Learning and representing molecular networks from data

Sushmita Roy

sroy@biostat.wisc.edu

Computational Network Biology

Biostatistics & Medical Informatics 826

https://compnetbiocourse.discovery.wisc.edu

Sep 18th 2018

Some of the material covered in this lecture is adapted from BMI 576

Plan for this section

- Overview of network inference (Sep 18th)
- Directed probabilistic graphical models Bayesian networks (Sep 18th, Sep 20th)
- Gaussian graphical models (Sep 20th)
- Dependency networks (Sep 25th)
- Integrating prior information for network inference (Sep 27th, Oct 2nd, 4th)

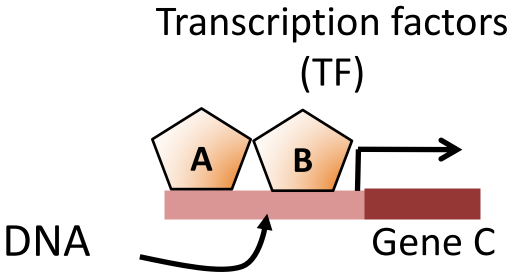
Readings

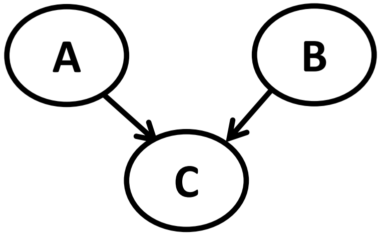
- Inferring cellular networks -- a review. http://dx.doi.org/10.1186/1471-2105-8-s6-s5
- Using bayesian networks to analyze expression data. http://dx.doi.org/10.1089/106652700750050961
- Learning module networks. http://www.jmlr.org/papers/volume6/segal05a/segal05a.pdf

Goals for today

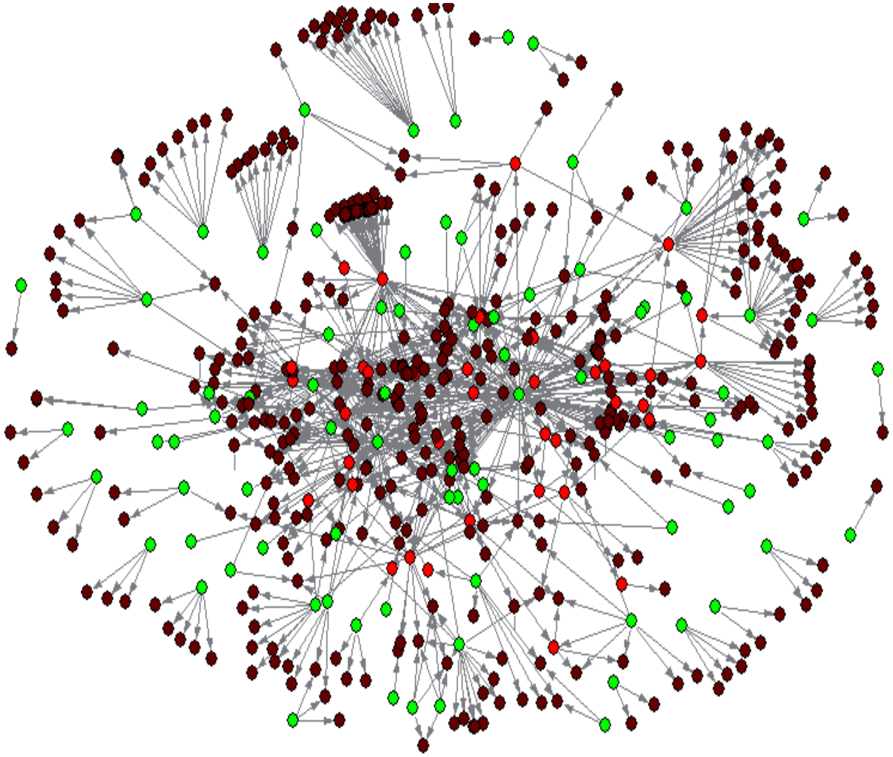
- Background on transcriptional networks
- Expression-based network inference
  - Per-gene and Per-module based methods
- Different types of probabilistic graphical models
- Learning Bayesian networks gene expression data

Transcriptional regulatory networks





- Directed, signed, weighted graph
- Nodes: TFs and Target genes
- Edges: A regulates C's expression level



Regulatory network of *E. coli*.
153 TFs (green & light red), 1319 targets

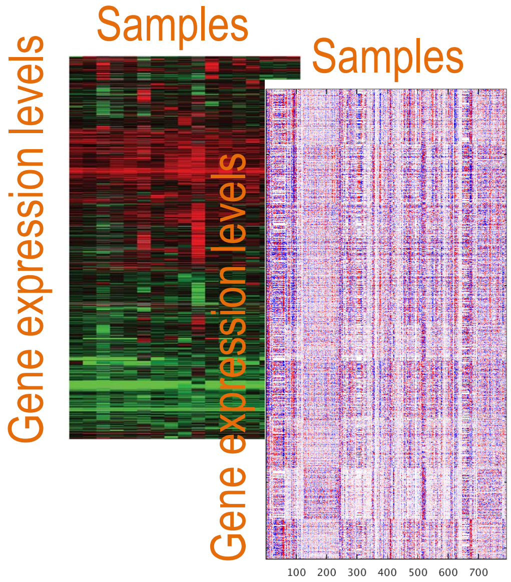
Vargas and Santillan, 2008

Why do we need to computationally infer transcriptional networks?

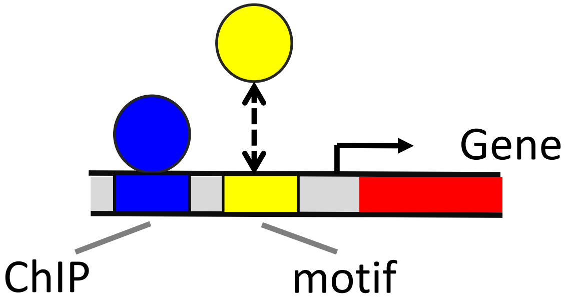
- Why infer transcriptional networks?
  - Control which genes are expressed when and where
  - Needed for accurate information processing in cells
  - Many diseases are associated with changes in transcriptional networks
- Why do so computationally?
  - Experimental detection of networks is hard, expensive
  - A first step towards having an *in silico* model of a cell
  - A model can be used to make predictions that can be tested and refine the model

Types of data for reconstructing transcriptional networks

- Node-specific datasets
  - Genome-wide gene expression (mRNA) levels
  - Can potentially recover genome-wide regulatory networks



- Edge-specific datasets
  - ChIP-chip and ChIP-seq
  - Sequence specific motifs
  - Factor knockout followed by whole-transcriptome profiling

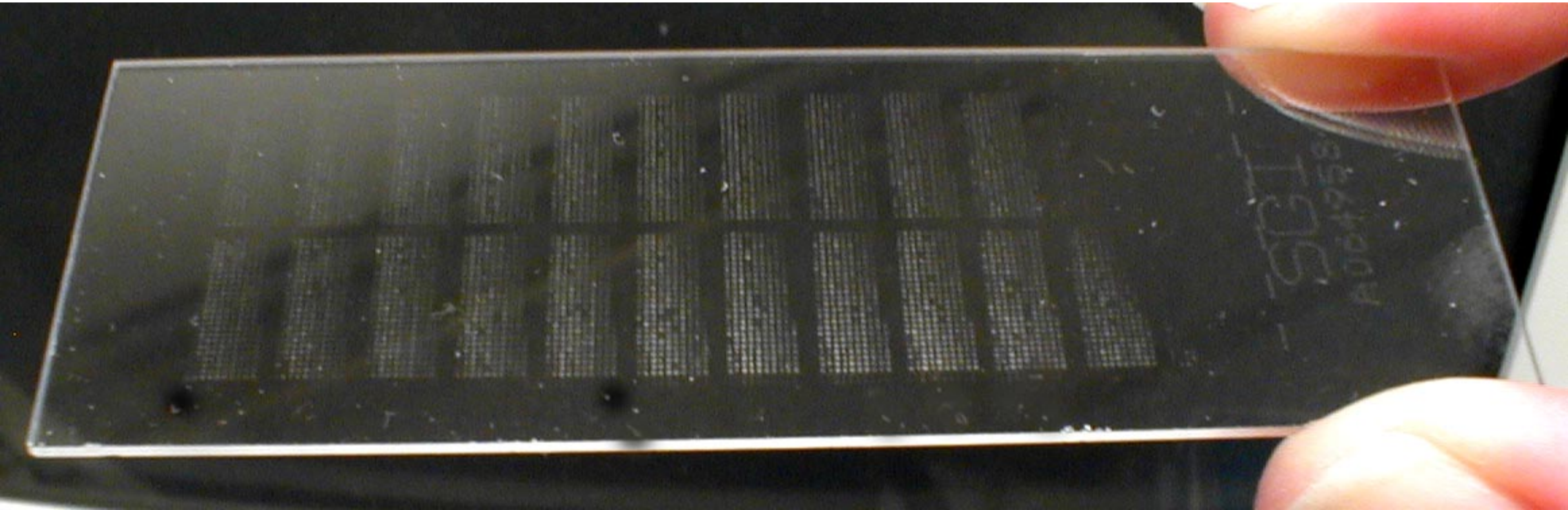


Experimental techniques to measure expression

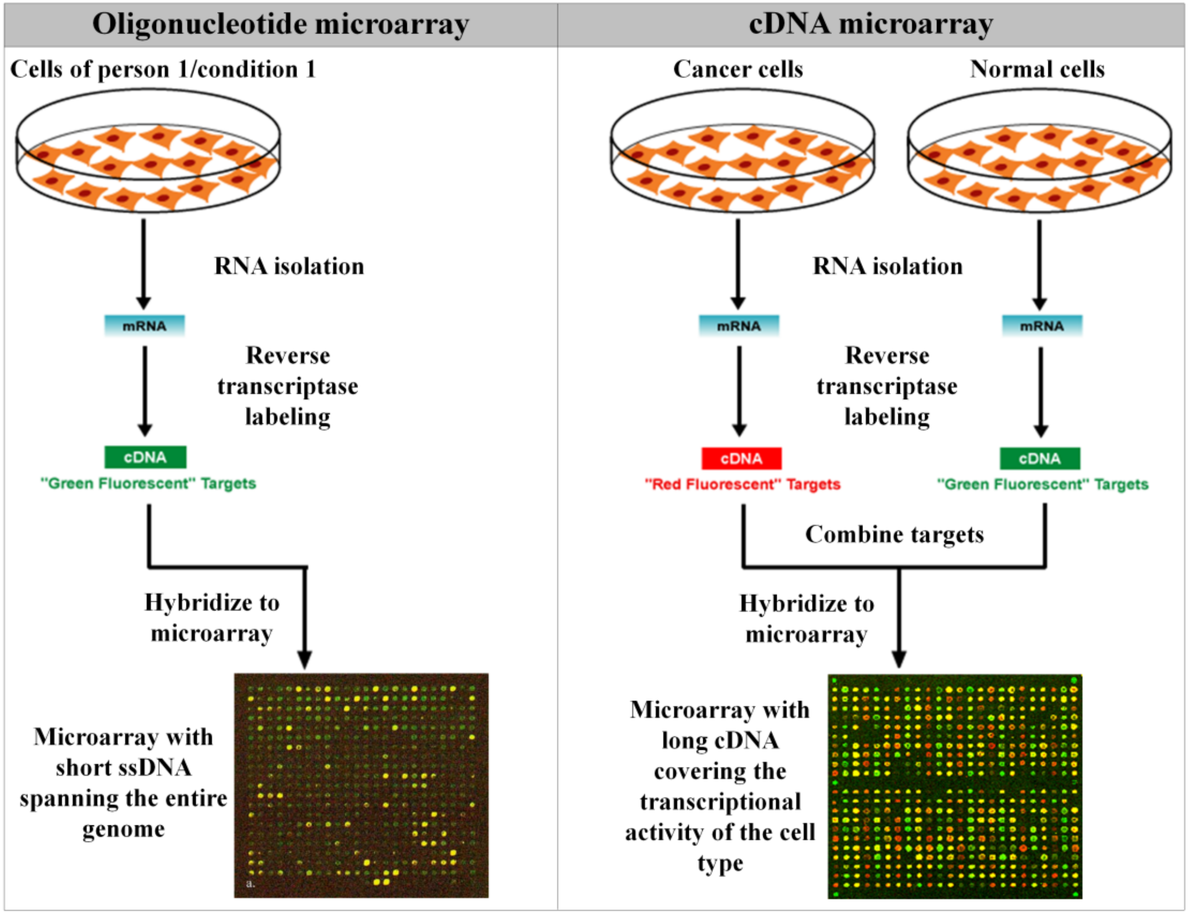
- Microarrays
  - cDNA/spotted arrays
  - Oligonucleotides arrays
- Sequencing
  - RNA-seq

Microarrays

- A microarray is a solid support, on which pieces of DNA are arranged in a grid-like array
  - Each piece is called a probe
- Measures RNA abundances by exploiting complementary hybridization
  - DNA from labeled sample is called target



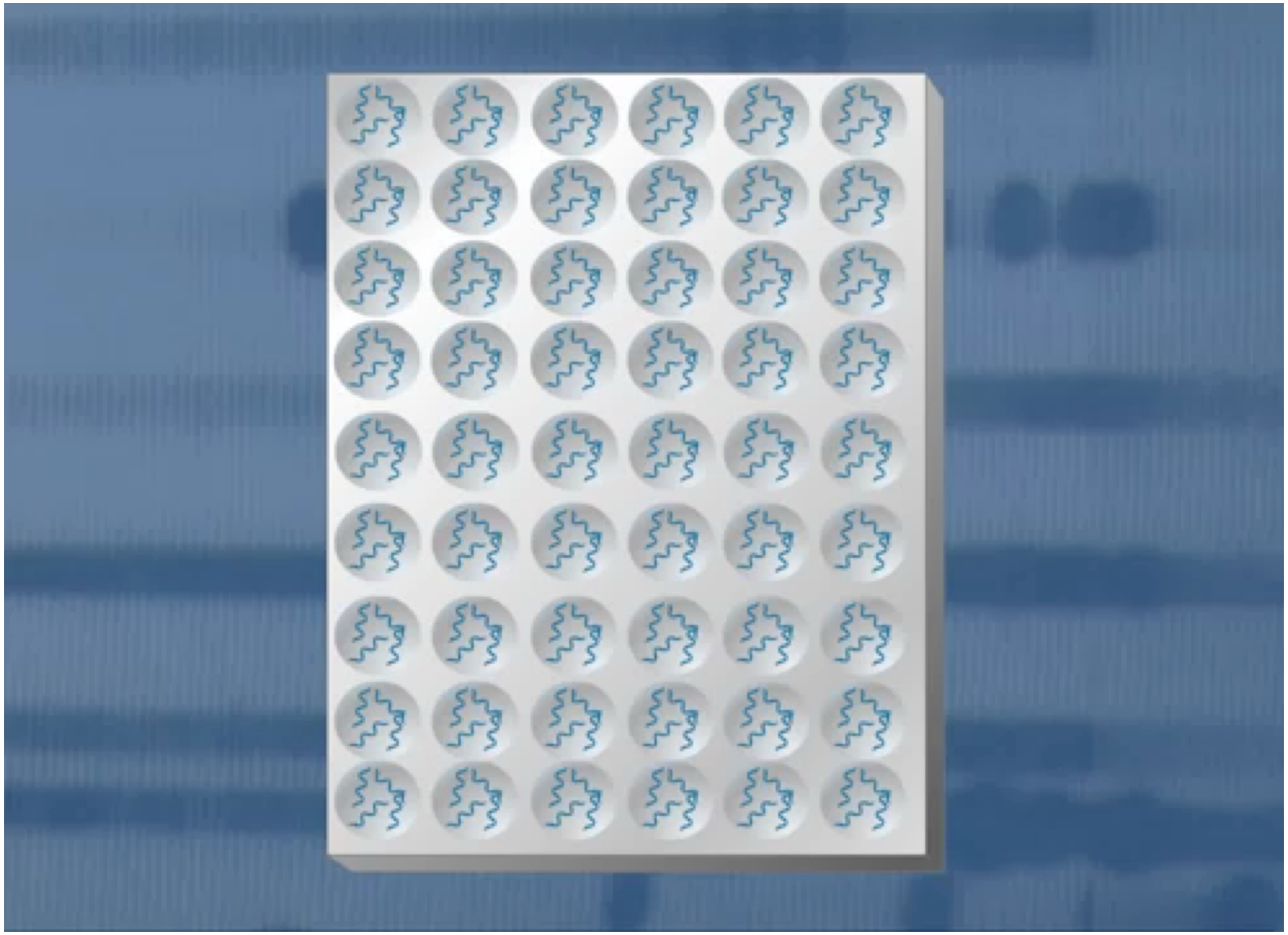
Spotted versus oligonucleotide array


Oligonucleotide microarray
Cells of person 1/condition 1
RNA isolation
mRNA
Reverse transcriptase labeling
cDNA
"Green Fluorescent" Targets
Hybridize to microarray
Microarray with short ssDNA spanning the entire genome
cDNA microarray
Cancer cells
Normal cells
RNA isolation
mRNA
mRNA
Reverse transcriptase labeling
cDNA
"Red Fluorescent" Targets
cDNA
"Green Fluorescent" Targets
Combine targets
Hybridize to microarray
Microarray with long cDNA covering the transcriptional activity of the cell type

Three key steps: Reverse transcription, labeling and hybridization

From Vermeeren and Luc Michiels 2011 *DOI: 10.5772/19432*

A video about DNA microarrays



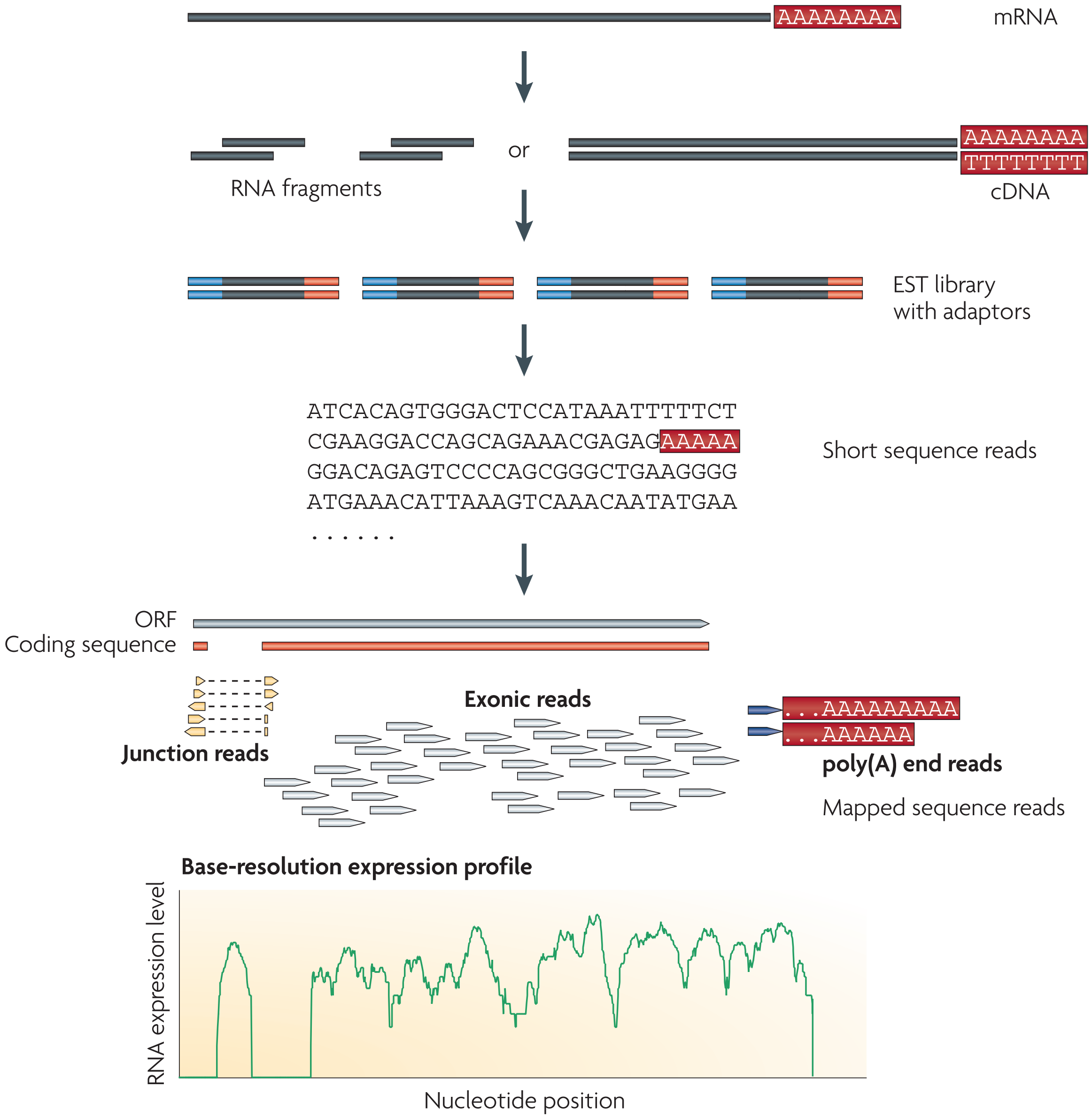
From: https://www.youtube.com/watch?v=_6ZMEZK-alM.

Also see: http://www.bio.davidson.edu/courses/genomics/chip/chip.html

A video about two color DNA microarrays

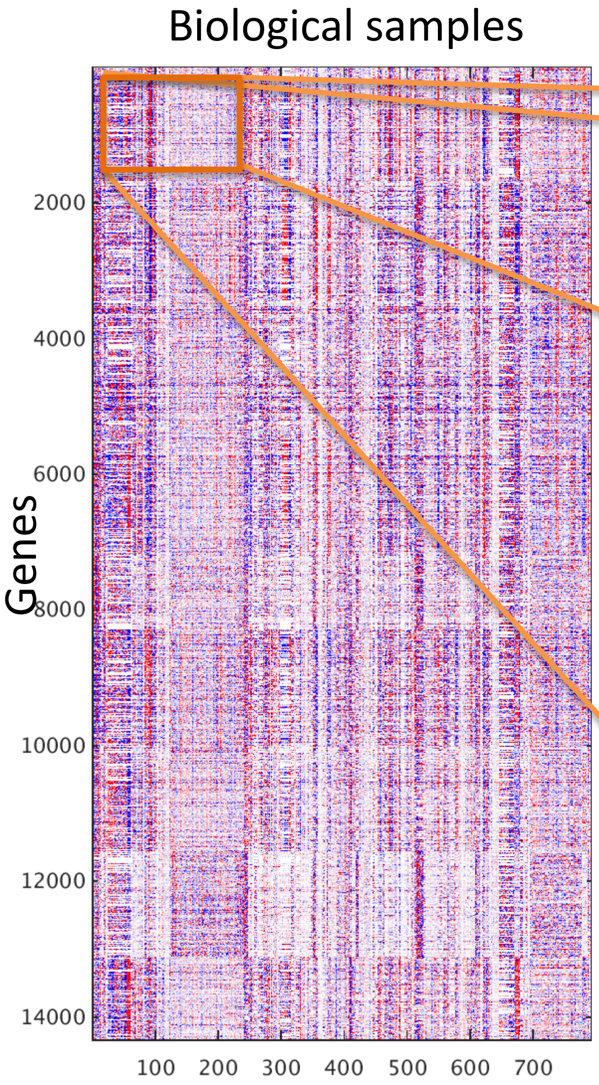
From: https://www.youtube.com/watch?v=VNsThMNjKhM

A typical RNA-seq pipeline



Wang et al, Nature Genetics 2009;

Gene expression profiling experiments produce expression matrices



Genes

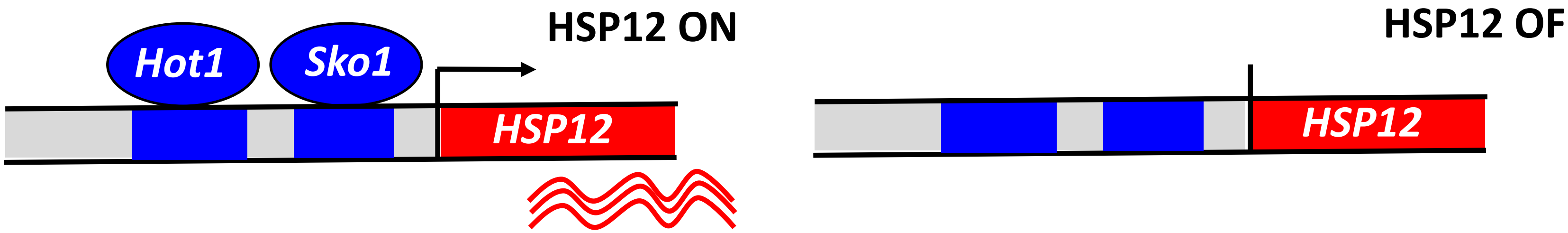
	E1	E2	E3	E4	E5	E6	E7	…
ENSMUSG00000028180	8.82	9.09	6.43	8.11	7.13	7.55	9.18	
ENSMUSG00000053211	0.0	0.15	0.07	0.0	0.08	0.0	0.0	
ENSMUSG00000028182	0.0	0.0	0.0	0.0	0.1	0.08	0.26	
ENSMUSG00000002017	2.83	1.92	2.33	0.86	2.17	2.53	3.19	
ENSMUSG00000028184	2.0	1.32	1.13	0.72	1.25	1.17	2.27	
ENSMUSG00000028187	12.41	10.72	10.23	8.59	8.08	8.92	11.61	
ENSMUSG00000028186	0.02	0.68	0.0	0.0	0.0	0.0	0.0	
ENSMUSG00000028189	0.69	0.95	1.09	0.97	0.71	0.44	0.76	
ENSMUSG00000028188	0.11	0.22	0.12	0.21	0.24	0.2	0.43	

⋮

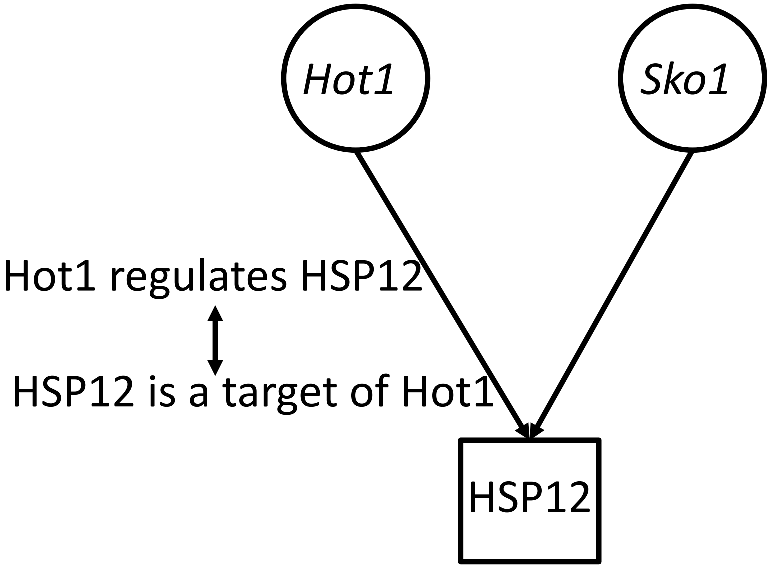
Goals for today

- Background on transcriptional networks
- Expression-based network inference
  - Per-gene and Per-module based methods
- Different types of probabilistic graphical models
- Learning Bayesian networks gene expression data

What do we want a model for a regulatory network to capture?

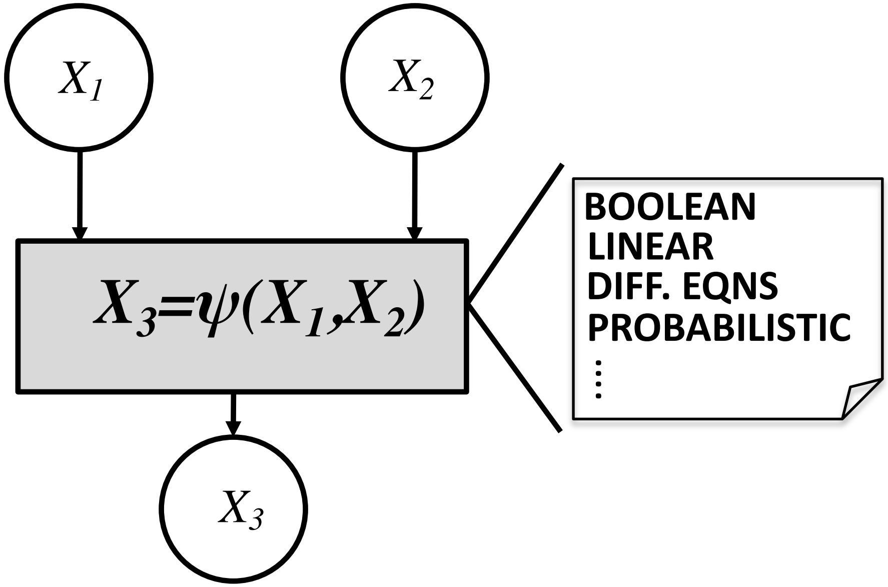


HSP12's expression is dependent upon Hot1 and Sko1 binding to HSP12's promoter



Hot1 regulates HSP12

HSP12 is a target of Hot1



BOOLEAN
LINEAR
DIFF. EQNS
PROBABILISTIC
⋮

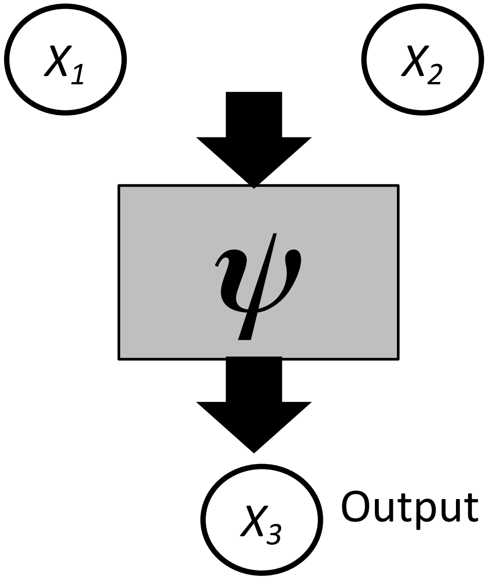
Structure

Who are the regulators?

Function

How they determine expression levels?

Mathematical representations of the “how” question



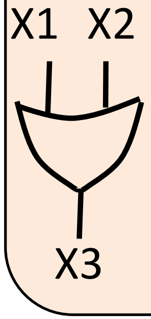
Input expression of neighbors

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

Output expression of target

Boolean Networks

Input Output



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

Differential equations

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

Probabilistic graphical models

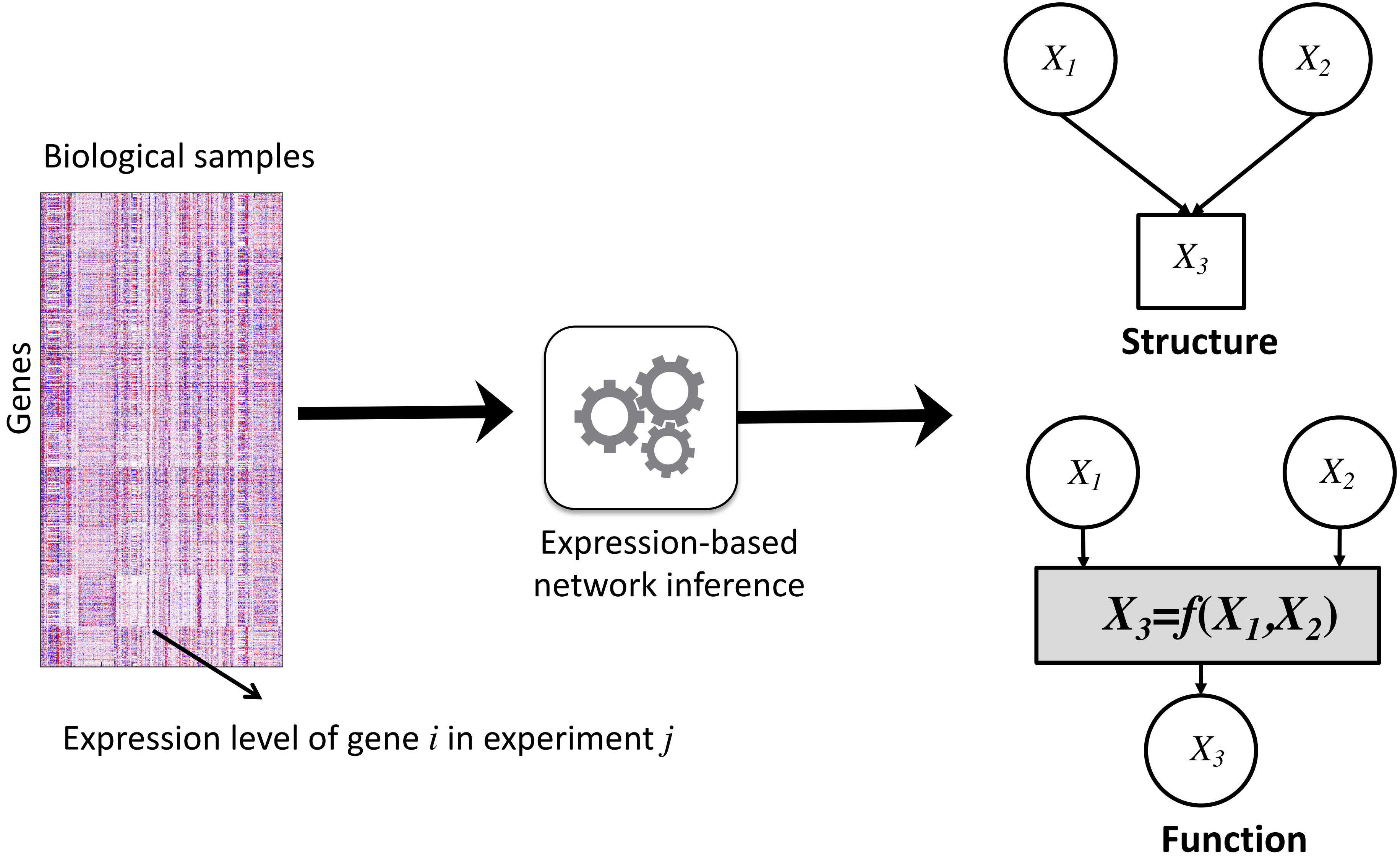
$$P(X_3|X_1, X_2) = N(X_1a + X_2b, \sigma)$$

Probability distributions

Expression-based regulatory network inference

- Given
  - A set of measured mRNA levels across multiple biological samples
- Do
  - Infer the regulators of genes
  - Infer how regulators specify the expression of a gene
- Algorithms for network reconstruction vary based on their meaning of interaction

Expression-based network inference



Biological samples

Genes

Expression level of gene i in experiment j

Expression-based network inference

X_1 X_2 X_3

Structure

X_1 X_2

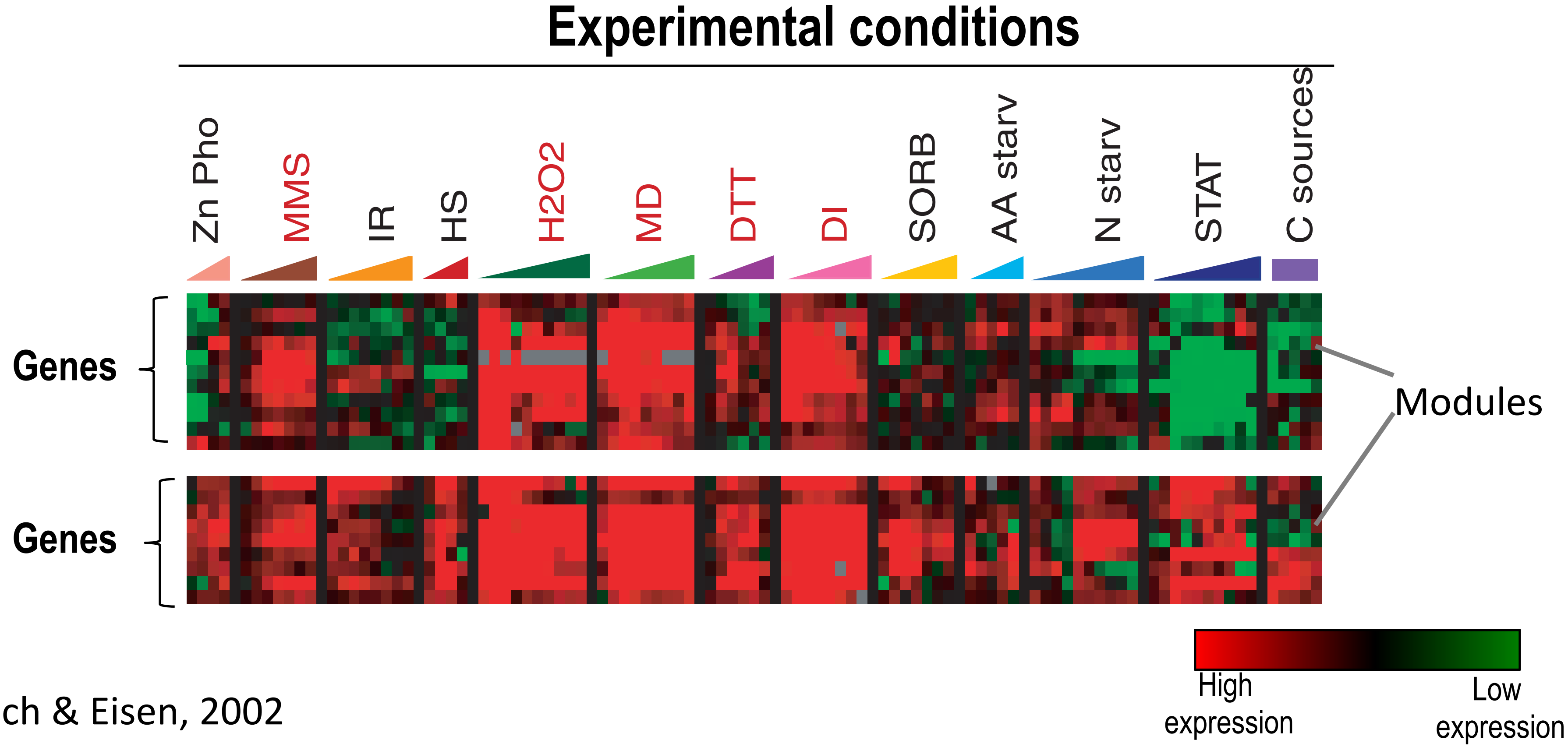
$X_3=f(X_1,X_2)$

X_3

Function

Regulatory gene modules

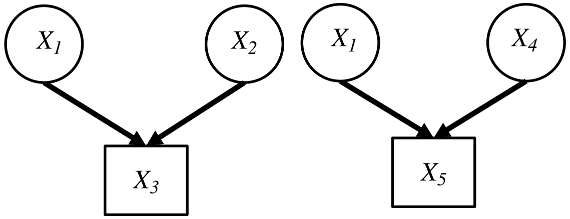
A regulatory module: set of genes with similar regulatory state



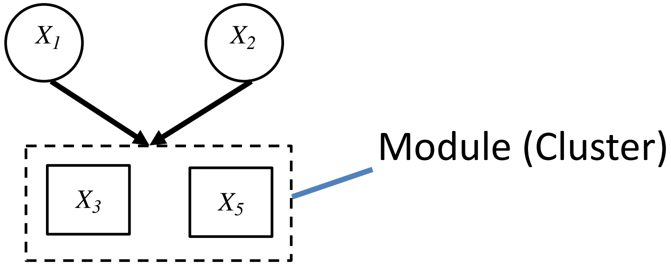
Gasch & Eisen, 2002

Two classes of expression-based network inference methods

- Per-gene/direct methods



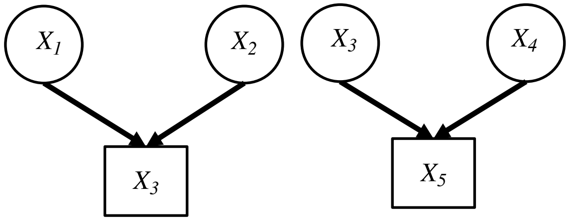
- Module based methods



A non-exhaustive list of expression-based network inference method

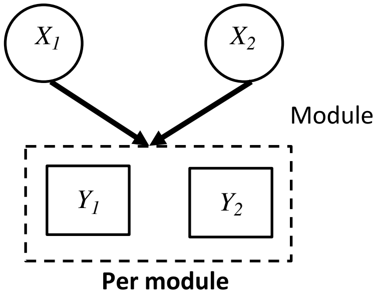
Method Name	Per-module	Per-gene	Model type
Sparse candidate		✓	Bayesian network
CLR		✓	Information theoretic
ARACNE		✓	Information theoretic
TIGRESS		✓	Dependency network
Inferelator		✓	Dependency network
GENIE3		✓	Dependency network
ModuleNetworks	✓		Bayesian network
LemonTree	✓		Dependency network
WGCNA		✓	Correlation

Per-gene methods



- Key idea: find the regulators that "best explain" expression level of a gene
- Probabilistic graphical methods
  - Bayesian network
    - Sparse Candidates
  - Dependency networks
    - GENIE3, TIGRESS
- Information theoretic methods
  - Context Likelihood of relatedness
  - ARACNE

Per-module methods



- Find regulators for an entire module
  - Assume genes in the same module have the same regulators
- Module Networks (Segal et al. 2005)
- Stochastic LeMoNe (Joshi et al. 2008)

Goals for today

- Background on transcriptional networks
- Expression-based network inference
  - Per-gene and Per-module based methods
- Different types of probabilistic graphical models
- Learning Bayesian networks gene expression data

Probabilistic graphical models (PGMs)

- A marriage between probability and graph theory
- Nodes on the graph represent random variables
- Graph structure specifies statistical dependency structure
- Graph parameters specify the nature of the dependency
- PGMs can be directed or undirected
- Examples of PGMs: Bayesian networks, Dependency networks, Markov networks, Factor graphs

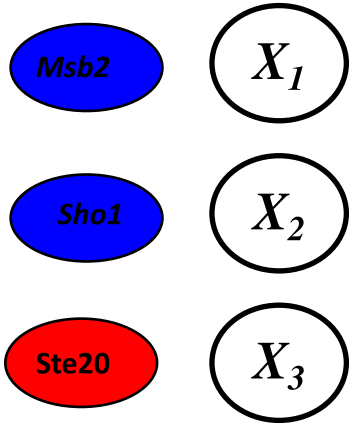
Different types of probabilistic graphs

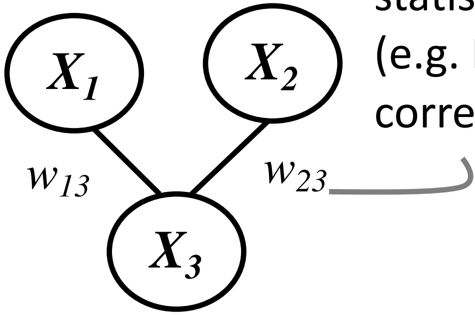
- In each graph type we can assert different conditional independencies
- Correlation networks
- Gaussian Graphical models
- Dependency networks
- Bayesian networks

Correlational networks

- An undirected graph
- Edges represent high correlation
  - Need to determine what "high" is
- Edge weights denote different values of correlation
- Cannot discriminate between direct and indirect correlations

Random variables represent gene expression levels





A measure of statistical correlation (e.g. Pearson's correlation)

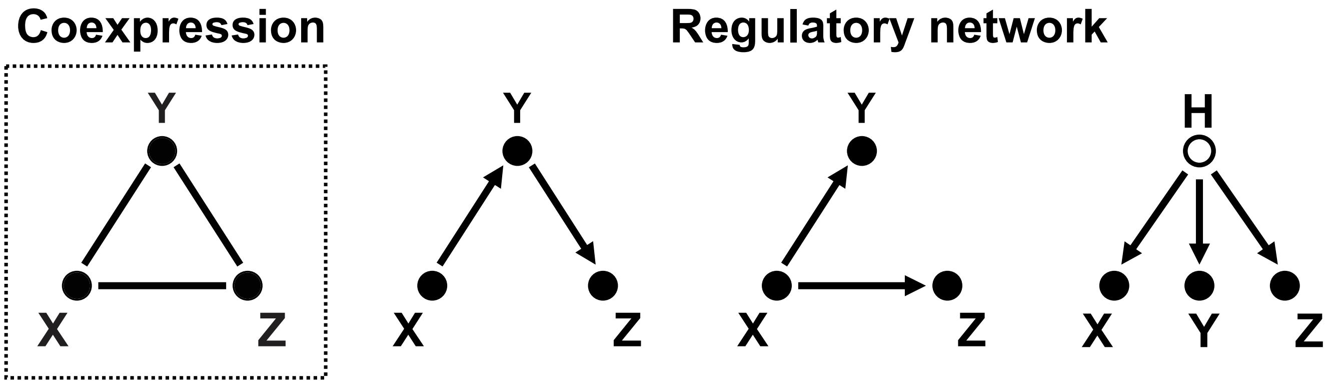
An undirected weighted graph.

Popular examples of correlational networks

- Weighted Gene Co-expression Network Analysis (WGCNA)
  - Zhang and Horvath 2005
- Relevance Networks
  - Butte & Kohane, 2000 Pacific symposium of biocomputing

Limitations of correlational networks

- Correlational networks cannot distinguish between direct and indirect dependencies
- This makes them less interpretable than other PGMs



- For any co-expression network, there are several possible regulatory networks that can explain these correlations.

- What we would like is to be able to discriminate between direct and indirect dependencies

- Here we need to review conditional independencies

Conditional independencies in PGMs

- The different classes of models we will see are based on a general notion of specifying statistical independence
- Suppose we have two genes X and Y. We add an edge between X and Y if X and Y are not independent given a third set Z.
- Depending upon Z we will have a family of different PGMs

Conditional independence and PGMs

- Correlational networks
  - Z is the empty set
- Markov networks
  - X and Y are not independent given all other variables
  - Gaussian Graphical models are a special case (later lectures)
- Dependency networks
  - Approximate Markov networks
  - May not be associated with a valid joint distribution (later lectures)
- First-order conditional independence models
  - Explain the correlation between two variables by a third variable
- Bayesian networks
  - Generalize first-order conditional independence models

Goals for today

- Background on transcriptional networks
- Expression-based network inference
  - Per-gene and Per-module based methods
- Different types of probabilistic graphical models
- Bayesian networks
- Learning Bayesian networks gene expression data

Bayesian networks (BN)

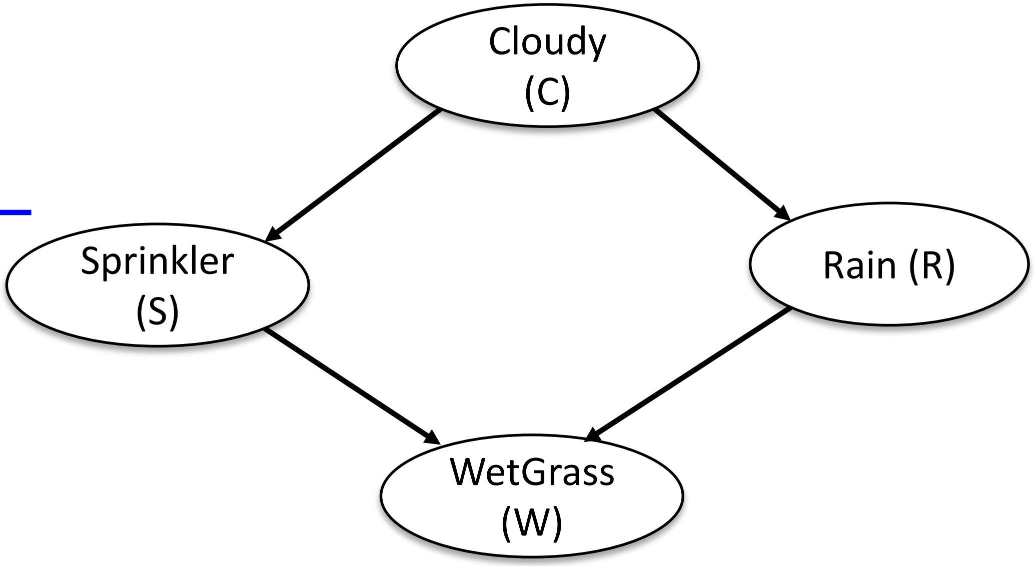
- A special type of probabilistic graphical model
- Has two parts:
  - A graph which is directed and acyclic
  - A set of conditional distributions
- Directed Acyclic Graph (DAG)
  - The nodes denote random variables $X_1 \ldots X_N$
  - The edges
    - encode statistical dependencies between the random variables
    - establish parent child relationships
- Each node X_i has a *conditional probability distribution* (CPD) representing $\mathrm{P}(X_i \mid Pa(X_i))$; Pa: Parents
- Provides a tractable way to represent large joint distributions

Key questions in Bayesian networks

- What do the CPDs look like?
- What independence assertions can be made in Bayesian networks?

An example Bayesian network

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1	0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Adapted from Kevin Murphy: Intro to Graphical models and Bayes networks:
http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

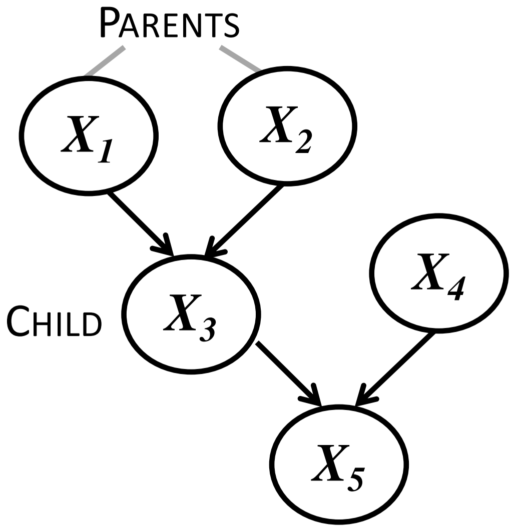
Notation

- $B=\{G, \mathbf{\Theta}\}$ A Bayesian network
- X_i: i^th random variable
- If there are few random variables, we will just use upper case letters. E.g. A, B, C..
- $\boldsymbol{X}=\{X_1,.., X_p\}$: set of p random variables
- x_i^k: An assignment of X_i in the k^{th} sample
- $\boldsymbol{Pa}(X_i)$: Parents of random variable X_i
- $\boldsymbol{D}=\{\boldsymbol{x}^1,.. ,\boldsymbol{x}^m\}$: Dataset of m observations/samples of $\boldsymbol{X}$
- $I(X_i;X_j/X_k,X_l)$: Conditional independence notation: X_i is independent of X_j given X_k and X_l

Bayesian networks compactly represent joint distributions

$$P(X_1, \cdots, X_p) = \prod_{i=1}^{p} P(X_i | Pa(X_i))$$

Example Bayesian network of 5 variables



Assume X_i is binary

Needs 2^5 measurements

No independence assertions

$$P(\mathbf{X}) = P(X_1, X_2, X_3, X_4, X_5)$$

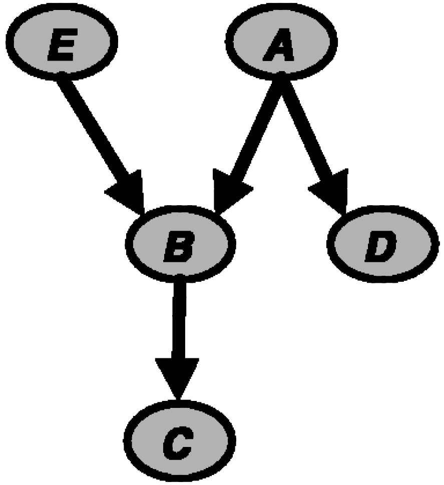
Needs 2^3 measurements

Independence assertions

$$P(\mathbf{X}) = P(X_1)P(X_2)P(X_4)P(X_3|X_1, X_2)P(X_5|X_3, X_4)$$

Conditional independencies in BN

- A variable X_i is independent of its non-descendants given its parents
- $I(X_i;X_j/X_k,X_l)$: X_i is independent of X_j given X_k and X_l



$$I(A;E)$$

$$I(B;D|A,E)$$

$$I(D;E,B,C|A)$$

$$I(C;E,A,D|B)$$

$$I(E;A,D)$$

Consider the example Bayesian network. What are the set of conditional independencies in this graph?

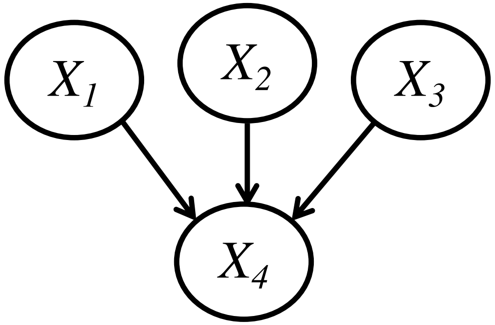
Friedman et al., 2000

CPD in Bayesian networks

- The CPD $P(X_i|Pa(X_i))$ specifies a distribution over values of X_i for each combination of values of $Pa(X_i)$
- CPD $P(X_i|Pa(X_i))$ can be parameterized in different ways
- X_i are discrete random variables
  - Conditional probability table or tree
- X_i are continuous random variables
  - CPD can be linear Gaussians, conditional Gaussians or regression trees

Representing CPDs as tables

- Consider four binary variables X_1, X_2, X_3, X_4



Pa($\boldsymbol{X_4}$): $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$

P(X_4 | X_1, X_2, X_3) **as a table**

			$\boldsymbol{X_4}$	
X_1	X_2	X_3	**t**	**f**
t	**t**	**t**	**0.9**	**0.1**
t	**t**	**f**	**0.9**	**0.1**
t	**f**	**t**	**0.9**	**0.1**
t	**f**	**f**	**0.9**	**0.1**
f	**t**	**t**	**0.8**	**0.2**
f	**t**	**f**	**0.5**	**0.5**
f	**f**	**t**	**0.5**	**0.5**
f	**f**	**f**	**0.5**	**0.5**

Estimating CPD table from data

- Assume we observe the following assignments for X_1, X_2, X_3, X_4

N=7

X_1	X_2	X_3	X_4
T	F	T	T
T	T	F	T
T	T	F	T
T	F	T	T
T	F	T	F
T	F	T	F
F	F	T	F

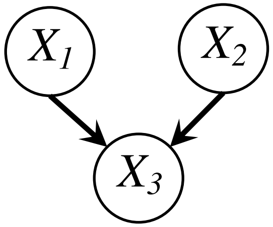
For each joint assignment to X_1, X_2, X_3, estimate the probabilities for each value of X_4

For example, consider $X_1=T, X_2=F, X_3=T$

$$P(X_4=T|X_1=T, X_2=F, X_3=T)=2/4$$
$$P(X_4=F|X_1=T, X_2=F, X_3=T)=2/4$$

Gaussians distribution for CPD

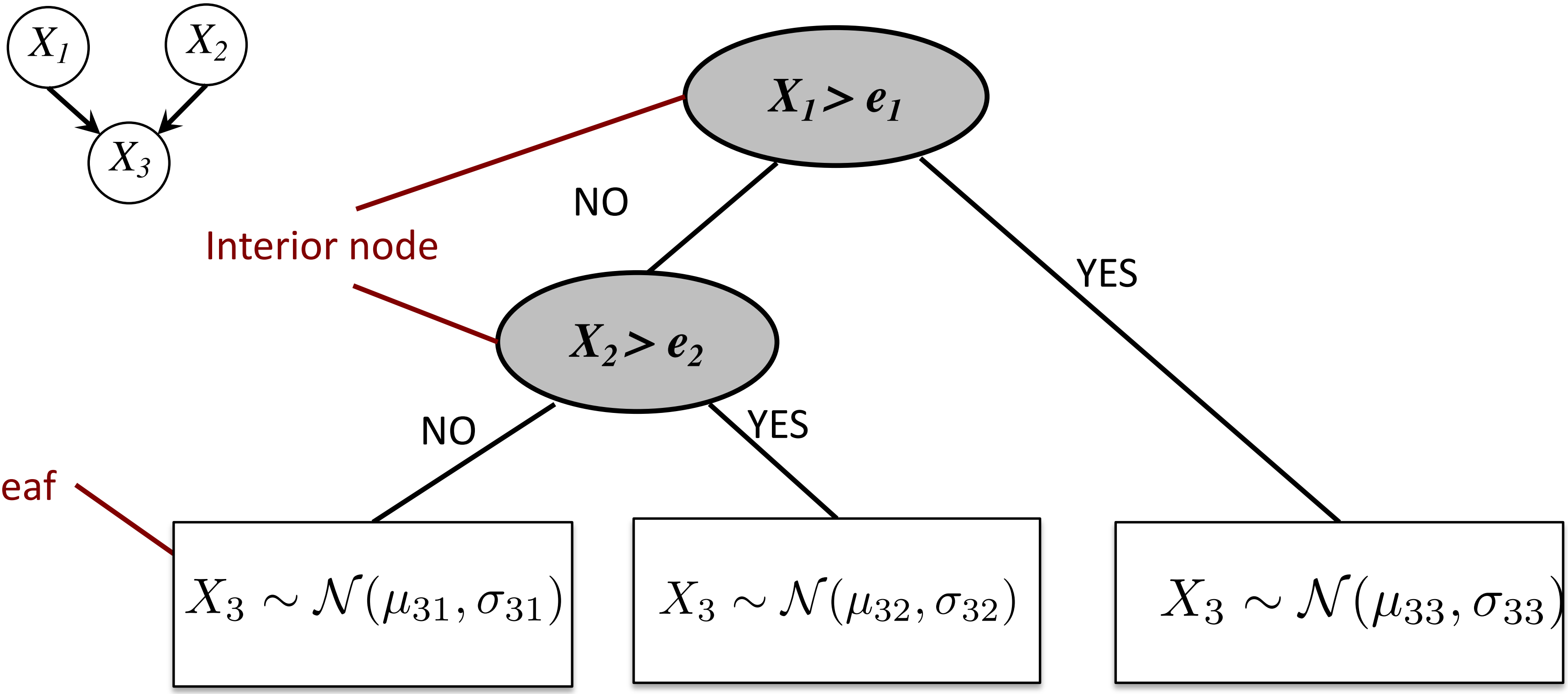
- For every joint assignment of the parent set, we have a Gaussian distribution on the child variable.



$$P(X_3|X_1 = x_1, X_2 = x_2) = \mathcal{N}(a_0 + a_1 x_1 + a_2 x_2, \sigma)$$

A regression tree to capture a CPD $P(X_3|X_1, X_2)$

e_1, e_2 are values seen in the data



Interior node

Leaf

$X_1 > e_1$

NO

YES

$X_2 > e_2$

NO

YES

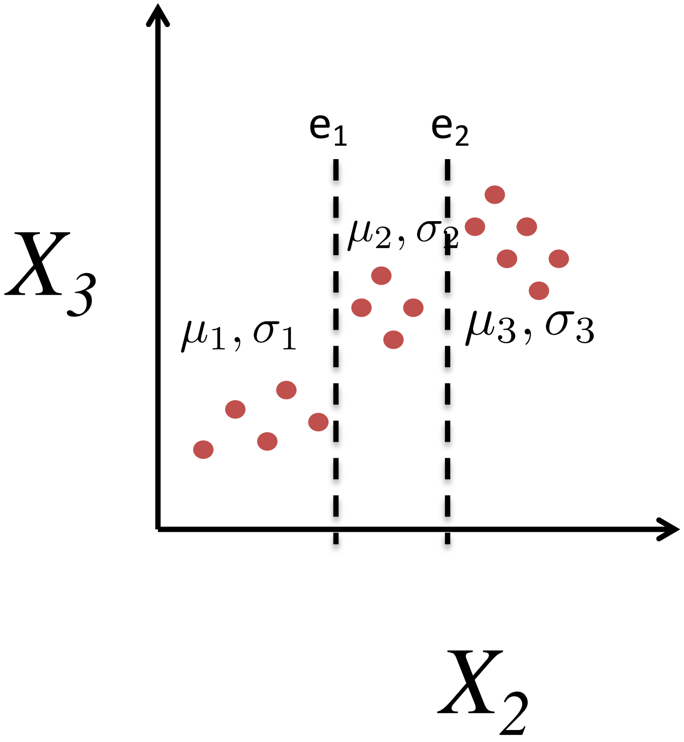
$X_3 \sim \mathcal{N}(\mu_{31}, \sigma_{31})$

$X_3 \sim \mathcal{N}(\mu_{32}, \sigma_{32})$

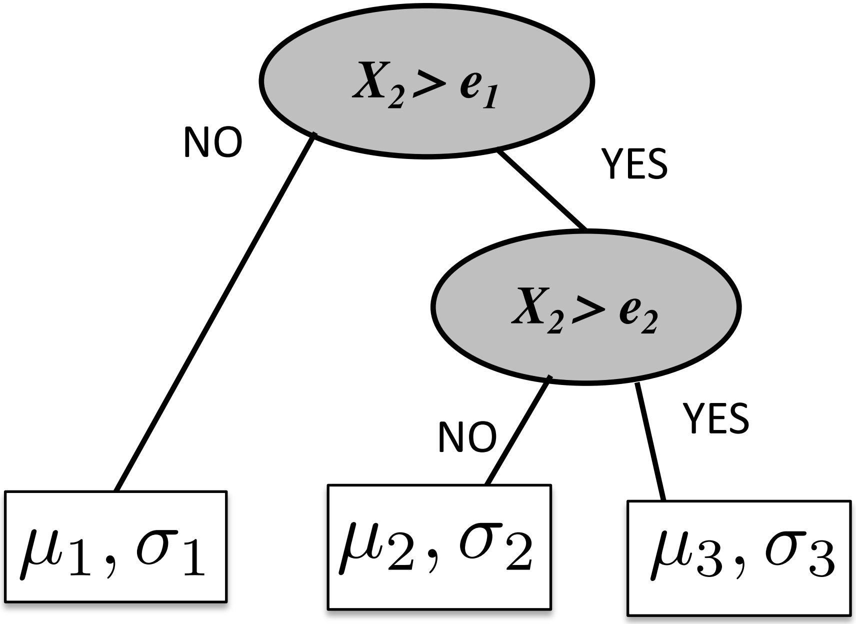
$X_3 \sim \mathcal{N}(\mu_{33}, \sigma_{33})$

Expression of gene represented by X_3 modeled using Gaussians at each leaf node

A regression tree captures non-linear dependencies



X_3



Compute probabilities using a Bayesian network

What is the probability of

$$P(C=F, R=T, S=F, W=T)$$

Bayes net allows us to write

$$P(W|S,R)P(S|C)P(R|C)P(C)$$

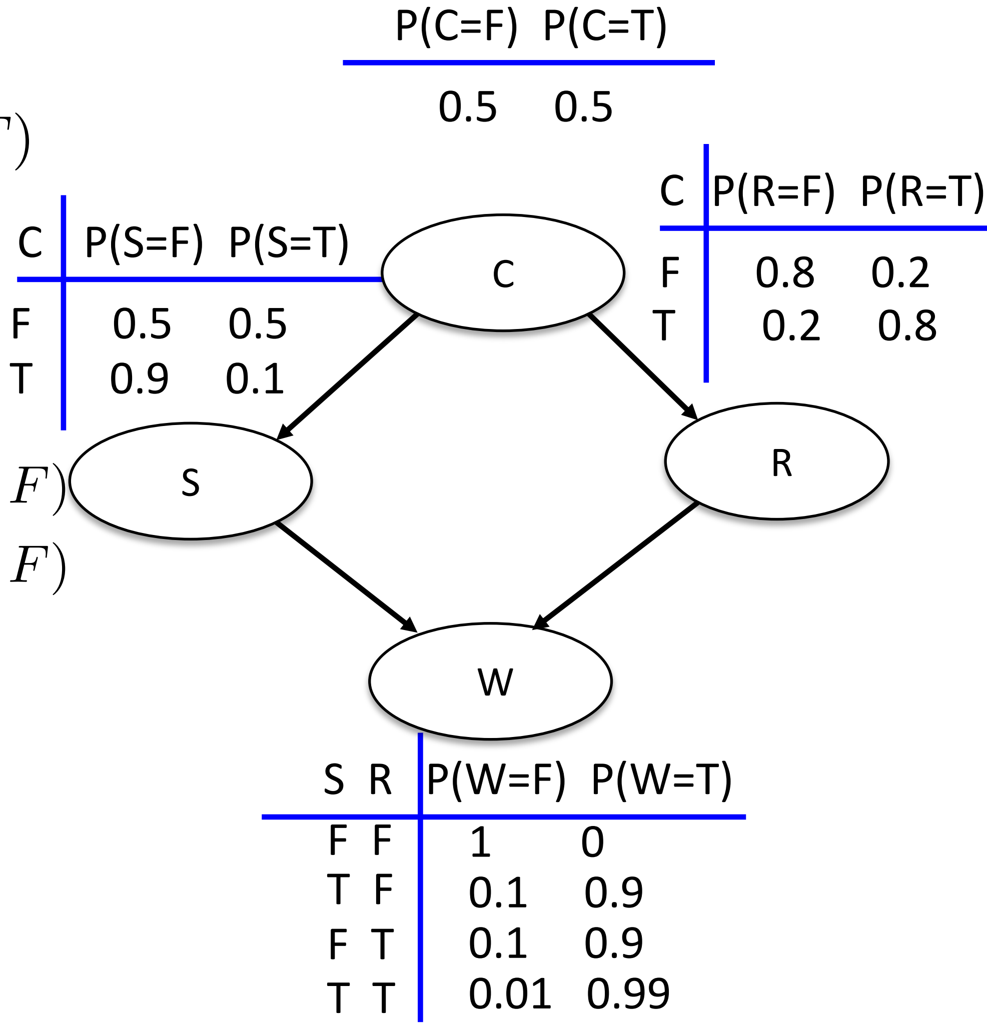
Plugging in the assignments for the variables:

$$P(W=T|S=F,R=T)P(S=F|C=F)$$
$$P(R=T|C=F)P(C=F)$$

Looking up in the CPD

0.9*0.5*0.2*0.5

=0.045



P(C=F)	P(C=T)
0.5	0.5

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

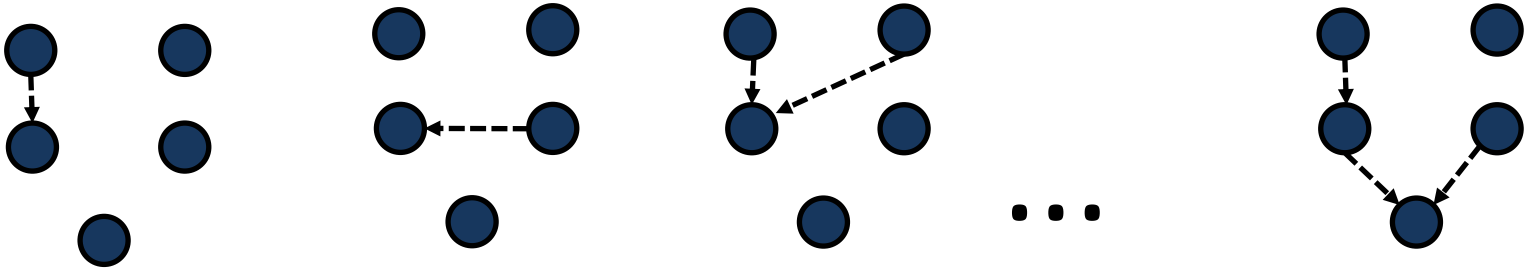
S	R	P(W=F)	P(W=T)
F	F	1	0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Learning problems in Bayesian networks

- Parameter learning on known graph structure
  - Given a set of joint assignments of the random variables, estimate the parameters of the model
- Structure learning
  - Given a set of joint assignments of the random variables, estimate the structure and parameters of the model
  - Structure learning subsumes parameter learning

Structure learning using score-based search

$\mathrm{Score}(B)$ Describes how well B describes the data



$\mathrm{Score}(B_1)$ $\mathrm{Score}(B_2)$ $\mathrm{Score}(B_3)$ … $\mathrm{Score}(B_m)$

Exhaustive search is not computationally tractable

Scoring a Bayesian network

- Maximum likelihood score

$$\text{Score}_{ML}(\mathbf{G}:\mathbf{D}) = \max_{\boldsymbol{\Theta}} P(\mathbf{D}|G,\boldsymbol{\Theta})$$

- Bayesian score

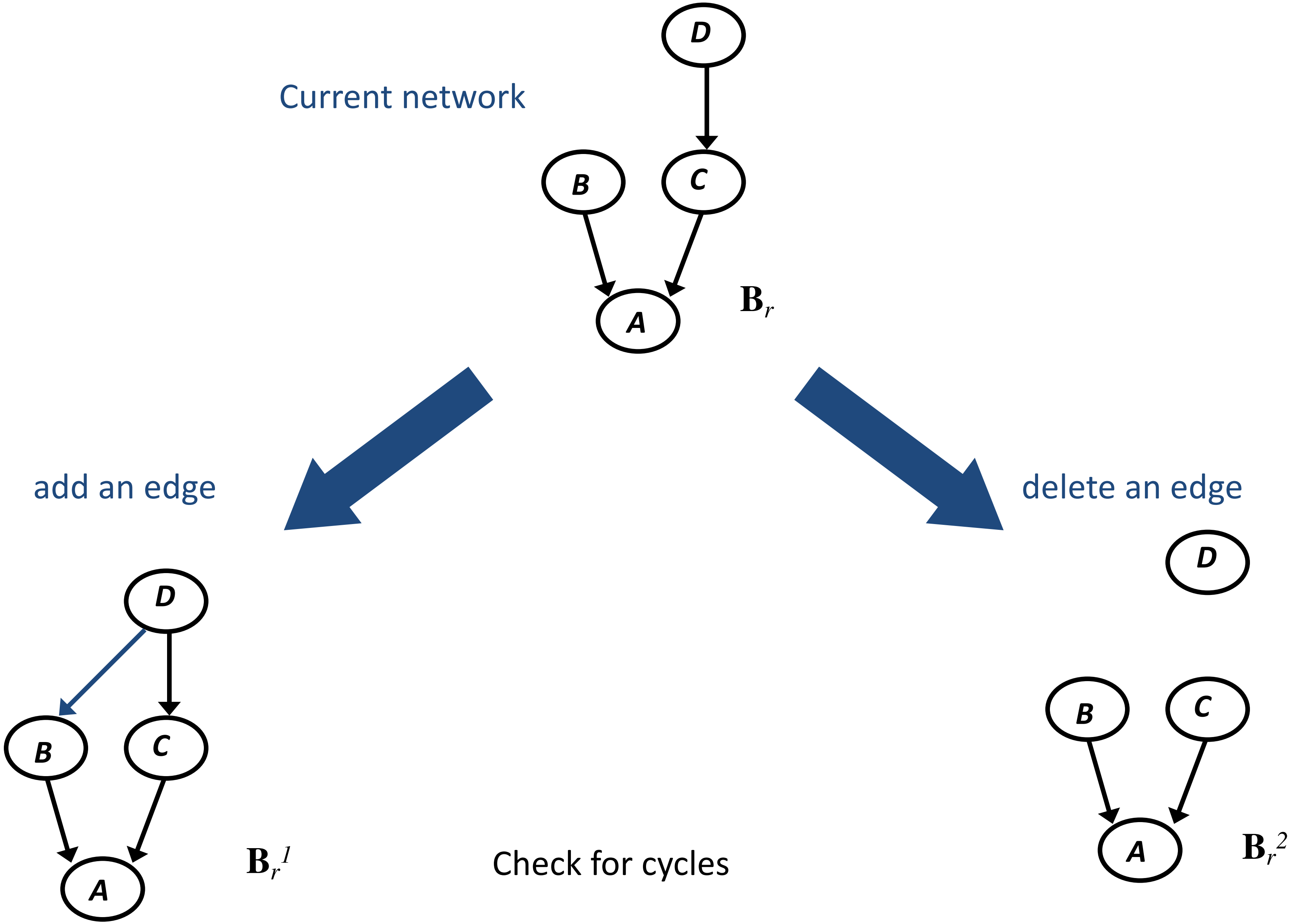
$$\text{Score}_{Bayes}(\mathbf{G}:\mathbf{D}) = P(\mathbf{G}|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{G})P(\mathbf{G})}{P(\mathbf{D})}$$

We typically ignore the denominator
as it is the same for all models

Greedy hill climbing to search Bayesian network space

- Input: Data $\mathbf{D}$, An initial Bayesian network, $\mathbf{B}_0$={$\mathbf{G}_0$, $\mathbf{\Theta}_0$}
- Output: $\mathbf{B}_{\text{best}}$
- Loop for $r=1, 2..$ until convergence:
  - $\{\mathbf{B}_r^{1}, .., \mathbf{B}_r^{\text{m}}\}$ = $Neighbors(\mathbf{B}_r)$ by making local changes to $\mathbf{B}_r$
  - $\mathbf{B}_{r+1}$: $arg\ max_{\text{j}}(\text{Score}(\mathbf{B}_r^{\text{j}}))$
- Termination:
  - $\mathbf{B}_{\text{best}}$= $\mathbf{B}_r$

Local changes to $\mathbf{B}_i$



Current network

$\mathbf{B}_r$

add an edge

delete an edge

$\mathbf{B}_r^1$

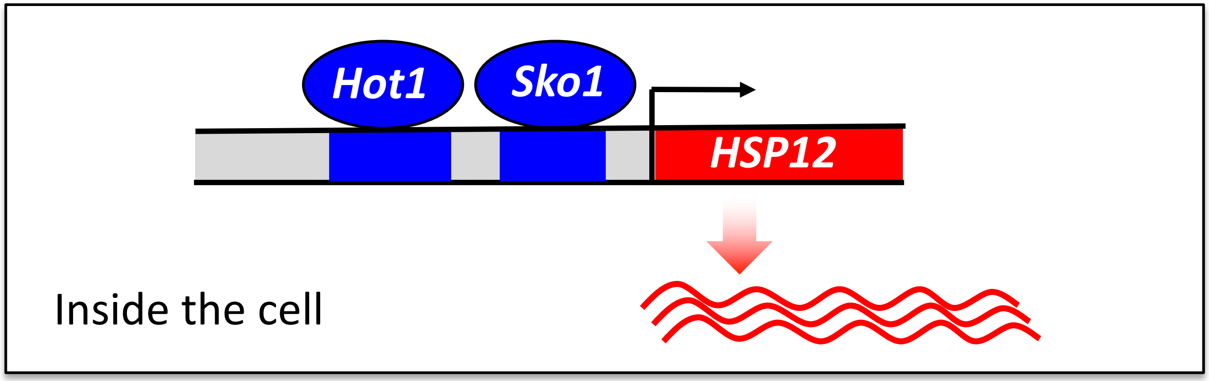
Check for cycles

$\mathbf{B}_r^2$

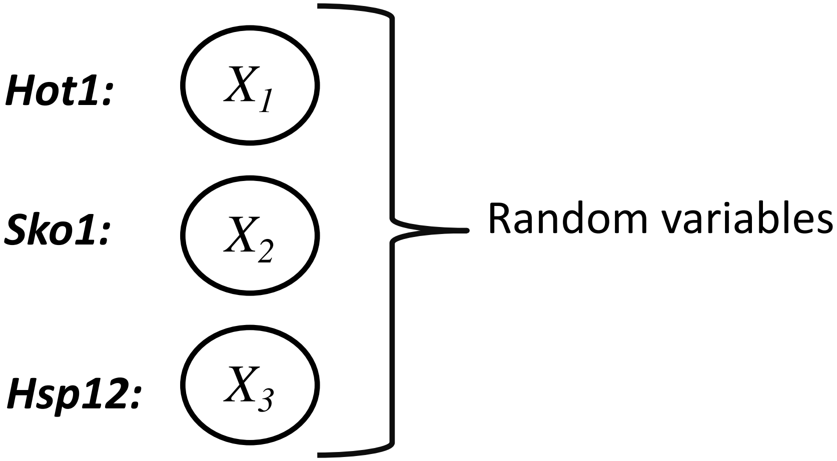
Goals for today

- Background on transcriptional networks
- Expression-based network inference
  - Per-gene and Per-module based methods
- Different types of probabilistic graphical models
- Bayesian networks
- Learning Bayesian networks gene expression data

Bayesian network representation of a regulatory network



Inside the cell

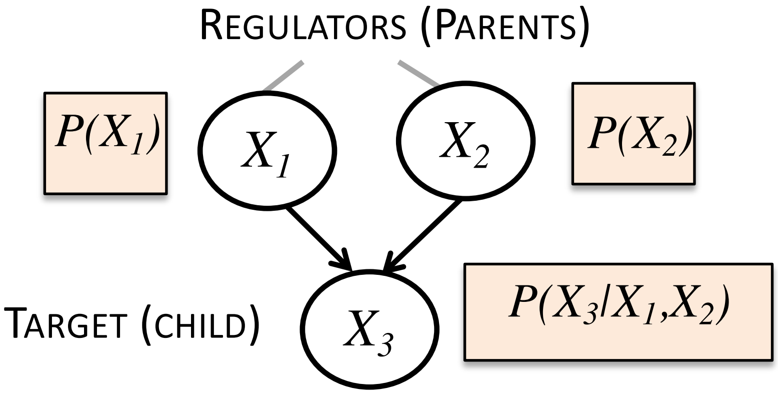


Hot1:

Sko1:

Hsp12:

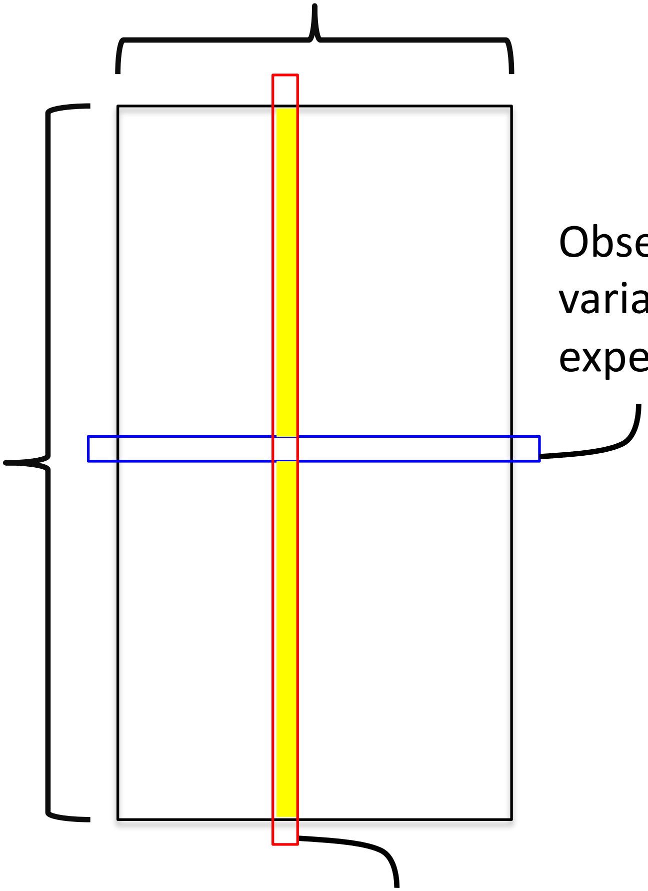
Random variables



Bayesian network

Expression data matrix

N Experiments/Time points etc



p Genes

Observations of variable X_j in all N experiments

Observations (expression levels) of all variables in sample i, $\boldsymbol{x}^{(i)}$

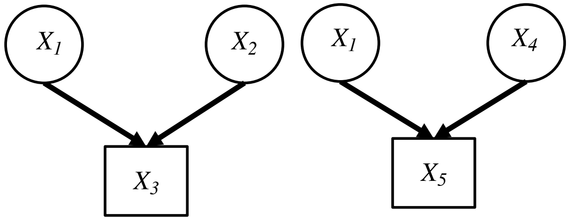
Challenges with applying Bayesian network to genome-scale data

- Number of variables, p is in thousands

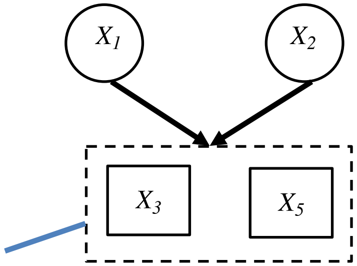
- Number of samples, N is in hundreds

Bayesian network-based methods to handle genome-scale networks

- Sparse candidate algorithm
  - Friedman, Nachman, Pe’er. 1999
  - Friedman, Linial, Nachman, Pe’er. 2000.
- Module networks
  - Segal, Pe’er, Regev, Koller, Friedman. 2005



Per-gene



Module (Cluster)

Per-module

The Sparse candidate Structure learning in Bayesian networks

- A fast Bayesian network learning algorithm
- Key idea: Identify k “promising” candidate parents for each X_i
  - $k<<p$, p: number of random variables
  - Candidates define a “skeleton graph” $\mathbf{H}$
- Restrict graph structure to select parents from $\mathbf{H}$
- Early choices in $\mathbf{H}$ might exclude other good parents
  - Resolve using an iterative algorithm

Sparse candidate algorithm

- Input:
  - A data set $\mathbf{D}$
  - An initial Bayes net $\mathbf{B}_0$
  - A parameter k: max number of parents per variable
- Output:
  - Final $\mathbf{B}_r$
- Loop for $r=1,2..$ until convergence
  - *Restrict*
    - Based on $\mathbf{D}$ and $\mathbf{B}_{r-1}$ select candidate parents C_i^r for X_i
    - This defines a skeleton directed network $\mathbf{H}_r$
  - *Maximize*
    - Find network $\mathbf{B}_r$ that maximizes the score $\mathrm{Score}(\mathbf{B}_r)$ among networks satisfying

$$Pa^r(X_i) \subseteq C_i^r$$

- Termination: Return $\mathbf{B}_r$

Information theory for measuring dependence

- *I(X;Y)* is the mutual information between two variables
  - Knowing *X*, how much information do we have for *Y*
- *P(Z)* is the probability distribution of *Z*

$$I(X;Y) = \Sigma_{x,y \in X,Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

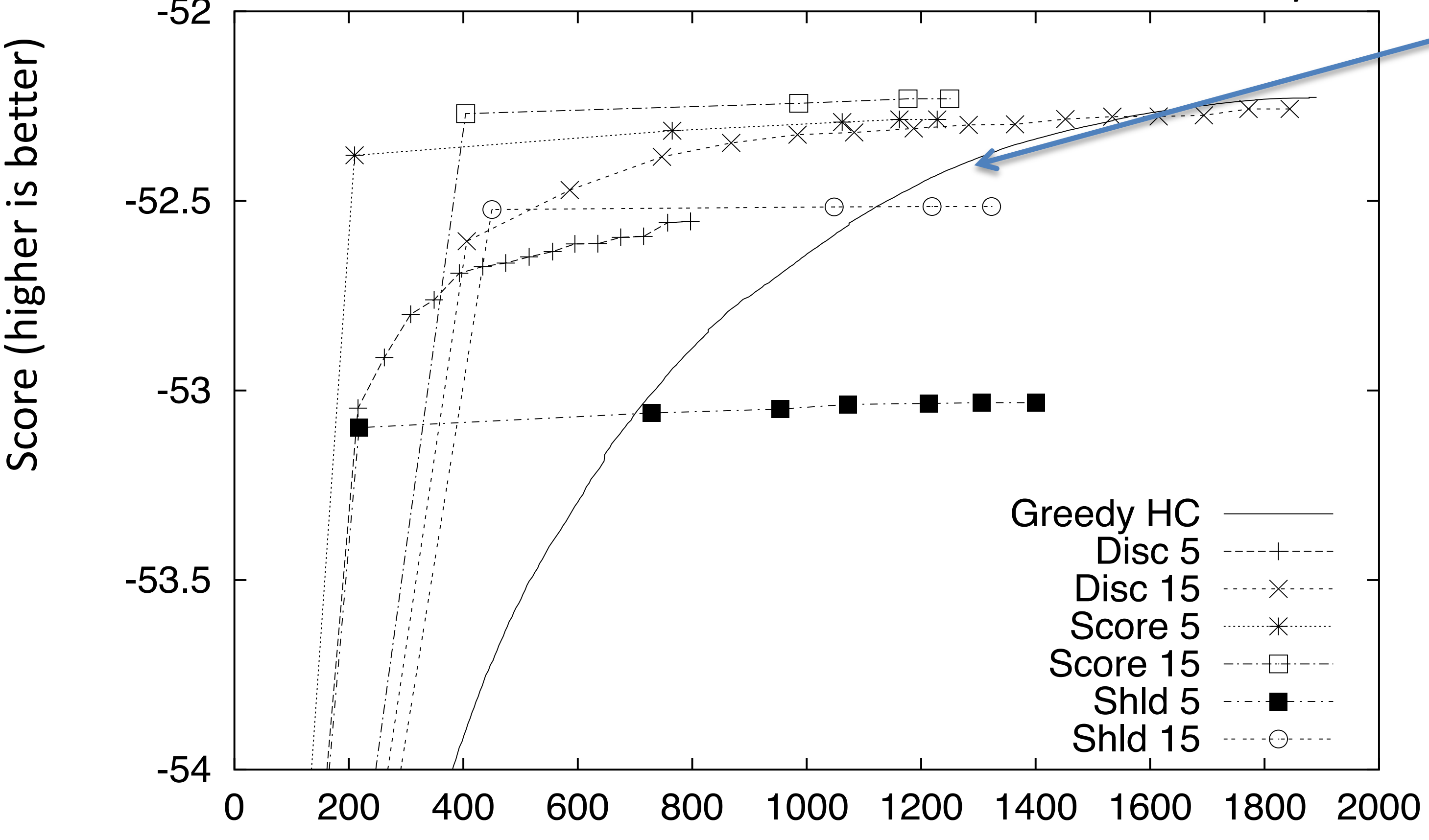
- Measures the difference between the two distributions: joint and product of marginals

Selecting candidate parents in the Restrict Step

- A good parent for X_i is one with strong statistical dependence with X_i
  - Mutual information provides a good measure of statistical dependence $I(X_i; X_j)$
  - Mutual information should be used only as a first approximation
    - Candidate parents need to be iteratively refined to avoid missing important dependences
- A good parent for X_i has the highest score improvement when added to $Pa(X_i)$

Sparse candidate learns good networks faster than hill-climbing

Greedy hill climbing takes much longer to reach a high scoring bayesian network



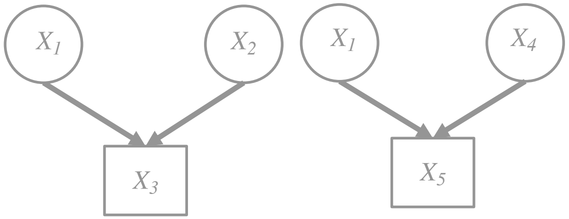
Size of Bayesian network: 100 variables

Some comments about choosing candidates

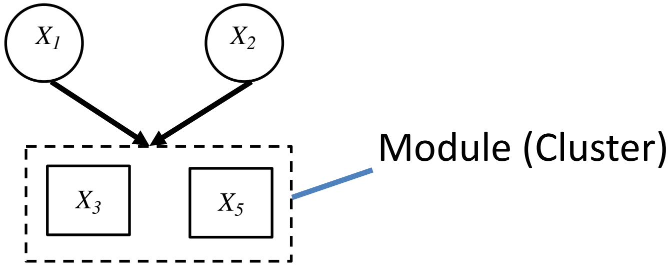
- How to select k in the sparse candidate algorithm?
- Should k be the same for all X_i ?
- Estimate an undirected *dependency network*
  - Learn a Bayesian network constrained on the dependency network structure
- Regularized regression approaches can be used to estimate the structure of an undirected graph
  - Schmidt, Niculescu-Mizil, Murphy 2007

Bayesian network-based methods to handle genome-scale networks

- Sparse candidate algorithm
  - Friedman, Nachman, Pe’er. 1999
  - Friedman, Linial, Nachman, Pe’er. 2000.
- Module networks
  - Segal, Pe’er, Regev, Koller, Friedman. 2005

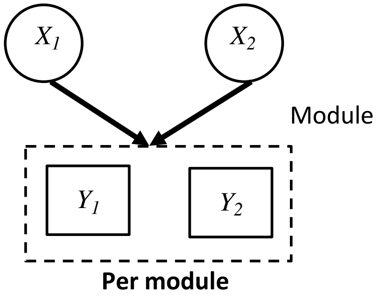


Per-gene



Per-module

Per-module methods



- Find regulators for an entire module
  - Assume genes in the same module have the same regulators
- Module Networks (Segal et al. 2005)
- Stochastic LeMoNe (Joshi et al. 2008)

Module Networks

- Motivation:
  - Most complex systems have too many variables
  - Not enough data to robustly learn networks
  - Large networks are hard to interpret
- Key idea: Group similarly behaving variables into "modules" and learn the same parents and parameters for each module
- Relevance to gene regulatory networks
  - Genes that are co-expressed are likely regulated in similar ways

Segal et al 2005, JMLR

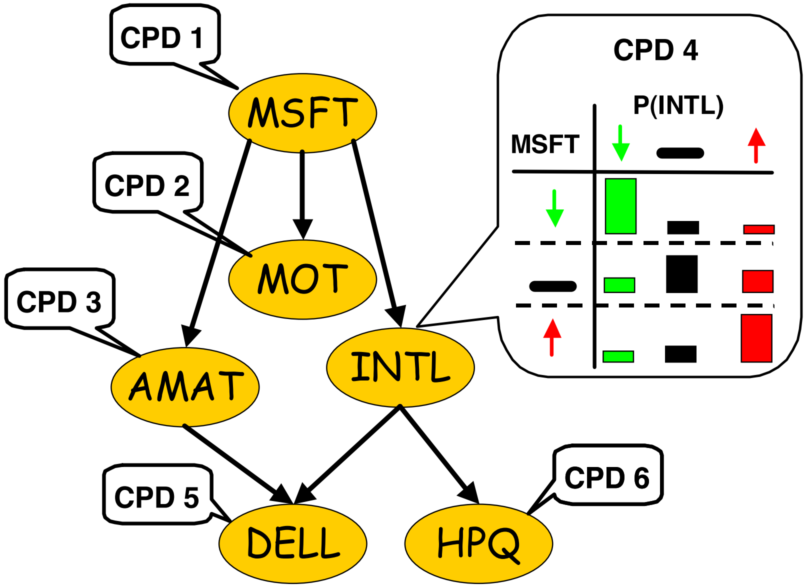
Definition of a module

- Statistical definition (specific to module networks by Segal 2005)
  - A set of random variables that share a statistical model
- Biological definition of a module
  - Set of genes that are co-expressed and co-regulated

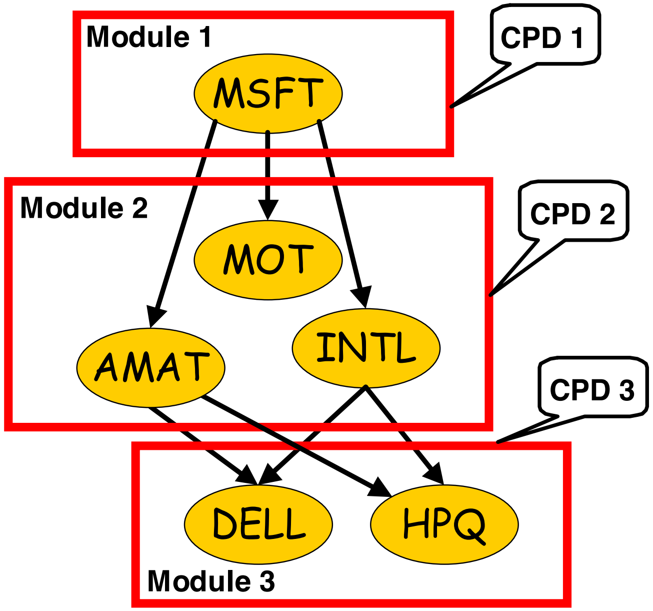
Bayesian network vs Module network

- Bayesian network
  - Different CPD per random variable
  - Learning only requires to search for parents
- Module network
  - CPD per module
    - Same CPD for all random variables in the same module
  - Learning requires parent search and module membership assignment

Bayesian network vs Module network



(a) Bayesian network



(b) Module network

Each variable takes three values: UP, DOWN, SAME

Modeling questions in Module Networks

- How to score and learn module networks?
- How to model the CPD between parent and children?
  - Regression Tree

Defining a Module Network

- A probabilistic graphical model over N random variables $\mathbf{X} = \{X_1, \cdots, X_N\}$
- Set of module variables $M_1 .. M_K$
- Module assignments A that specifies the module (1-to-K) for each X_i
- CPD per module $P(M_j|Pa_{Mj})$, Pa_{Mj} are parents of module M_j
  - Each variable X_i in M_j has the same conditional distribution

Learning a Module Network

- Given training dataset $\mathbf{D} = \{\mathbf{x}^1, \cdots, \mathbf{x}^m\}$, fixed number of modules, K
- Learn
  - Module assignments A of each variable to a module
  - The parents of each module to give structure S

Score of a module network

- Module network makes use of a Bayesian score

$$P(\mathcal{S},\mathcal{A}\mid\mathcal{D}) \propto \underbrace{P(\mathcal{A})P(\mathcal{S}\mid\mathcal{A})}_{\text{Priors}}\underbrace{P(\mathcal{D}\mid\mathcal{S},\mathcal{A})}_{\text{Data likelihood}}$$

$$\text{score}(\mathcal{S},\mathcal{A} : \mathcal{D}) = \underbrace{\log P(\mathcal{A}) + \log P(\mathcal{S}\mid\mathcal{A})}_{\text{Priors}} + \underbrace{\log P(\mathcal{D}\mid\mathcal{S},\mathcal{A})}_{\text{Data likelihood}}.$$

Score of a module network continued

Integrate parameters out

$$\log P(\mathcal{D}|\mathbf{S},\mathbf{A}) = \log \int P(\mathcal{D}|\mathbf{S},\mathbf{A},\theta)P(\theta|\mathbf{S},\mathbf{A})d\theta$$

Decomposes over each module

$$\log \prod_{j=1}^{k} \int L_j(\mathbf{U},\mathbf{X},\theta_{\mathbf{M}_j|\mathbf{U}} : \mathcal{D})P(\theta_{\mathbf{M}_j}|\mathbf{U})d\theta_{\mathbf{M}_j|\mathbf{U}}$$

Decomposes over each module

$$\sum_{j=1}^{K} \log \int L_j(\mathbf{U},\mathbf{X},\theta_{\mathbf{M}_j|\mathbf{U}} : \mathcal{D})P(\theta_{\mathbf{M}_j}|\mathbf{U})d\theta_{\mathbf{M}_j|\mathbf{U}}$$

$\mathbf{U}$: Set of parents defined by $\mathbf{S}$

$\mathbf{X}$: Set of variables.

For computing each L_j term we would need only the variables and parents associated with module j

Defining the data likelihood

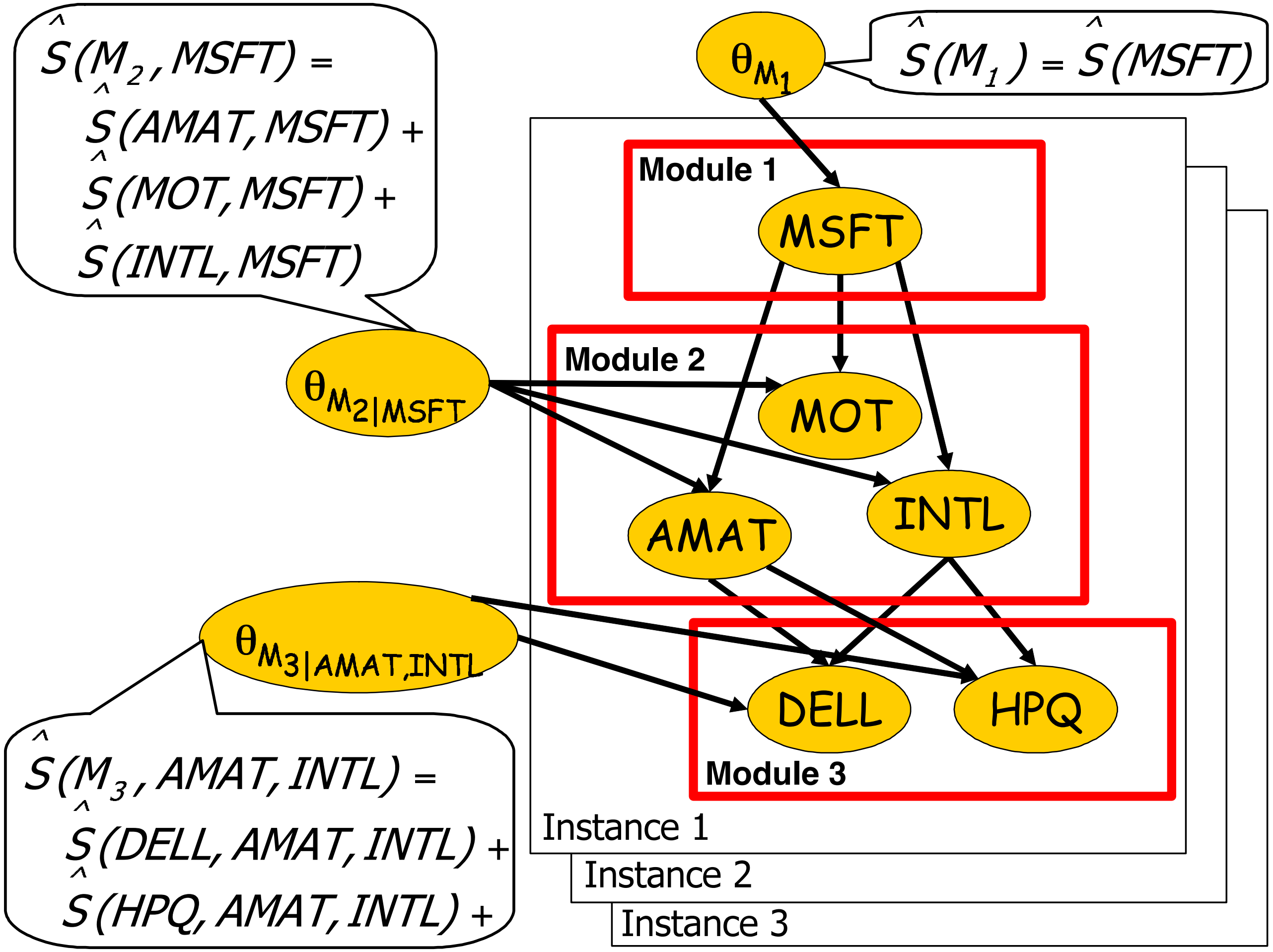
$$\boxed{\mathbf{X}^j = \{X_i \in \mathbf{X} | A(X_i) = j\}}$$

Likelihood of module j $L_j(\mathbf{Pa}_{M_j}, \mathbf{X}^j, \theta_j : \mathcal{D})$

$$L_j = \prod_{m=1}^{|\mathcal{D}|} \prod_{X_i \in \mathbf{X}^j} P(x_i[m] | \mathbf{pa}_{M_j}[m], \theta_j)$$

K: number of modules, $\boldsymbol{X^j}$: j^{th} module $\boldsymbol{Pa}_{Mj}$ Parents of module M_j

Data likelihood example



$$\hat{S}(M_2, MSFT) = \hat{S}(AMAT, MSFT) + \hat{S}(MOT, MSFT) + \hat{S}(INTL, MSFT)$$

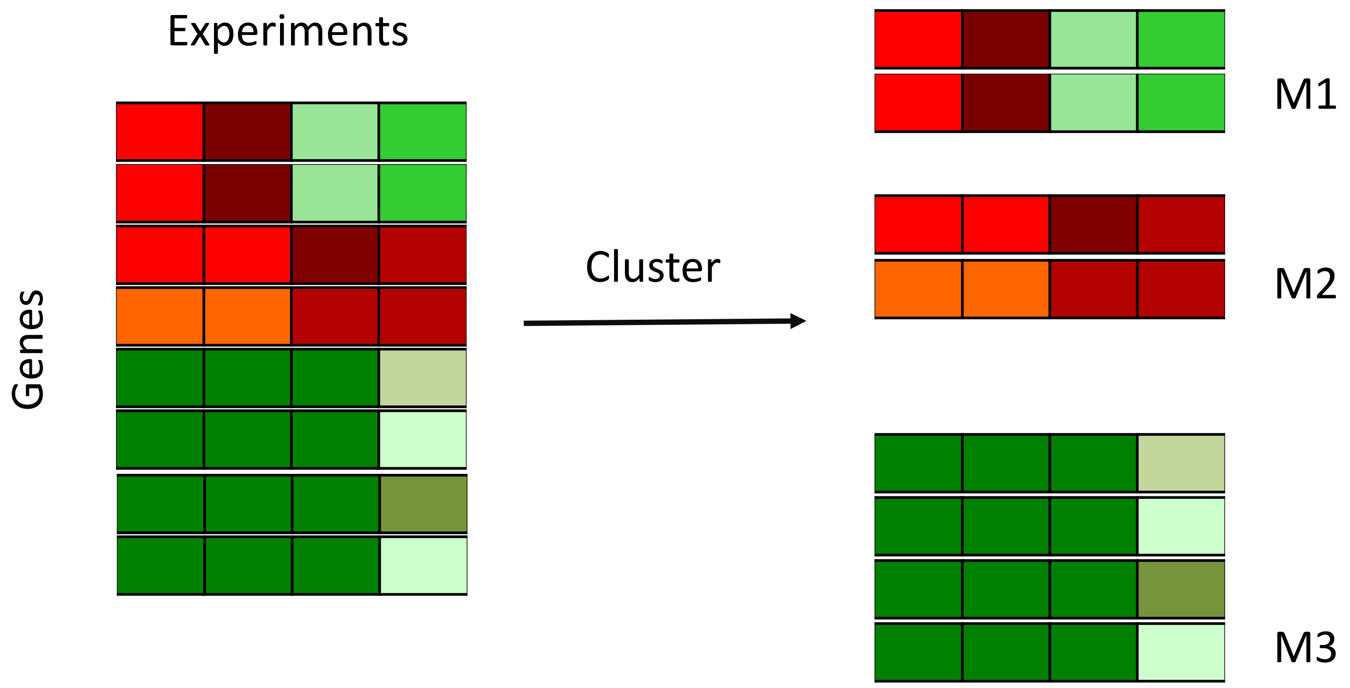
$$\hat{S}(M_1) = \hat{S}(MSFT)$$

$$\hat{S}(M_3, AMAT, INTL) = \hat{S}(DELL, AMAT, INTL) + \hat{S}(HPQ, AMAT, INTL) +$$

Module network learning algorithm

```
Input:
    $D$ // Data set
    $K$ // Number of modules
Output:
    M // A module network
Learn-Module-Network
    $\mathcal{A}_0$ = cluster $\mathcal{X}$ into $K$ modules
    $\mathcal{S}_0$ = empty structure
    Loop $t = 1, 2, \ldots$ until convergence
        $\mathcal{S}_t$ = Greedy-Structure-Search($\mathcal{A}_{t-1}, \mathcal{S}_{t-1}$)
        $\mathcal{A}_t$ = Sequential-Update($\mathcal{A}_{t-1}, \mathcal{S}_t$);
    Return $\mathbf{M} = (\mathcal{A}_t, \mathcal{S}_t)$
```

Initial modules identified by expression clustering



Experiments

Genes

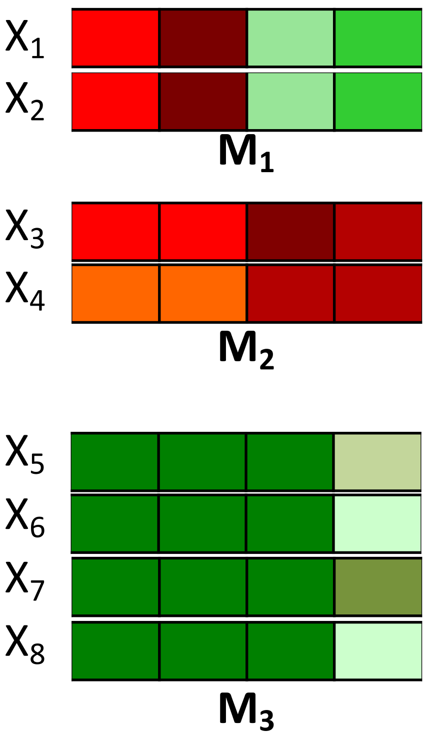
Cluster

M1

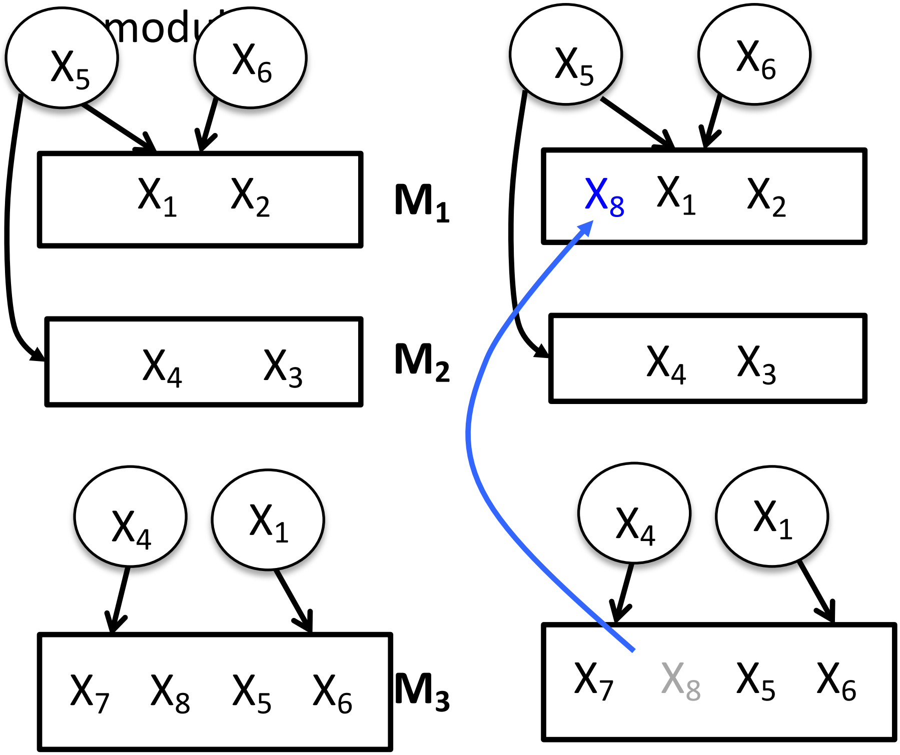
M2

M3

Iterations in learning Module Networks



Learn regulators/CPD modu



Revisit the modules

Module M_1 and M_3 get updated

Module re-assignment

- Must preserve the acyclic graph structure
- Must improve score
- Module re-assignment happens using a *sequential update* procedure:
  - Update only one variable at a time
  - The change in score of moving a variable from one module to another while keeping the other variables fixed

Module re-assignment via sequential update

```
Input:
    $D$ // Data set
    $\mathcal{A}_0$ // Initial assignment function
    $\mathcal{S}$ // Given dependency structure
Output:
    $\mathcal{A}$ // improved assignment function
Sequential-Update
    $\mathcal{A} = \mathcal{A}_0$
    Loop
        For $i = 1$ to $n$
            For $j = 1$ to $K$
                $\mathcal{A}' = \mathcal{A}$ except that $\mathcal{A}'(X_i) = j$
                If $\langle \mathcal{G}_{\mathcal{M}}, \mathcal{A}' \rangle$ is cyclic, continue
                If score$(\mathcal{S}, \mathcal{A}' : \mathcal{D}) >$ score$(\mathcal{S}, \mathcal{A} : \mathcal{D})$
                    $\mathcal{A} = \mathcal{A}'$
    Until no reassignments to any of $X_1, \ldots X_n$
    Return $\mathcal{A}$
```

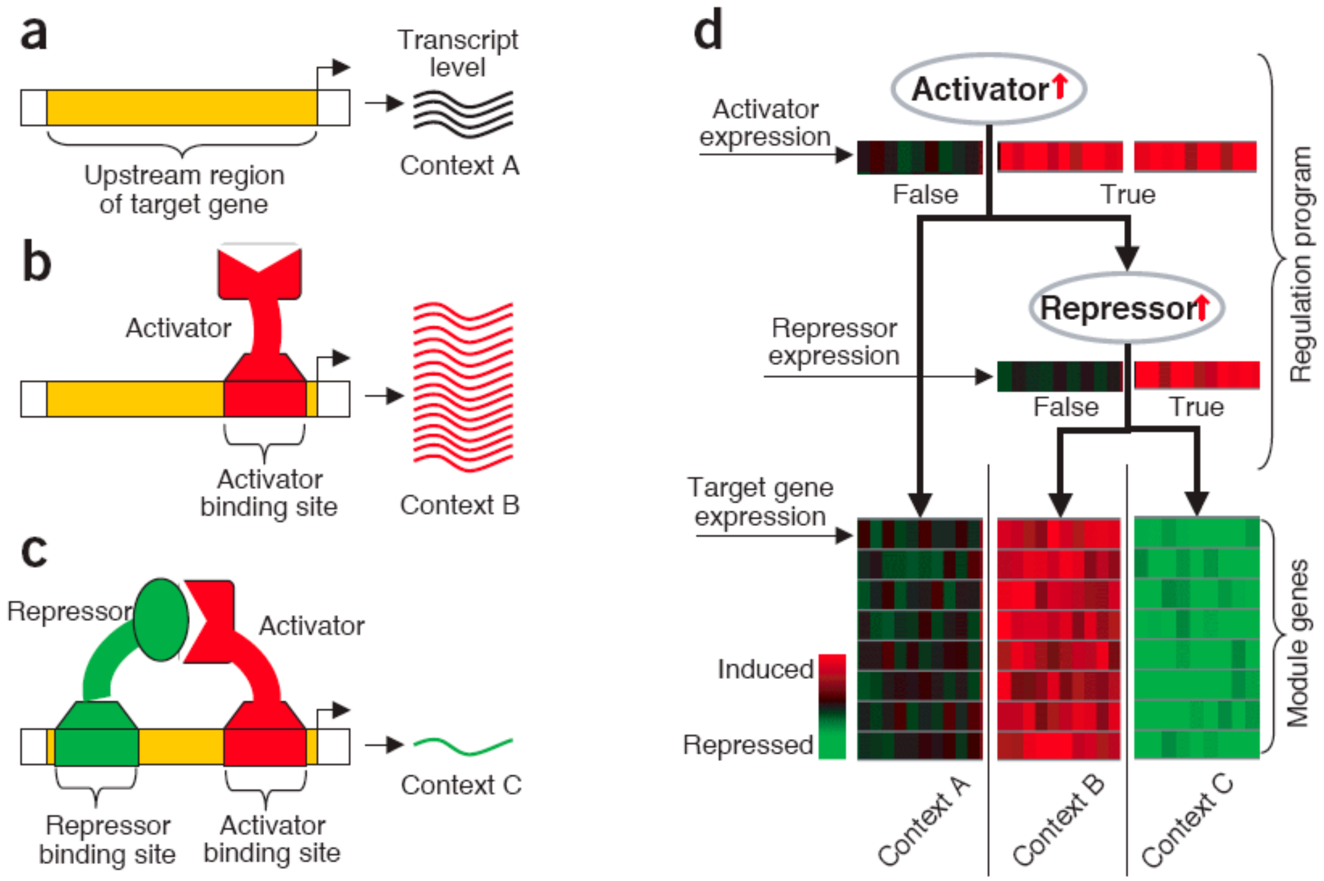
Modeling questions in Module Networks

- How to score and learn module networks?
- How to model the CPD between parent and children?
  - Regression Tree

Representing the Conditional probability distribution

- X_i are continuous variables
- How to represent the distribution of X_i given the state of its parents?
- How to capture context-specific dependencies?
- Module networks use a **regression tree**

Modeling the relationship between regulators and targets



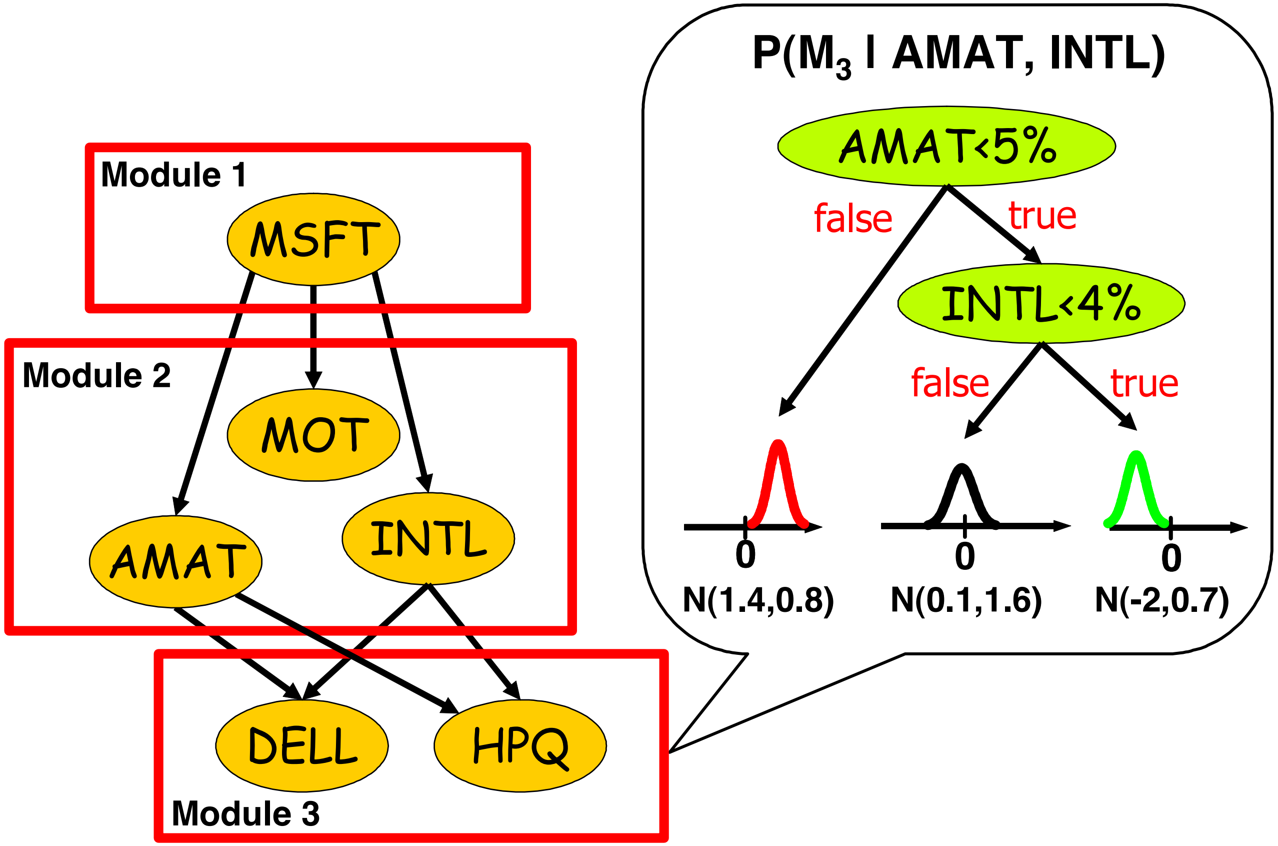
- suppose we have a set of (8) genes that all have in their upstream regions the same activator/repressor binding sites

Segal et al., Nature Genetics 2003

A regression tree

- A rooted binary tree T
- Each node in the tree is either an interior node or a leaf node
- Interior nodes are labeled with a binary test $X_i < u$, u is a real number observed in the data
- Leaf nodes are associated with univariate distributions of the child

An example regression tree for a Module network

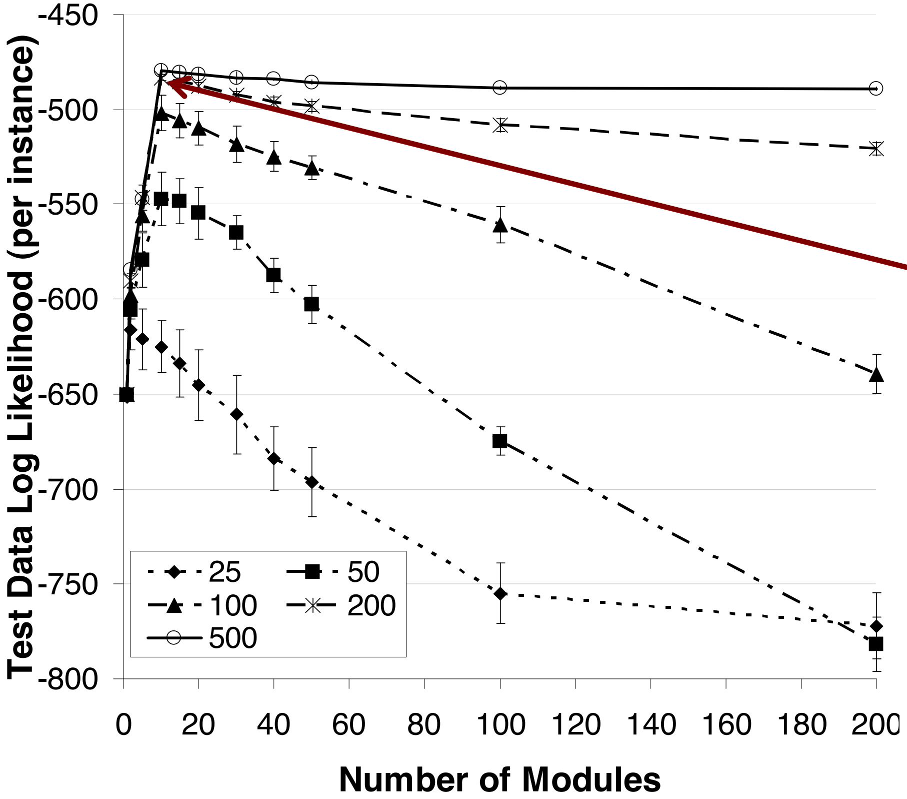


Module 3 values are modeled using Gaussians at each leaf node

Assessing the value of using Module Networks

- Using simulated data
  - Generate data from a known module network
  - Known module network was in turn learned from real data
    - 10 modules, 500 variables
  - Evaluate using
    - Test data likelihood
    - Recovery of true parent-child relationships are recovered in learned module network
- Using gene expression data
  - External validation of modules (Gene ontology, motif enrichment)
  - Cross-check with literature

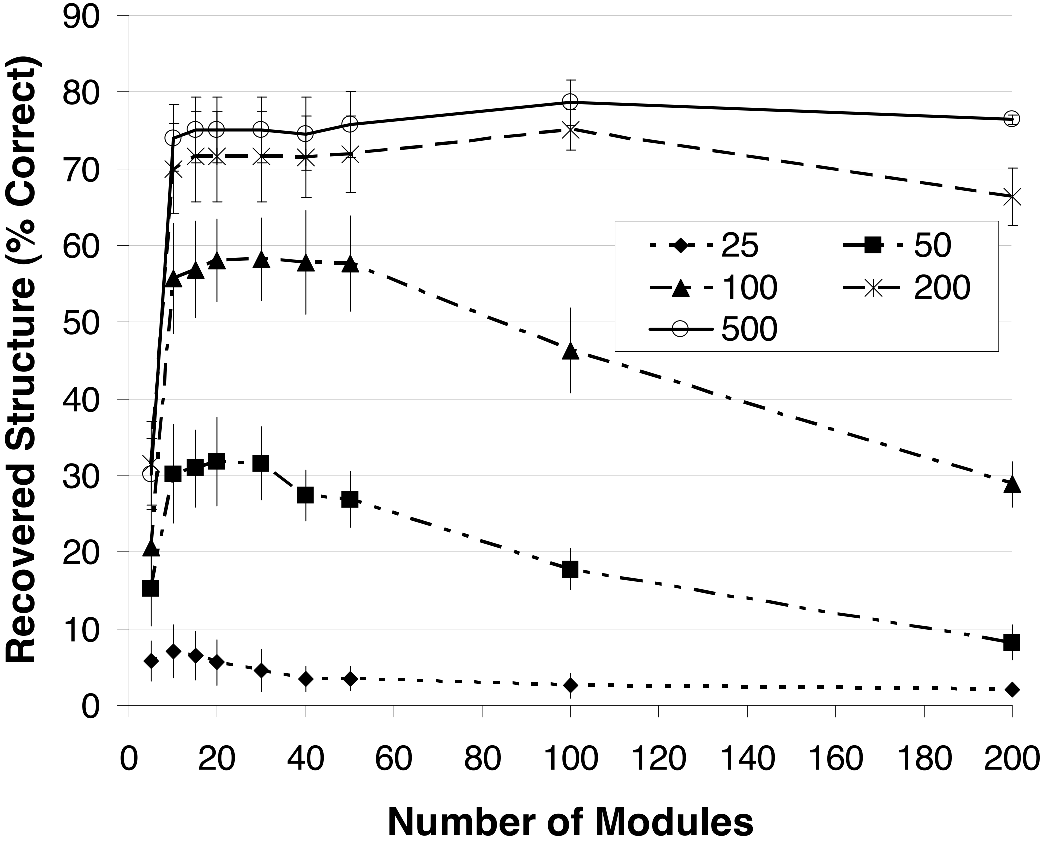
Test data likelihood



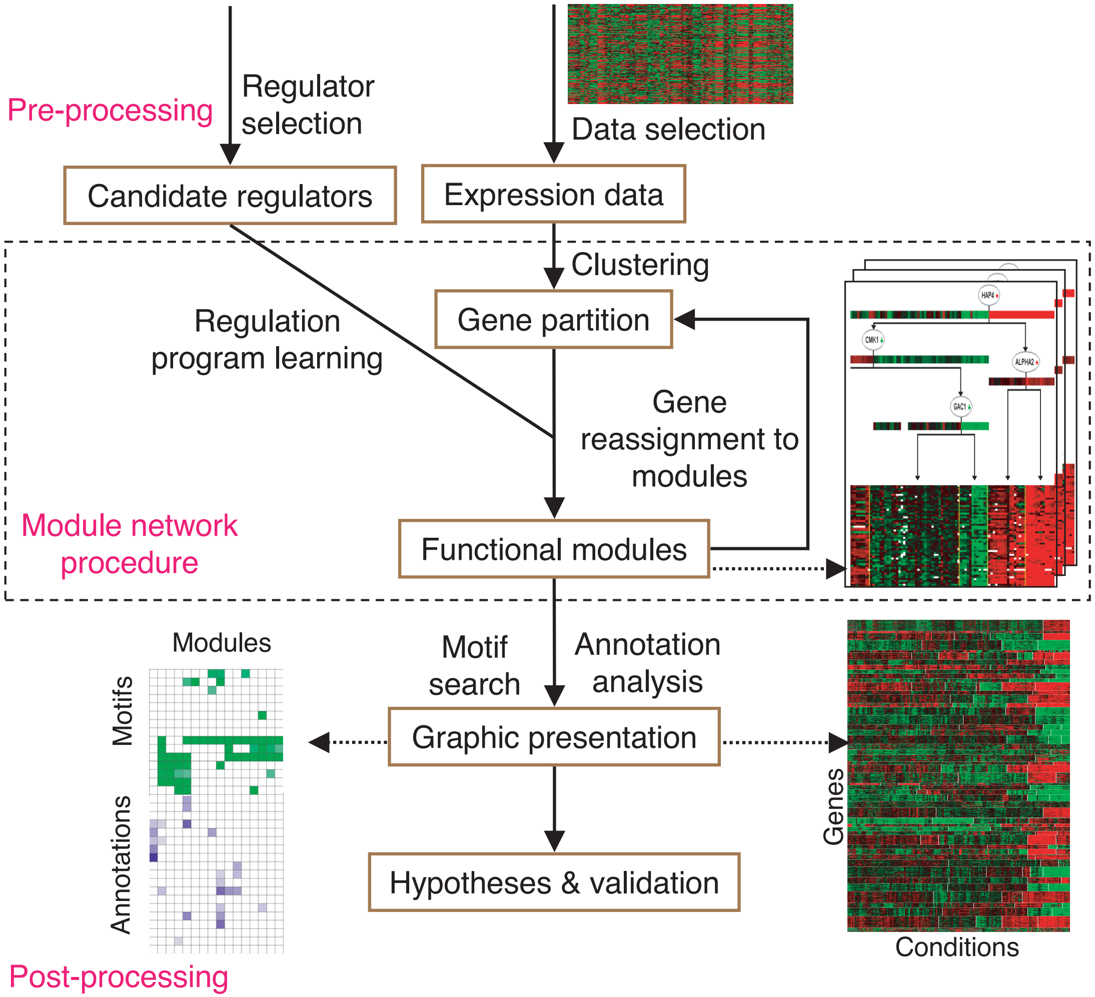
10 Modules is the best for almost all training data set sizes

Each line type represents size of training data

Recovery of graph structure



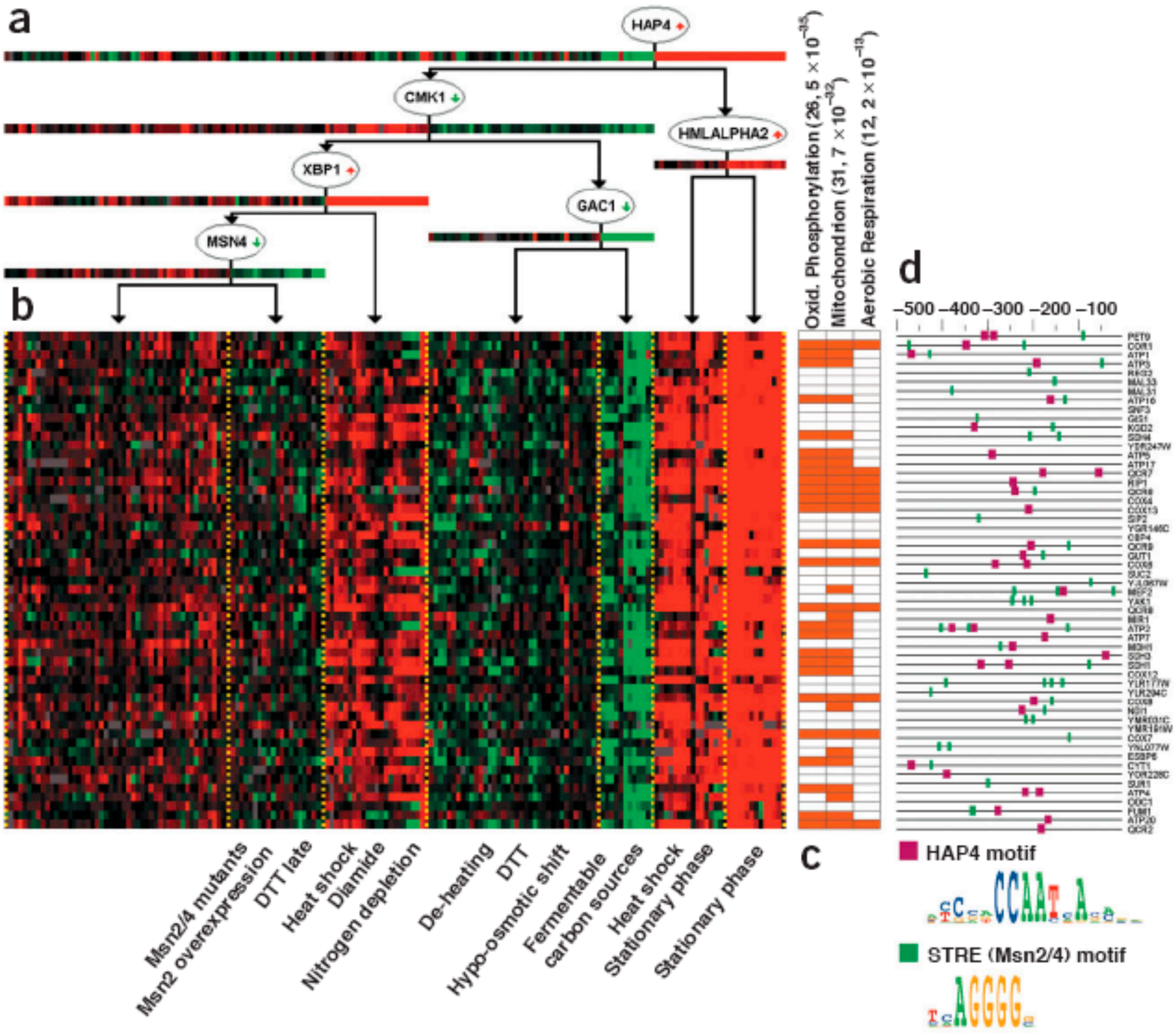
Application of Module networks to yeast expression data



Segal et al, Regev, Pe'er, Gasch, Nature Genetics 2003

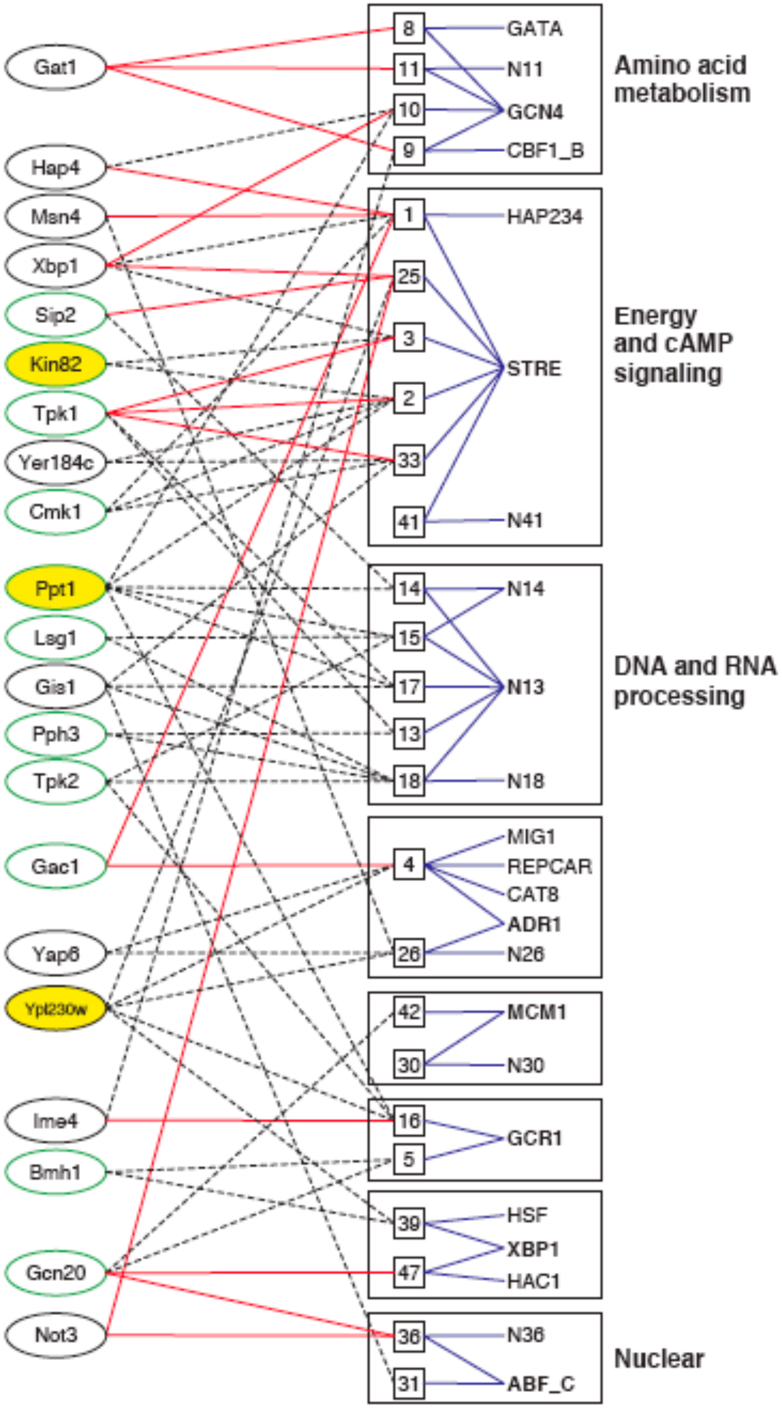
The Respiration and Carbon Module

Regression tree representing rules of regulation



Global View of Modules

- modules for common processes often share common
  - regulators
  - binding site motifs



□ Module (number)

○ Regulator (signaling molecule)

○ Regulator (transcription factor)

┆ Inferred regulation

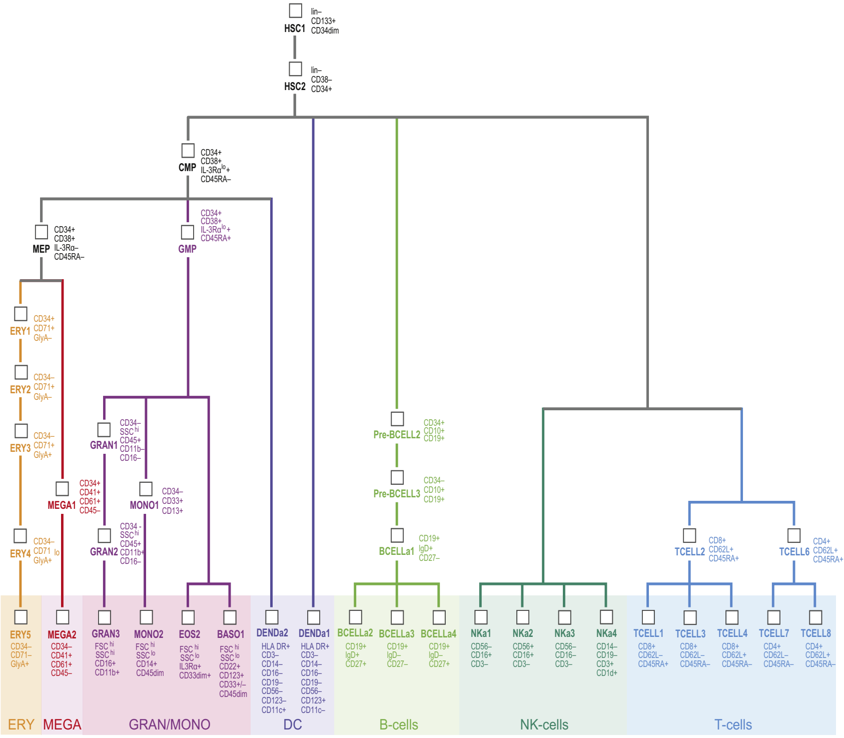
| Regulation supported in literature

| Enriched *cis*-regulatory motif

● Experimentally tested regulator

Application of Module networks to mammalian data

- Module networks have been applied to mammalian systems as well
- We will look at a case-study in the human blood cell lineage
- Dataset
  - Genome-wide expression levels in 38 hematopoietic cell types (211 samples)
  - 523 candidate regulators (Transcription factors)

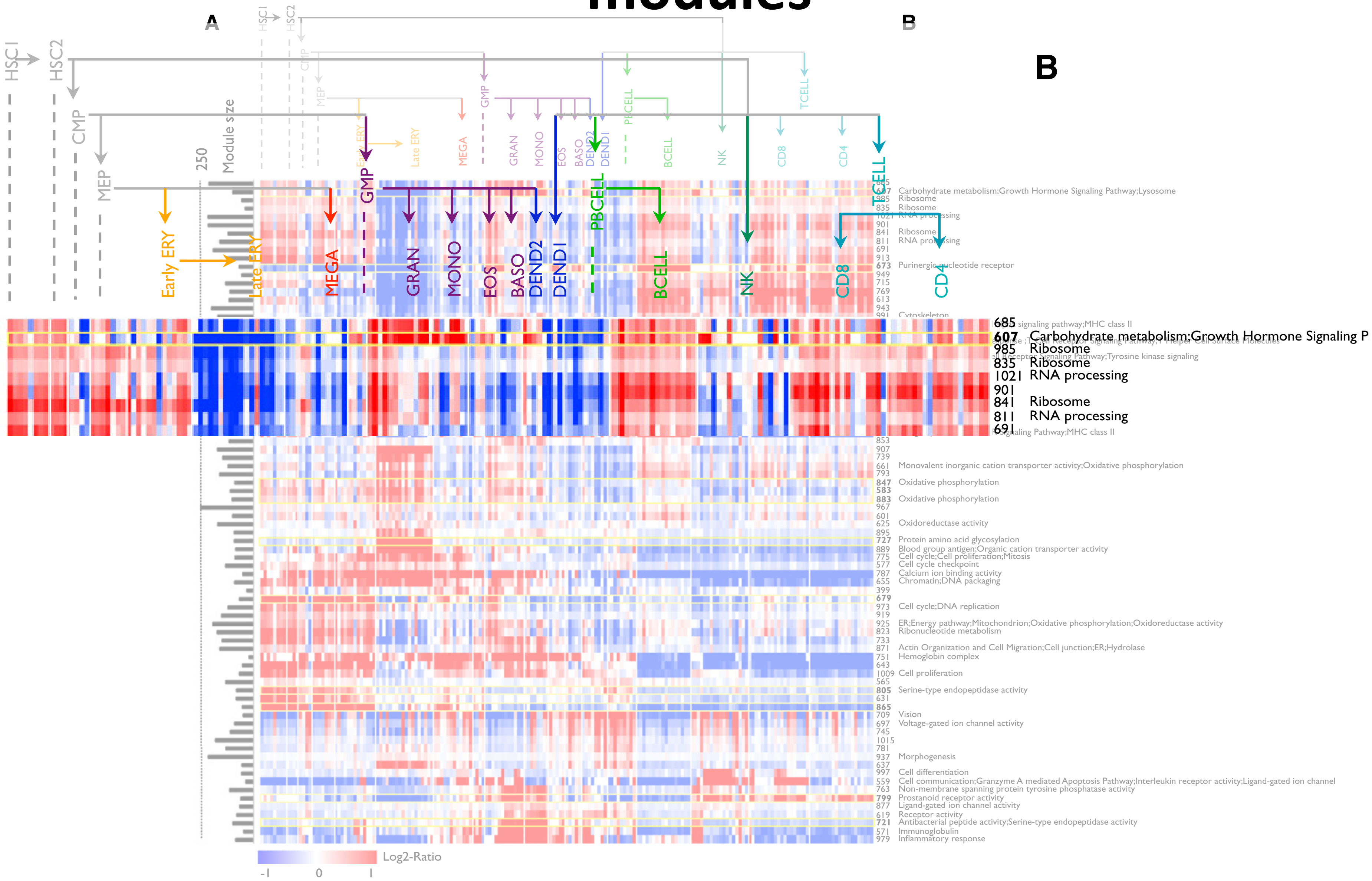


HSC1 lin– CD133+ CD34dim
HSC2 lin– CD38– CD34+
CMP CD34+ CD38+ IL-3Rα lo + CD45RA–
MEP CD34+ CD38+ IL-3Rα– CD45RA–
GMP CD34+ CD38+ IL-3Rα lo + CD45RA+
ERY1 CD34+ CD71+ GlyA–
ERY2 CD34– CD71+ GlyA–
ERY3 CD34– CD71+ GlyA+
ERY4 CD34– CD71 lo GlyA+
MEGA1 CD34+ CD41+ CD61+ CD45–
GRAN1 CD34– SSC hi CD45+ CD11b– CD16–
GRAN2 CD34 - SSC hi CD45+ CD11b+ CD16–
MONO1 CD34– CD33+ CD13+
Pre-BCELL2 CD34+ CD10+ CD19+
Pre-BCELL3 CD34– CD10+ CD19+
BCELLa1 CD19+ IgD+ CD27–
TCELL2 CD8+ CD62L+ CD45RA+
TCELL6 CD4+ CD62L+ CD45RA+
ERY5 CD34– CD71– GlyA+
MEGA2 CD34– CD41+ CD61+ CD45–
GRAN3 FSC hi SSC hi CD16+ CD11b+
MONO2 FSC hi SSC lo CD14+ CD45dim
EOS2 FSC hi SSC lo IL3Rα+ CD33dim+
BASO1 FSC hi SSC lo CD22+ CD123+ CD33+/– CD45dim
DENDa2 HLA DR+ CD3– CD14– CD16– CD19– CD56– CD123– CD11c+
DENDa1 HLA DR+ CD3– CD14– CD16– CD19– CD56– CD123+ CD11c–
BCELLa2 CD19+ IgD+ CD27+
BCELLa3 CD19+ IgD– CD27–
BCELLa4 CD19+ IgD– CD27+
NKa1 CD56– CD16+ CD3–
NKa2 CD56+ CD16+ CD3–
NKa3 CD56– CD16– CD3–
NKa4 CD14– CD19– CD3+ CD1d+
TCELL1 CD8+ CD62L– CD45RA+
TCELL3 CD8+ CD62L– CD45RA–
TCELL4 CD8+ CD62L+ CD45RA–
TCELL7 CD4+ CD62L– CD45RA–
TCELL8 CD4+ CD62L+ CD45RA–
ERY
MEGA
GRAN/MONO
DC
B-cells
NK-cells
T-cells

Human hematopoetic lineage

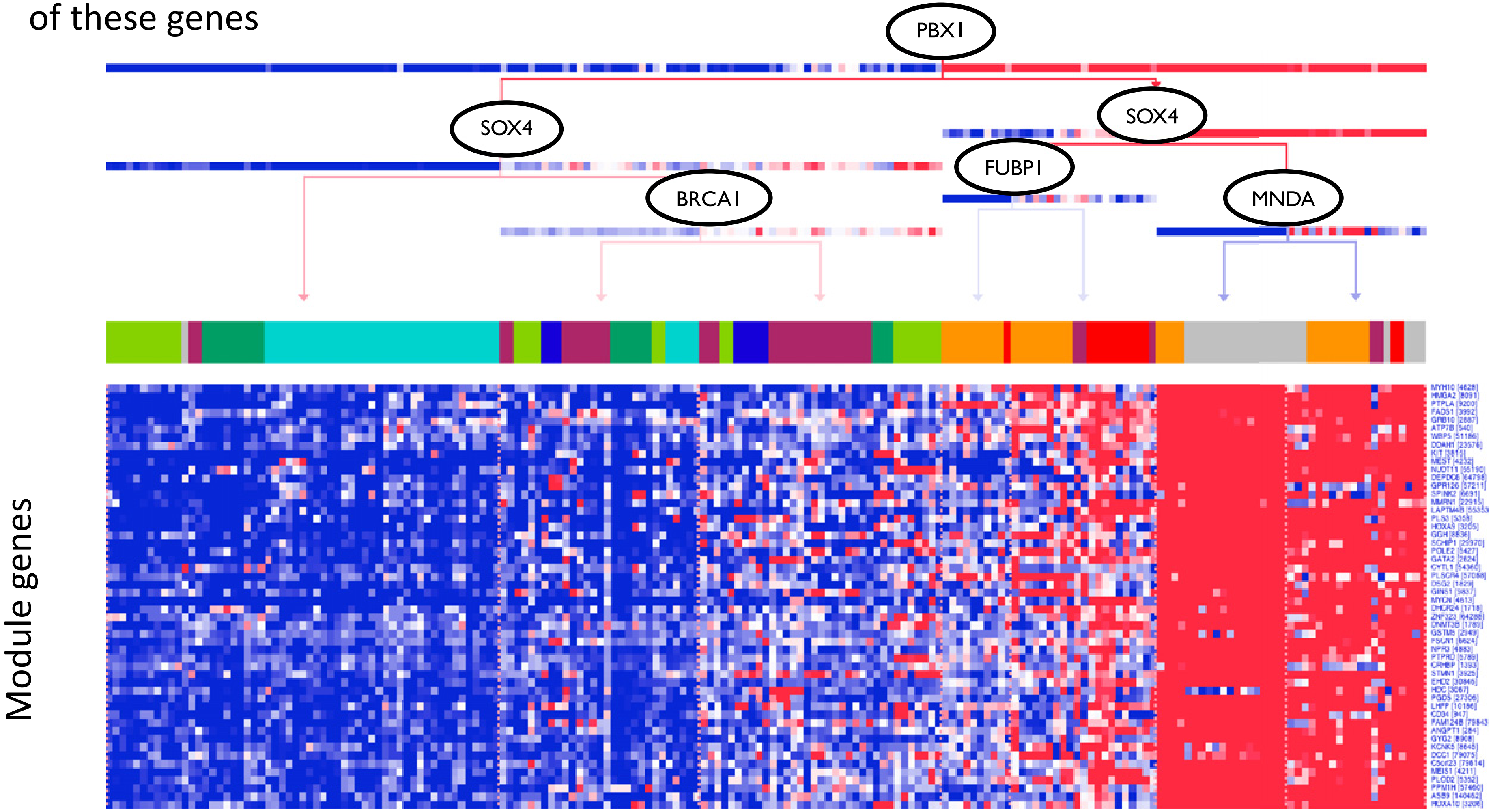
Novershtern et al., Cell 2011

Expression profiles of 80 transcriptional modules



An HSCs, MEPs, and Early Erythroid-Induced Module

PBX1, SOX4 need to be high and MNDA need to be low for the highest expression of these genes



Other key points from this analysis

- Many novel regulators associated with the hematopoietic lineage
- Several regulators were validated based on shRNA and ChIP-seq analysis

Extensions to module networks

- Physical module networks
  - – Novershtern et al., Bioinformatics 2011
- Integrating sequence variants with expression modules
  - – Lee et al., PLOS Genetics 2009
- Combining module networks with per-gene methods
  - – Roy et al., PLOS computational biology 2013

Limitations with Bayesian networks

- Cannot model cyclic dependencies
- In practice have not been shown to be better than dependency networks
  - However, most of the evaluation has been done on structure not function
- Directionality is often not associated with causality
  - Too many hidden variables in biological systems

Take away points

- Network inference from expression provides a promising approach to identify cellular networks
- Graphical models are one representation of networks that have a probabilistic and graphical component
  - Network inference naturally translates to learning problems in these models
- Bayesian networks were among the first type of PGMs for representing networks
- Applying Bayesian networks to expression data required several additional considerations
  - Too few samples: Sparse candidates, Module networks
  - Too many parents: Sparse candidates
  - Imposing modularity: Module networks

Plan for next lectures

- Gaussian graphical models
- Dependency networks
  - GENIE3

References

- Kim, Harold D., Tal Shay, Erin K. O'Shea, and Aviv Regev. "Transcriptional Regulatory Circuits: Predicting Numbers from Alphabets." *Science* 325 (July 2009): 429-432.
- De Smet, Riet and Kathleen Marchal. "Advantages and limitations of current network inference methods.." *Nature reviews. Microbiology* 8 (October 2010): 717-729.
- Markowetz, Florian and Rainer Spang. "Inferring cellular networks-a review." *BMC bioinformatics* 8 Suppl 6 (2007): S5+.
- N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601-620, Aug. 2000. [Online]. Available: http://dx.doi.org/10.1089/106652700750050961
- E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman, "Learning module networks," *Journal of Machine Learning Research*, vol. 6, pp. 557-588, Apr. 2005. [Online]. Available: http://www.jmlr.org/papers/volume6/segal05a/segal05a.pdf
- E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, no. 2, pp. 166-176, May 2003. [Online]. Available: http://dx.doi.org/10.1038/ng1165
- N. Novershtern, et al., "Densely interconnected transcriptional circuits control cell states in human hematopoiesis." *Cell*, vol. 144, no. 2, pp. 296-309, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2011.01.004