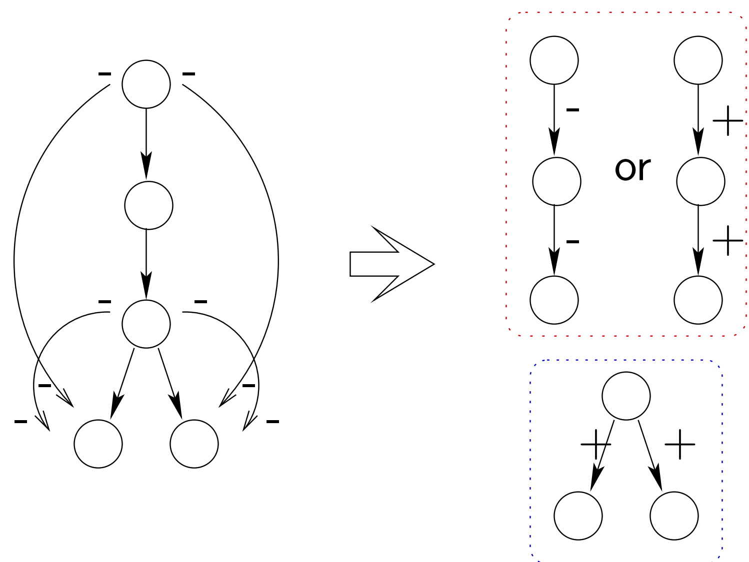
$$\psi(\vec{x}_p, \vec{s}_p, \vec{d}_p) =$$

$$\underbrace{\prod_{i=1}^{4} \delta(x_i, 1)}_{\text{and}} \cdot \underbrace{\delta(\prod_{i=1}^{4} s_i, 1)}_{\text{parity}} \cdot \underbrace{\prod_{i=1}^{3} \delta(d_i, 1)}_{\text{and}}$$



Characterization of inference results

- The available data is often limited (especially knock-outs) and thus there are many (approximately) equally likely model configurations
- We can summarize the inference results by decomposing the model into subnetworks that are uniquely determined by data and those that have few remaining degrees of freedom (e.g., overall sign)



Example results

- Yeast mating pathway
  - are knock-out effects predictable?
  - are predictions robust to various modeling choices?
  - are the results consistent with known biology?
- Genomic scale analysis

Example results: yeast mating pathway

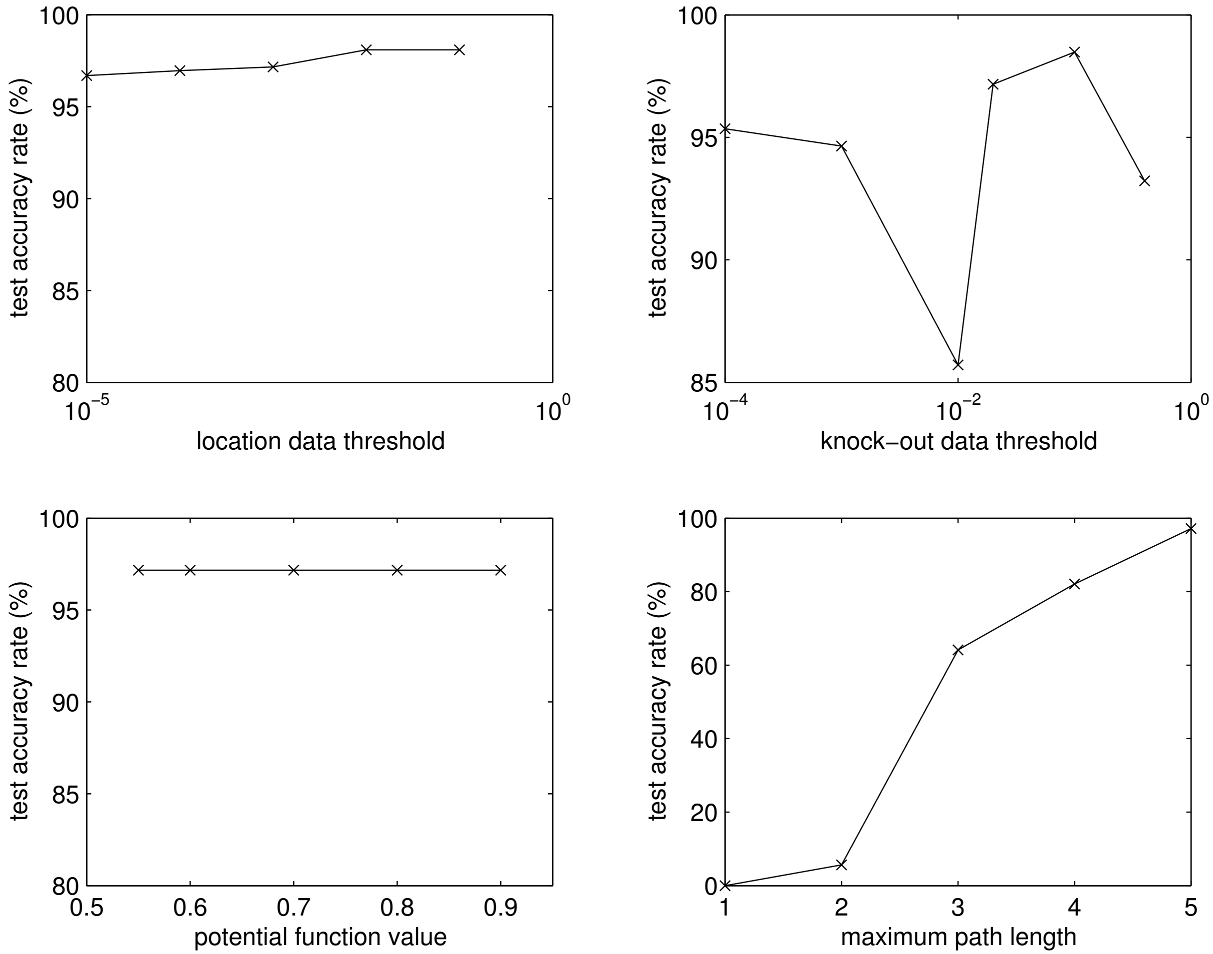
- Model estimated from 30 protein-protein, 34 protein-DNA, and 13 knock-out experiments (106 knock-out pairs)
- Cross-validation accuracy

# hold-outs	# trials	% accuracy
1	106	97.2 %
5	200	96.5 %
20	200	94.1 %

  – only significant knock-out effects included
  – prediction of a knock-out effect is deemed correct only when all the candidate paths (explanations) are consistent

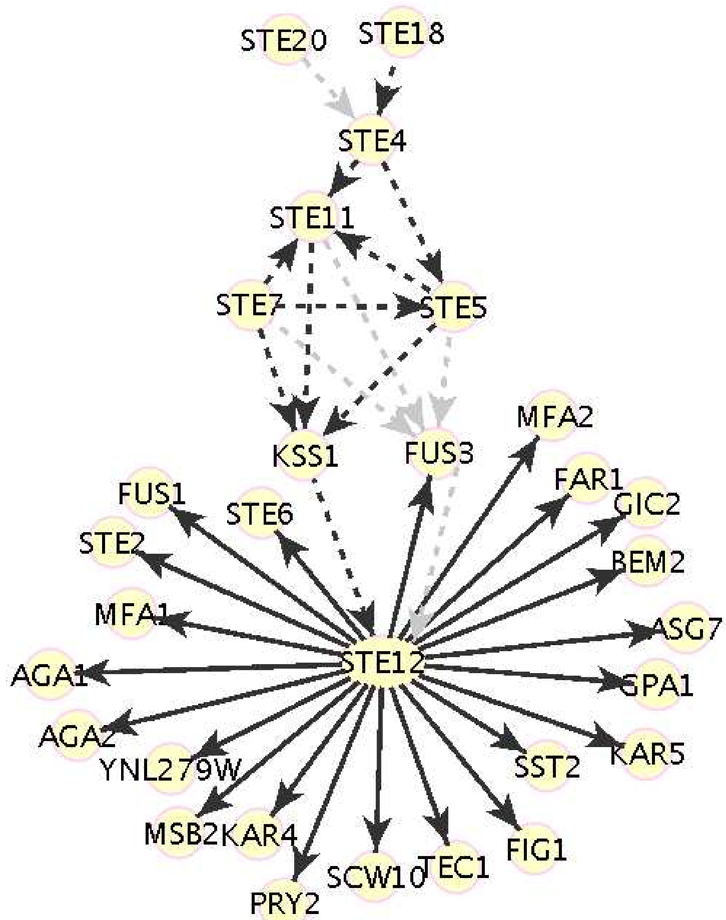
Example results: yeast mating pathway

- Robustness



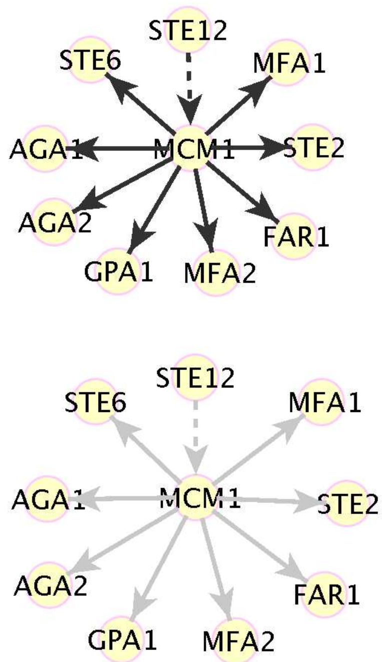
Example results: yeast mating pathway

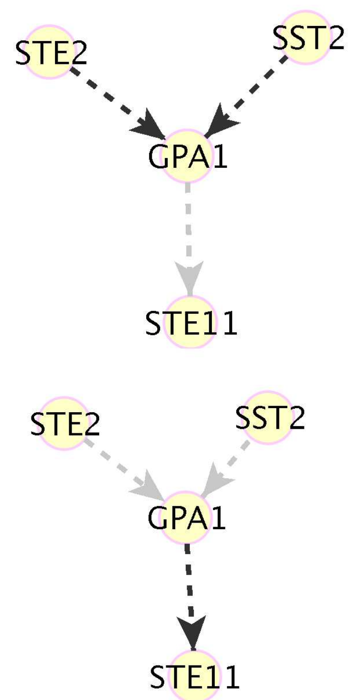
- Uniquely identified subnetwork



Example results: yeast mating pathway

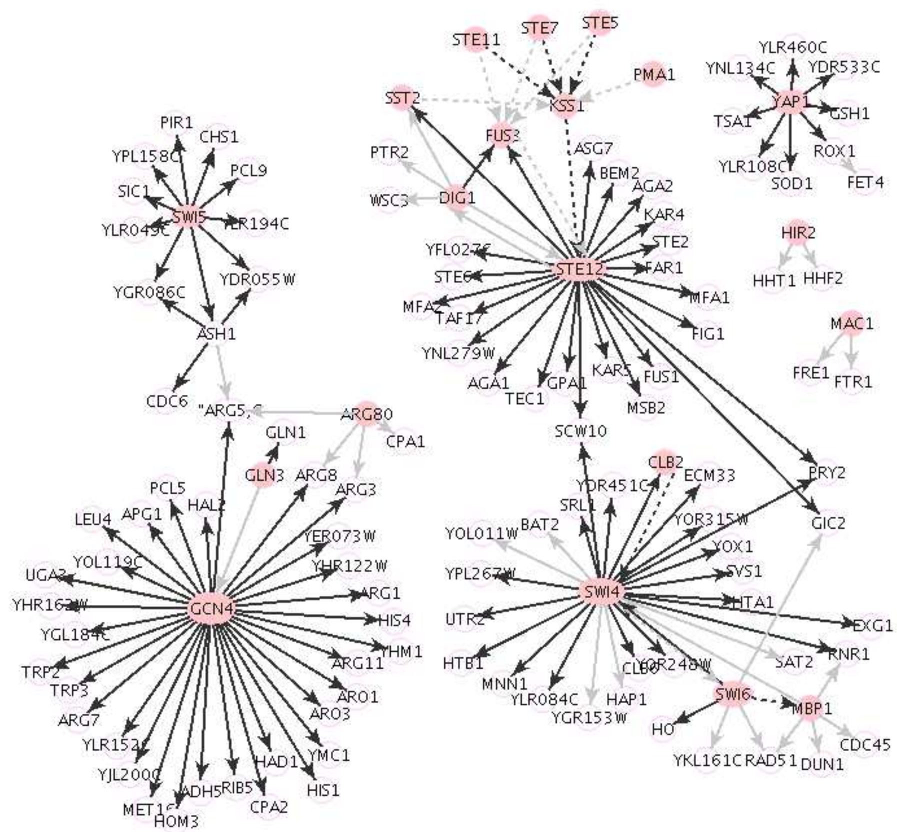
- Subnetworks with one degree of freedom (overall sign)





Example results: yeast genomic scale analysis

- 5485 protein-DNA interactions (Lee et al.), 14876 protein-protein interactions (DIP database), 23766 knock-out effects (Hughes et al)
- Uniquely identified subnetwork:

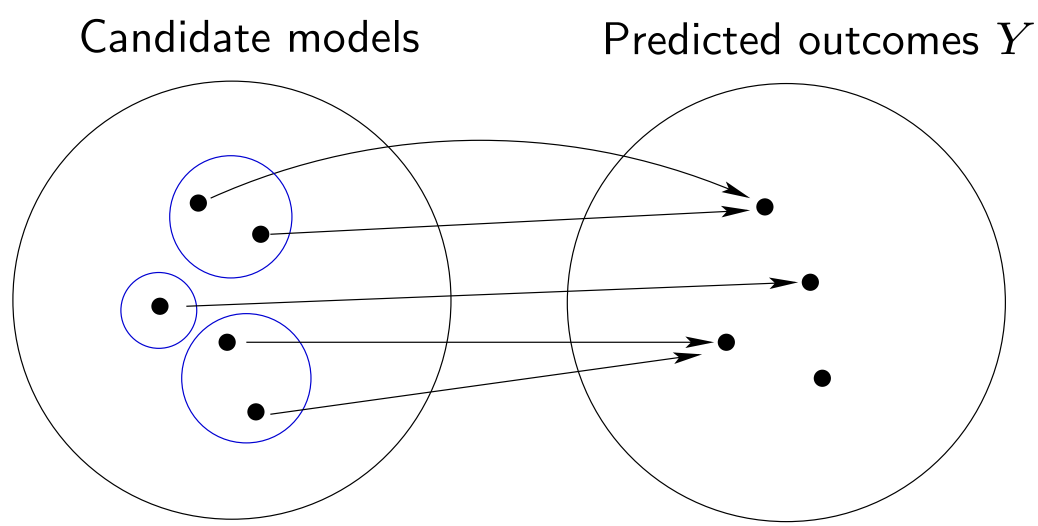


Remaining steps

- Experiment design
  - select new experiments to maximally reduce the uncertainty about the model
- Model revision
  - expand the set of possible explanations and molecular interactions to maximize explanatory power
- Model verification
  - derive and test “unique” predictions from the model

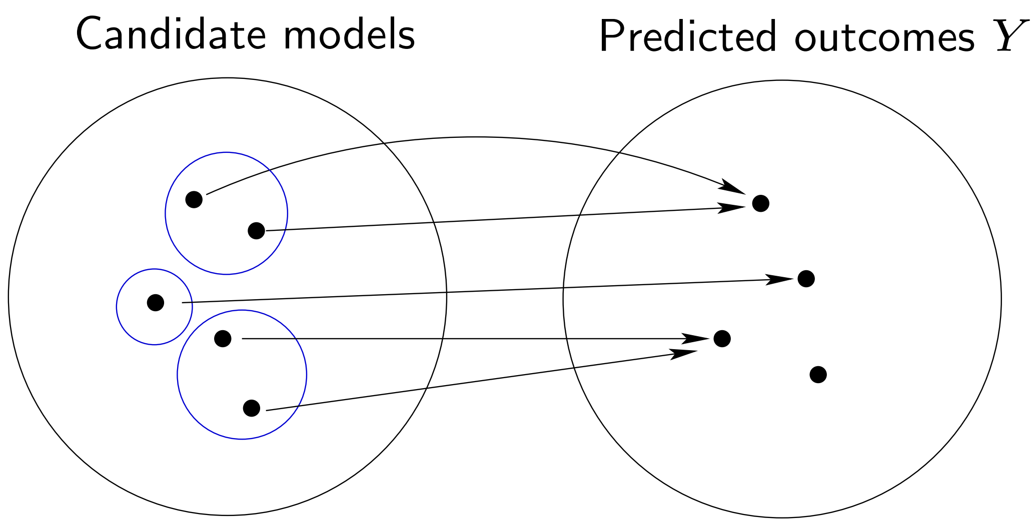
Experiment design

- The goal here is to identify experiments such as knock-outs that would best discriminate among possible models (value of information calculations)
- Each potential knock-out divides the space of possible models into sets that make identical predictions about up/down regulation of downstream genes



- The information gain from the knock-out reduces to $H(Y)$, the entropy of possible outcomes.

Experiment design cont'd

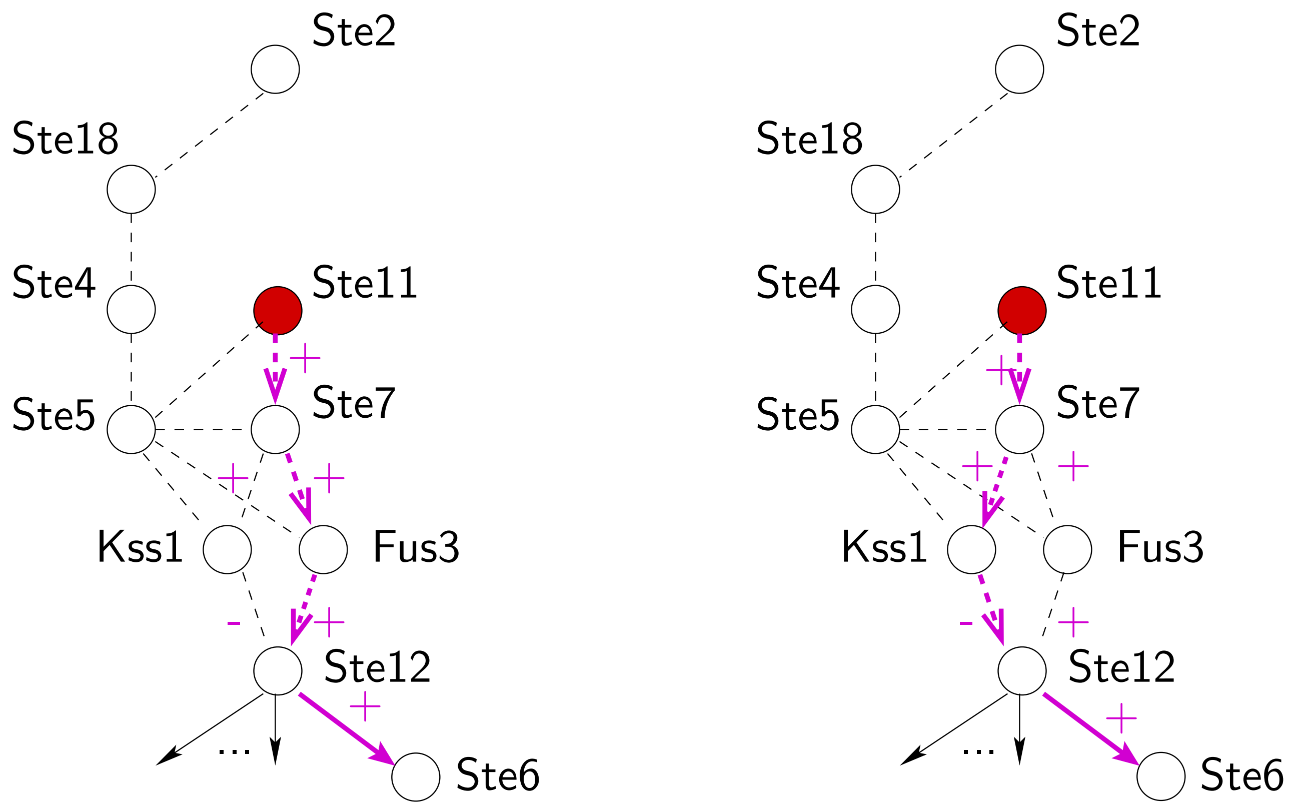


- We could evaluate $H(Y)$ via sampling but instead rely on a local approximation (upper bound) $\sum_i H(Y_i)$ that can be evaluated efficiently through propagation algorithms.

  Criterion: rank potential knock-outs according to $\sum_i H(Y_i)$, where $Y_i \in \{-1, 0, 1\}$.

Experiment design cont'd

- Paths and predictions:



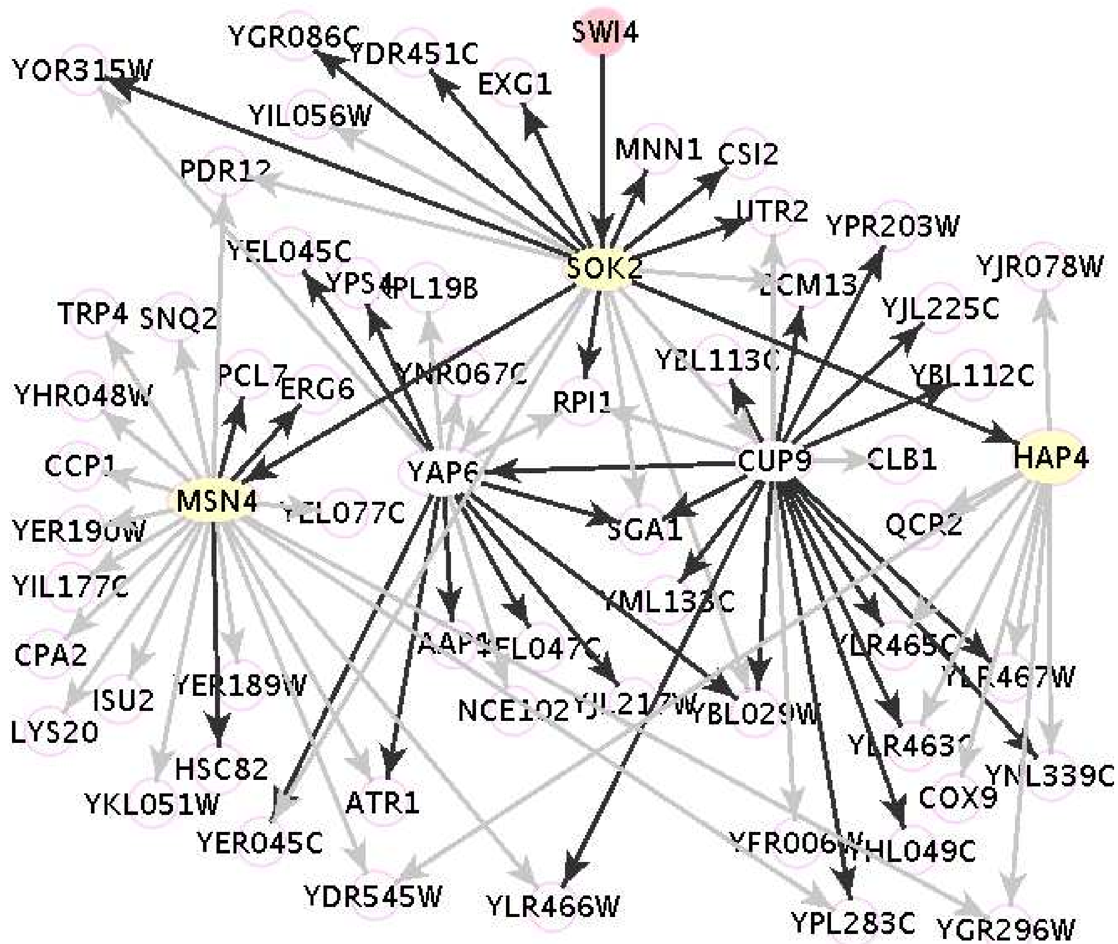
Experiment design cont'd

- List of high ranking knock-outs

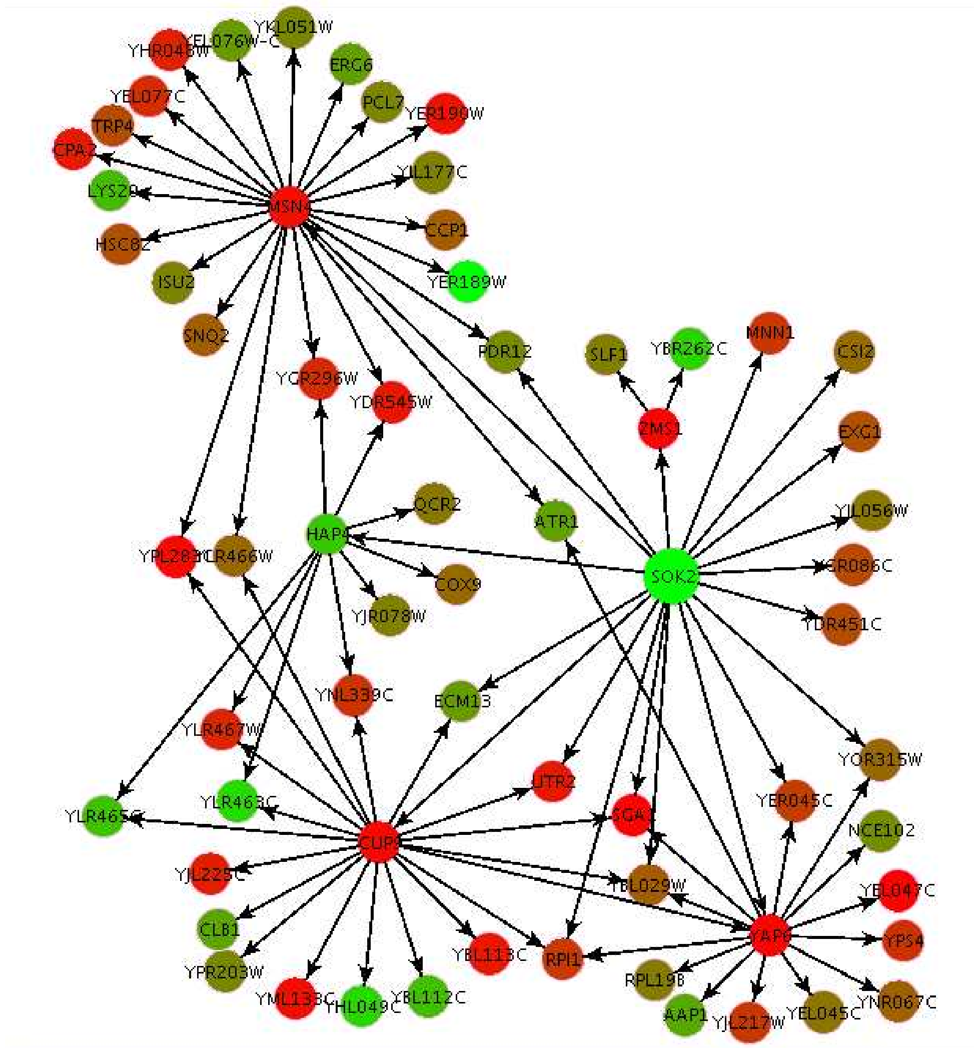
Gene	Function
ABF1	ARS-binding factor
SRB4	RNA pol II holoenzyme
CKA1	casein kinase II
NRG1	transcriptional repressor
SOK2	growth and meiosis regulator
MCM1	cell cycle and mating regulator
FHL1	RNA pol III transcription factor
FKH2	cell cycle and morphology regulator
RAP1	silencing telomeres and mating type
REB1	transcription factor for RNA pol II.

Experiment design cont'd

- A subnetwork involving SOK2 with one degree of freedom (overall sign)



SOK2 knock-out



Discussion

- We find readily interpretable and testable models by relying on networks of molecular interactions
- By articulating causal mechanisms for observed behavior we can maintain coherent models even when the models are constrained by heterogeneous data sources
- Extensions:
  - mechanisms underlying combinatorial control
  - model verification (competitive formulation)