Biostatistics and Medical Informatics 776
Computer Sciences 776
Advanced Bioinformatics (Spring 2012)

Transcriptional regulatory networks: inference and evolution

Sushmita Roy

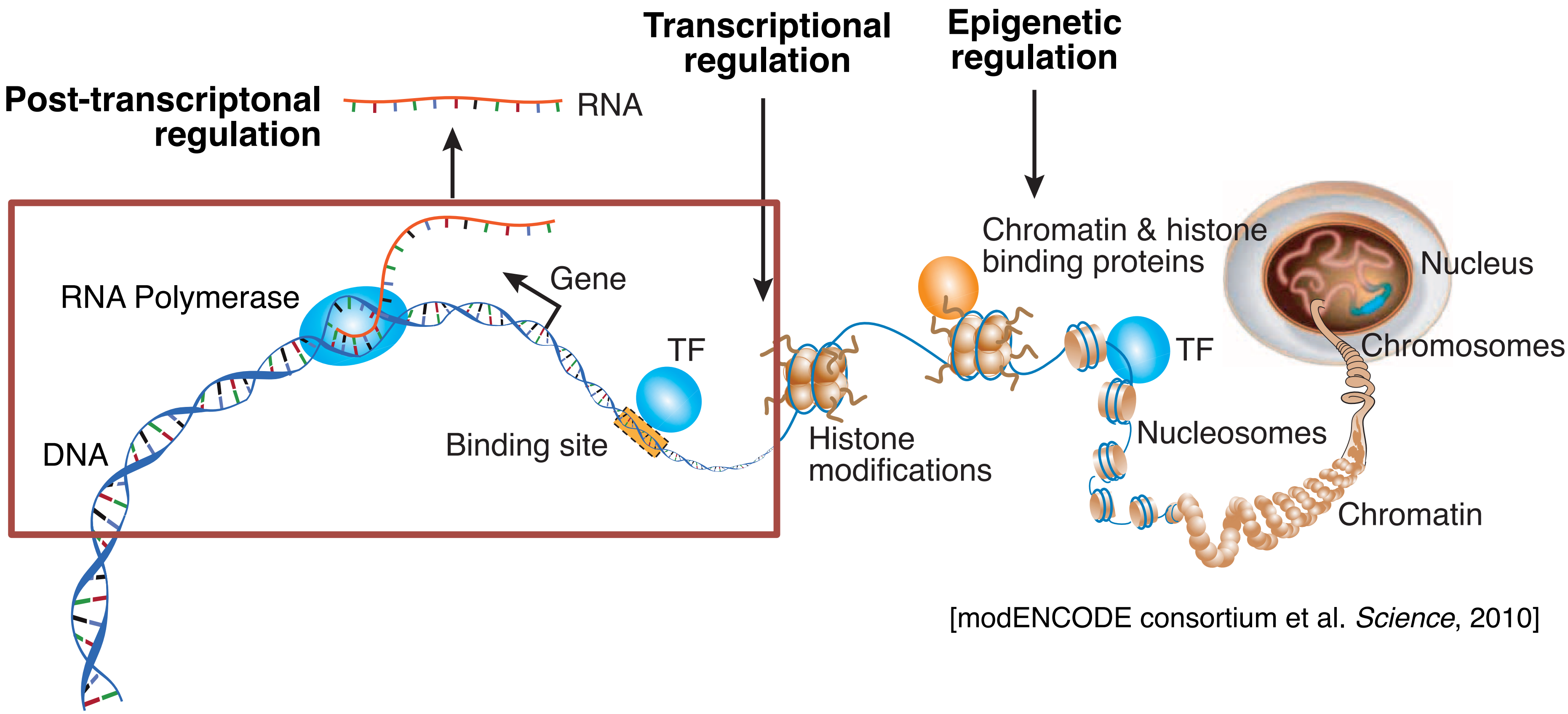
Goals for today

- Background
  - Components of the regulation machinery
  - Transcriptional gene regulation
- Challenges in regulatory networks
  - Element identification
  - Network identification
    - Extensions to inference
  - Network structure analysis
- Evolution of regulatory networks
  - Comparative functional genomics

Gene Regulation

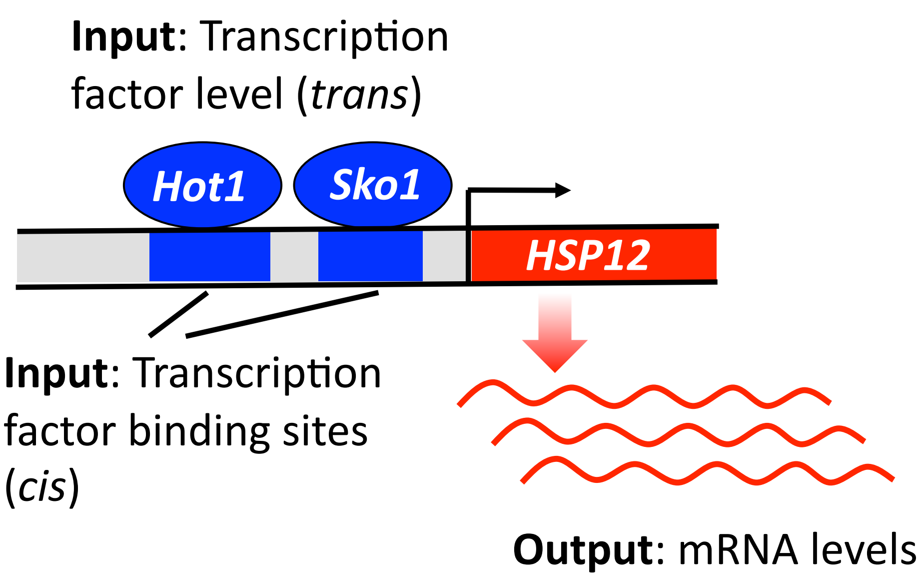
Collection of biological processes that determine what set of genes get expressed when and where.

What regulates gene expression?



[modENCODE consortium et al. *Science*, 2010]

Transcriptional gene regulation



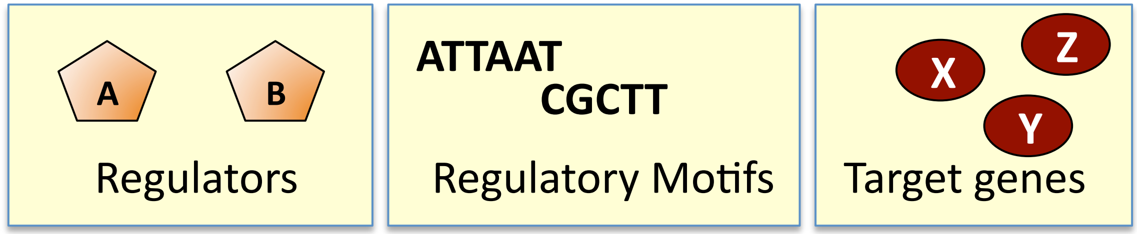
Transcriptional regulatory network connects TFs to target genes

Goals for today

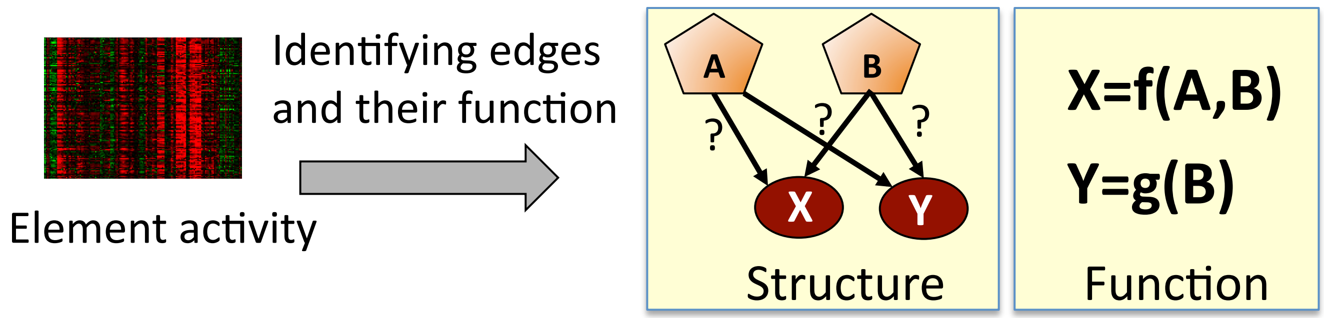
- Background
  - Components of the regulation machinery
  - Transcriptional gene regulation
- Challenges in regulatory networks
  - Element identification
  - Network identification
    - Extensions to inference
  - Network structure analysis
- Evolution of regulatory networks
  - Comparative functional genomics

Challenges in regulatory networks

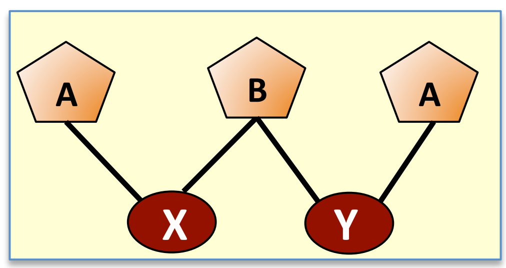
1. **Parts Identification**



2. **Network identification**



3. **Network Structure Analysis**



Hubs, degree-distributions, Network motifs

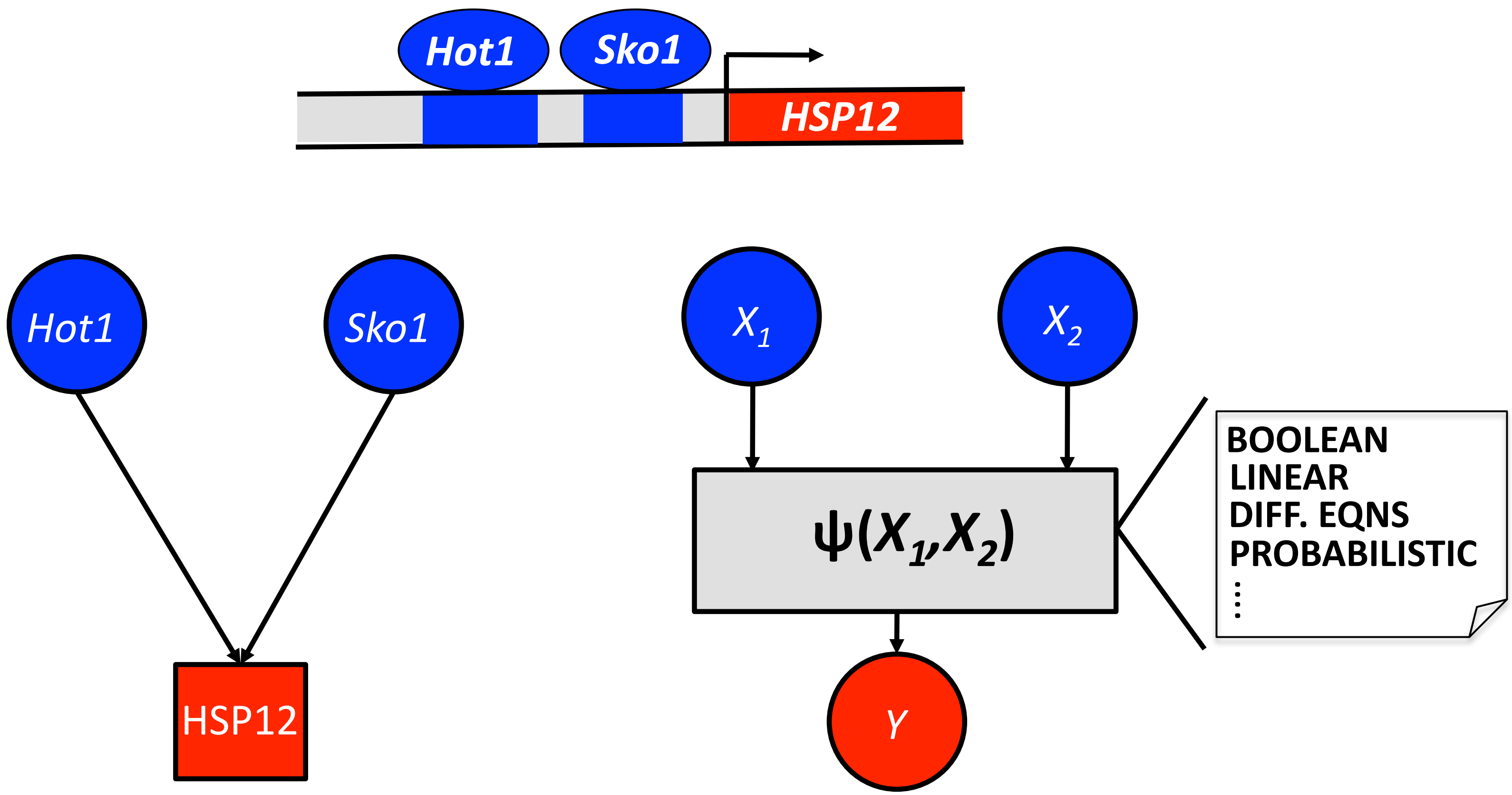
Element identification

- Elements
  - Regulators: Transcription factor proteins
  - Targets: Sequence-specific binding sites
- Computational approaches
  - Regulators: Sequence alignment
  - Motifs: *De novo* motif discovery
  - Targets: Sequence specific motif scanning

Goals for today

- Background
  - Components of the regulation machinery
  - Transcriptional gene regulation
- Challenges in regulatory networks
  - Element identification
  - Network identification
    - Extensions to inference
  - Network structure analysis
- Evolution of regulatory networks
  - Comparative functional genomics

Network identification



Structure

Who are the regulators?

Function

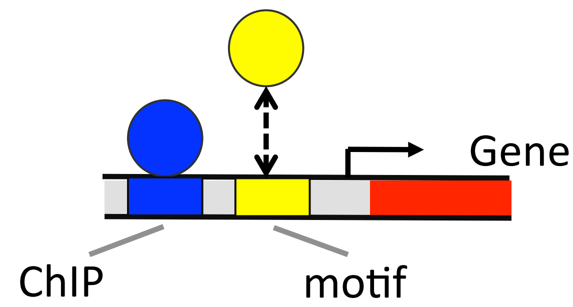
How they determine expression levels?

Approaches to Network identification

- Wet-lab approaches
  - ChIPseq/ChIP-chip
  - Genetic perturbations
- Computational approaches
  - What data to learn networks?
    - Motifs, ChIP binding assays, Expression
  - How to learn networks?
    - Supervised network inference
    - Unsupervised network inference
  - How to evaluate network usefulness?

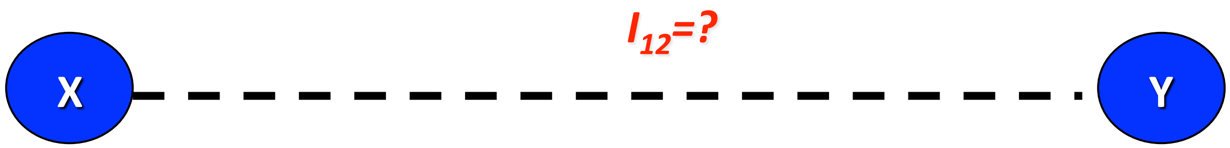
Types of data

- Physical
  - ChIP-chip and ChIP-seq
  - Sequence specific motifs
  - Measure static information



- Functional
  - Gene co-expression
  - Measure dynamic information

Supervised learning of TF-target interactions



Define:

$$I_{12} = \begin{cases} 1 & \text{if X and regulates Y} \\ 0 & \text{otherwise} \end{cases}$$

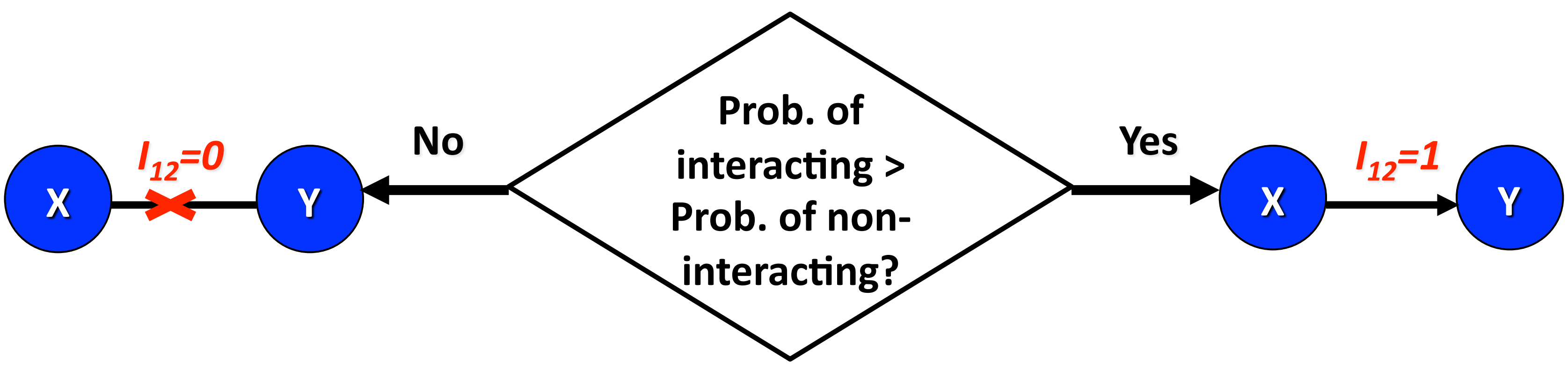
Given:

XY.features: Attributes of X and Y

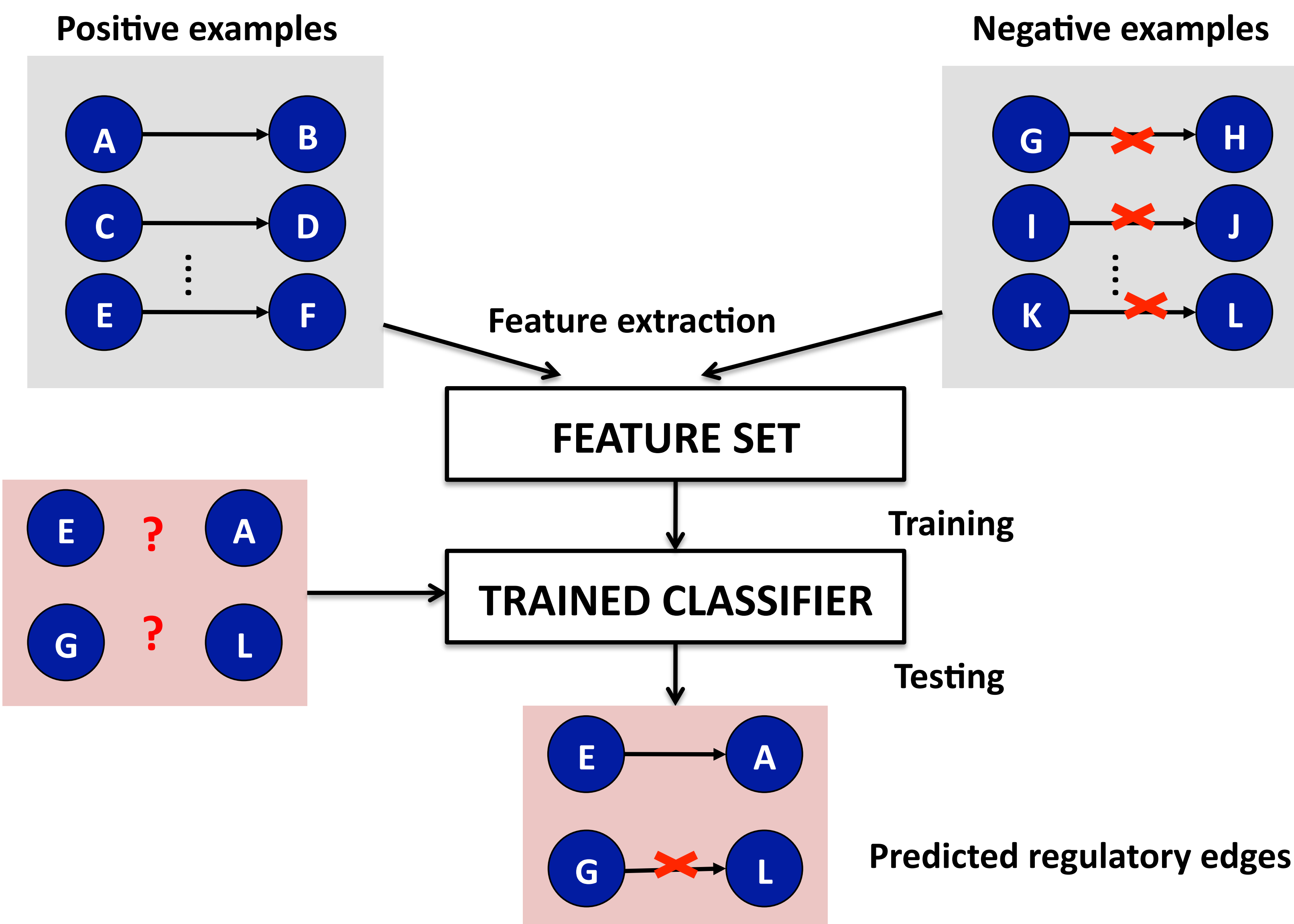
We need:

Prob. of regulating: P(I_{12}=1|*XY.features*)

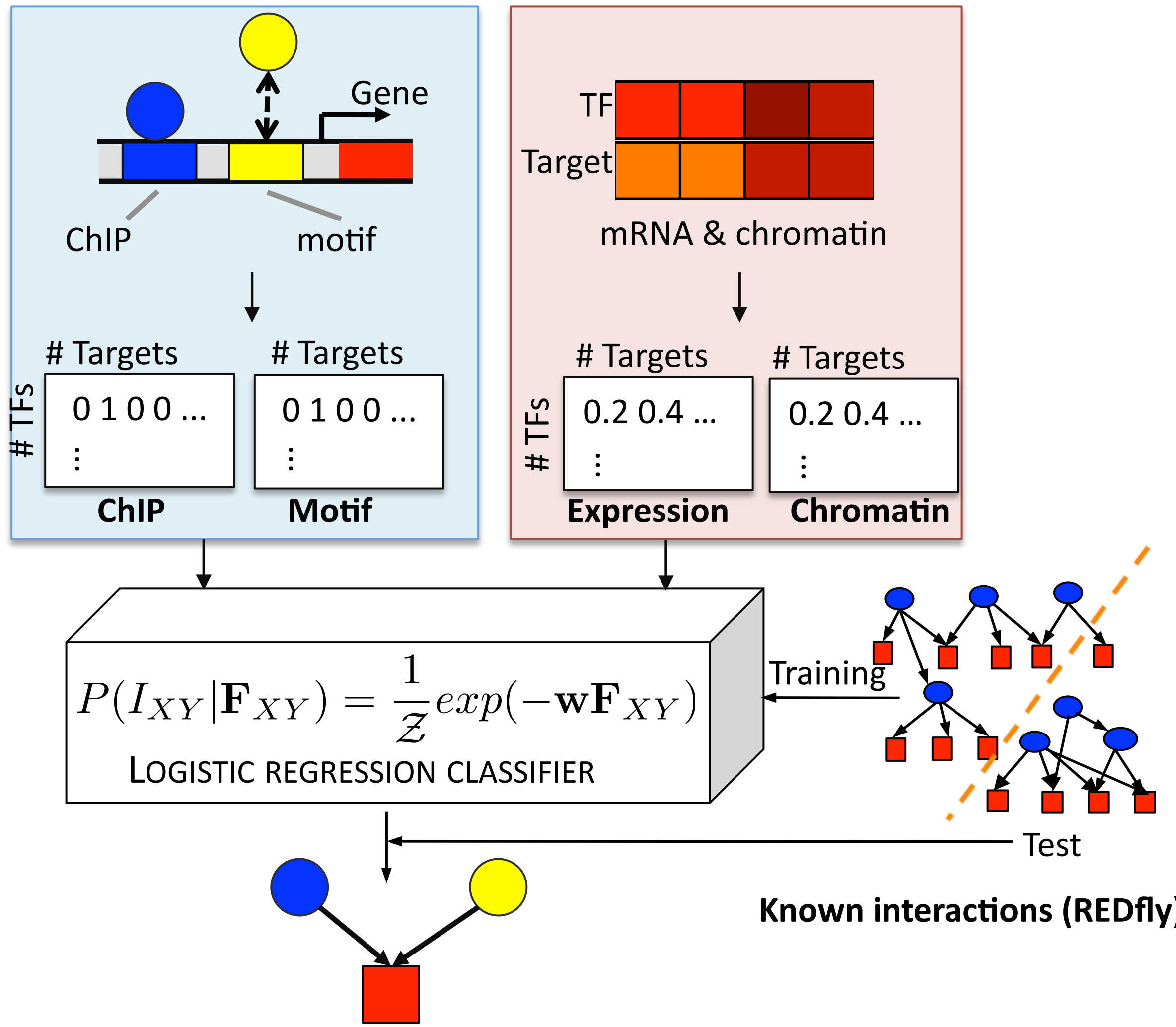
Prob. of not regulating: P(I_{12}=0|*XY.features*)



Supervised learning of TF-target interactions

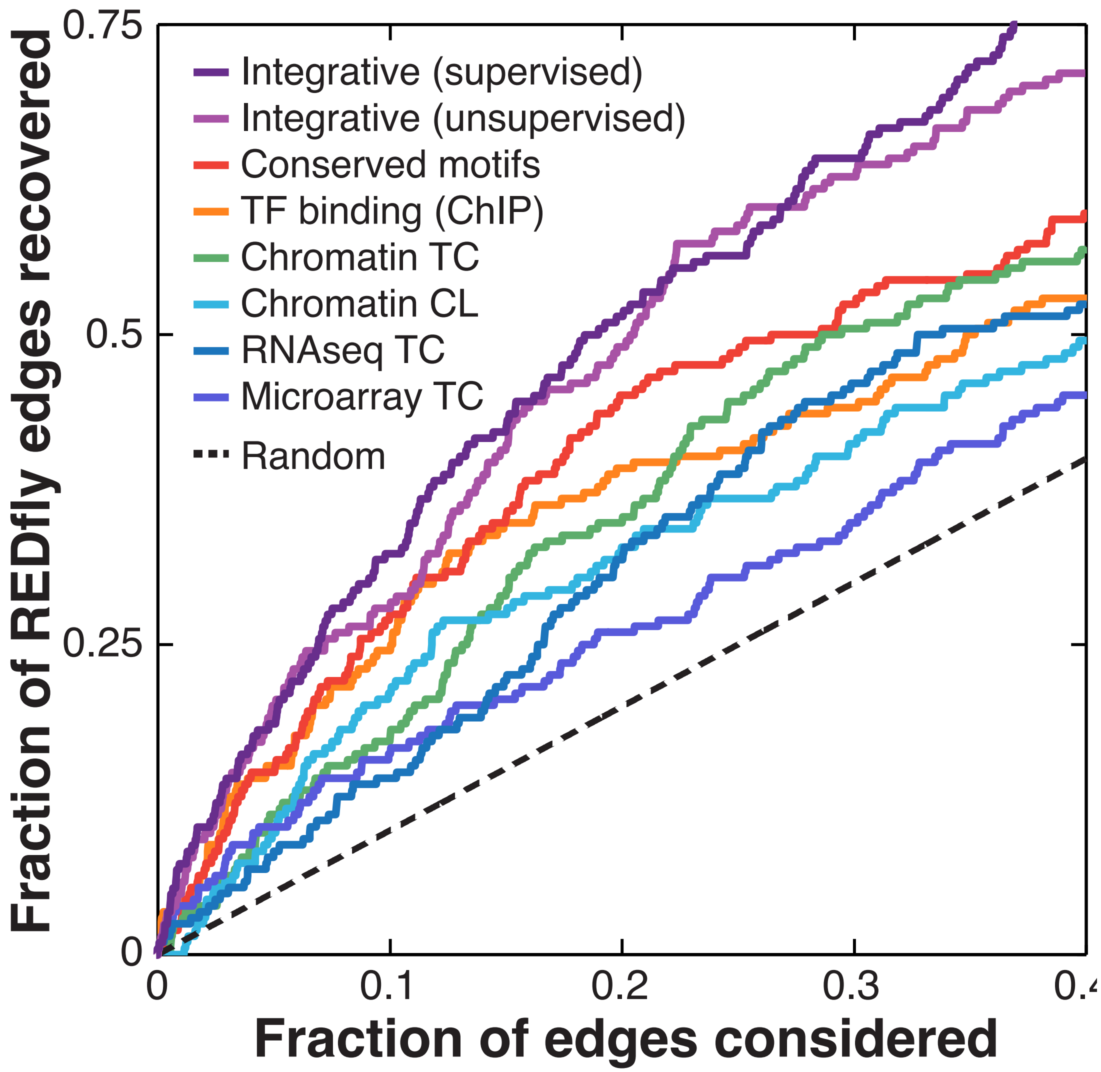


Inferring the regulatory network of the fly

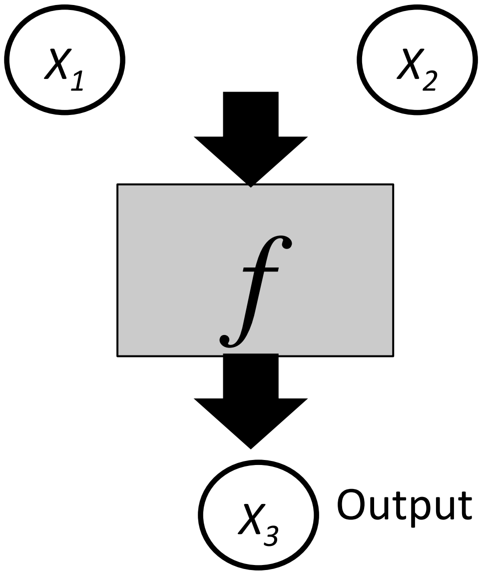


Marbach, Roy *et al.*, 2012

Supervised, integrative approach recovers more ground truth edges



Unsupervised network inference



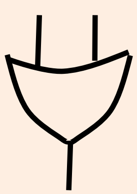
Input expression of neighbors

Output expression of node

Models differ in the function that maps input system state to output state

Boolean Networks

Input Output



X1	X2	X3
0	0	0
0	1	1
1	0	1
1	1	1

Differential equations

$$\frac{dX_3(t)}{dt} = \kappa \ g(X_1(t), X_2(t)) - rX_3(t)$$

Rate equations

Probabilistic graphical models

$$P(X_3|X_1, X_2) = N(X_1 a + X_2 b, \sigma)$$

Probability distributions

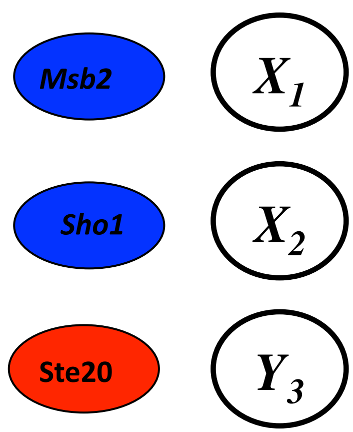
Probabilistic graphical models (PGMs)

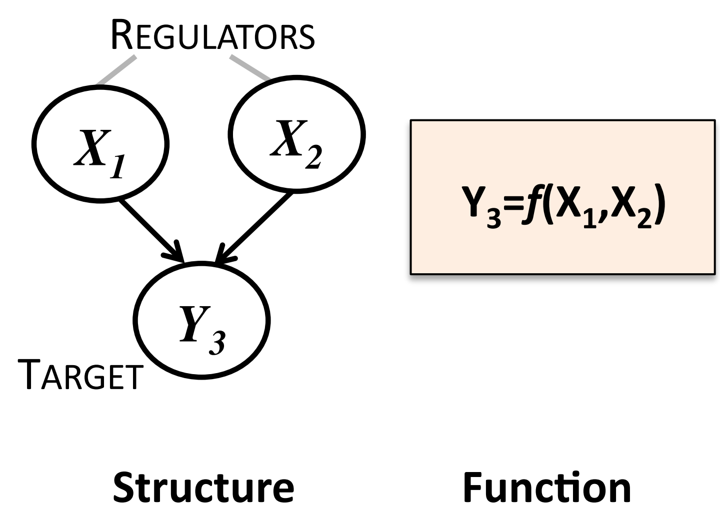
- A marriage between graph and probability theory
  - Handle noise and uncertainty
  - Nodes: Random variables
  - Edges: statistical dependency among random variables
- Model the joint probability distribution
  - Parameters: mathematical description of relations
- Enable incorporation of prior knowledge

Graphical models for unsupervised network inference

- Bayesian networks
- Dependency networks

Random variables encode expression levels





Goal: learn the structure and function of these networks

Some notation

- Random variables

$$\mathbf{X} = X_1, \cdots, X_N$$

- Joint assignment

$$\mathbf{x}_d = x_{1d} \cdots, x_{Nd}$$

- Dataset

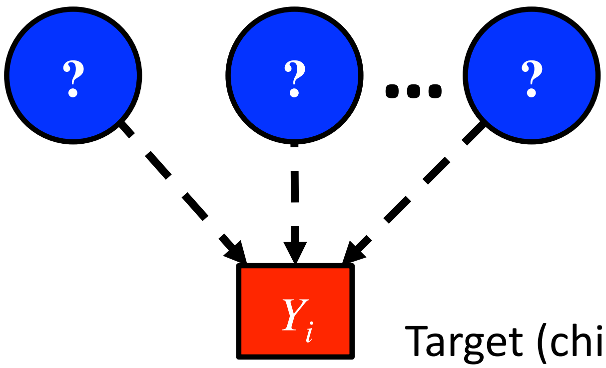
$$D = \{\mathbf{x}_1, \cdots, \mathbf{x}_d\}$$

- Joint probability distribution

$$P(\mathbf{X} = \mathbf{x}_d)$$

Bayesian networks: estimate a set of conditional probability distributions

Regulators (parents)



Target (child)

$$P(Y_i|\mathrm{Pa}(X_1, \cdots, X_p))$$

Function: Conditional probability distribution (CPD)

JPD: product of conditionals per variable

The learning problems

- Parameter learning on known structure
  - Estimate θ_i of the conditionals
- Structure learning
  - Find the statistical dependency structure
  - Subsumes parameter learning

Parameter learning

Maximum likelihood parameter estimation

$$\widehat{\theta} = \arg\max_{\theta} P(D|\theta, \mathcal{G})$$

Known graph structure

Data

Data likelihood

$$P(D|\theta, \mathcal{G}) = \prod_{d=1}^{|D|} P(\mathbf{X} = \mathbf{x}_d|\theta, \mathcal{G})$$

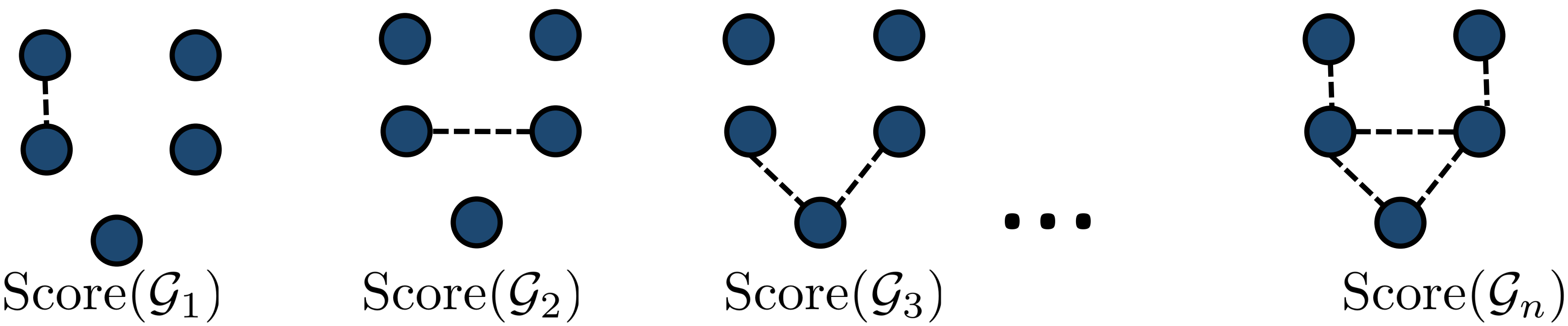
Structure learning

- Maximum likelihood framework

$$\widehat{\mathcal{G}} = \arg\max_{\mathcal{G}} \quad \max_{\theta} P(D|\theta, \mathcal{G})$$

Structure learning using score-based search

$$\text{Score}(\mathcal{G}) = P(D|\mathcal{G}, \theta)$$



$$\widehat{\mathcal{G}} = \arg\max_{\mathcal{G}} \underbrace{\max_{\theta} P(\mathbf{X}|\theta, \mathcal{G})}_{\text{Maximum likelihood}}$$

Best graph

Maximum likelihood

Learning network structure is computationally expensive

- For N variables there are $2^{\binom{N}{2}}$ possible networks:
- Set of possible networks grows super exponentially

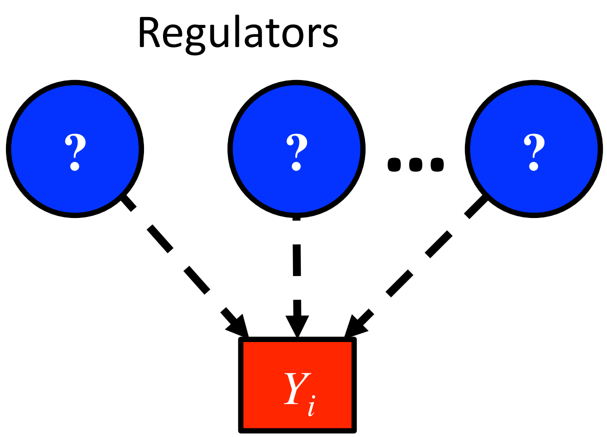
N	Number of networks
3	8
4	64
5	1024
6	32768

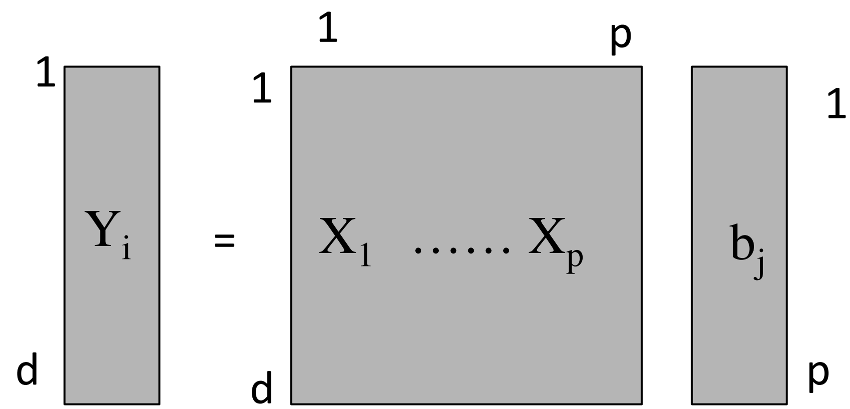
Need approximate methods to search the space of networks

Approximation strategies

- Search the parent set independently
- Restrict the size of the parent set
- Assume linear relationships

Dependency networks: a set of regression problems





Function: Linear regression

$$\mathbf{b}_i^* = \arg\min_{b_i} ||\mathbf{Y}_i - \mathbf{X}_i * \mathbf{b}_i|| + f(\mathbf{b}_i, \lambda)$$

Regularization term

$$1 \leq i \leq m$$

Number of genes

Regularized linear regression

- Lasso: sparsity

$$b_i^* = \arg\min_{b_i} ||\mathbf{Y}_i - \mathbf{X}_i * \mathbf{b}_i|| + \lambda|\mathbf{b}_i|$$

- Ridge regression: smoothness

$$b_i^* = \arg\min_{b_i} ||\mathbf{Y}_i - \mathbf{X}_i * \mathbf{b}_i|| + \lambda||\mathbf{b}_i||$$

- Elastic net: sparsity + smoothness

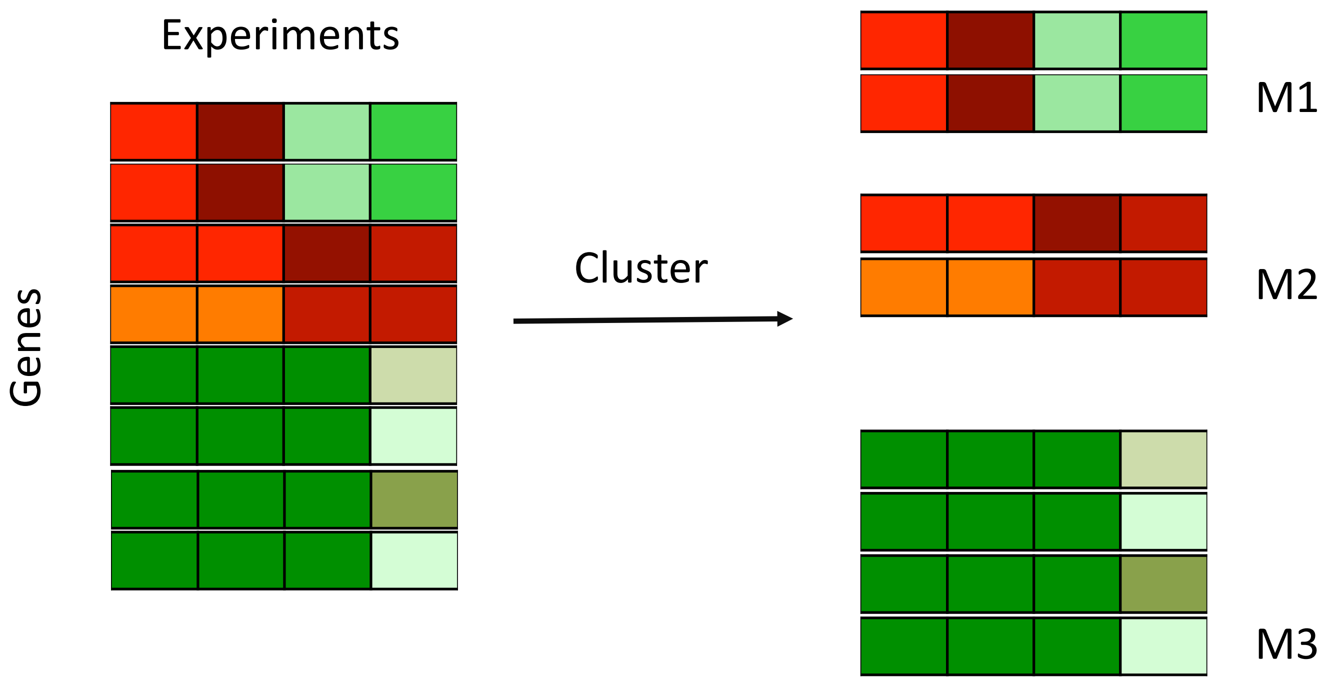
Goals for today

- Background
  - Components of the regulation machinery
  - Transcriptional gene regulation
- Challenges in regulatory networks
  - Element identification
  - Network identification
    - Extensions to inference
  - Network structure analysis
- Evolution of regulatory networks
  - Comparative functional genomics

Extensions to vanilla network inference approaches

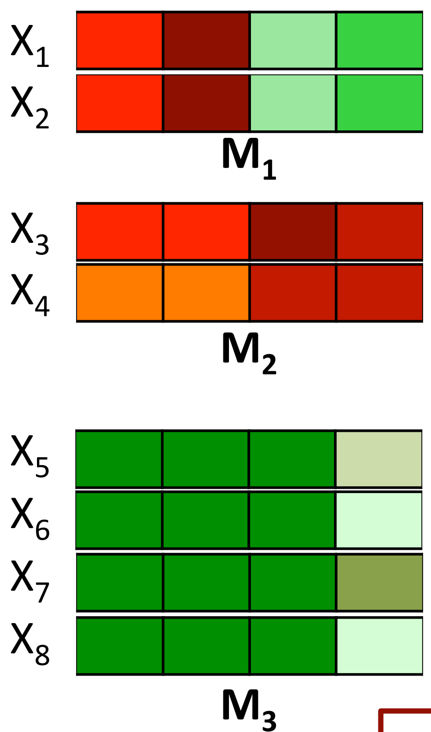
- Making methods more scalable
- Imposing biological constraints
- Integrating other types of data

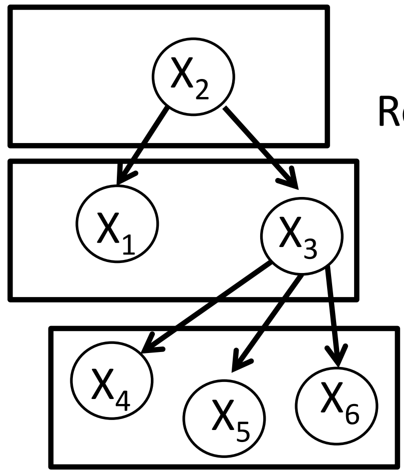
Concept: Expression modules



$$\min_C \ \sum_{k=1}^{|C|} \sum_{i,j} d(X_i, X_j)$$

Learning regulatory programs of modules instead of genes





Regulatory program

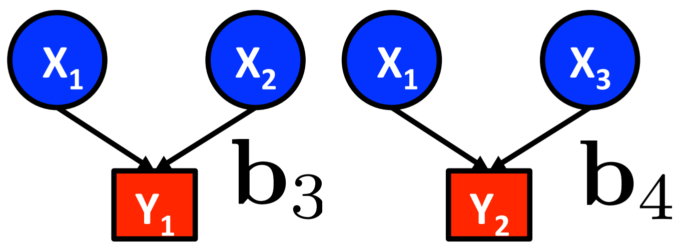
- Bayesian Network formalism
- No cyclic dependencies
- Target genes share CPDs
- Modules are re-visited

Module networks

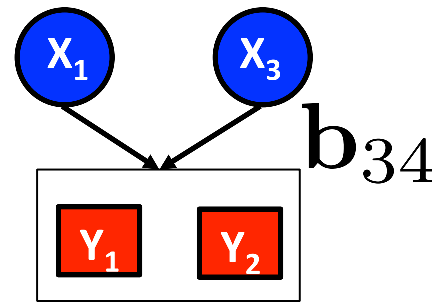
But every gene has the same set of parameters

Lee *et al* 2009, Segal *et al* 03

Combine per-gene and per-module network inference methods



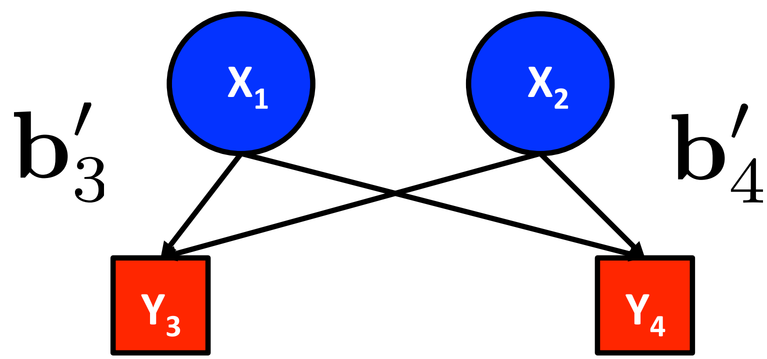
Per gene



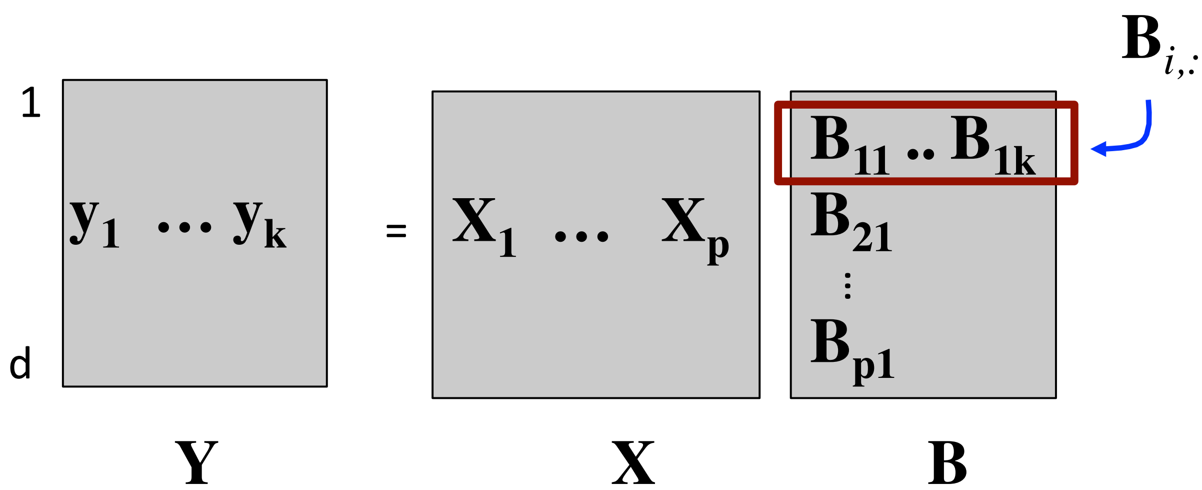
Per module

How to impose module constraints?

Keep regulators same but params different



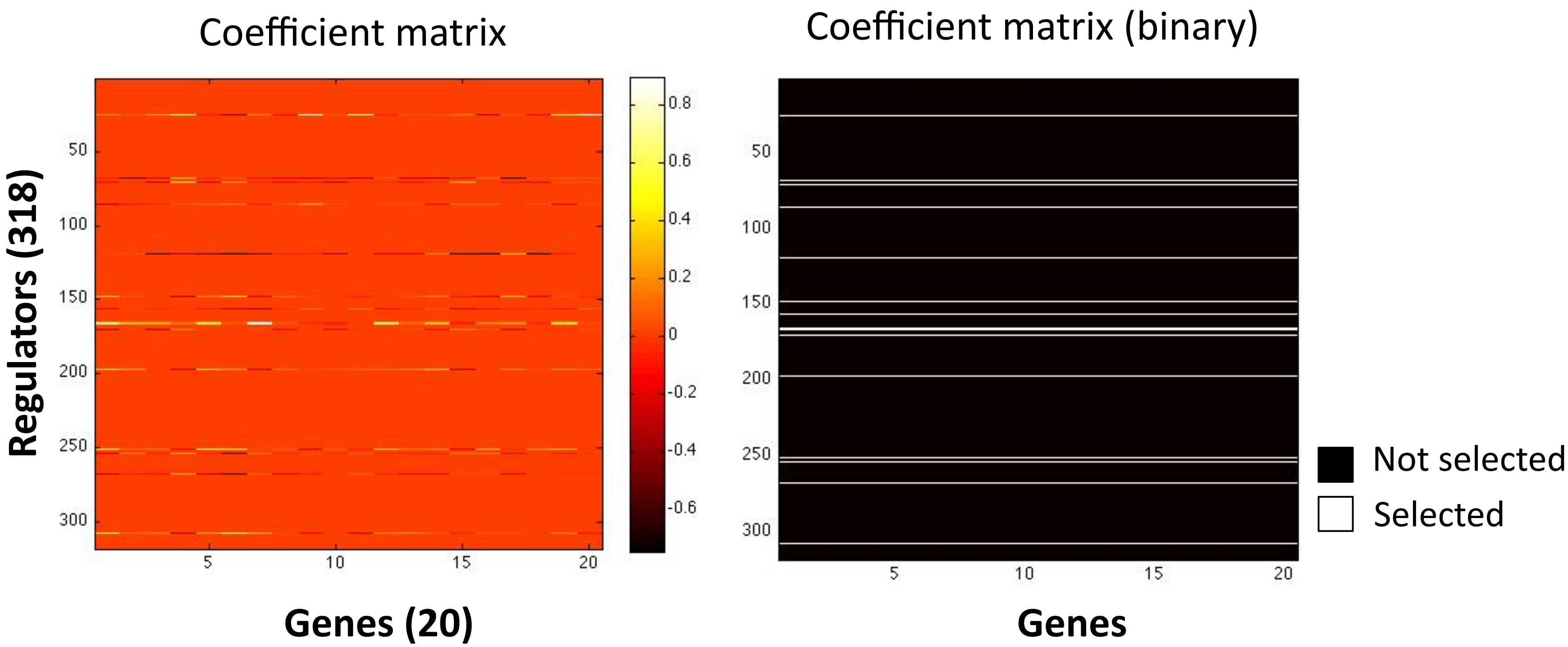
Group lasso for module constrained per gene model



$$\mathbf{B}^* = \arg\min_{B} ||\mathbf{Y} - \mathbf{XB}||_2^2 + \lambda \Sigma_{i=1}^{p} ||\mathbf{B}_{i,:}||_2$$

With Nikhil Rao and Rob Nowak

Example coefficient matrix



A regulator is selected for all or no genes

Integrating data as structure priors

Revisiting Structure learning

- Bayesian framework
- $\mathcal{G}$ is an unknown random variable
- Optimize posterior distribution of graph given data

Graph prior

$$P(\mathcal{G}|D) = P(D|\mathcal{G})P(\mathcal{G})$$

$$P(\mathcal{G}|D) \propto P(\mathcal{G}) \int P(D, \theta|\mathcal{G}) d\theta$$

$$P(\mathcal{G}|D) = P(D|\mathcal{G}, \theta_{MAP})P(\mathcal{G})$$

Maximum *a posteriori* estimate

A structure prior to integrate data

- Let $P(G)$ distributes independently over edges

$$P(\mathcal{G}) = \left[\prod_{X_i \to X_j} P(X_i \to X_j)\right] \left[\prod_{X_i \not\to X_j} (1 - P(X_i \to X_j))\right]$$

Present edges

Absent edges

- Define prior probability of edge presence/absence

$$P(X_i \to X_j) = \frac{1}{1 + exp(-(\beta_1 + \beta_2 w_{ij}))}$$

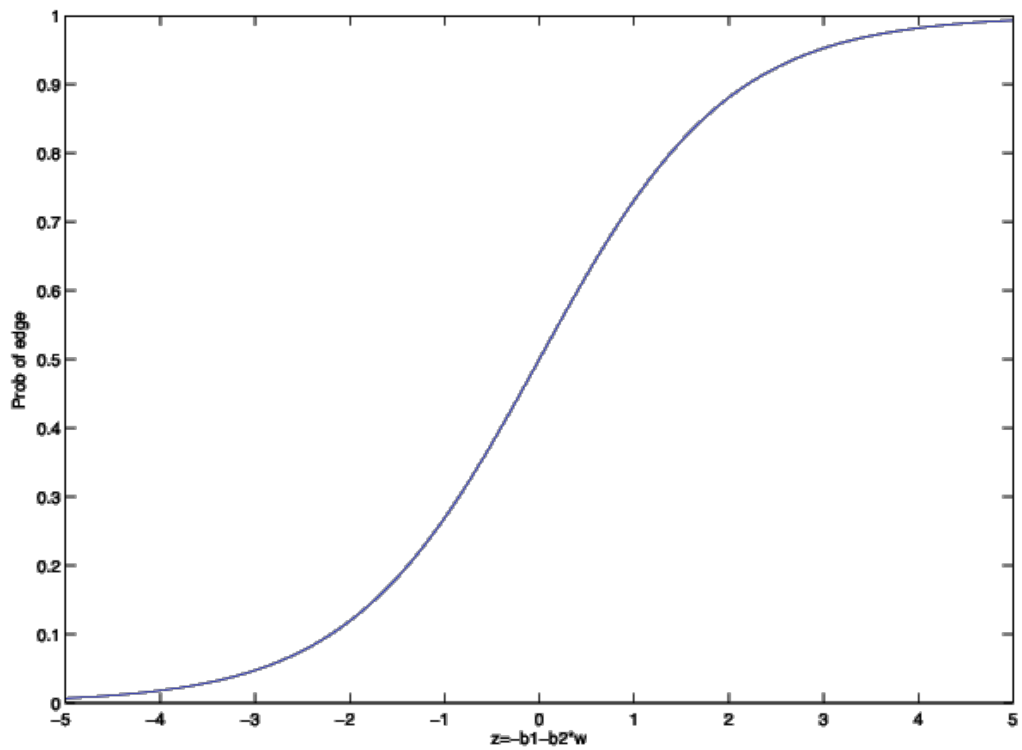
Graph structure complexity

Prior strength

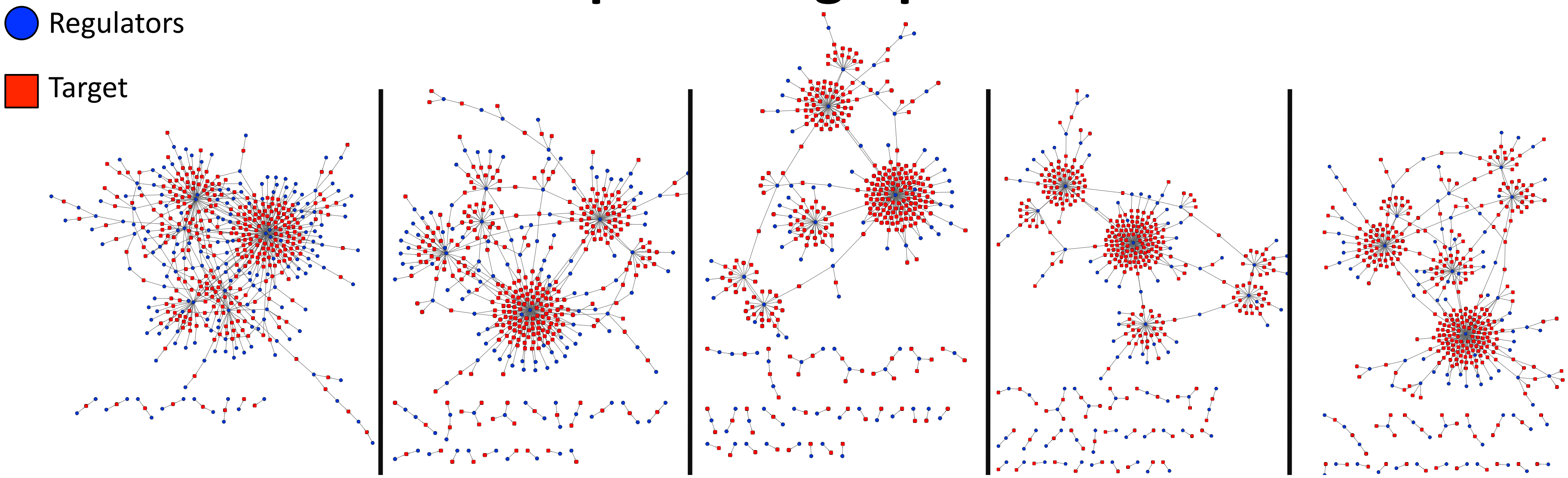
Edge prior strength

Behavior of graph structure prior

$$P(X_i \rightarrow X_j) = \frac{1}{1 + exp(-(\beta_1 + \beta_2 w_{ij}))}$$



Effect of prior on graph structure



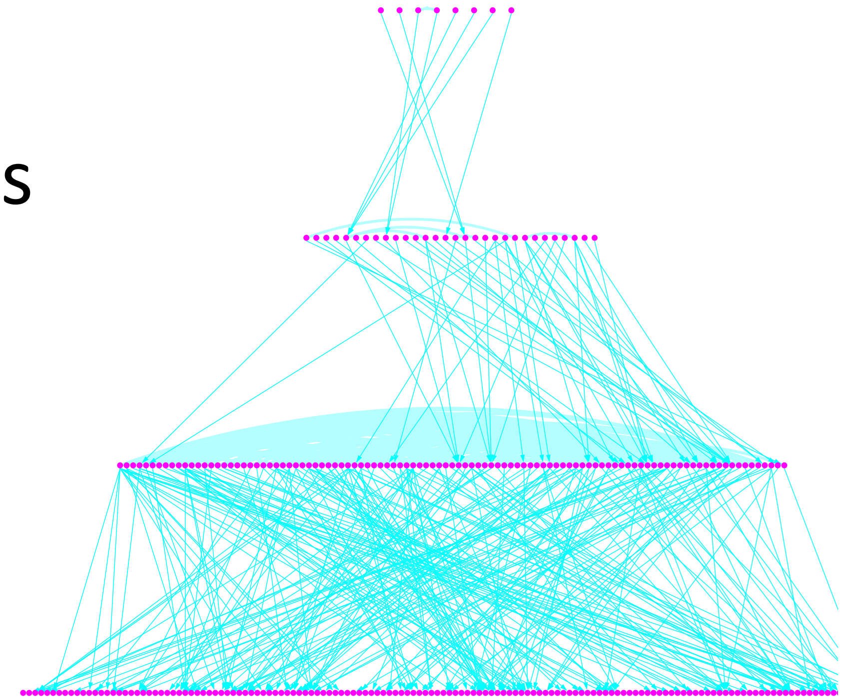
β_1	0.5	-2	-4	-4	-4
β_2	0.4	0.4	0.4	2	4
TFs	199	141	92	96	108
Known Edges	3%	3.2%	3.25%	6.4%	18.5%
Score	-6890	-8319.53	-9216	-9187	-9055

Goals for today

- Background
  - Components of the regulation machinery
  - Transcriptional gene regulation
- Challenges in regulatory networks
  - Element identification
  - Network identification
    - Extensions to inference
  - Network structure analysis
- Evolution of regulatory networks
  - Comparative functional genomics

Hierarchical nature

- Regulators are hierarchically organized with different roles per level
  - Top: Master regulators influence many genes
  - Middle: Bottle necks directly targeting most genes
  - Bottom: Essential regulators

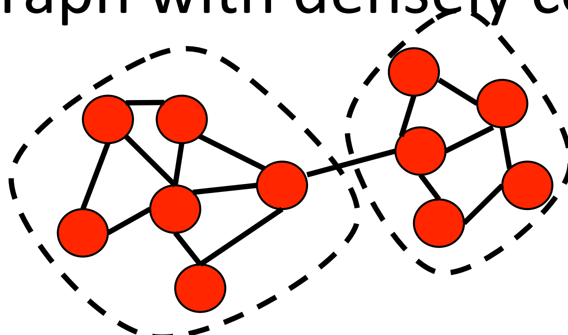


Hierarchical structure of *S. cerevisiae* regulatory network

Yu & Gerstein 2006, Jothi *et al.* 2009

Modularity of regulatory networks

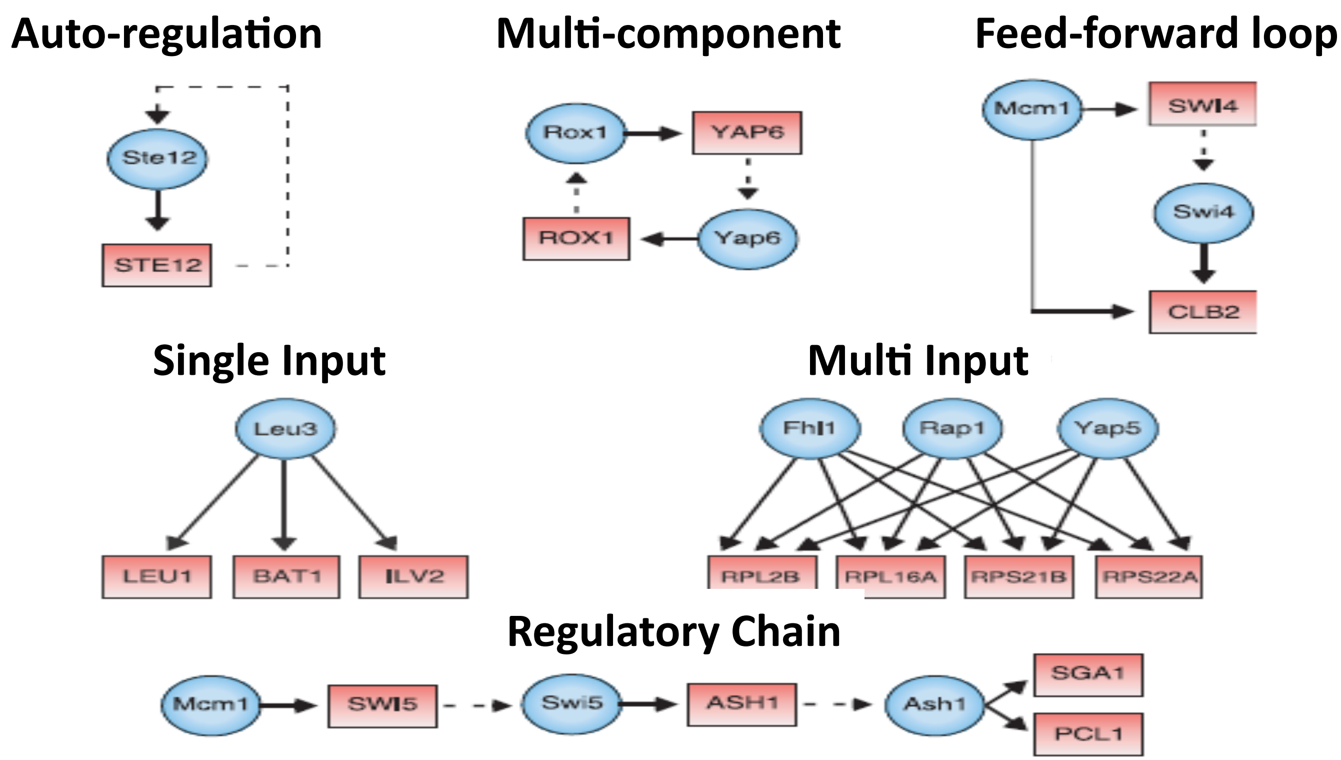
- Modular: Graph with densely connected subgraphs



- Genes in modules involved in similar functions and co-regulated
- Modules can be identified using graph partitioning algorithms
  - Markov Clustering Algorithm
  - Girvan-Newman Algorithm

Newman 2002, van Dongen 2008

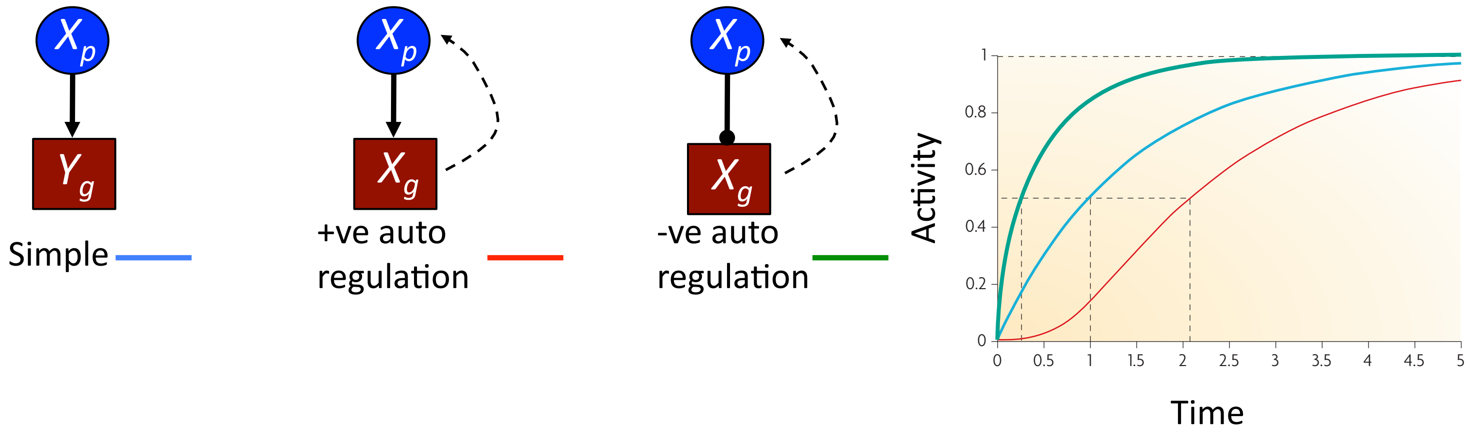
Structural network motifs

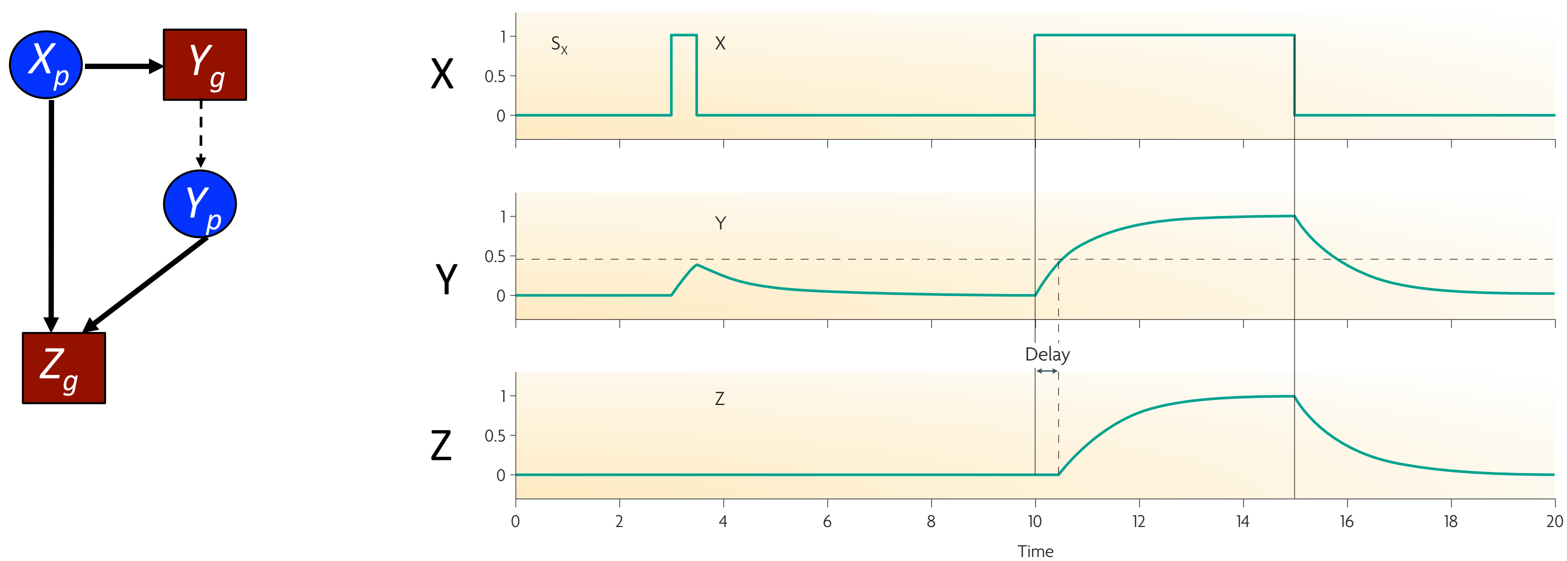


Feed-forward loops involved in speeding up in response of target gene

Lee *et.al. 2002,* Mangan & Alon, 2003

Network motifs often have specific functions





Uri Alon, Nature Review Genetics 2007

Goals for today

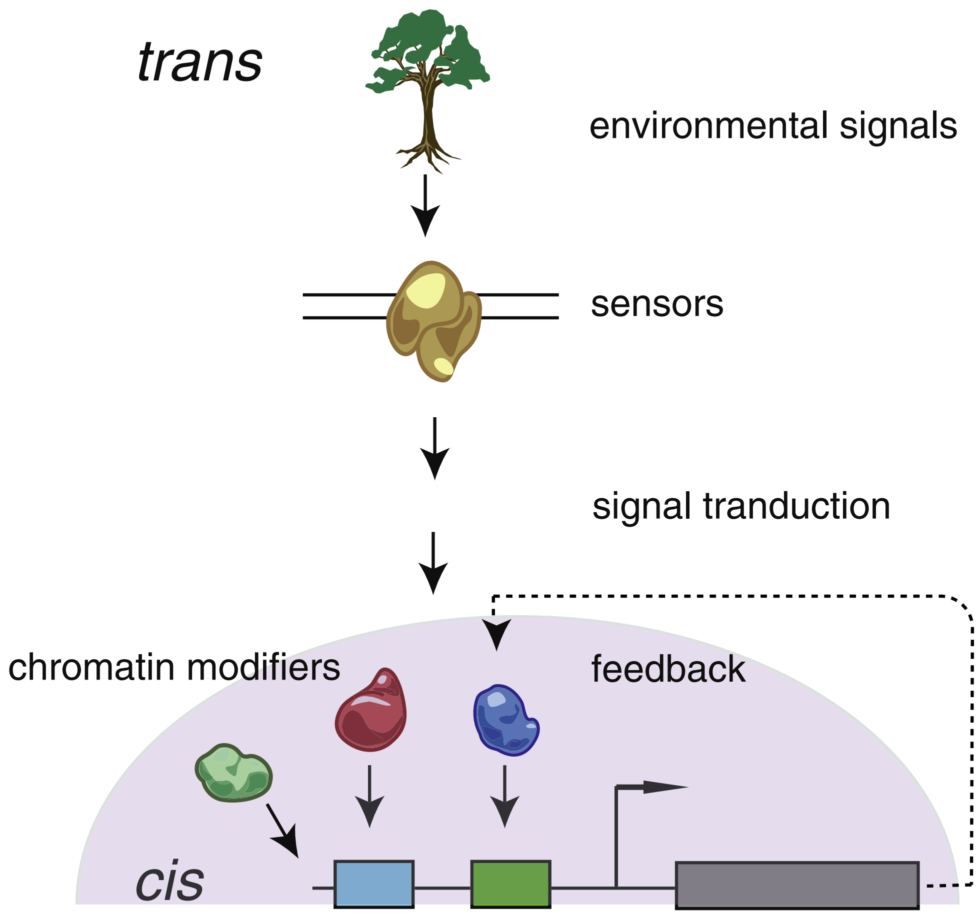
- Background
  - Transcriptional gene regulation
  - *cis* and *trans* elements
- Challenges in regulatory networks
  - Element identification
  - Network identification
    - Extensions to inference
  - Structural properties of networks
- Evolution of regulatory networks
  - Comparative functional genomics

Why understand evolution of regulatory networks

Importance in evolution of complex body plan:

" Although a variety of ways of thinking about evolution have been proposed, the evolution of the body plan is fundamentally a system-level problem to which GRN structure/function provides the most compelling direct access" Peter & Davidson, 2011

Factors affecting regulatory network evolution



***trans*-factors**

1. Transcription factors
2. Chromatin modelers
3. Signaling proteins
4. Environment

***cis*-factors**

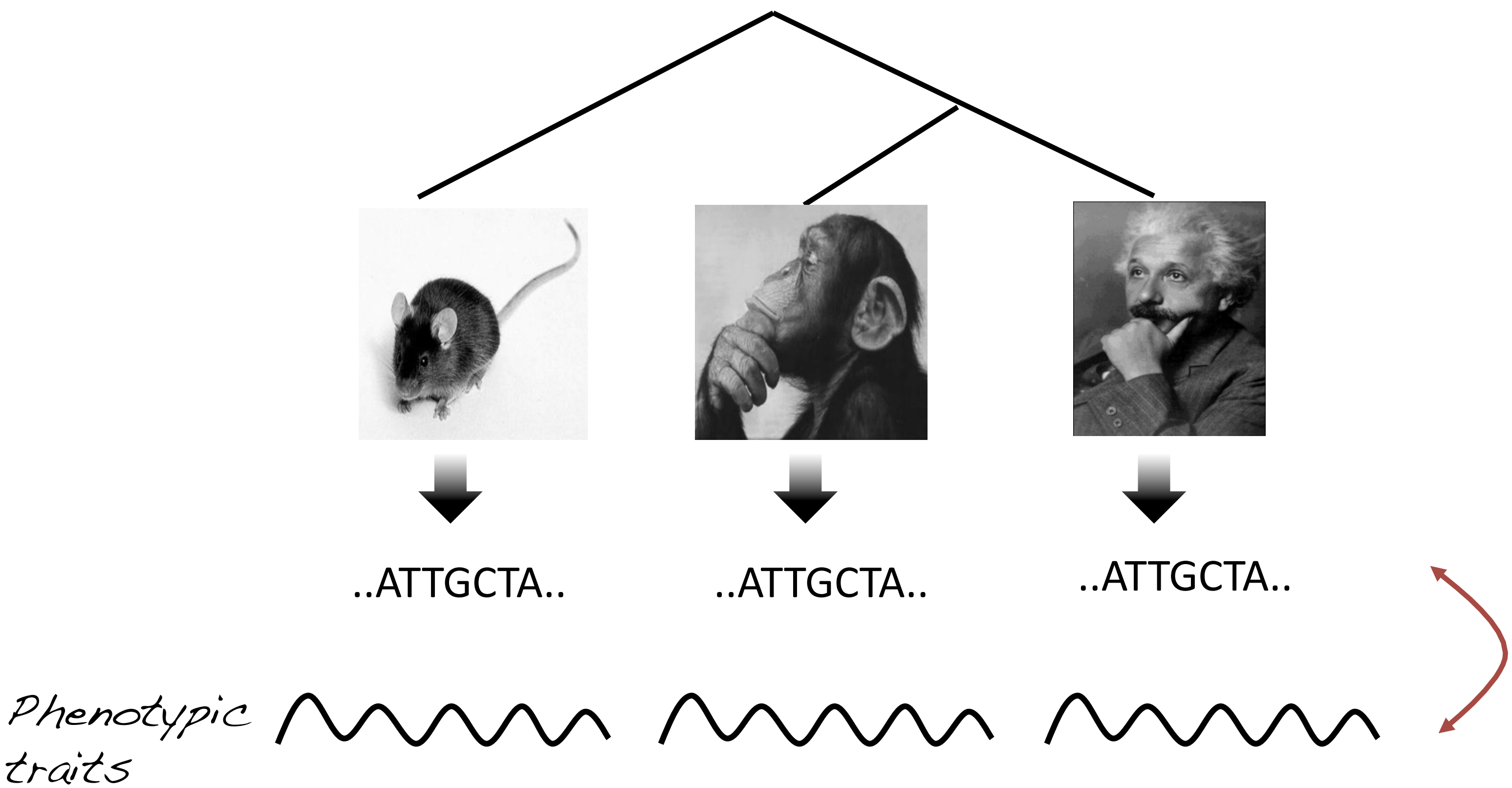
1. Binding sites
2. Nucleosomes
3. Histone marks

Thompson and Regev, FEBS letters, 2009.

Key questions

- How conserved are regulatory networks?
  - Elements
  - Connections
- How are different conservation/divergence scenarios implemented?
- What is the ancestral state?
- Do regulatory differences explain functional innovation?

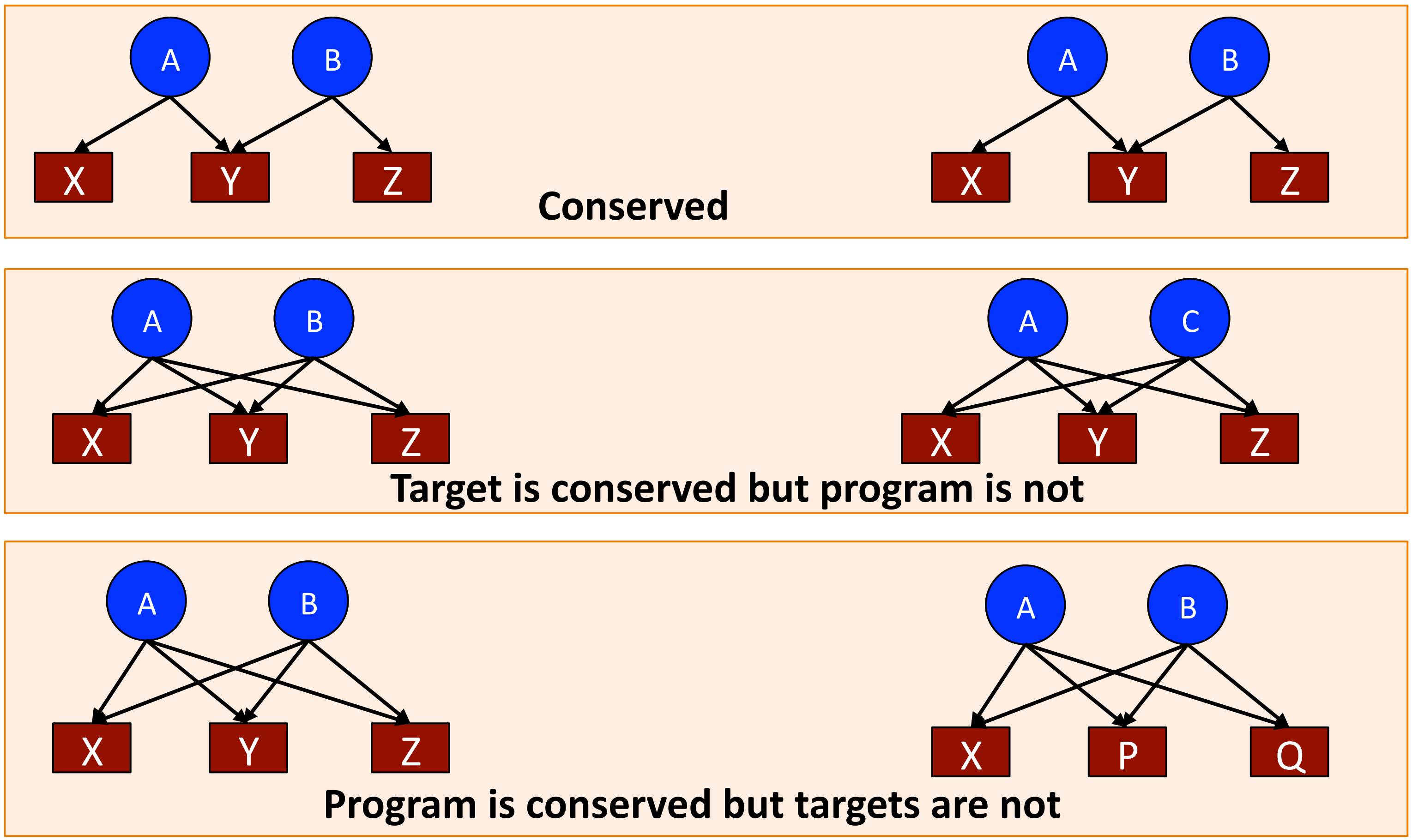
Comparative genomics approaches to understanding regulation evolution



Phylogenetic relationships to compare sequence differences, and relate to phenotypic traits.

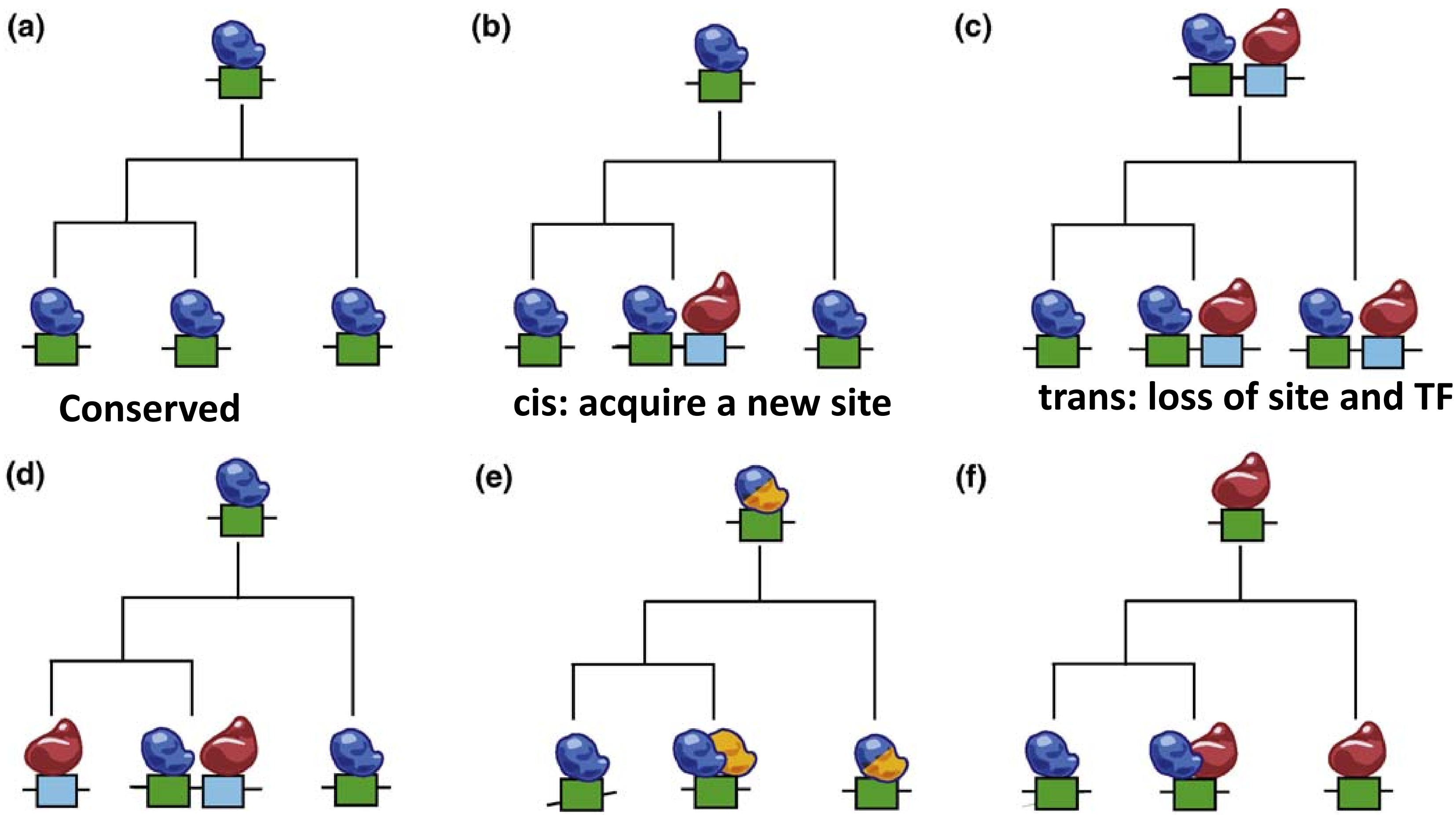
Scenarios of conservation & divergence

Network of Organism 1 Network of Organism 2



Hao Li and Johnson, 2010

How do regulatory networks rewire?



Cis: combinatorics & divergence

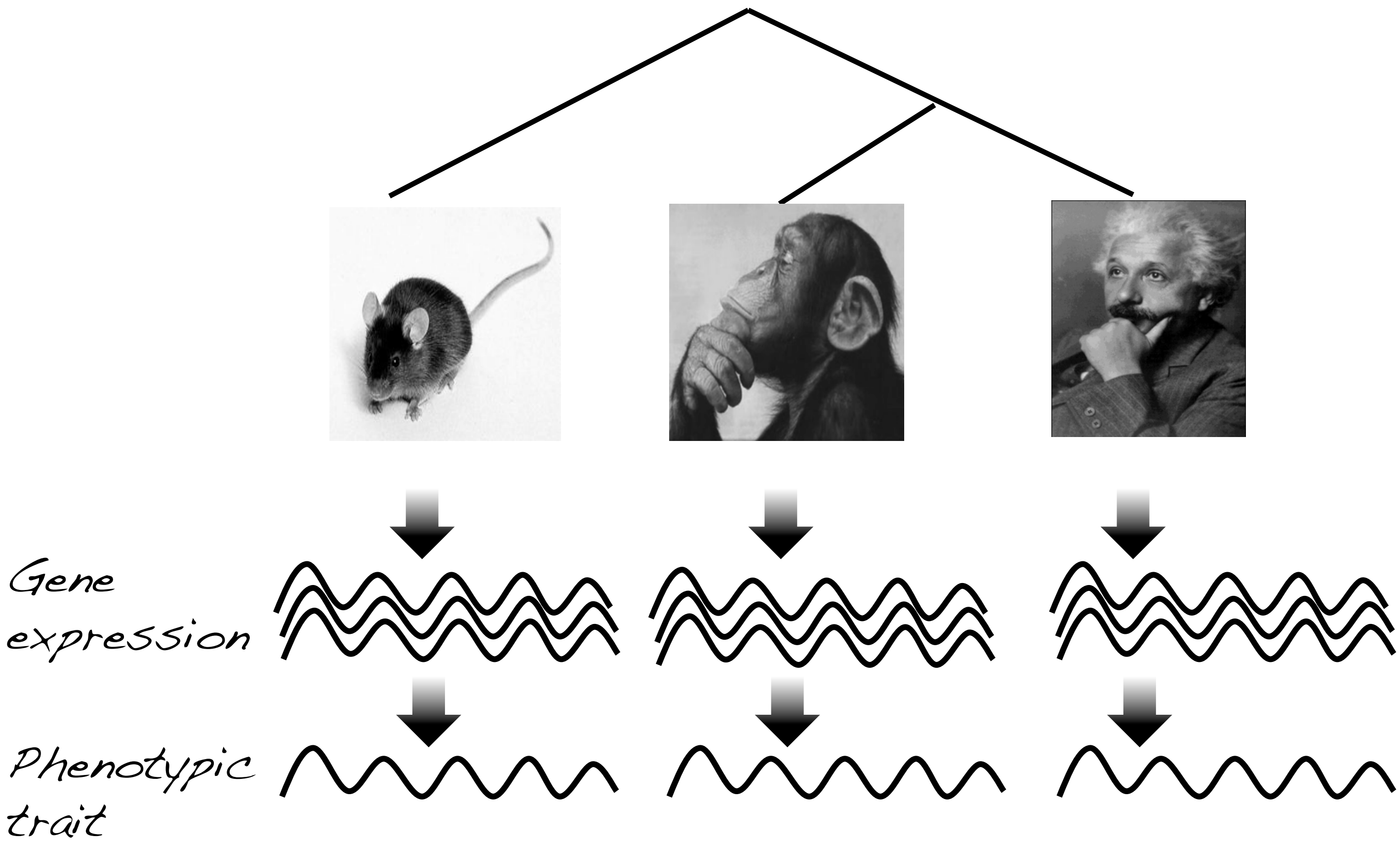
Trans: Duplication & divergence

Trans: Independent protein DNA binding evolution

Wolbach, Thompson *et al.,* Current Opinions in Genetics & Dev. 2009

But, we know only a handful of examples from the pre-mRNA era.

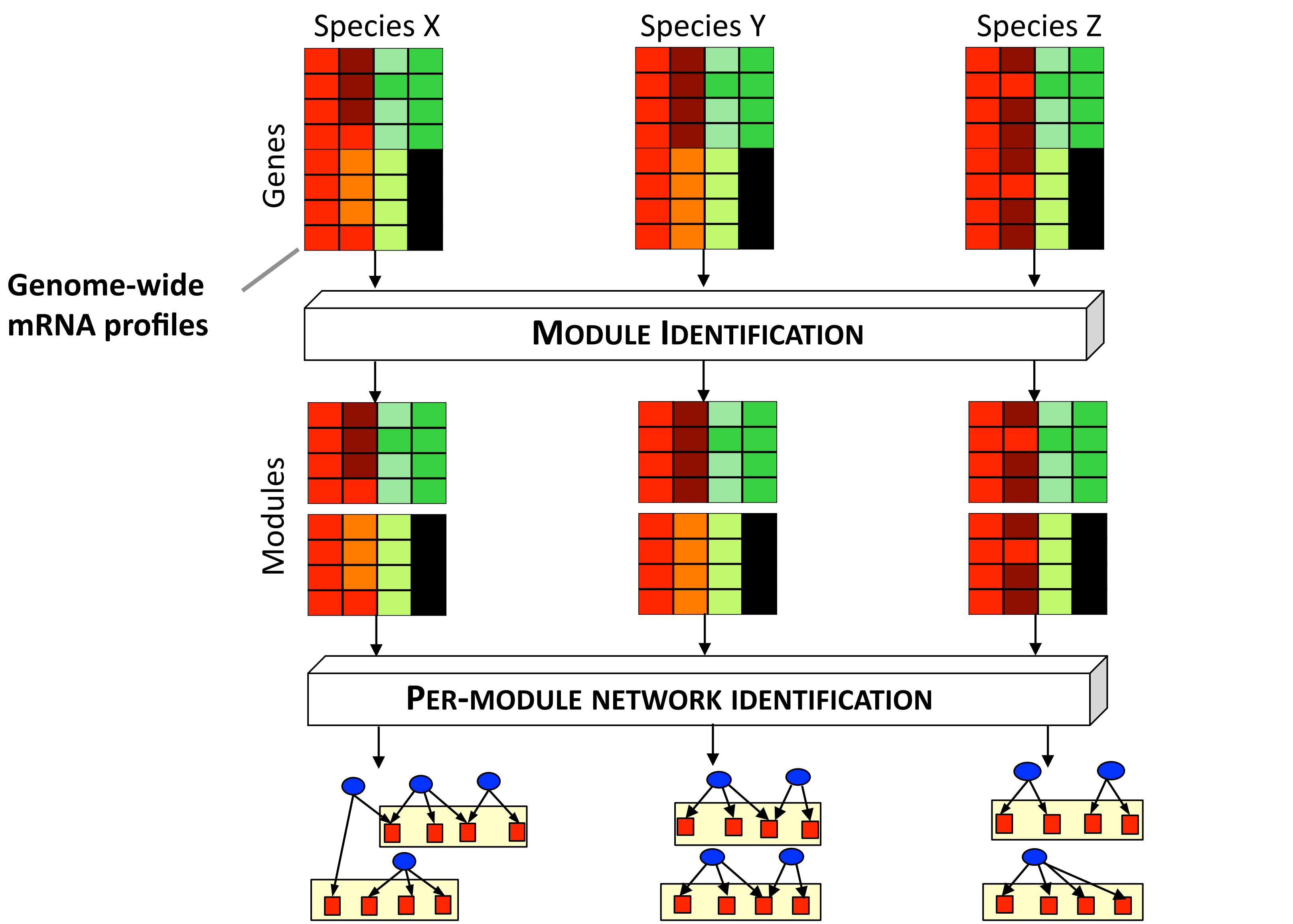
Comparative functional genomics approaches to understanding regulation evolution



Systematic approaches to compare regulatory networks

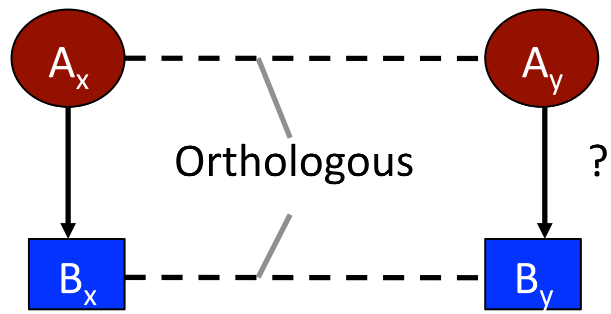
- One species at a time
  - Infer a regulatory network per species
  - Compare networks across species
- Learn multiple networks simultaneously
  - Use phylogenetic relationships to constrain the network structure

Learning networks one species at a time

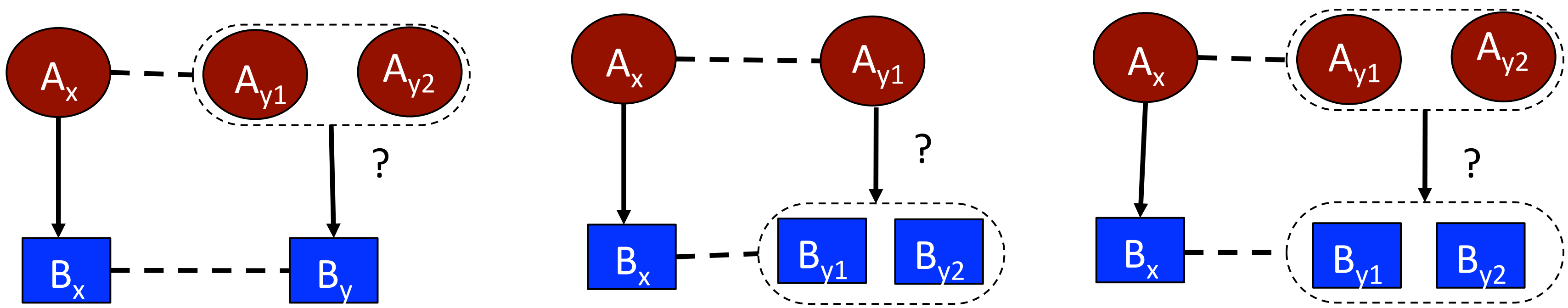


Comparing networks across species

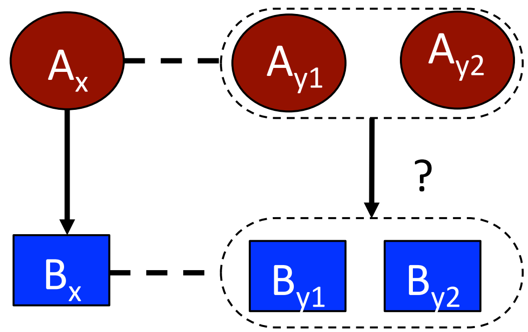
Easy case: One to one orthologs:



Not so easy cases: One to many orthologs:



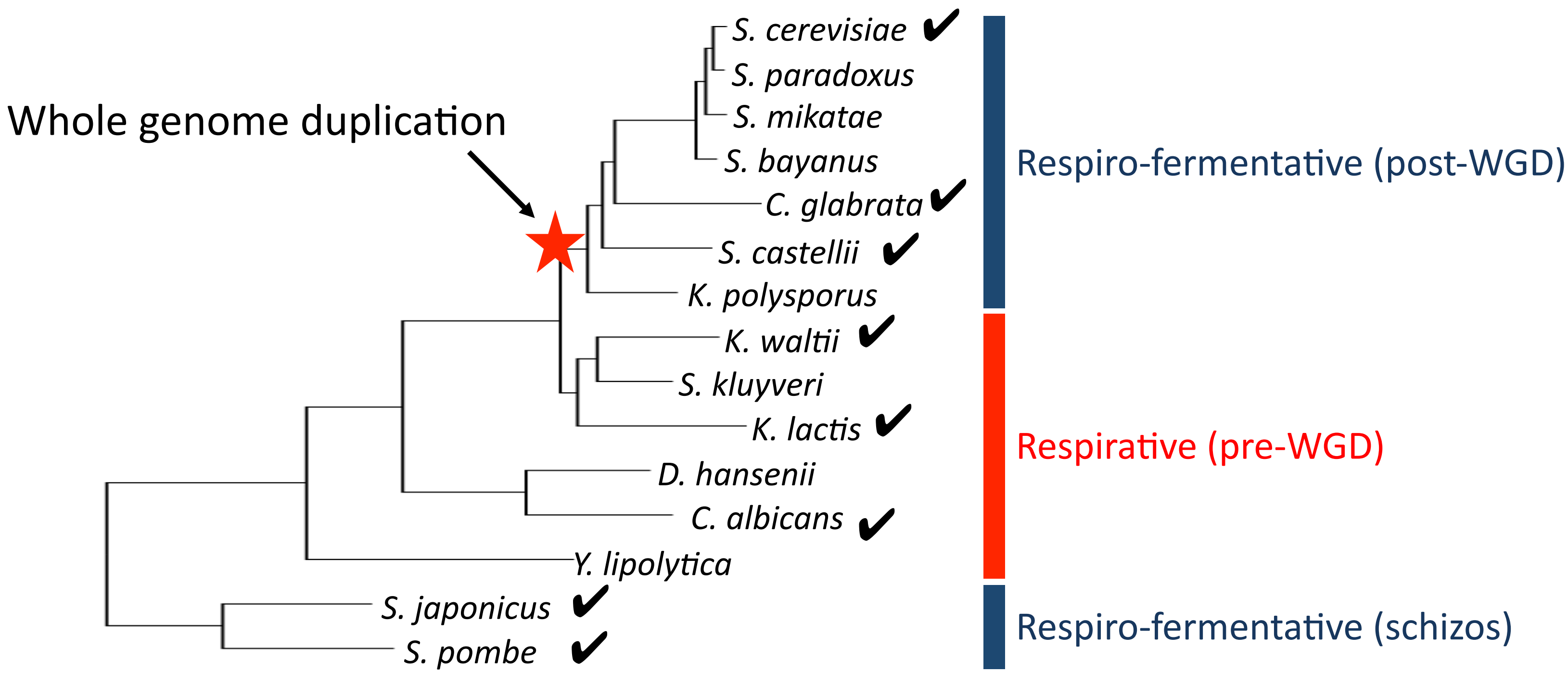
Defining an edge match



$$E_Y^{AB} : \{(i,j) \in \{A_{y1}, A_{y2}\} \times \{B_{y1}, B_{y2}\}\}$$

$$A_X \rightarrow B_X \text{ is conserved in } Y \text{ if } E_Y^{AB} \neq \emptyset$$

Using yeast Ascomycetes to understand regulatory evolution



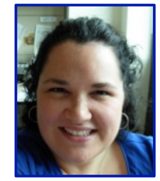
Respiro-fermentative: use fermentation (ethanol production) when grown on glucose

Respirative: use respiration when grown on glucose

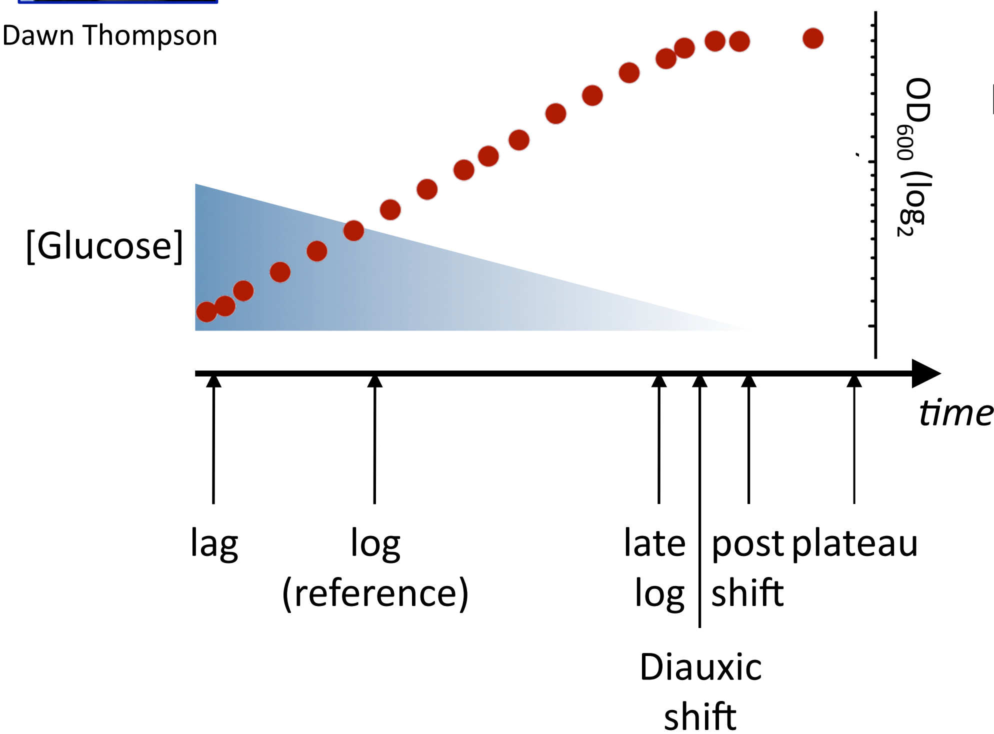
300 million years of evolution

Experiments for capturing functional response

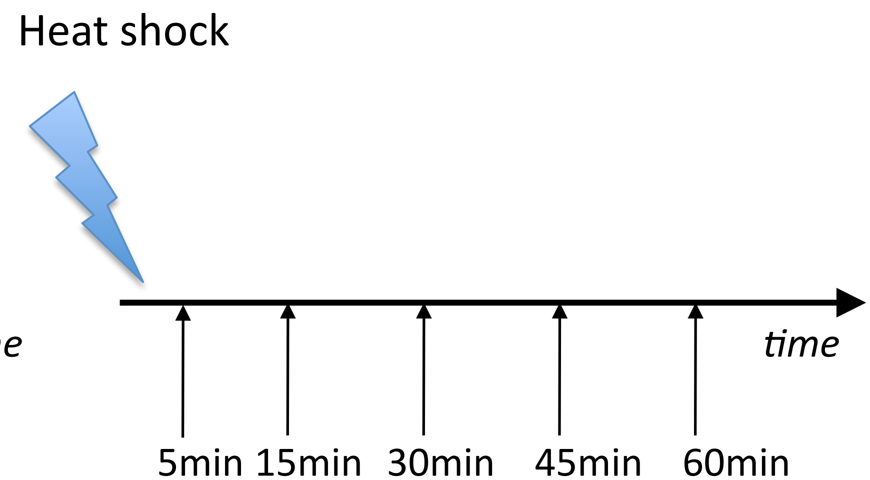

Dawn Thompson


Jenna Pfiffner


Ilan Wapinski



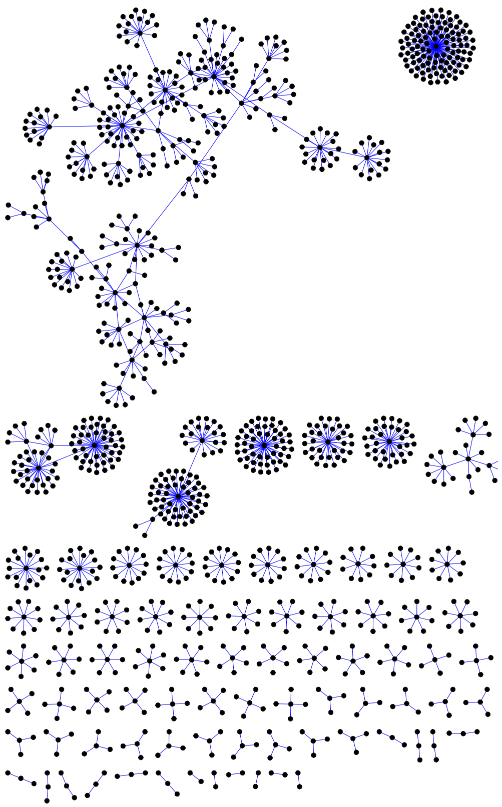
Glucose depletion (mild stress)



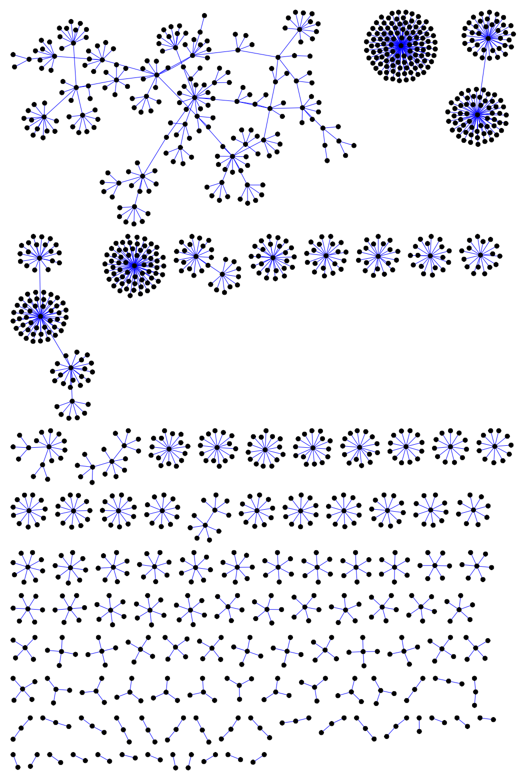
Heat shock (severe stress)

Wapinski *et al. 2010,* Thompson *et al.* In prep.

Topologically networks look similar, but have very few common edges



C. glabrata network



S. cerevisiae network

	Scer	Cgla	Scas	Kwal	Klac	Calb	Sjap	Spom
Scer	1	0.04	0.02	0.01	0.03	0.01	0.01	0.01
Cgla	0.04		0.03	0.09	0.01	0.02	0.01	0.01
Scas	0.02	0.03		0.02	0.03	0.01	0.01	0.01
Kwal	0.01	0.09	0.02		0.03	0.02	0.01	0.01
Klac	0.03	0.01	0.03	0.03		0.02	0.01	0.01
Calb	0.01	0.01	0.01	0.02	0.02		0.01	0.01
Sjap	0.01	0.01	0.01	0.01	0.01	0.01		0.08
Spom	0.01	0.01	0.01	0.01	0.01	0.01	0.08	

Conservation

0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.5	0.5

Pairwise network similarity

Take-away messages

- Transcriptional regulatory networks determine context specific gene expression
  - Important in development and disease
- Most of the regulatory network is not known
- Machine learning approaches to network inference
  - Supervised
  - Unsupervised
- Extensions to existing inference algorithms
  - Incorporate biological intuition
  - Integrate different types of datasets
- Evolution of regulatory networks
  - Major player for diversifying phenotypic diversity of organism
  - Comparative functional genomics brings new opportunities
    - Need phylogenetically-aware network analysis algorithms

For further reading, discussions, chats

sroy@biostat.wisc.edu