Probabilistic Graphical Models in Systems Biology

Nir Friedman
Hebrew University

Includes slides by:

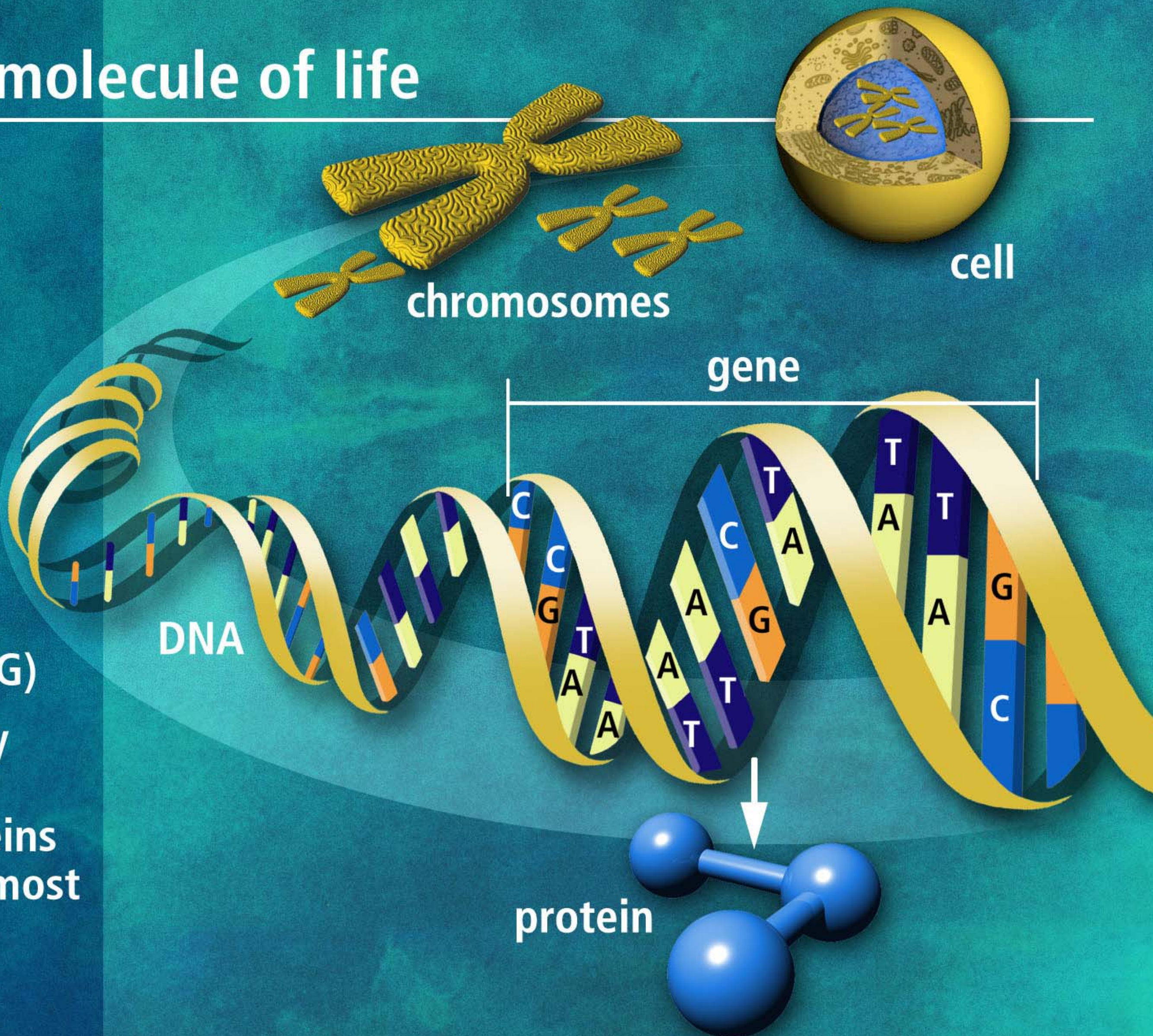
Yoseph Barash, Nebojsa Jojic, Ariel Jaimovich, Tommy Kaplan, Daphne Koller, Iftach Nachman, Dana Pe’er, Tal Pupko, Aviv Regev, Eran Segal

DNA the molecule of life

Trillions of cells

Each cell:

- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- Approximately 30,000 genes code for proteins that perform most life functions

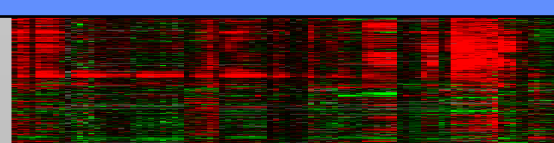


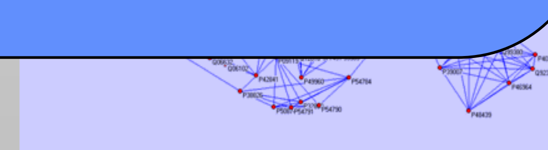
Y-GG 01-0085

Challenges of The Post-Genome Era

High-throughput assays:

- Observations about **one** aspect of the system
- Often **noisy** and less reliable than traditional assays
- Provide partial account of the system





Challenges of The Post-Genome Era

Issues:

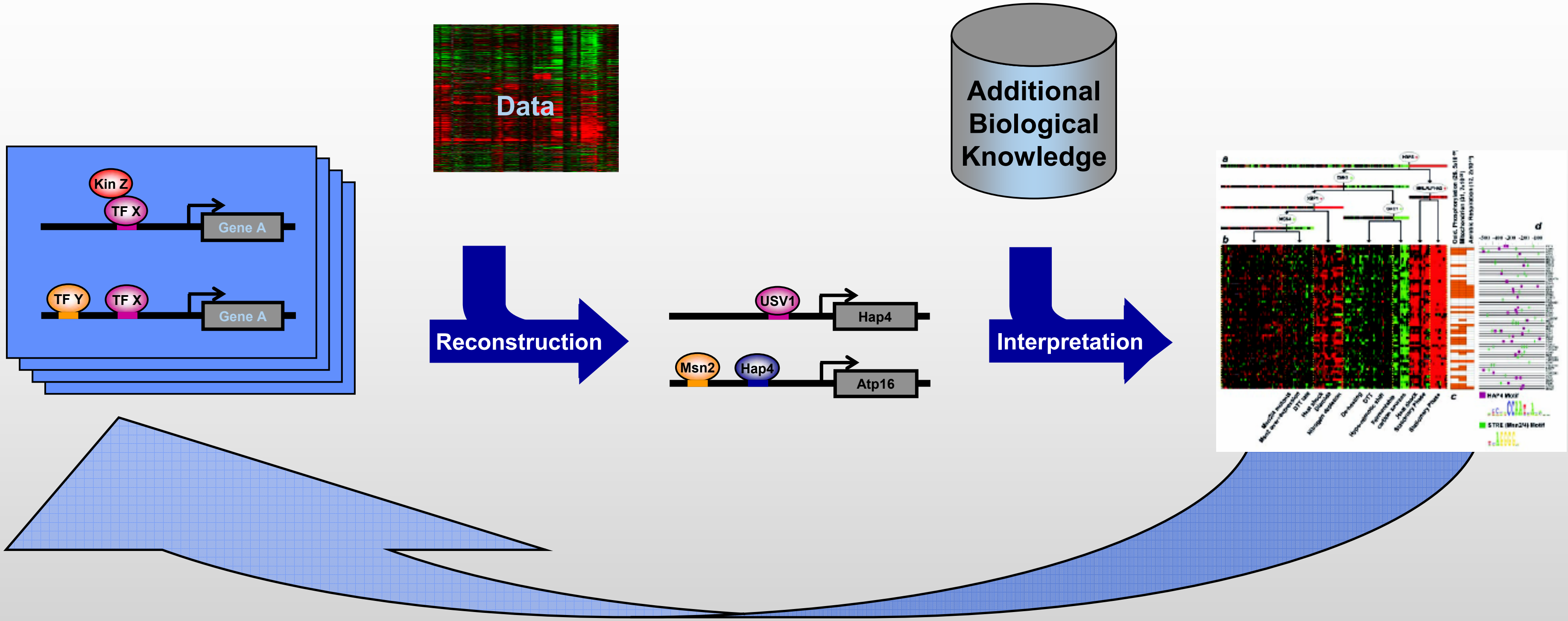
- Measurement noise
  - $\Rightarrow$ Conclusions supported by more than one assay
- Each assay provides a view of a single aspect
  - $\Rightarrow$ Combine multiple types of assays for more coherent reconstruction
- Combinatorial explosion of assay combinations
  - $\Rightarrow$ Principles for integrating results from new assays

Solution Strategies

Procedural

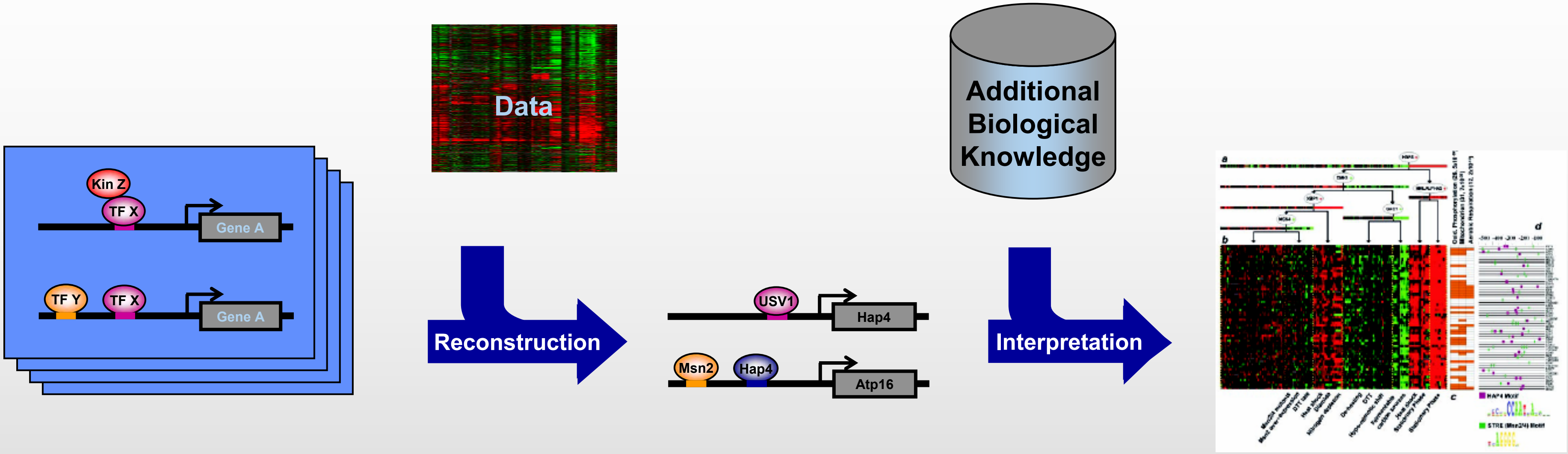
- Specify a set of steps for reaching biological conclusions from experimental data
  - Example
    - Cluster gene expression profiles
    - Search for enriched motif in each cluster
    - …
- emphasis on the computational procedure and the order of data manipulation steps

Model Based Approach



- **Step 1:** define class of potential models
- **Step 2**: reconstruct a specific model
- **Step 3**: visualization & testable hypotheses
- Emphasis on the choice of model and how to use it
  - The data manipulation steps are derived from the model

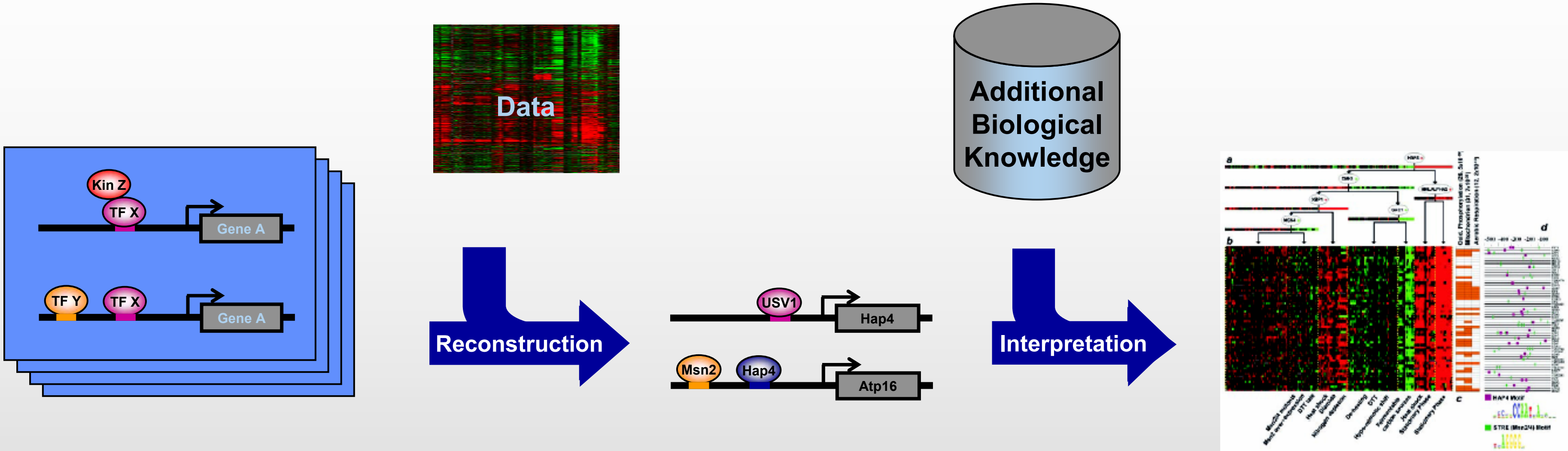
Model Based Approach



Representation – defining the class of models

- What entities to involve
- Model granularity
- Identifiably

Model Based Approach



Interpretation – what do they tell us about system

- Relation between components in the model to biological entities/mechanisms
- What predictions can be made with the model

Why Model-Based?

Declarative

- Explicit statement of the assumptions made
- Closer connection between biological principles and the solution
- Decouple the “what” (model) from the “how”

Flexibility

- Can use different computational approaches to perform the task specified by the model

Reusability

- Modifications & extensions are specified the level of the model

Stochastic Models

Use the probability theory to describe the system

- **State of the system**: assignment of value to all the attributes of all the relevant entities
- A **distribution** over these states describe which states are achievable and which ones are abnormal

Extensions:

- Modeling inputs: interventions, conditions
- Modeling outputs: phenotype, behavior, assays

Why Stochastic Models?

- Inherent noise in the system
- Uncertainty due to granularity of the model
- Noise in sensors
- Imperfect modeling --- noise as slack variable

What Can We Do with a Model?

- **Inference**
  - Set some evidence, compute posterior over unobserved variables

- **Estimation/Learning**
  - “Fill in the gaps” in the model based on empirical data

The Representation Hurdle

- Joint distributions grow large
  - Exponential in the number of attributes
  - Problem for inference & learning

We need to find **compact** representation

Strategy:

- Impose constraints
- Exploit these constraints for compact representation

Probabilistic Graphical Models

- Language(s) for representing complex joint distributions
- Generic methods for performing tasks with these representations

In this tutorial we will examine these in the context of modeling biological systems.

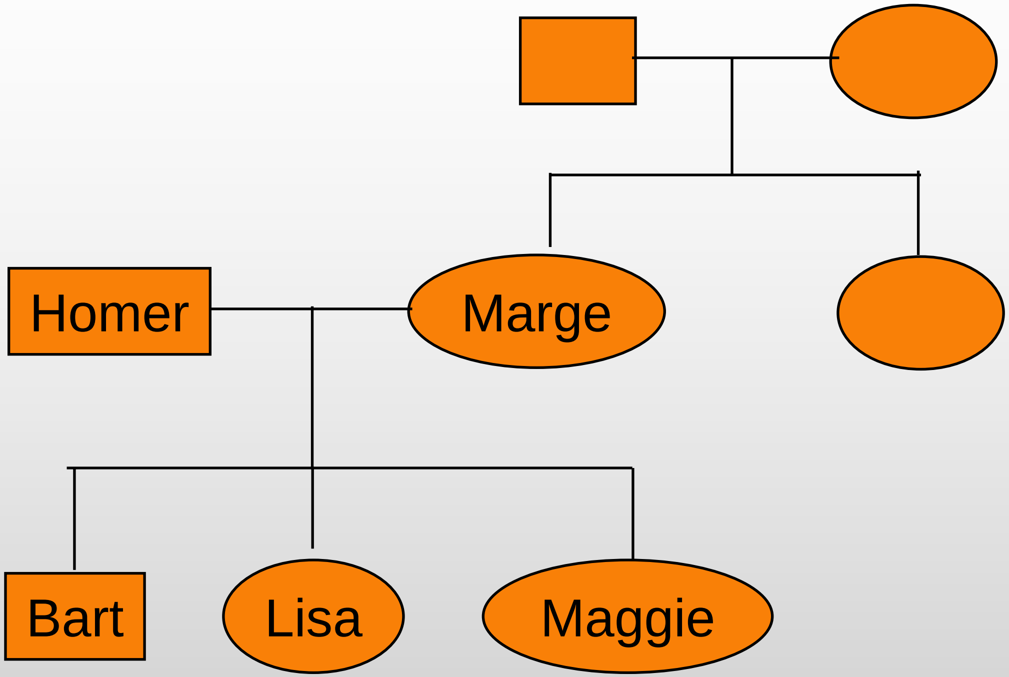
Outline

- Introduction
- Bayesian Networks
- Learning Bayesian Networks
- Transcriptional regulation
- Gene expression
- Markov Networks
- Protein-Protein Interactions
- Discussion

Bayesian Networks by Example

Example: Pedigree

- A node represents an individual's genotype



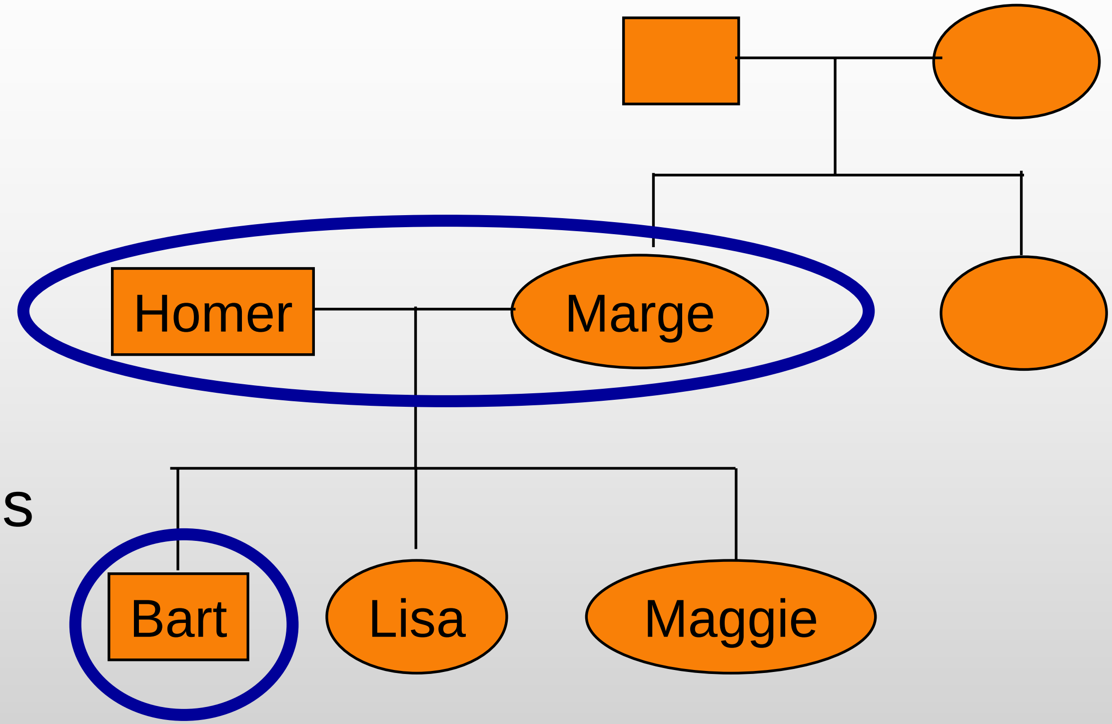
Joint distribution

$$P(G_{Bart}, G_{Lisa}, G_{Maggie}, G_{Homer}, G_{Marge}, \ldots)$$
$$= P(G_{Bart} \mid G_{Lisa}, G_{Maggie}, G_{Homer}, G_{Marge}, \ldots)$$
$$P(G_{Lisa} \mid G_{Maggie}, G_{Homer}, G_{Marge}, \ldots)$$
$$\ldots$$

Bayesian Networks by Example

Modeling assumption:

- Ancestors can effect descendants' genotype only by passing genetic materials through intermediate generations

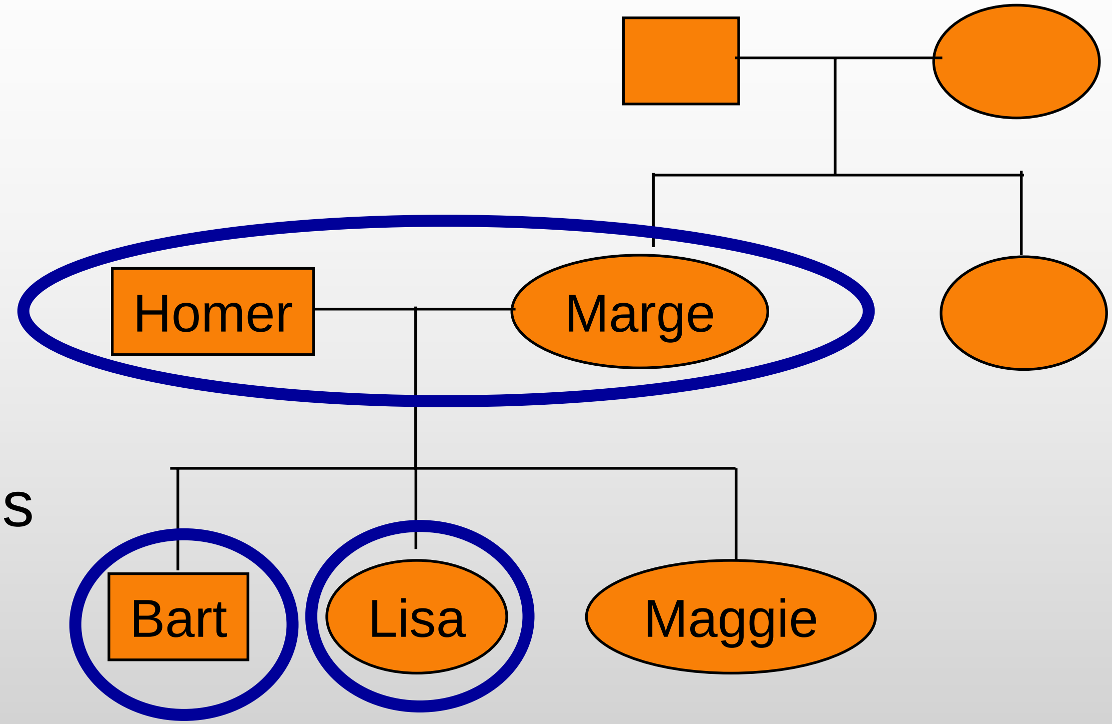


$$
\begin{aligned}
&P(G_{Bart}, G_{Lisa}, G_{Maggie}, G_{Homer}, G_{Marge}, \ldots) \\
&\quad = P(G_{Bart} \mid G_{Lisa}, G_{Maggie}, G_{Homer}, G_{Marge}, \ldots) \\
&\quad\quad P(G_{Lisa} \mid G_{Maggie}, G_{Homer}, G_{Marge}, \ldots) \\
&\quad\quad \ldots
\end{aligned}
$$

Bayesian Networks by Example

Modeling assumption:

- Ancestors can effect descendants' genotype only by passing genetic materials through intermediate generations



$$P(G_{\text{Bart}}, G_{\text{Lisa}}, G_{\text{Maggie}}, G_{\text{Homer}}, G_{\text{Marge}}, \ldots)$$

$$= P(G_{\text{Bart}} \mid G_{\text{Homer}}, G_{\text{Marge}})$$

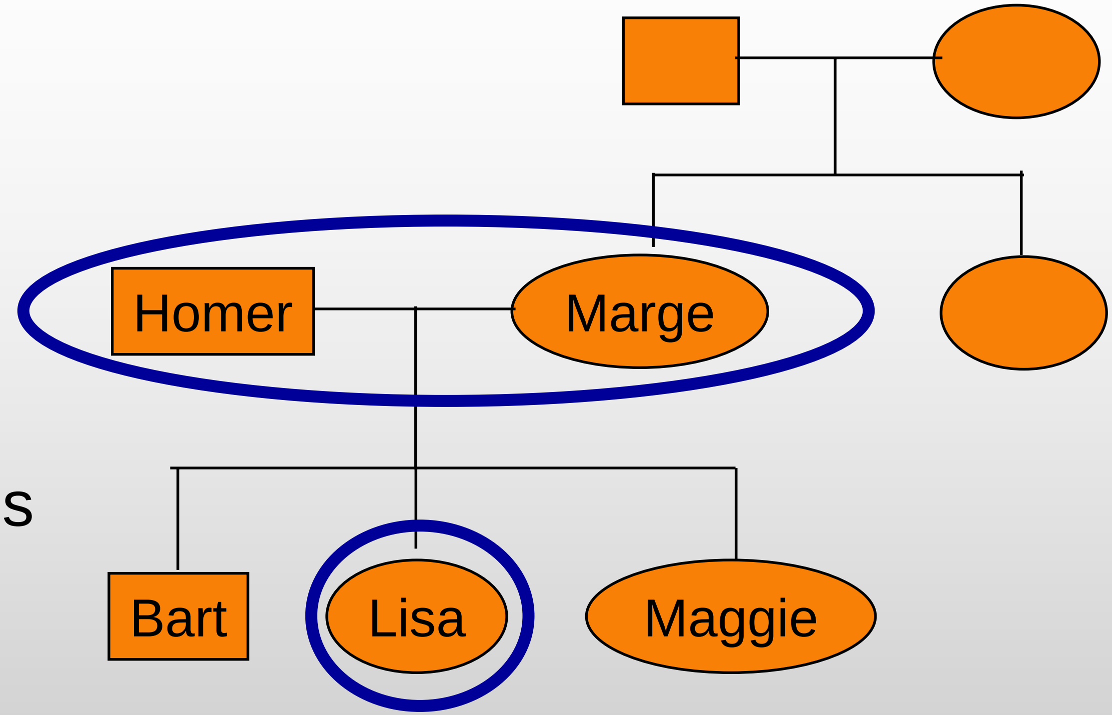
$$P(G_{\text{Lisa}} \mid G_{\text{Maggie}}, G_{\text{Homer}}, G_{\text{Marge}}, \ldots)$$

$$\ldots$$

Bayesian Networks by Example

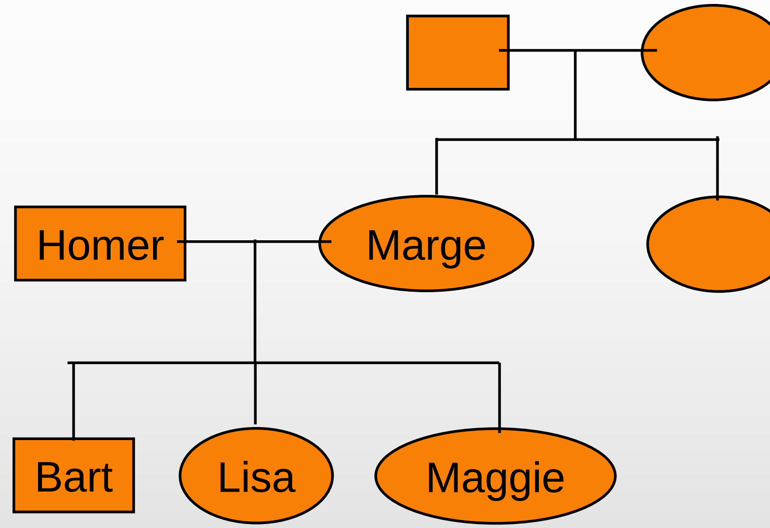
Modeling assumption:

- Ancestors can effect descendants' genotype only by passing genetic materials through intermediate generations



$$P(G_{Bart}, G_{Lisa}, G_{Maggie}, G_{Homer}, G_{Marge}, \ldots)$$
$$= P(G_{Bart} \mid G_{Homer}, G_{Marge})$$
$$P(G_{Lisa} \mid G_{Homer}, G_{Marge})$$
$$\ldots$$

Bayesian Networks by Example



Extending this argument, we can derive a functional form for general pedigrees

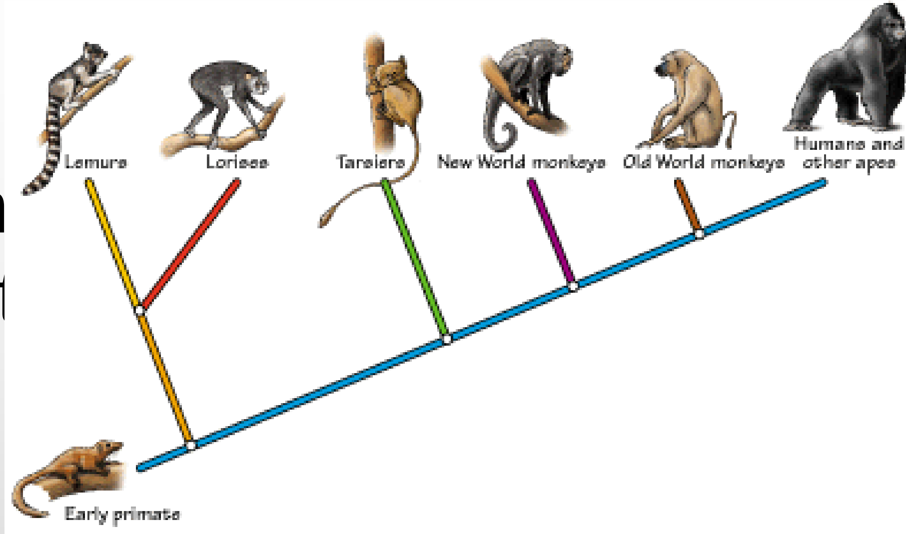
Descendants

$$P(G_1, G_2, \ldots) = \left(\prod_{j \in Ancestors} P(G_j) \right) \left(\prod_{i \in Descendants} P(G_i \mid G_{father(i)}, G_{mother(i)}) \right)$$

Probability of genetic transmission within family

Probability of random genotype in population

Bayesian Networks by Example II



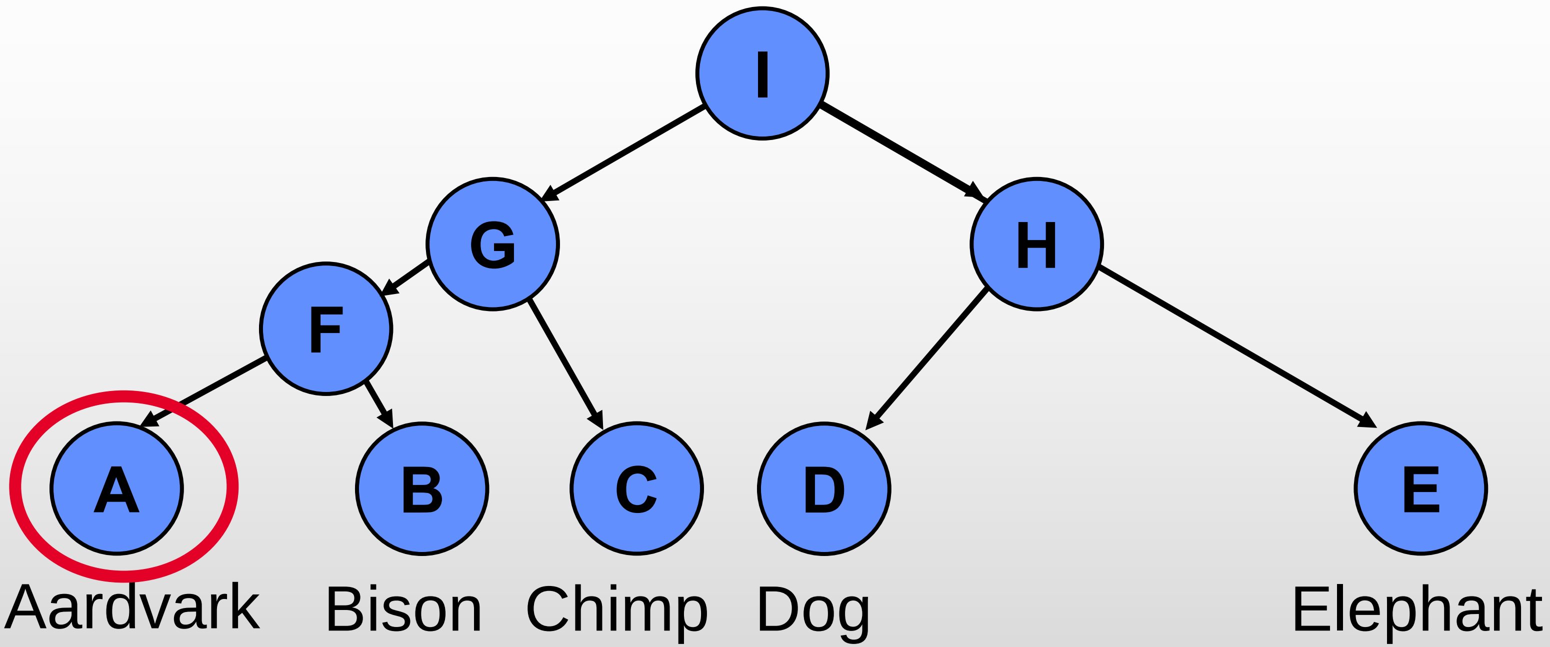
Sequence evolution

- Each random variable is th sequence of a taxa (ancest or current day)

Assumption (neutral changes):

- Past history does not affect how the sequence will change in the future

Bayesian Networks by Example II



$$P(S_A, S_B, S_C, \ldots, S_I) = P(S_A \mid S_B, S_C, \ldots, S_I)$$

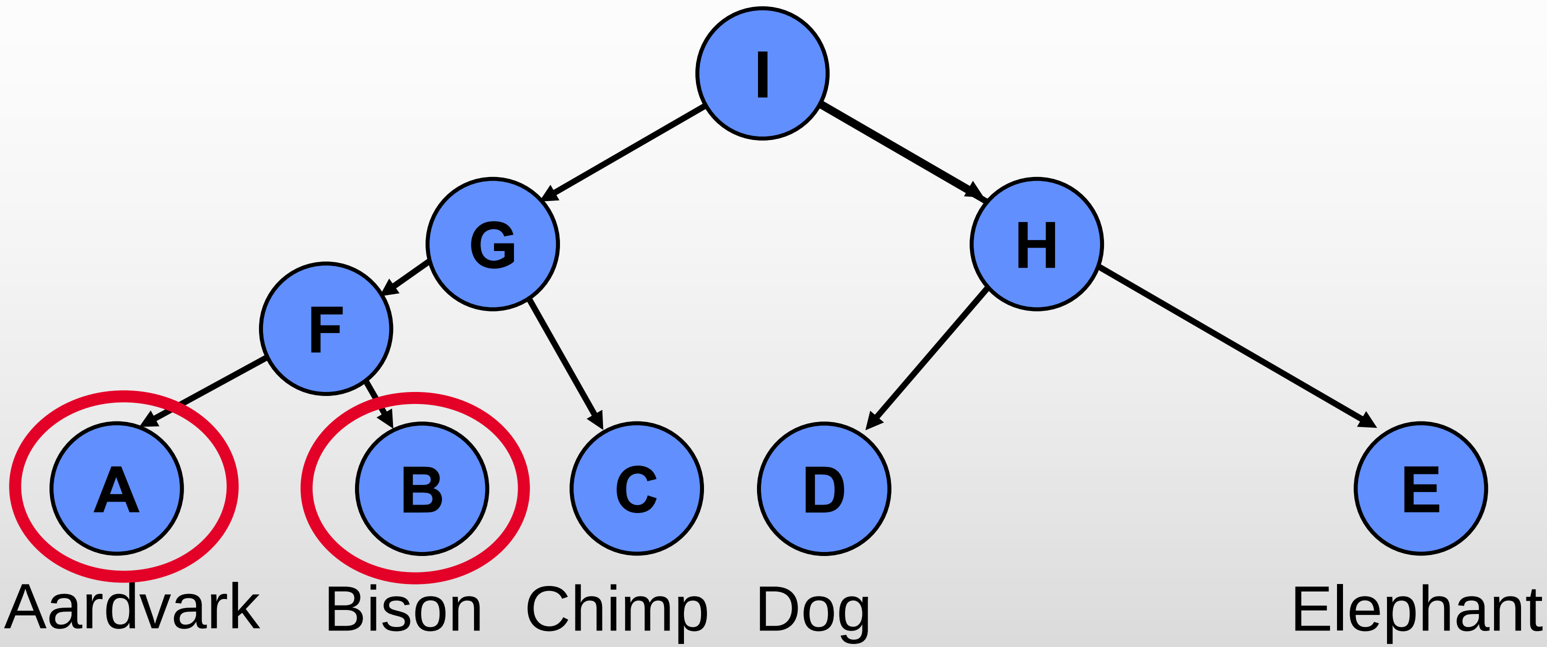
$$P(S_B \mid S_C, \ldots, S_I)$$

$$\ldots$$

$$P(S_F \mid S_G, S_H, S_I)$$

$$\ldots$$

Bayesian Networks by Example II



$$P(S_A, S_B, S_C, \ldots, S_I) = P(S_A \mid S_F)$$

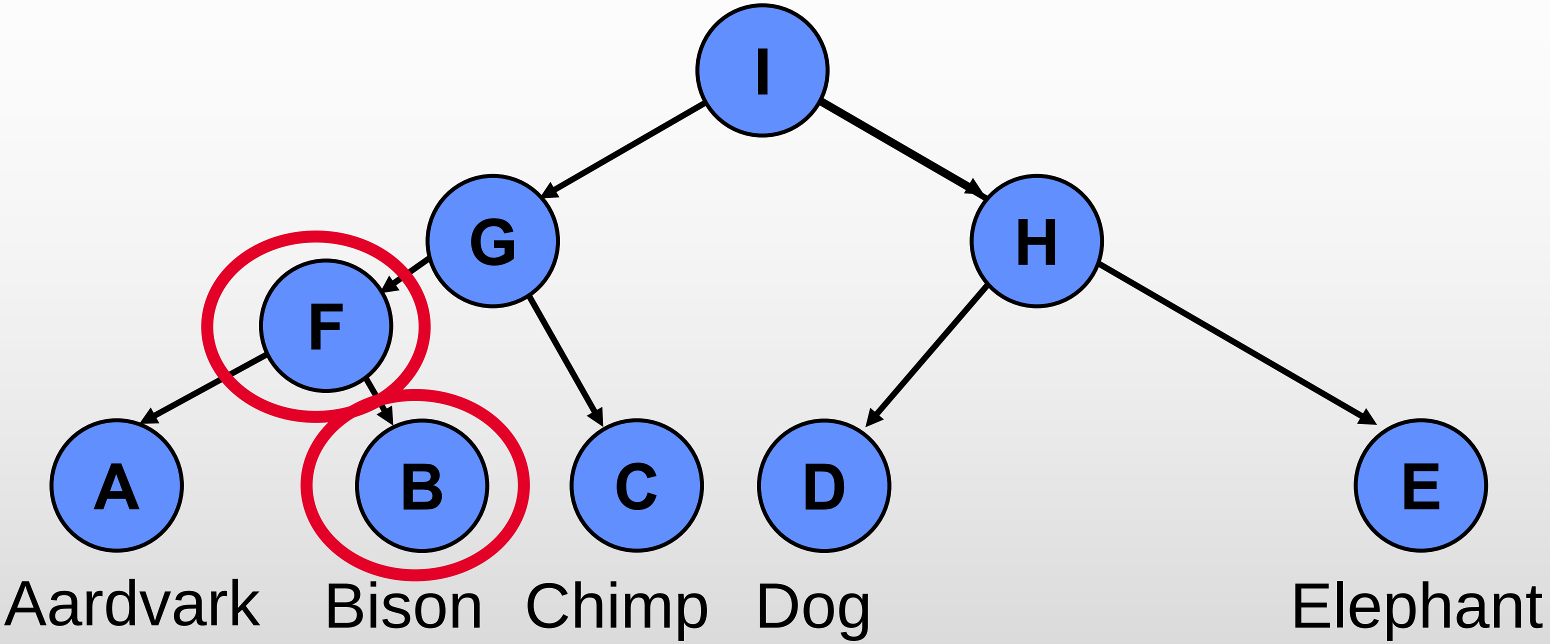
$$P(S_B \mid S_C, \ldots, S_I)$$

$$\ldots$$

$$P(S_F \mid S_G, S_H, S_I)$$

$$\ldots$$

Bayesian Networks by Example II



$$P(S_A, S_B, S_C, \ldots, S_I) = P(S_A \mid S_F)$$

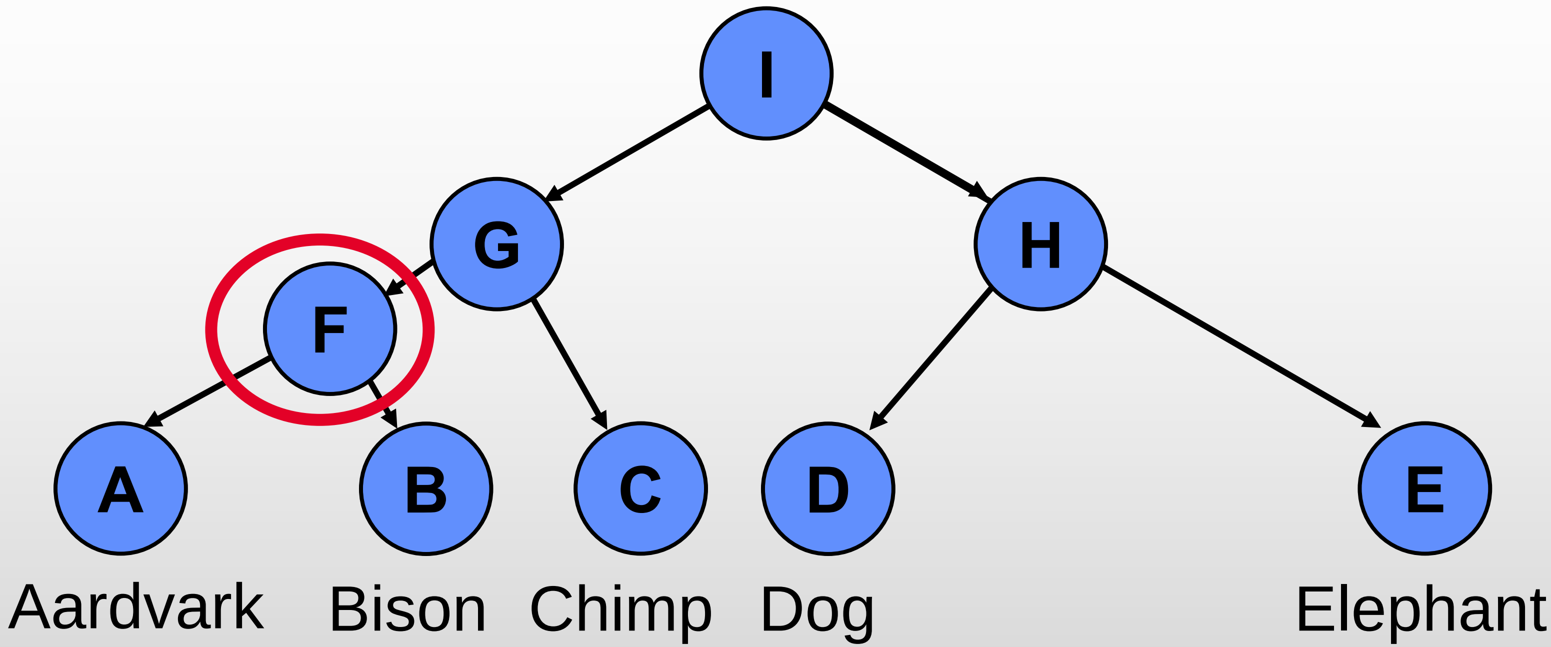
$$P(S_B \mid S_F)$$

$$\ldots$$

$$P(S_F \mid S_G, S_H, S_I)$$

$$\ldots$$

Bayesian Networks by Example II



$$P(S_A, S_B, S_C, \ldots, S_I) = P(S_A \mid S_F)$$
$$P(S_B \mid S_F)$$
$$\ldots$$
$$P(S_F \mid S_G)$$
$$\ldots$$

Probability of mutations over the given time period

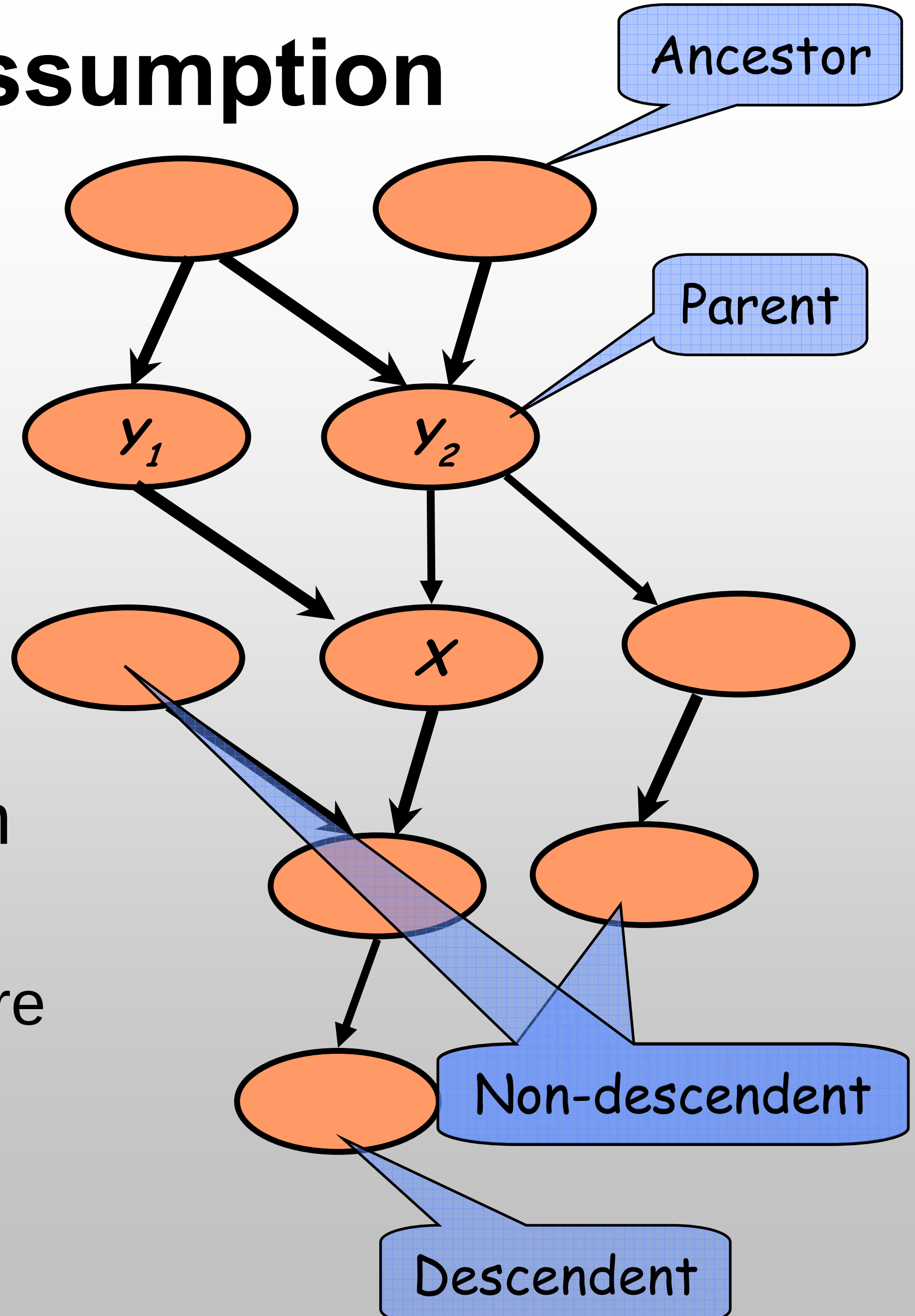
Markov Assumption

Generalizing to DAGs:

- A child is **conditionally independent** from its non-descendents, given the value of its parents

Often a natural assumption for **causal** processes

- if we believe that we capture the relevant state of each intermediate stage



Bayesian Networks

$$P(A,B,C,D,E) = \begin{array}{l} P(A) \\ P(B \mid A) \\ P(C \mid A, B) \\ P(D \mid A, B, C) \\ P(E \mid A, B, C, D) \end{array}$$

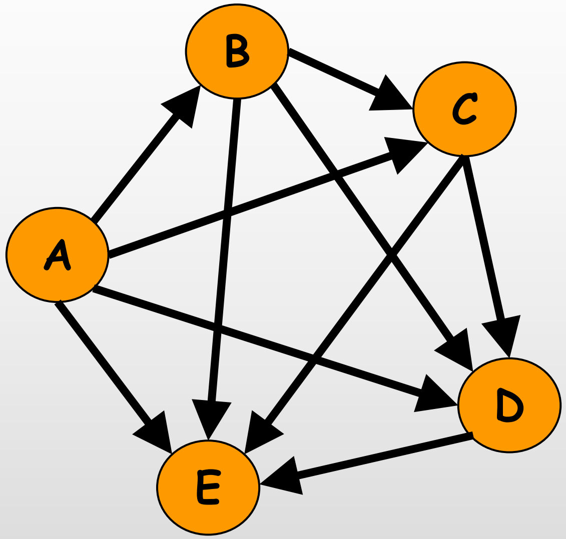
Bayesian Networks

Ind(C ; B | A)

Ind(D ; A, B | C)

Ind(E ; B, C | A, D)

…



$$P(A,B,C,D,E) = P(A)\, P(B \mid \cancel{A})\, P(C \mid \cancel{A}, \cancel{B})\, P(D \mid \cancel{A}, \cancel{B}, \cancel{C})\, P(E \mid \cancel{A}, \cancel{B}, \cancel{C}, \cancel{D})$$

Bayesian Networks

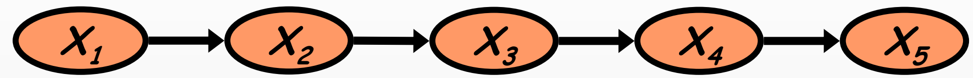
- Flexible language to capture a range

  Maximal independence → Full dependence

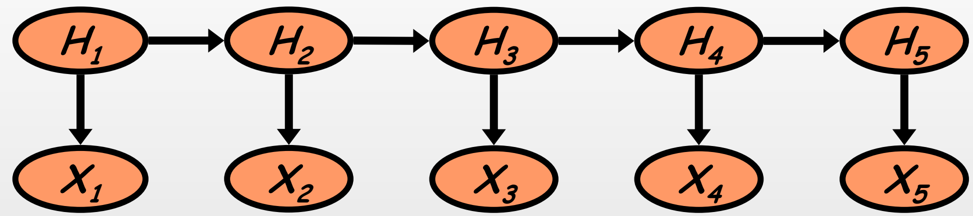
- Formal correspondence between
  - Acyclic directed graph structure
  - Factorization of joint distribution as a product of conditional probabilities
  - A set of (conditional) independence statements

Example Structures

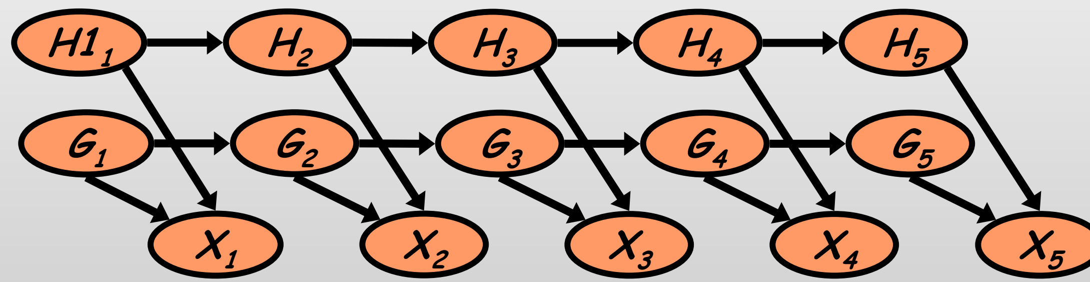
- Markov chain



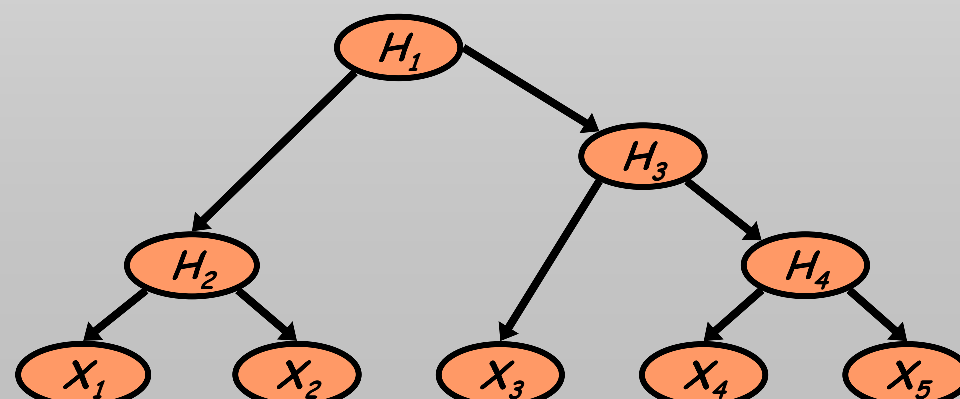
- Hidden Markov Model (HMM)



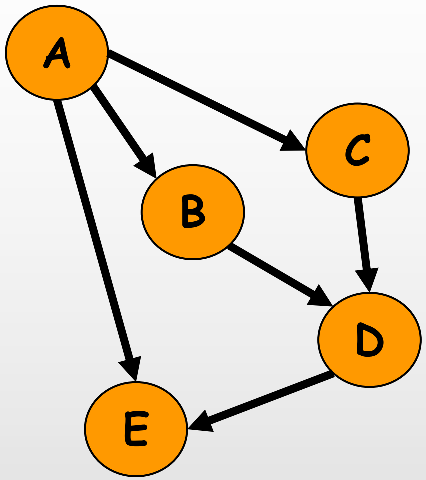
- Factorial HMM



- Tree



Local Probability Models



Bayesian Network Structure

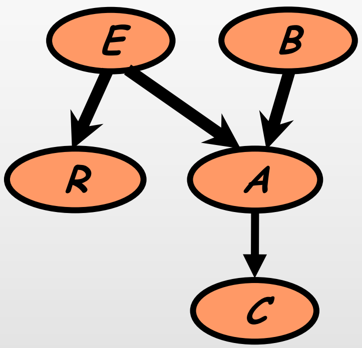
$\Rightarrow$Simpler product form

$$P(A,B,C,D,E) = P(A)\,P(B \mid A)P(C \mid A)P(D \mid B, C)P(E \mid A, D)$$

To specify a distribution we need to supply these conditional probabilities

- Describe to “local” stochastic effects

Bayesian Network Semantics



Qualitative part
conditional independence statements in BN structure

\+

Quantitative part
local probability models

=

Unique joint distribution over domain

Compact & efficient representation:

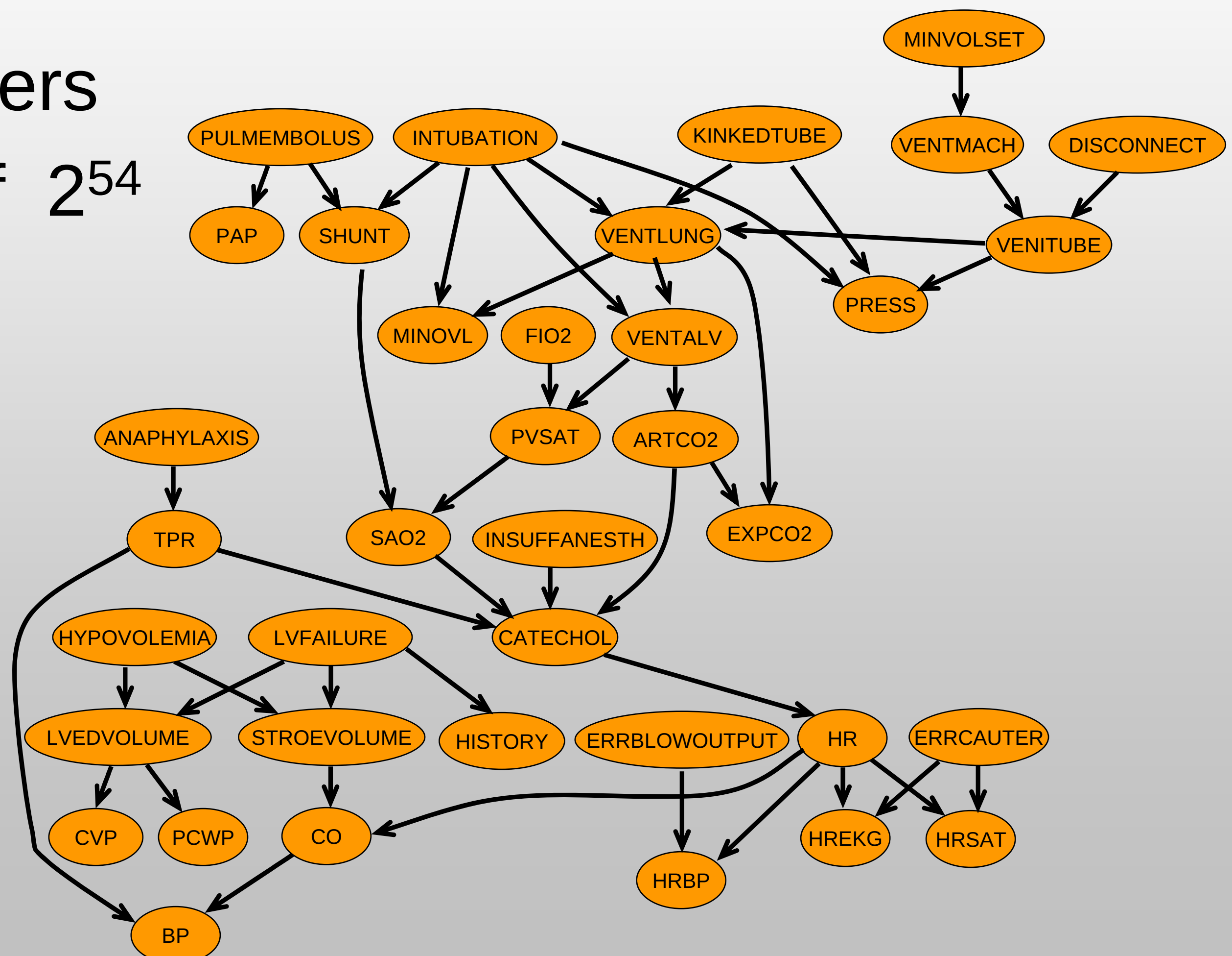
- nodes have $\leq$ k parents $\Rightarrow$ $O(2^k n)$ vs. $O(2^n)$ params

Example: “ICU Alarm” network

Domain: Monitoring Intensive-Care Patients

- 37 variables
- 509 parameters

  …instead of 2^{54}



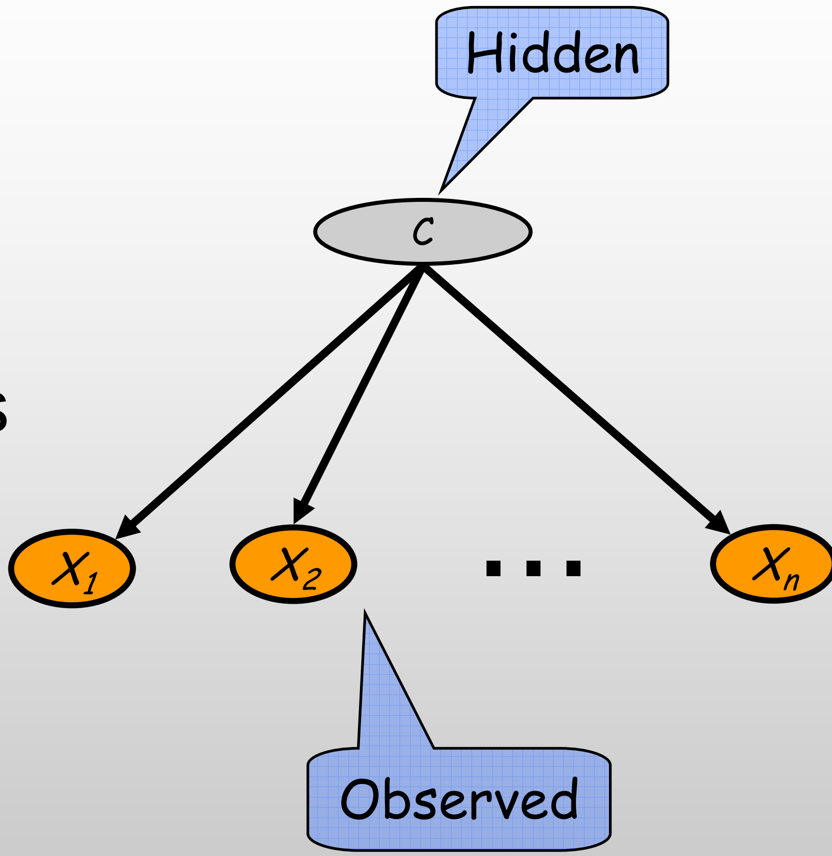
Hidden Variable(s)

A simple model of clustering

- C - gene's cluster
- $X_1, \ldots, X_n$ - expression of the gene in different experiments

Independence assumption:

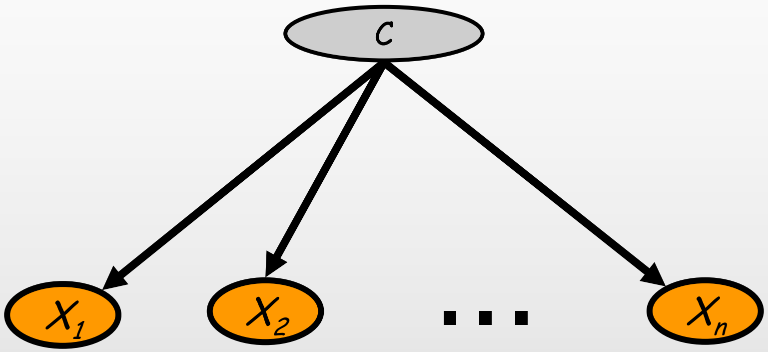
- $I(X_i; X_j \mid C)$



Hidden Variable(s)

Marginal distribution:

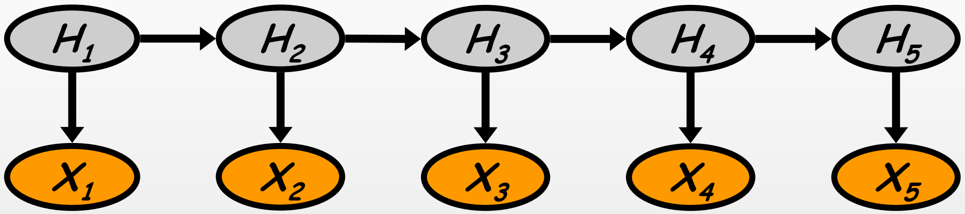
$$P(X_1,\ldots,X_n) = \sum_c P(X_1 \mid c)\cdots P(X_n \mid c)P(c)$$



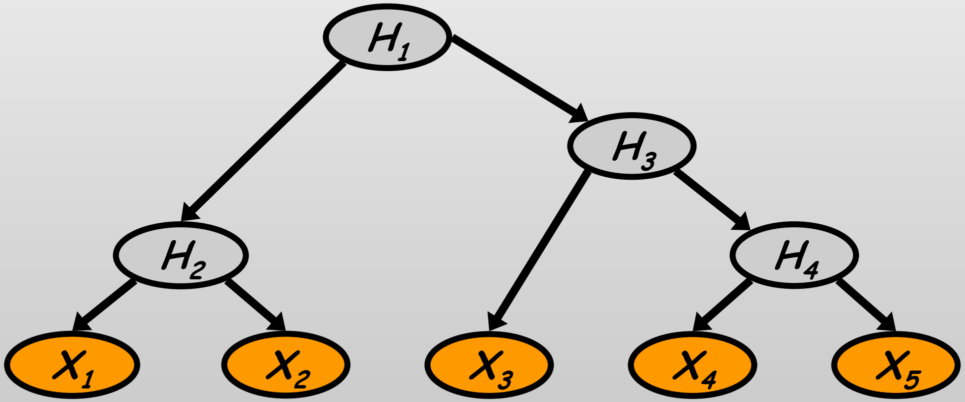
- Compact representation $(n+1)k$ vs. 2^n params
- No conditional independencies in the **marginal** distribution
- The variable *C* “channels” the dependencies between observed variables

Hidden Variables

Hidden Markov Model



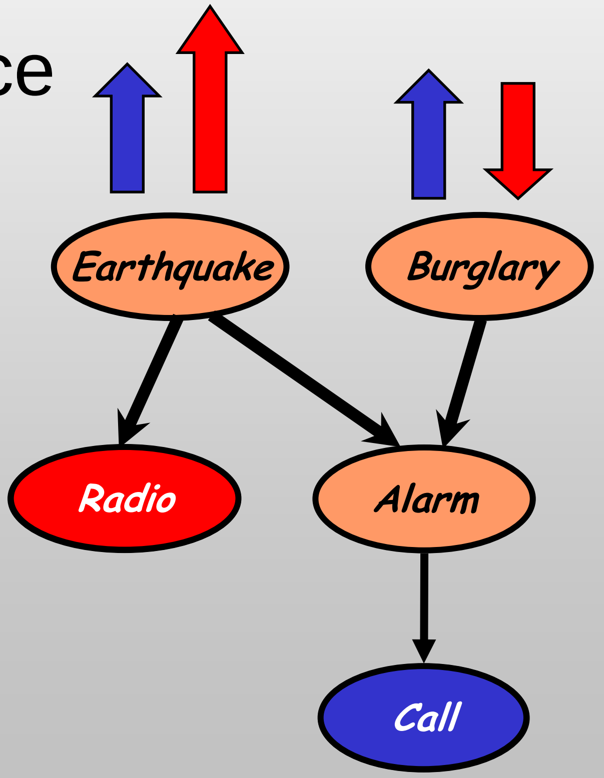
Phylogentic Trees



- The topology of hidden variables poses different constraints on the marginal distribution

Inference - Queries

- **Posterior probabilities**
  - Probability of any event given any evidence
- Most likely explanation
  - Scenario that explains evidence
- Rational decision making
  - Maximize expected utility
  - Value of Information
- Effect of intervention



Inference - Algorithms

Complexity:

- Worst case - exponential cost

Yet,

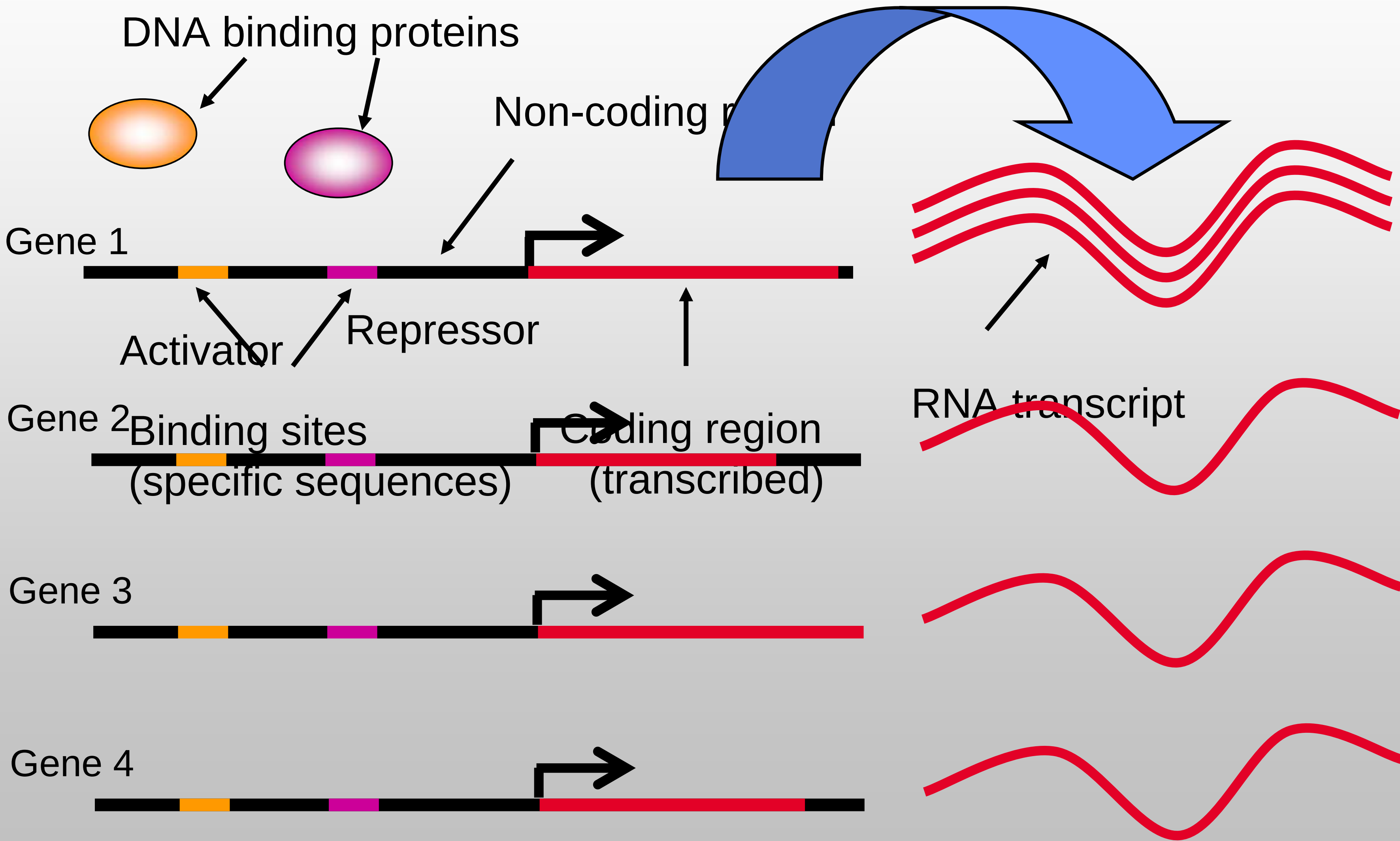
- Generic exact inference algorithms based on dynamic programming
    - Efficient in some network topologies
- Approximate inference algorithms
    - With the appropriate “dark art” they perform well

For the purposes of this tutorial we assume we can solve queries in networks.

Outline

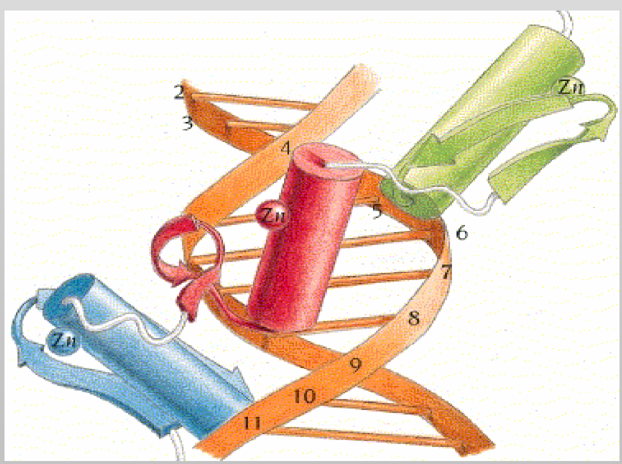
- Introduction
- Bayesian Networks
- Learning Bayesian Networks
- Transcriptional regulation
- Gene expression
- Markov Networks
- Protein-Protein Interactions
- Discussion

Transcriptional Regulation

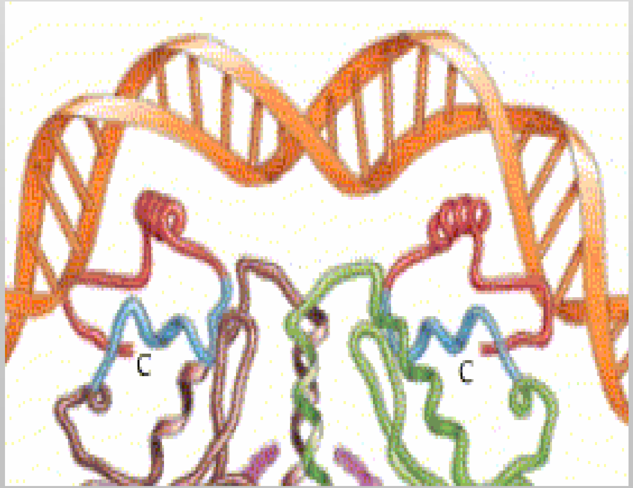


Transcription Factor Binding Sites

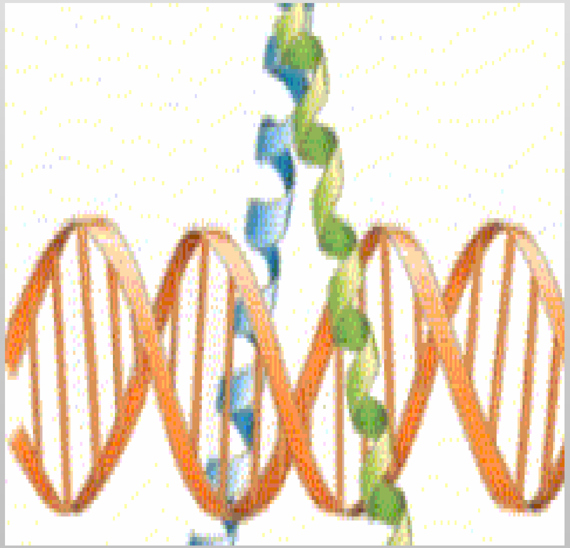
- Gene regulatory proteins contain structural elements that can “read” DNA sequence “motifs”
- The amino acid – DNA recognition is not straightforward
- Experiments can pinpoint binding sites on DNA



Zinc finger



Helix-Turn-Helix



Leucine zipper

Modeling Binding Sites

Given a set of (aligned) binding sites …

- Consensus sequence
- Probabilistic model (profile of a binding site)

NNGGGGCNGGGC

GCGGGGCCGGGC
TGGGGGCGGGGT
AGGGGGCGGGGG
TAGGGGCCGGGC
TGGGGGCGGGGT
TGGGGGCCGGGC
ATGGGGCGGGGC
GTGGGGCGGGGC
AAAGGGCCGGGC
GGGAGGCCGGGA

A	4	3	1	1	0	0	1	0	0	1	0	1
C	1	3	0	0	0	0	13	6	0	0	1	9
G	5	5	13	13	14	14	0	8	14	12	13	1
T	4	3	0	0	0	0	0	0	0	1	0	3

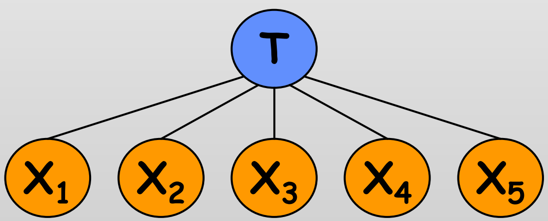
Is this sufficient?

How to model binding sites ?

$P(X_1\ X_2\ X_3\ X_4\ X_5) = ?$ **represents a distribution of binding sites**

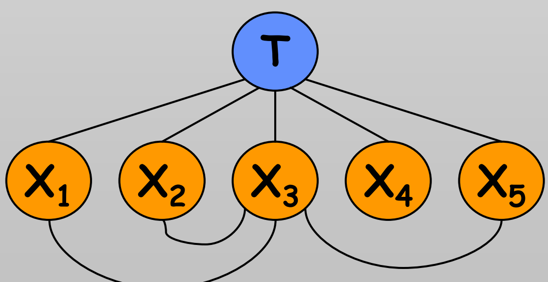
Profile: Independency model

Tree: Direct dependencies



Mixture of Profiles:

Global dependencies



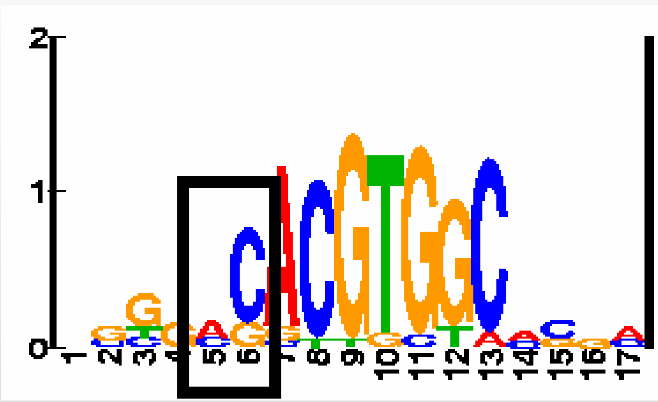
Mixture of Trees:

Both types of dependencies

Barash et al, *RECOMB* 2002

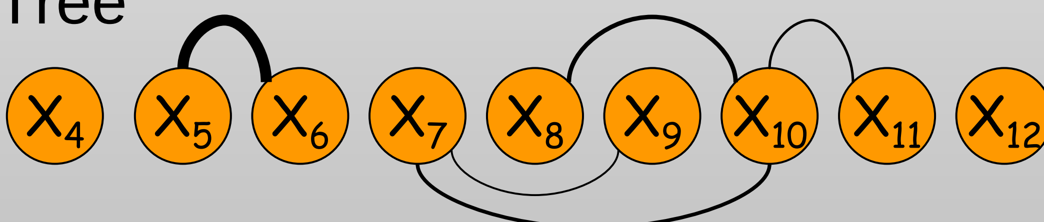
Arabidopsis ABA binding factor 1

Profile



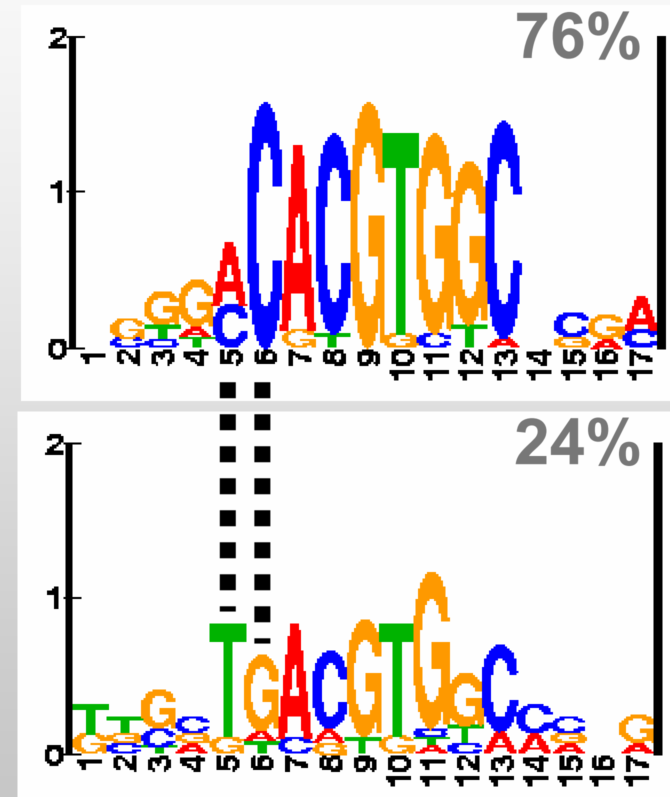
Test LL per instance **-19.93**

Tree



Test LL per instance **-18.47** **(+1.46)**
(improvement in likelihood > 2.5-fold)

Mixture of Profiles



Test LL per instance **-18.70** **(+1.23)**
(improvement in likelihood > 2-fold)

Barash et al, *RECOMB* 2002

The Knowledge Acquisition Bottleneck

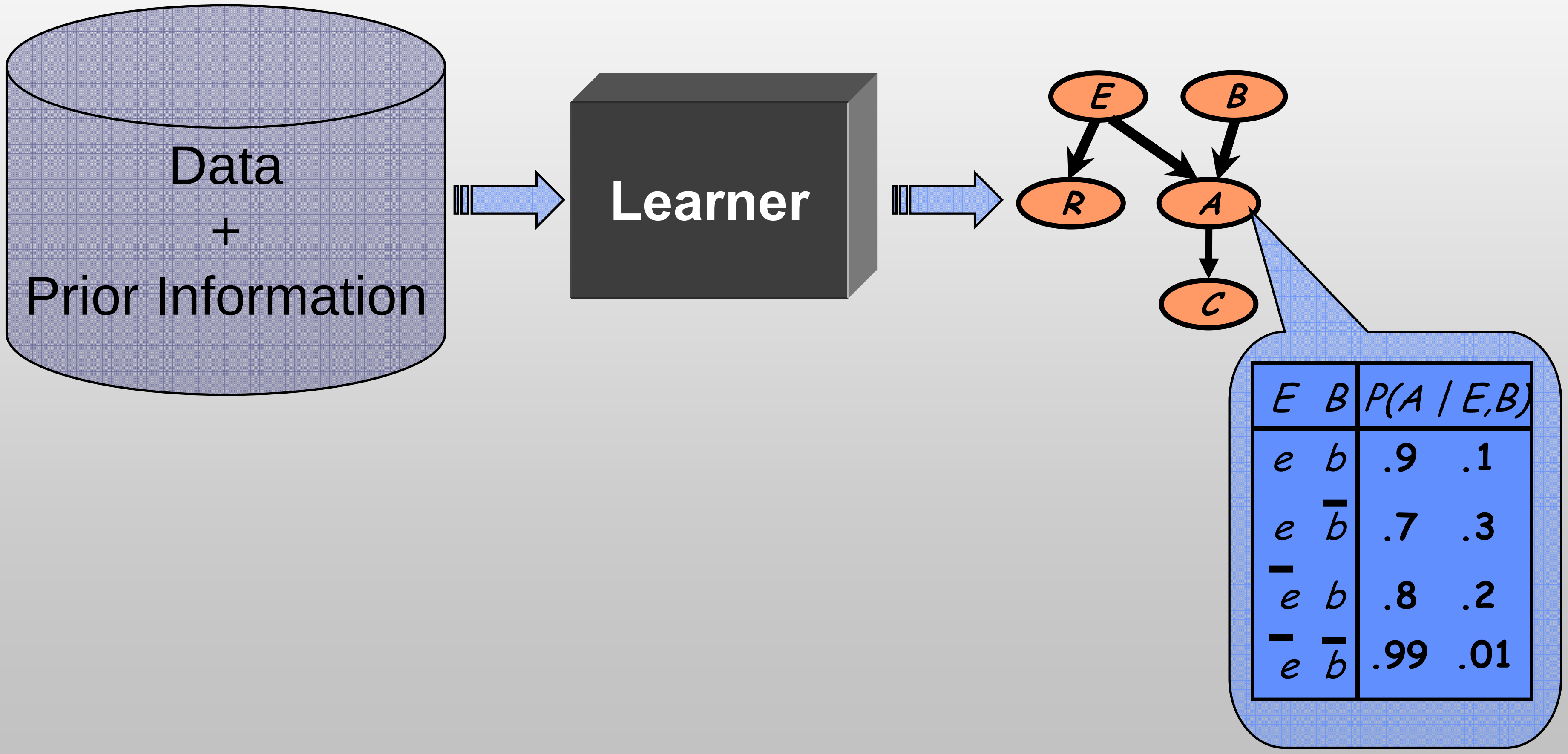
How do we construct these models?

- Knowledge acquisition is an expensive process
- Often we don’t have an expert

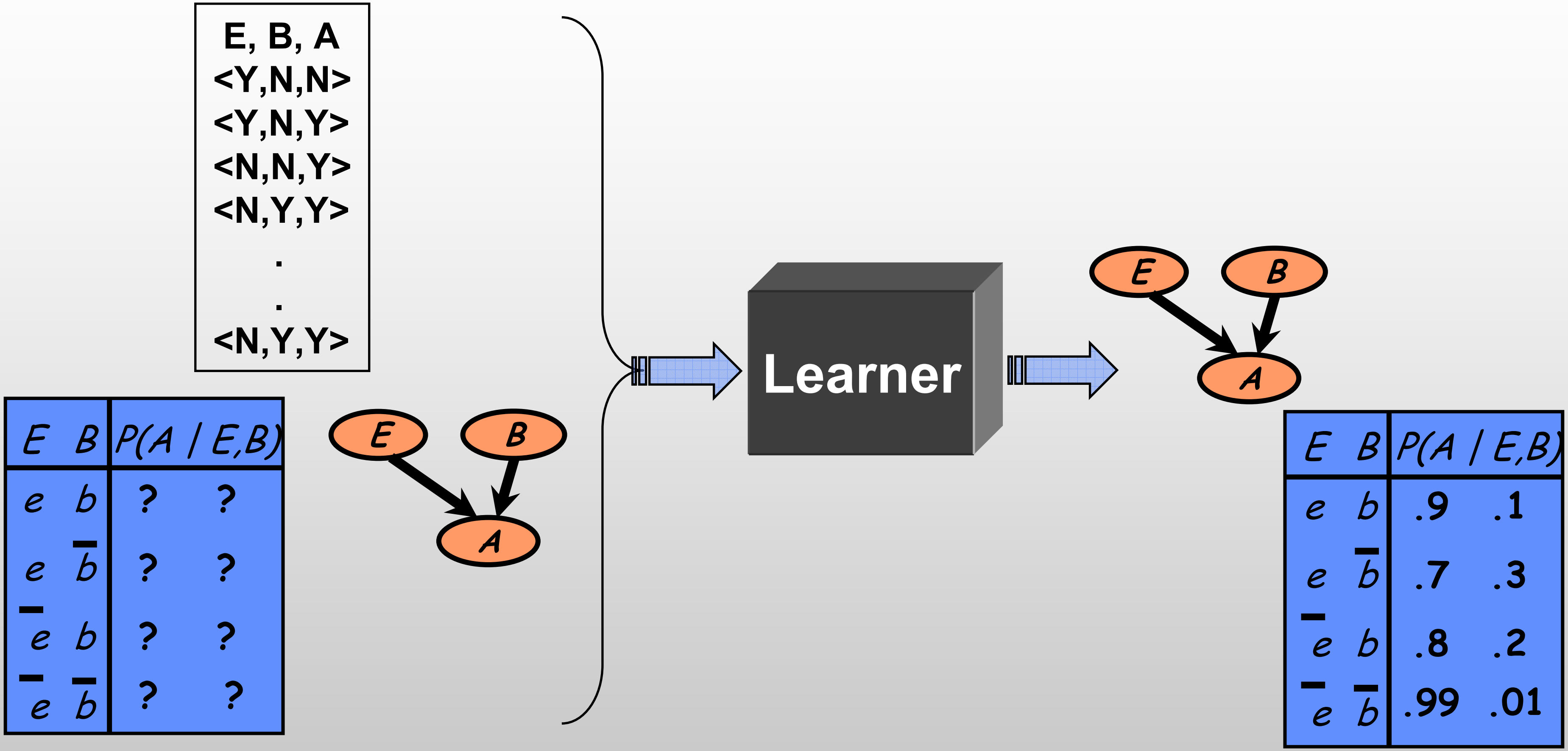
Harnessing Data

- Amount of available information growing rapidly
- Learning allows us to construct models from raw data
- The the details of learned models provide insights about the data

Learning Bayesian networks



Known Structure, Complete Data

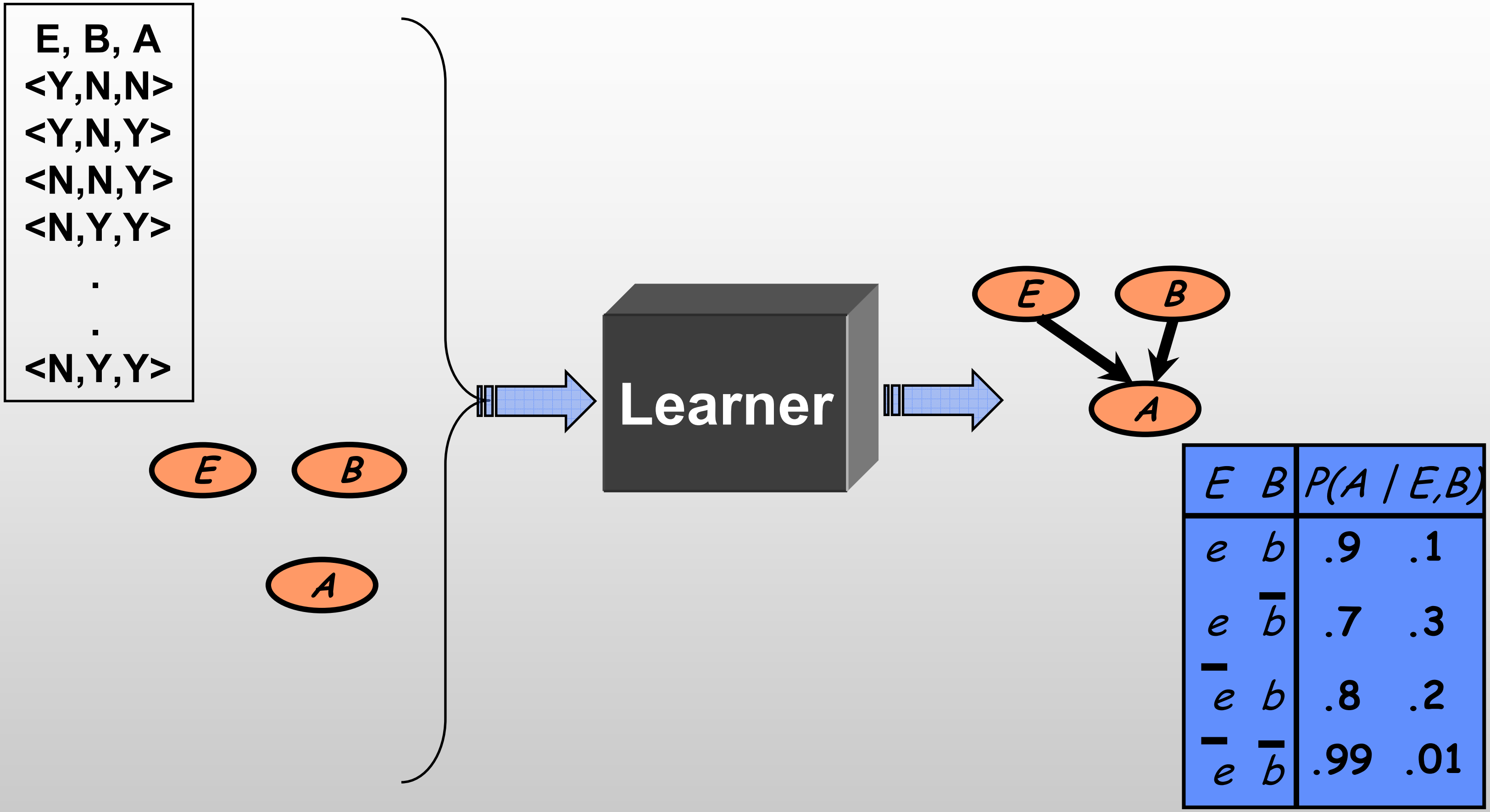


E	B	P(A | E,B)	
e	b	?	?
e	$\bar{b}$	?	?
$\bar{e}$	b	?	?
$\bar{e}$	$\bar{b}$	?	?

E	B	P(A | E,B)	
e	b	.9	.1
e	$\bar{b}$	.7	.3
$\bar{e}$	b	.8	.2
$\bar{e}$	$\bar{b}$	.99	.01

- Network structure is specified
  - Learner needs to estimate parameters
- Data does not contain missing values

Unknown Structure, Complete Data

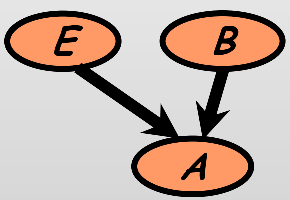


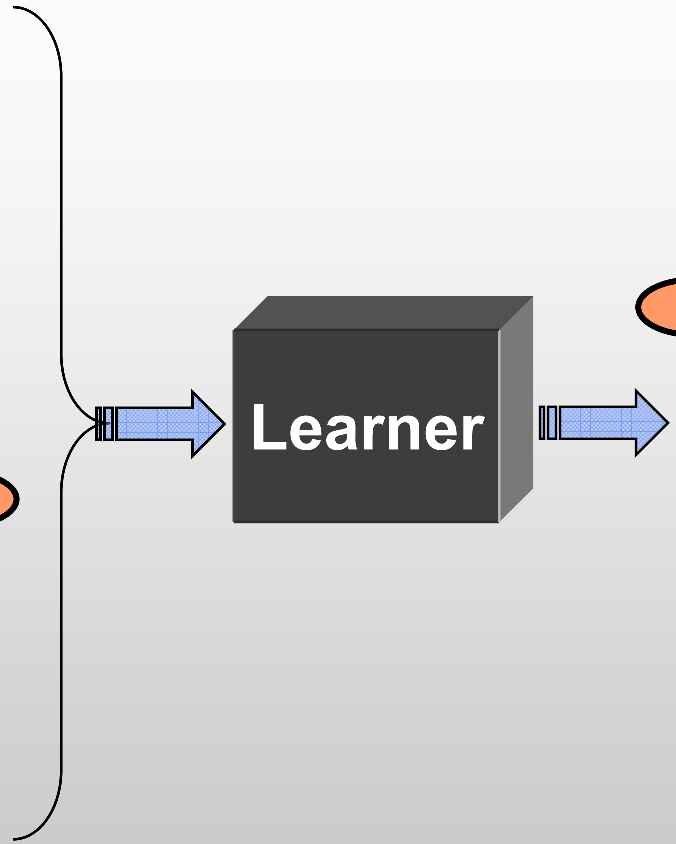
- Network structure is not specified
  - Inducer needs to select arcs & estimate parameters
- Data does not contain missing values

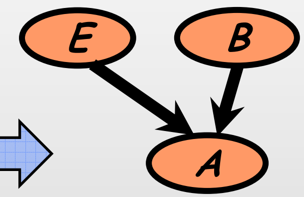
Known Structure, Incomplete Data

E, B, A
<Y,N,N>
<Y,?,Y>
<N,N,Y>
<N,Y,?>
.
.
<?,Y,Y>

E	B	P(A | E,B)	
e	b	?	?
e	$\bar{b}$	?	?
$\bar{e}$	b	?	?
$\bar{e}$	$\bar{b}$	?	?



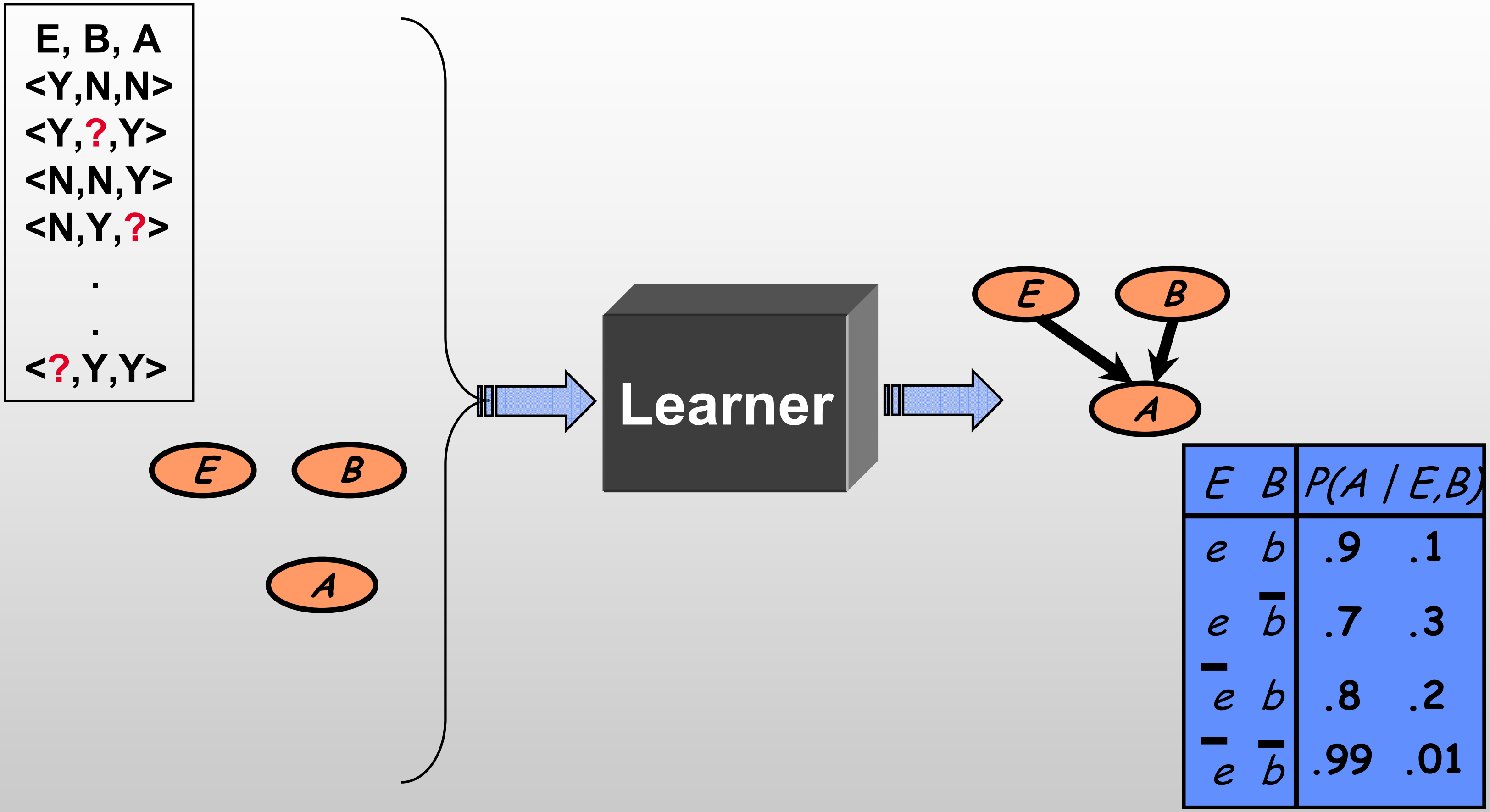




E	B	P(A | E,B)	
e	b	.9	.1
e	$\bar{b}$	.7	.3
$\bar{e}$	b	.8	.2
$\bar{e}$	$\bar{b}$	.99	.01

- Network structure is specified
- Data contains missing values
  - Need to consider assignments to missing values

Unknown Structure, Incomplete Data



- Network structure is not specified
- Data contains missing values
  - Need to consider assignments to missing values

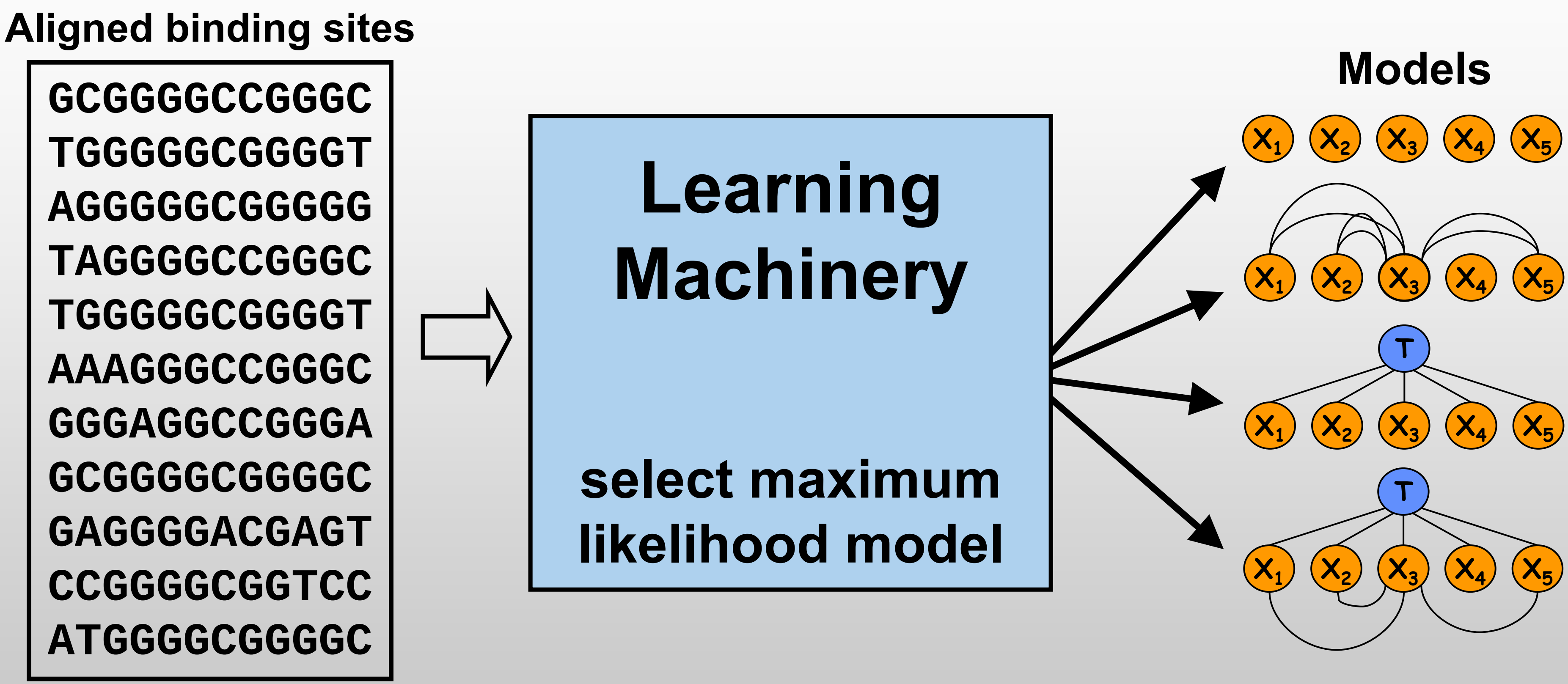
The Learning Problem

	Known Structure	**Unknown Structure**
Complete Data	Statistical parametric estimation (closed-form eq.)	Discrete optimization over structures (discrete search)
Incomplete Data	Parametric optimization (EM, gradient descent...)	Combined (Structural EM, mixture models…)

Outline

- Introduction
- Bayesian Networks
- Learning Bayesian Networks
- Transcriptional regulation
- Gene expression
- Markov Networks
- Protein-Protein Interactions
- Discussion

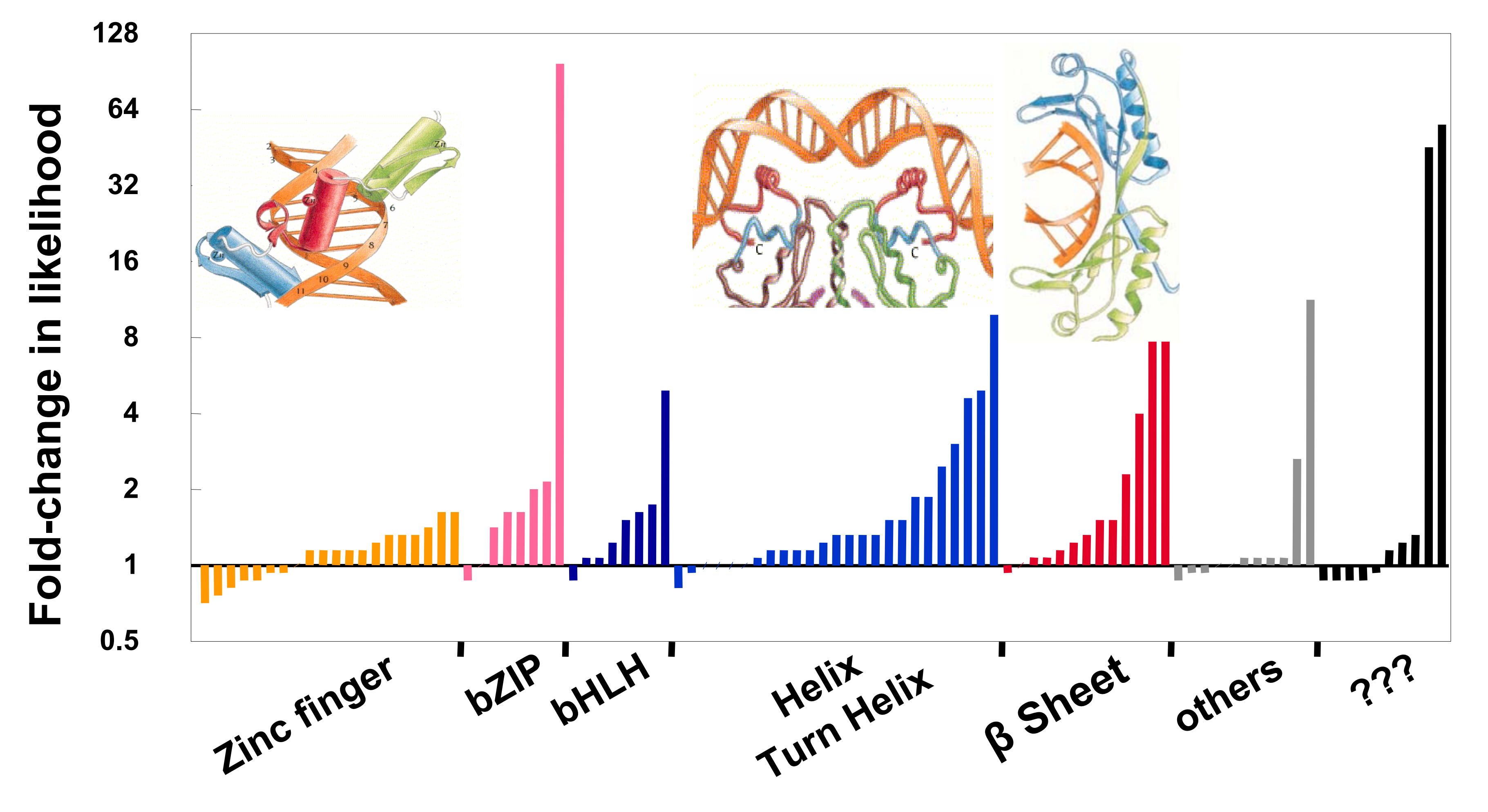
Learning models: Aligned binding sites



Learning based on methods for probabilistic graphical models (*Bayesian networks*)

Likelihood improvement over profiles

TRANSFAC: 95 aligned data sets

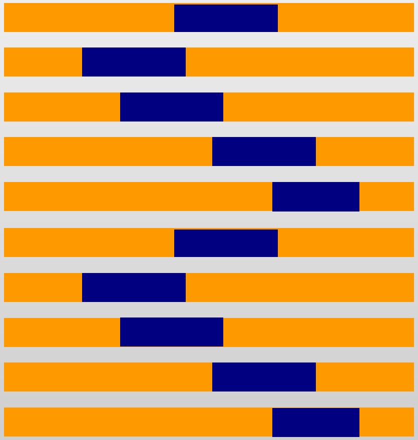


Barash et al, *RECOMB* 2002

Motif finding problem

Input: A set of potentially co-regulated genes

Output: A common motif in their promoters

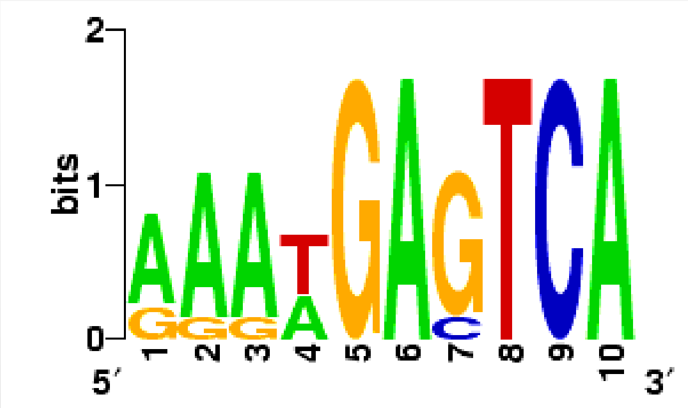


Sources of data:

- Gene annotation (e.g. **Hughes et al**, 2000)
- Gene expression (e.g. **Spellman et al**, 1998**; Tavazoie et al**, 2000)
- ChIP (e.g. **Simon et al**, 2001**; Lee et al**, 2002)

Example

- Upstream regions from yeast *Sacharomyces cerevisiae* genes (300-600bp)



5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …*ILV6*

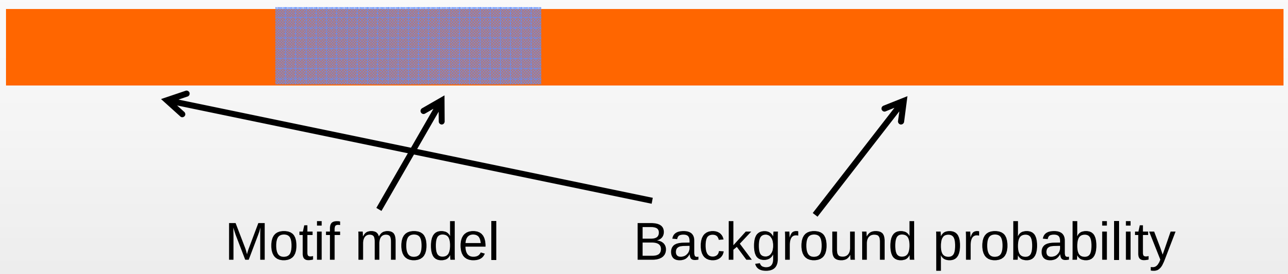
5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …*ARO1*

5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …*HOM2*

5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …*PRO3*

Probabilistic Model



- Background probability: given
- Motif model – parameters being learned

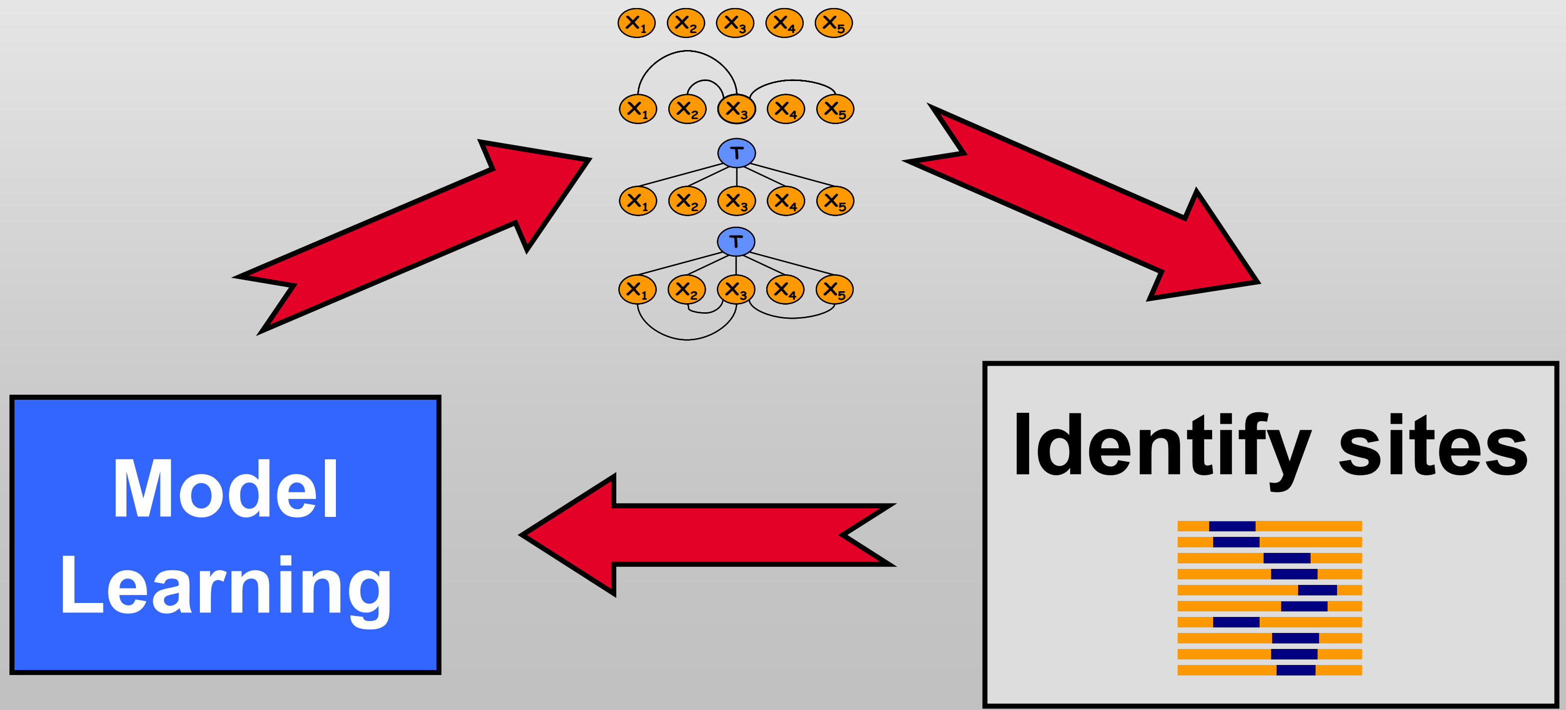
Hidden variable:

- Location of motif within each sequence

Learning models: unaligned data

EM (MEME-like)

- Identify binding site positions
- Learn a dependency model



ChIP location analysis

Yeast genome-wide location analysis

Target genes annotation for 106 TFs

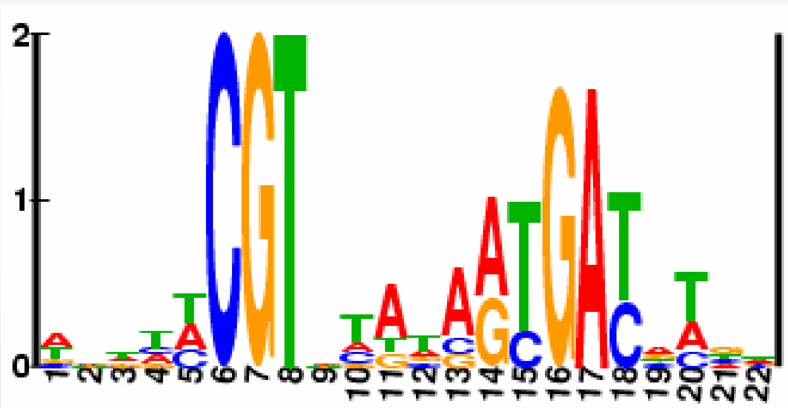
genes ~ 6000

Gene	ABF1 Targets	....	ZAP1 Targets
YAL001C	+		−
YAL002W	−		+
YAL003W	+		−
YAL005C	−		−
.	.		.
.	.		.
.	.		.
YAL010C	+		−
YAL012C	−		+
YAL013W	−		+
YPR201W	−		−

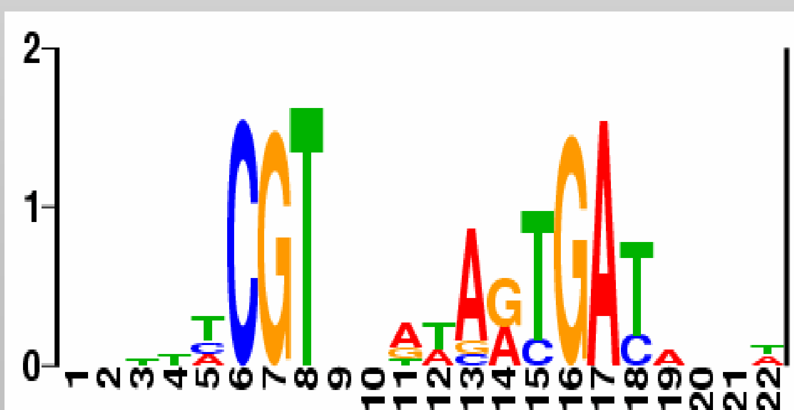
Example: Models learned for ABF1 (YPD)

Autonomously replicating sequence-binding factor 1

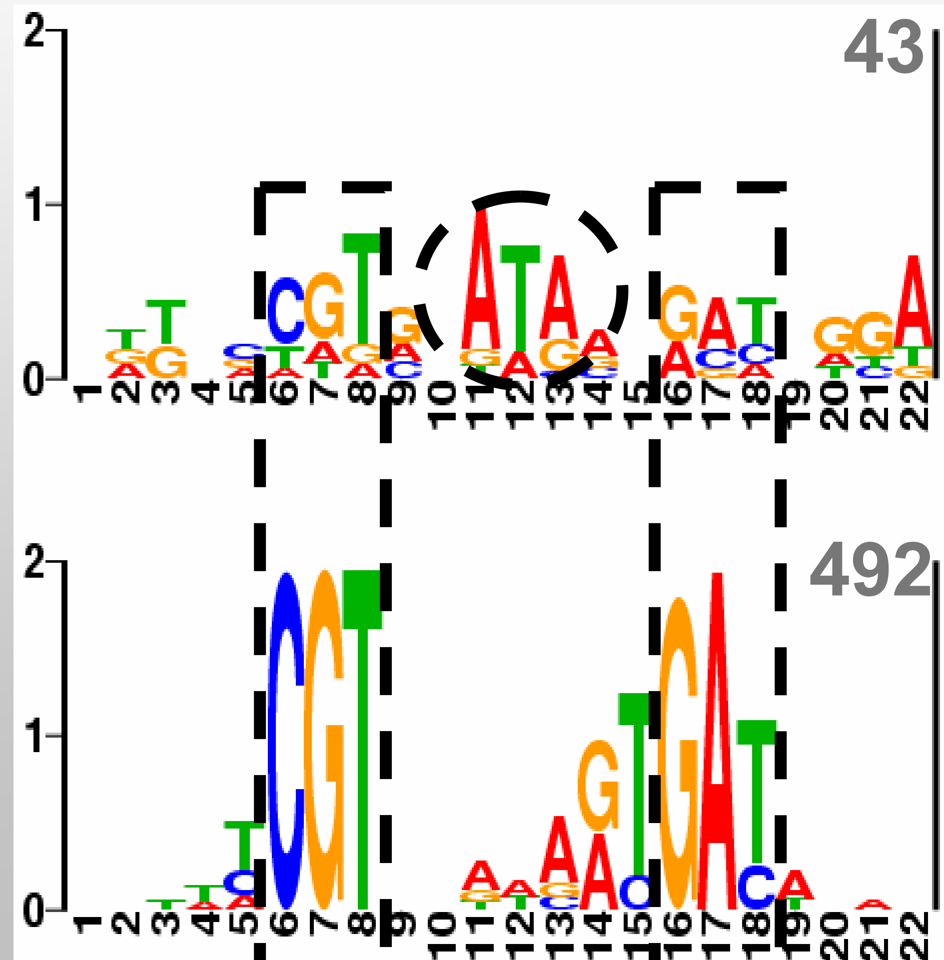
Known profile
(from TRANSFAC)



Learned profile



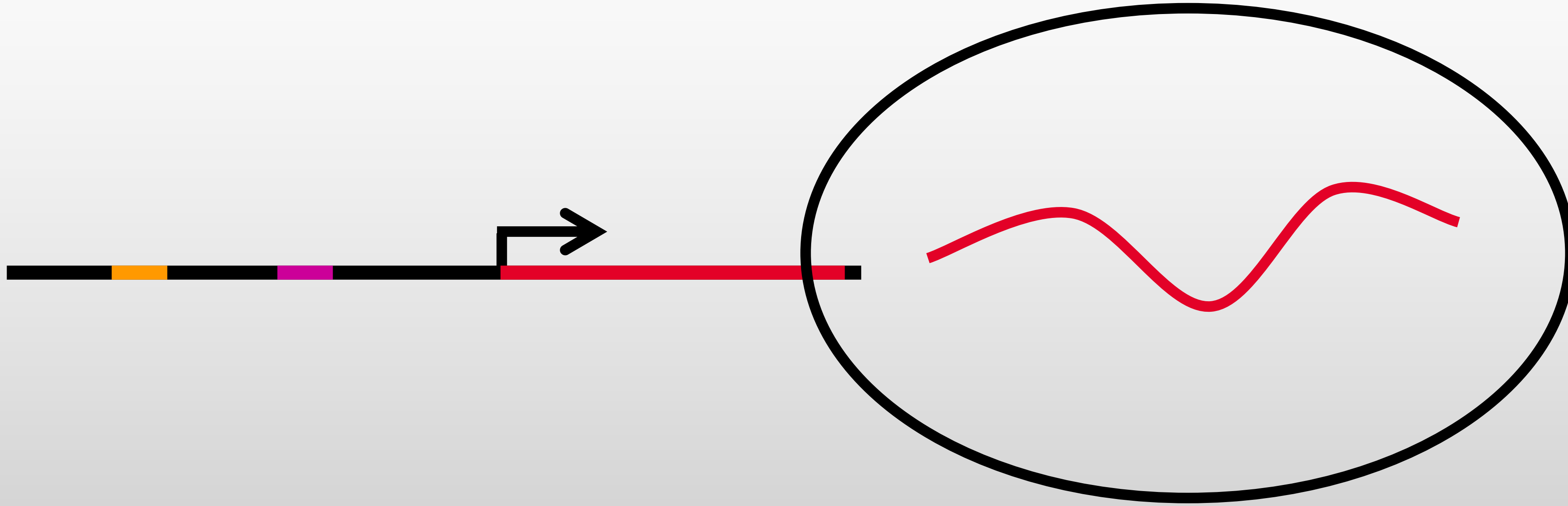
Learned Mixture of Profiles



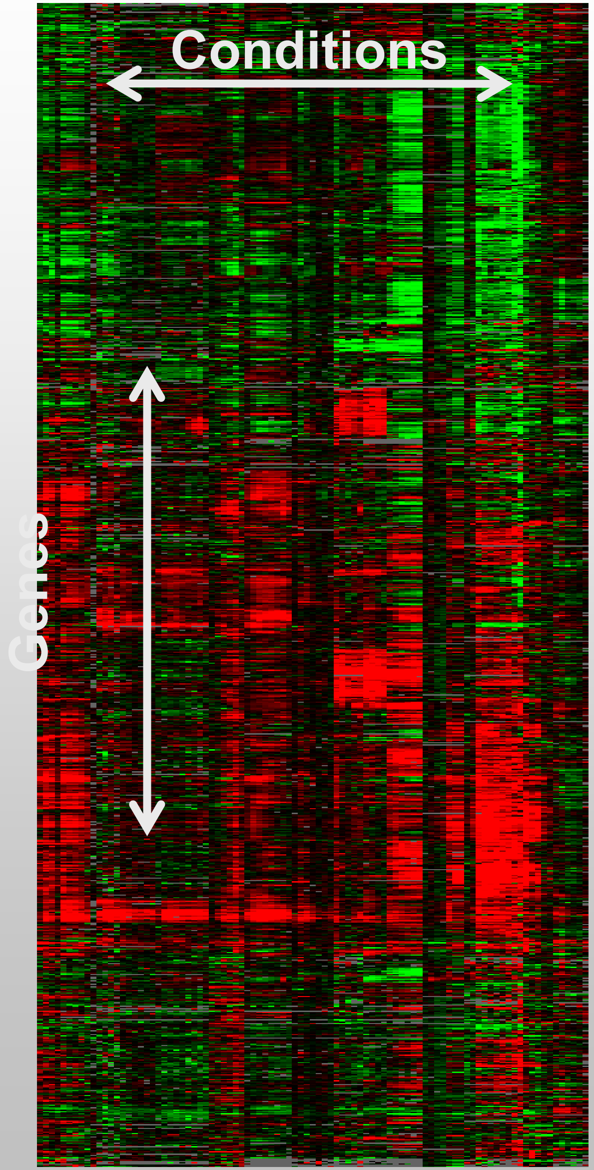
Outline

- Introduction
- Bayesian Networks
- Learning Bayesian Networks
- Transcriptional regulation
- Gene expression
- Markov Networks
- Protein-Protein Interactions
- Discussion

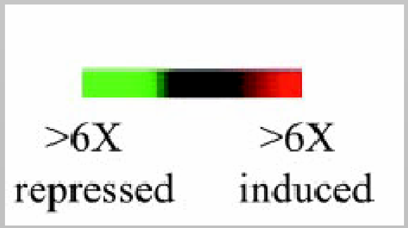
Transcriptional Regulation



Expression Data



Gasch *et al.* Mol. Cell 2001



- 1000s of genes
- 10-100s of arrays

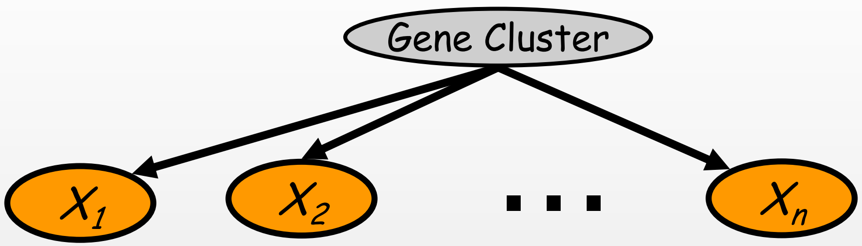
Possible designs

- Biopsies from different patient populations
- Time course
- Different perturbations

Clustering Gene Expression Profiles

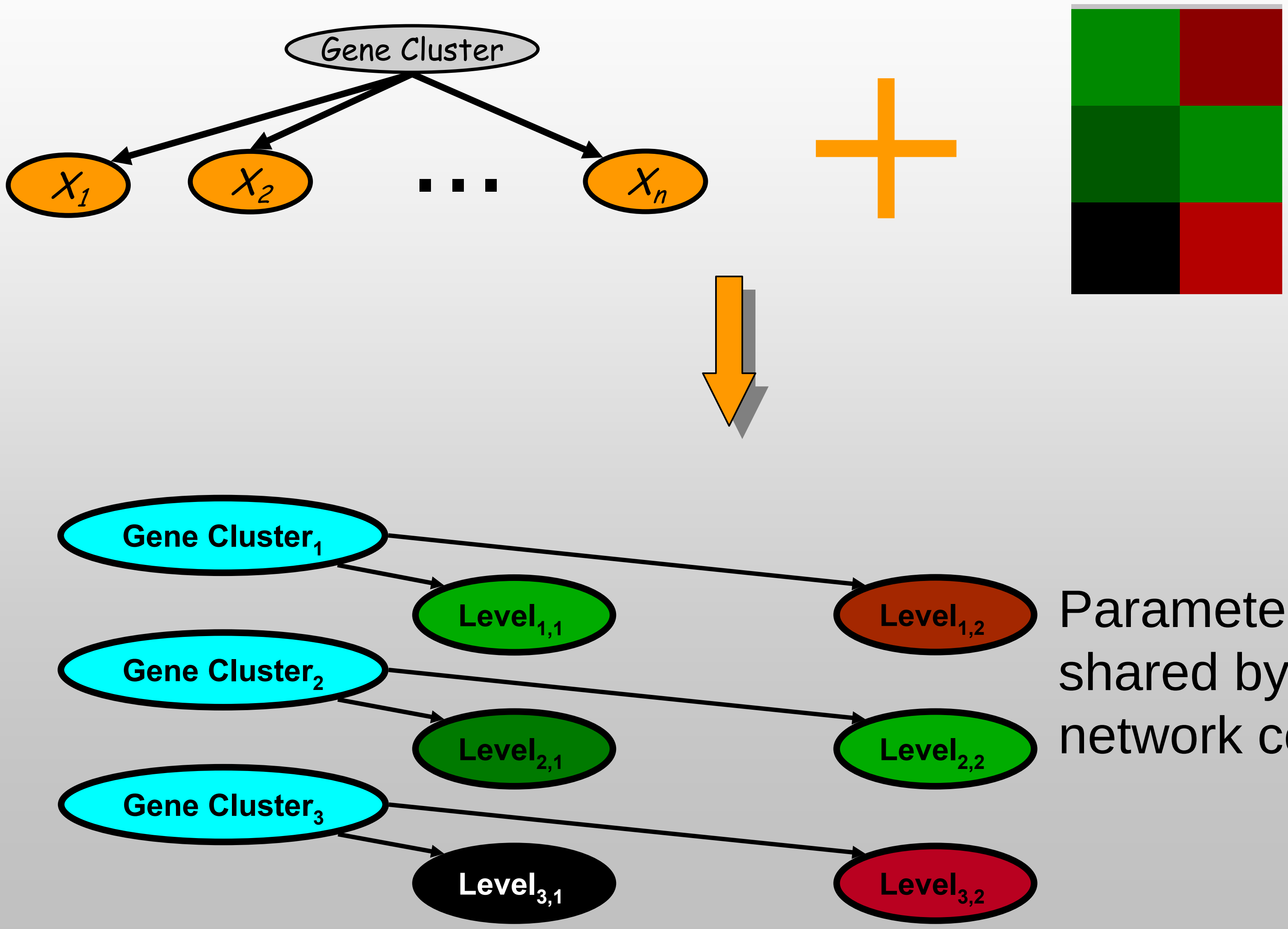
Clustering model

- "Cluster" hidden variable explains dependencies among measurement of a gene in different conditions



- Each gene is viewed as a sample from the same distribution

Clustering Genes



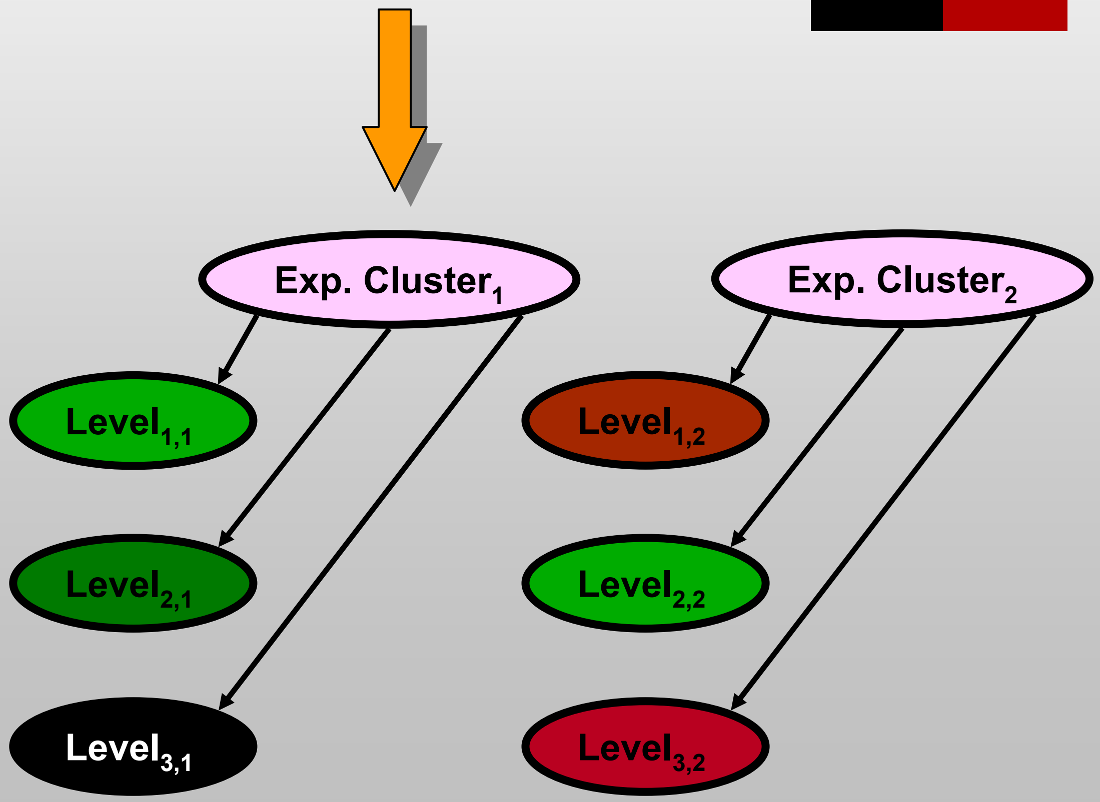
Parameters are shared by network copies

Clustering Conditions

Can we cluster both genes and conditions?

Now each condition is an instance

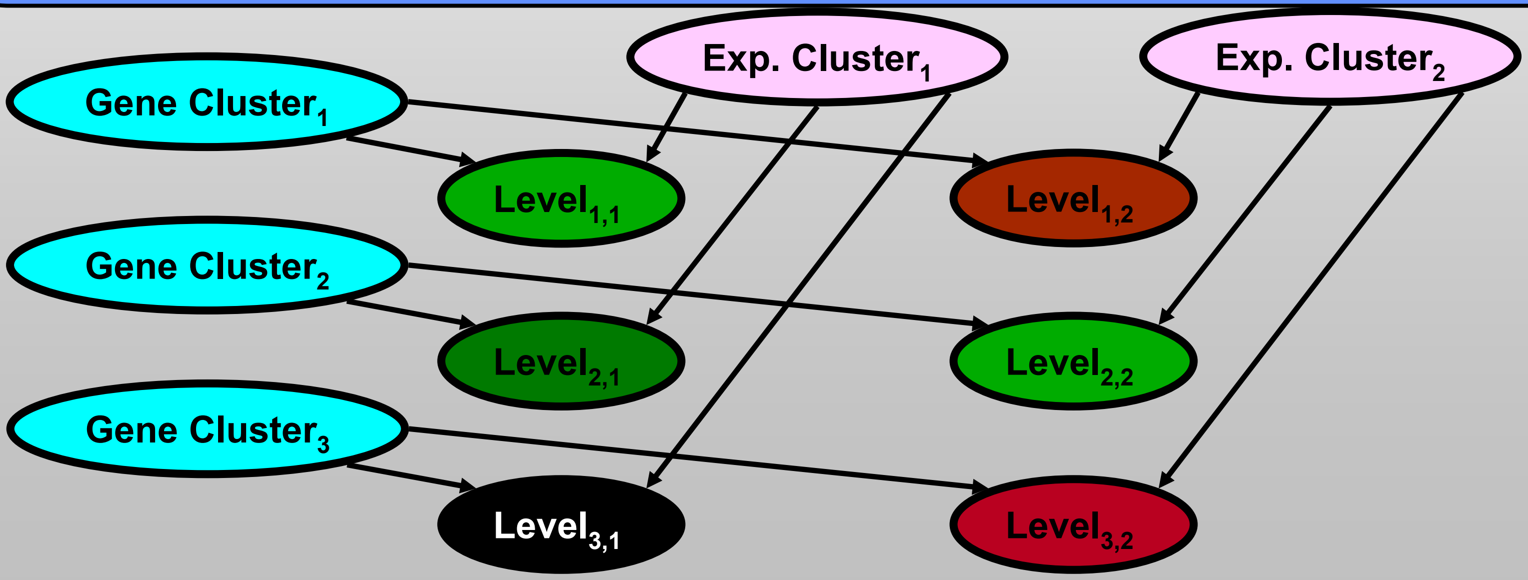
Parameters are shared by network copies



Joint Clustering?

A single network that spans the whole data

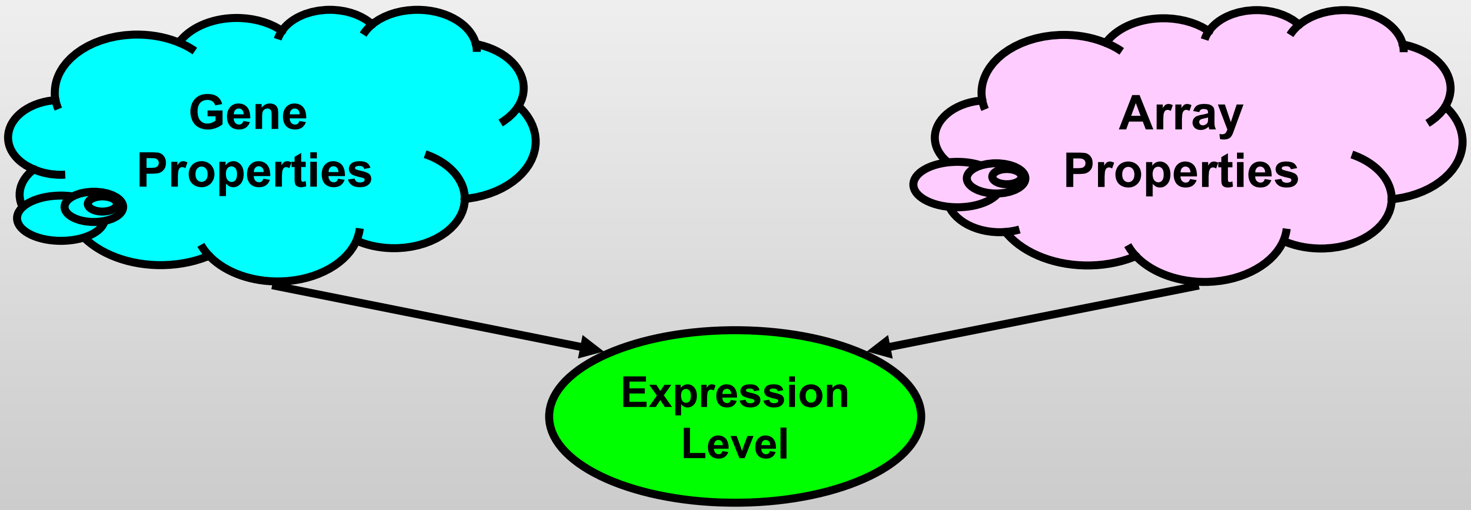
- Each expression variable has its own parameters
- # parameters >> # observations



Relational Approach

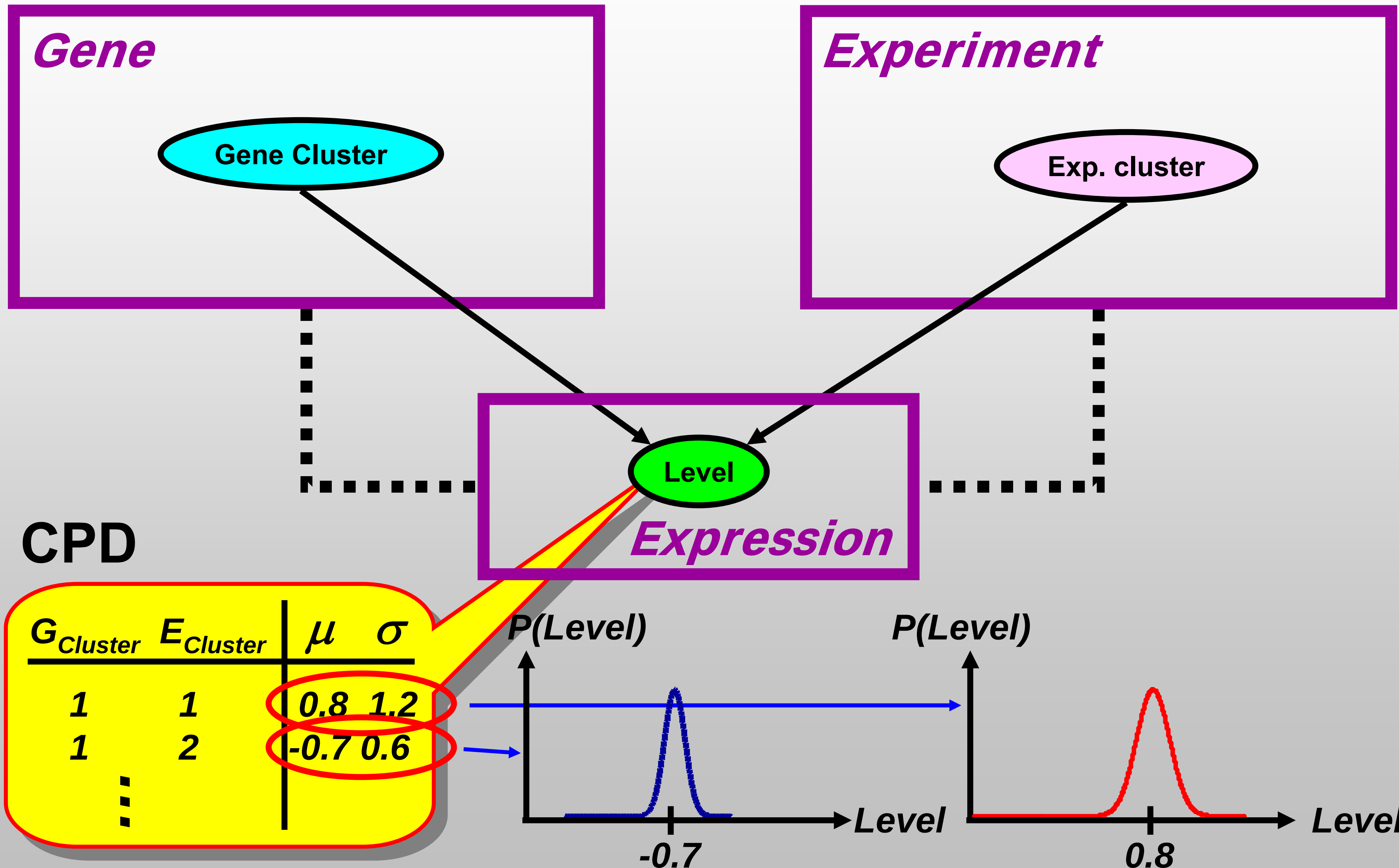
Key Idea:

Expression level is “explained” by properties of gene and properties of experiment

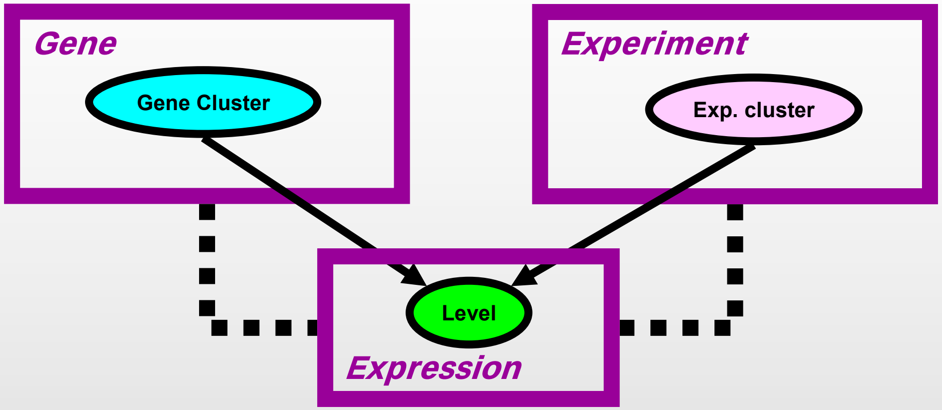


Segal *et al*, *ISMB* 2001

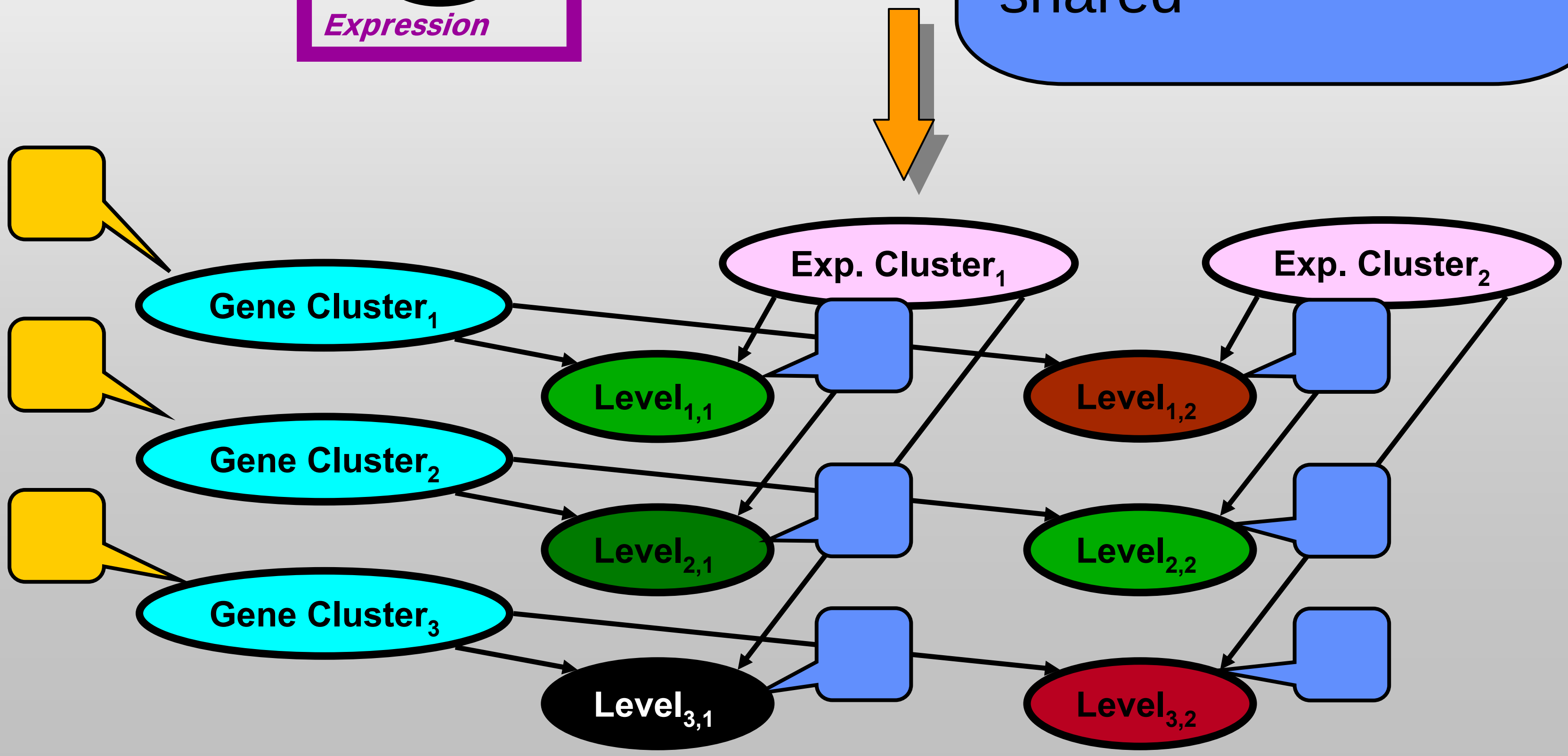
Probabilistic Relational Models



Unrolling a Relational Network



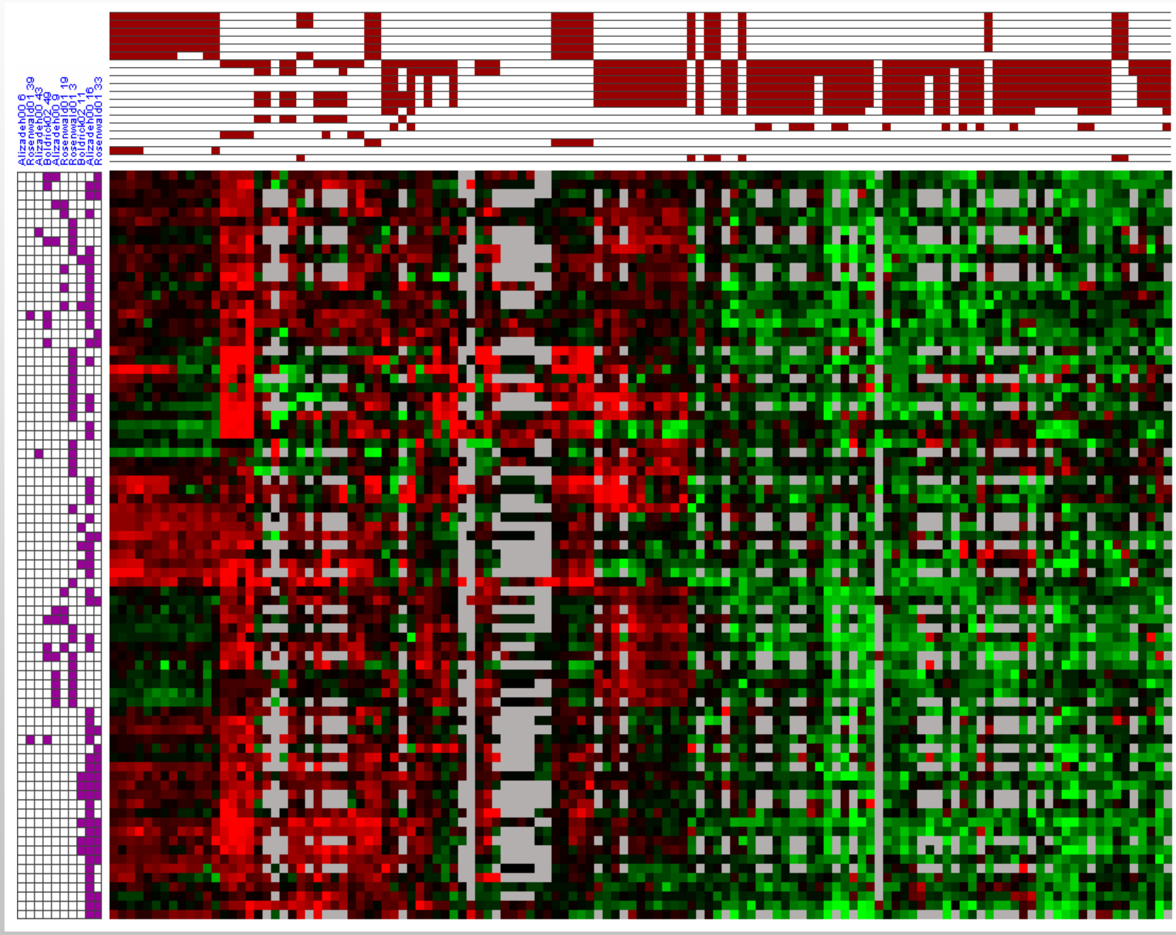
Parameters of variables from the same template are shared



Expression +Annotations

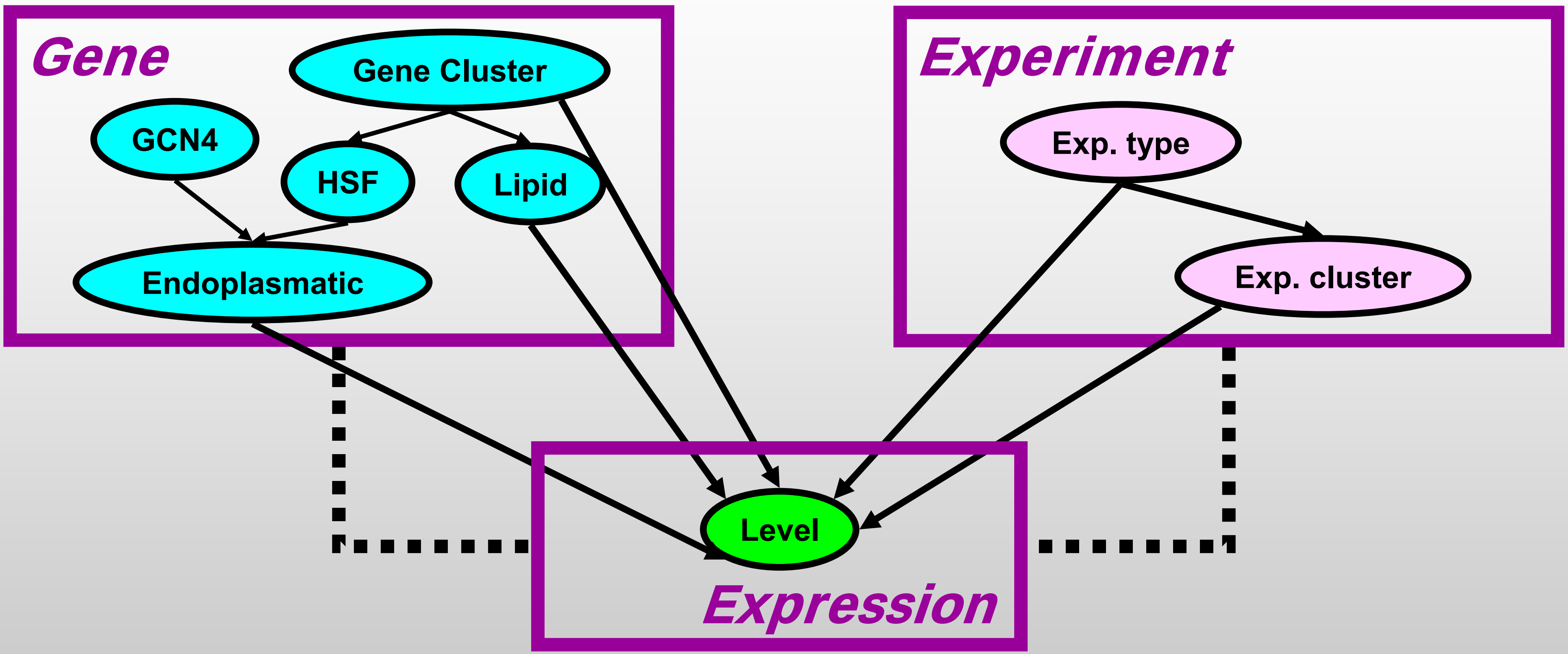
Array annotations:
- Tissue type, Clinical conditions, Treatments, Prognosis

Gene annotations:
- Function, Process, Regulatory regions Cellular location, protein family



Relational models!

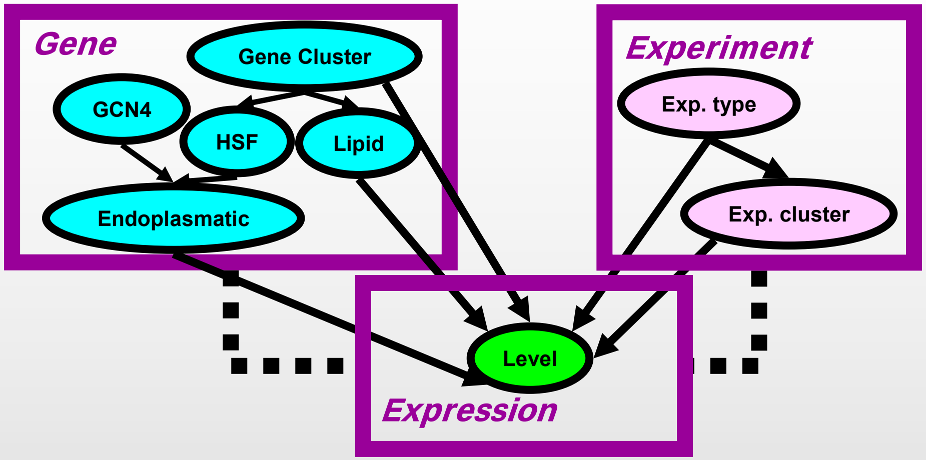
Adding Additional Data



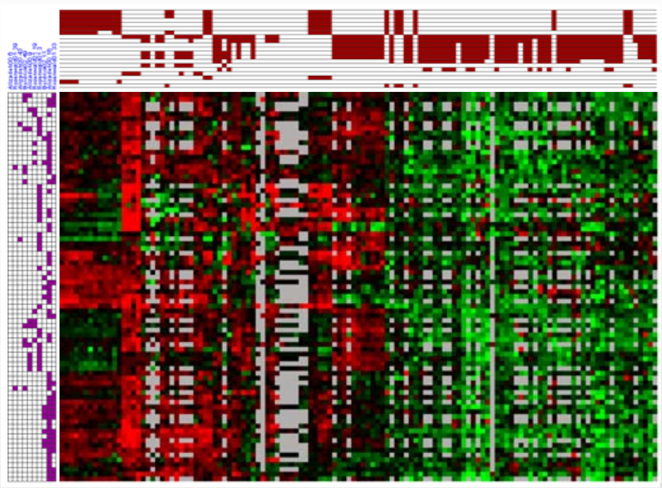
- Annotations
- Binding sites
- Experimental details

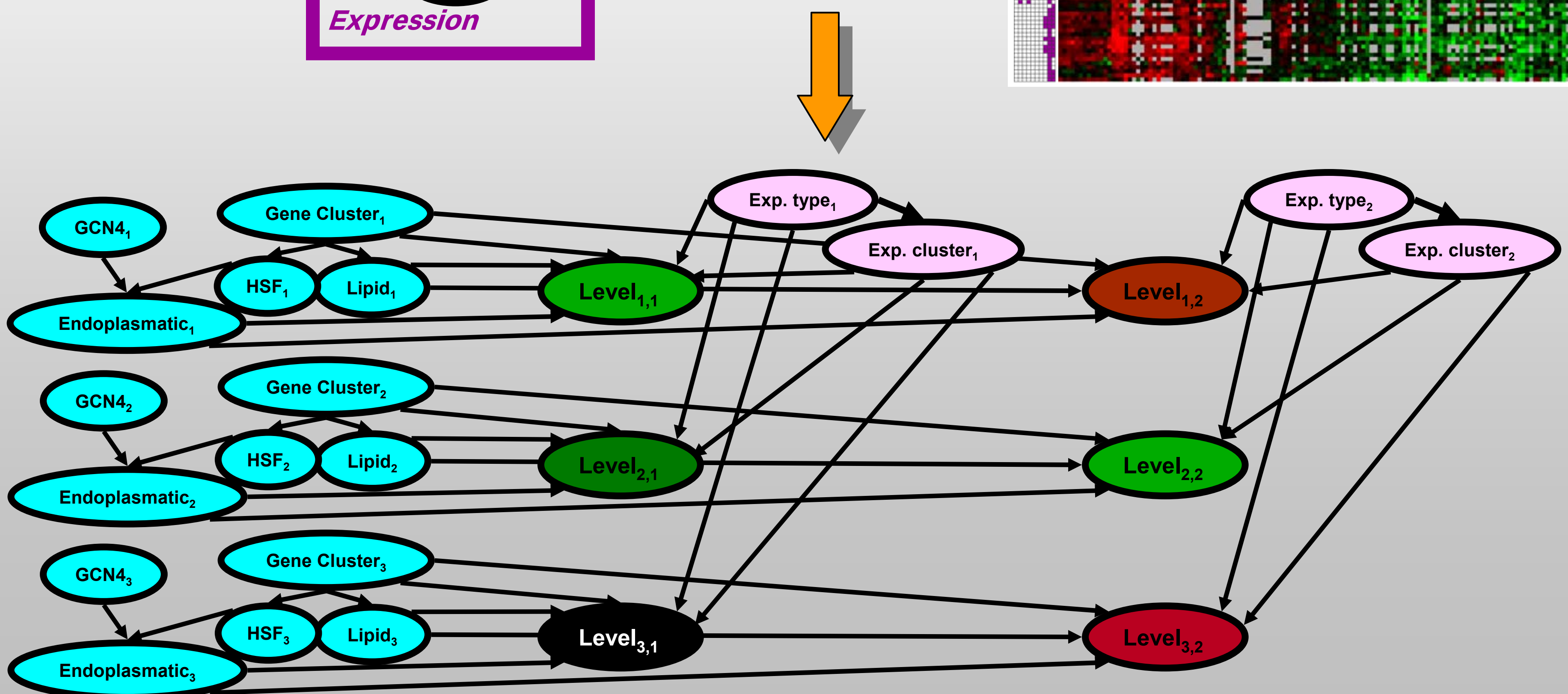
Segal *et al*, *ISMB* 2001

Semantics



Gene
Gene Cluster
GCN4
HSF
Lipid
Endoplasmatic
Experiment
Exp. type
Exp. cluster
Level
Expression





Exp. type$_1$
Exp. cluster$_1$
Exp. type$_2$
Exp. cluster$_2$
GCN4$_1$
Gene Cluster$_1$
HSF$_1$
Lipid$_1$
Endoplasmatic$_1$
Level$_{1,1}$
Level$_{1,2}$
GCN4$_2$
Gene Cluster$_2$
HSF$_2$
Lipid$_2$
Endoplasmatic$_2$
Level$_{2,1}$
Level$_{2,2}$
GCN4$_3$
Gene Cluster$_3$
HSF$_3$
Lipid$_3$
Endoplasmatic$_3$
Level$_{3,1}$
Level$_{3,2}$

TF to Expression

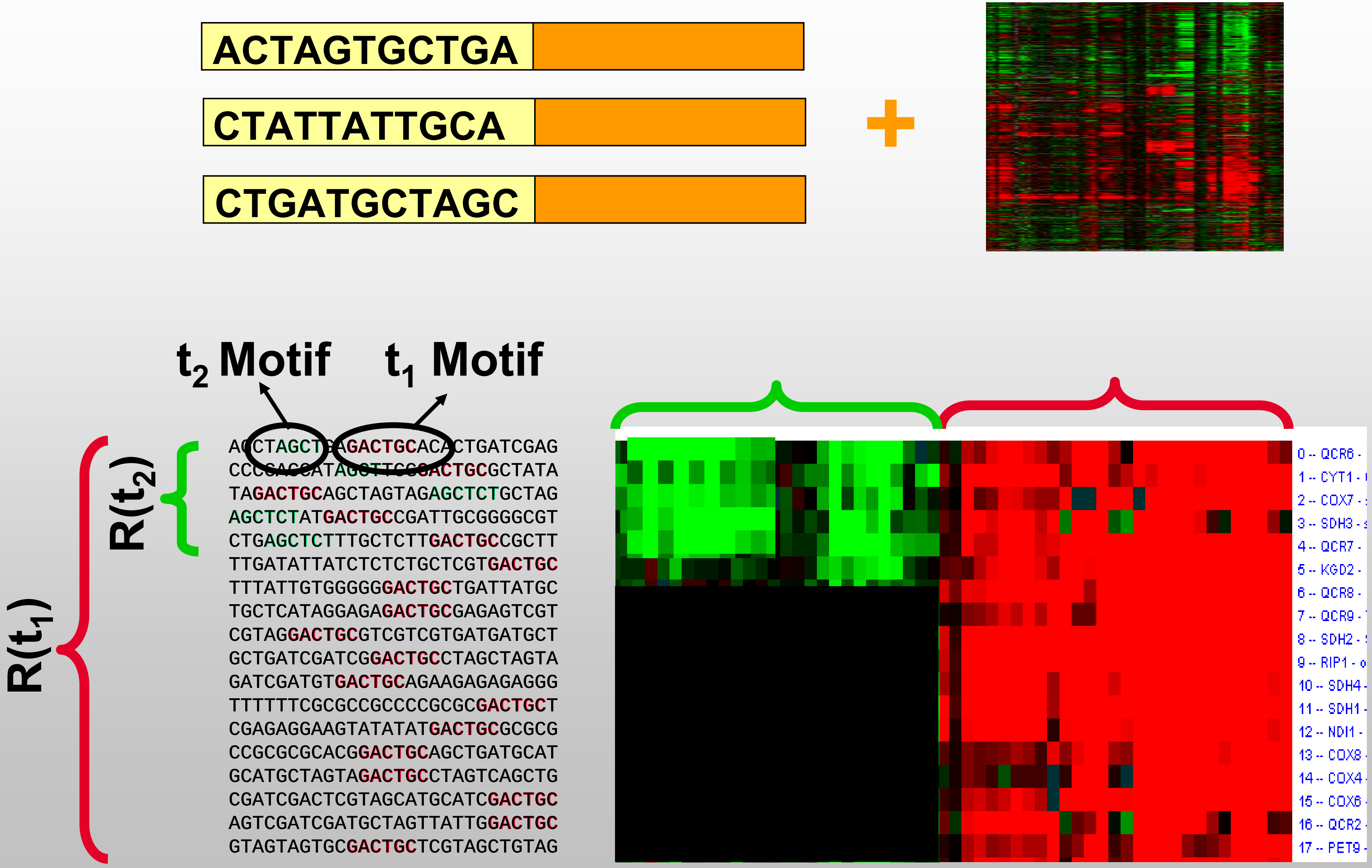
Key Question:

- Can we explain changes in expression?

General model:

- Transcription factor binding sites in promoter region should "explain" changes in transcription

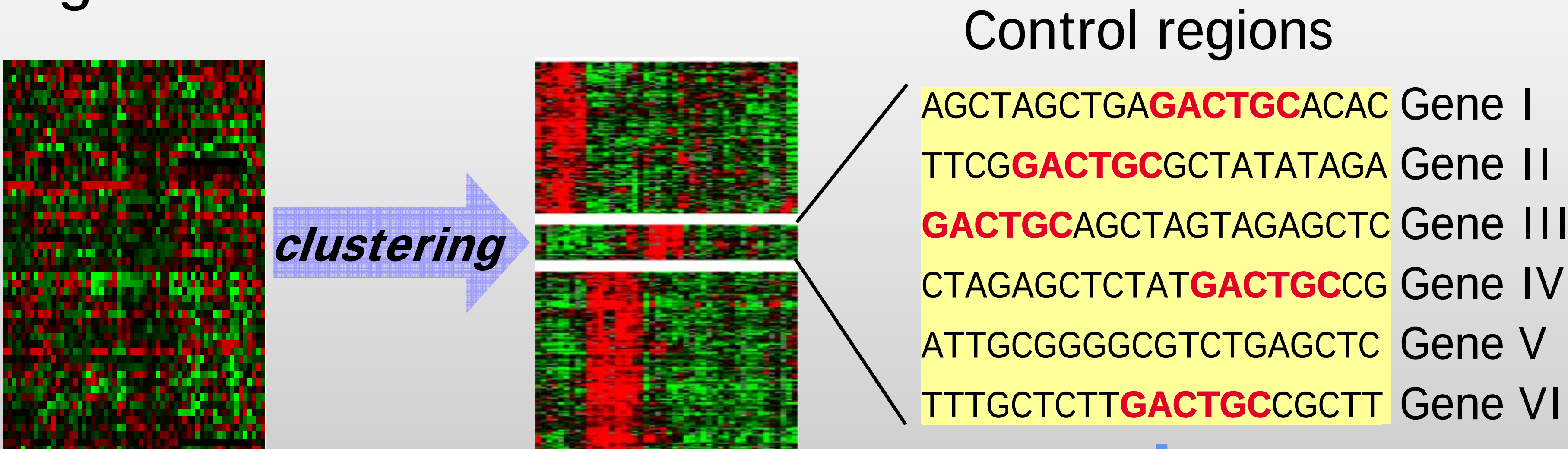
Goal



Segal *et al*, *RECOMB* 2002, ISMB 2003

"Classical" Approach

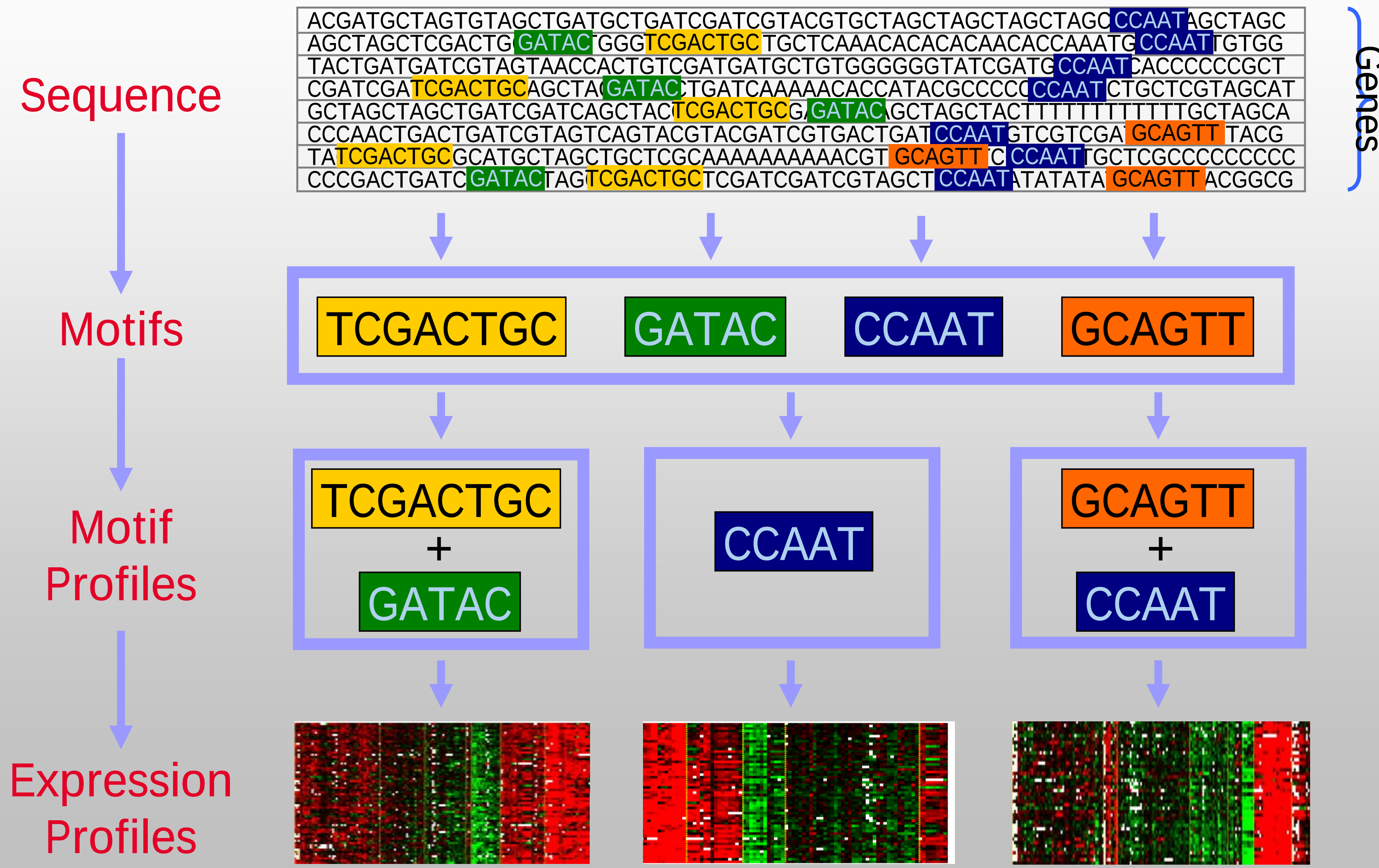
- Cluster gene expression profiles
- Search for motifs in control regions of clustered genes



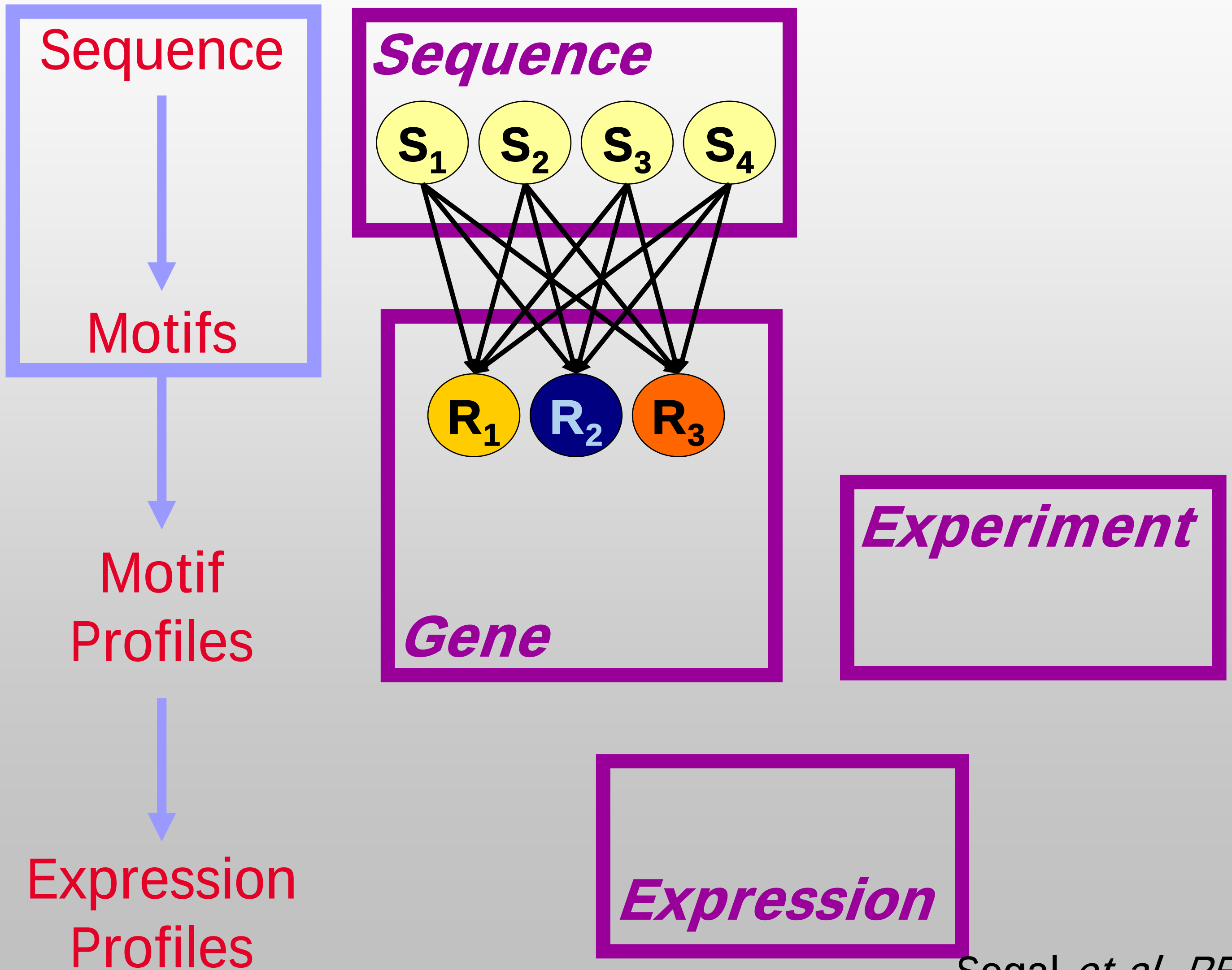
Procedural

- Apply separate method to each type of data
- Use output of one method as input to the next
- Unidirectional information flow

Flow of Information

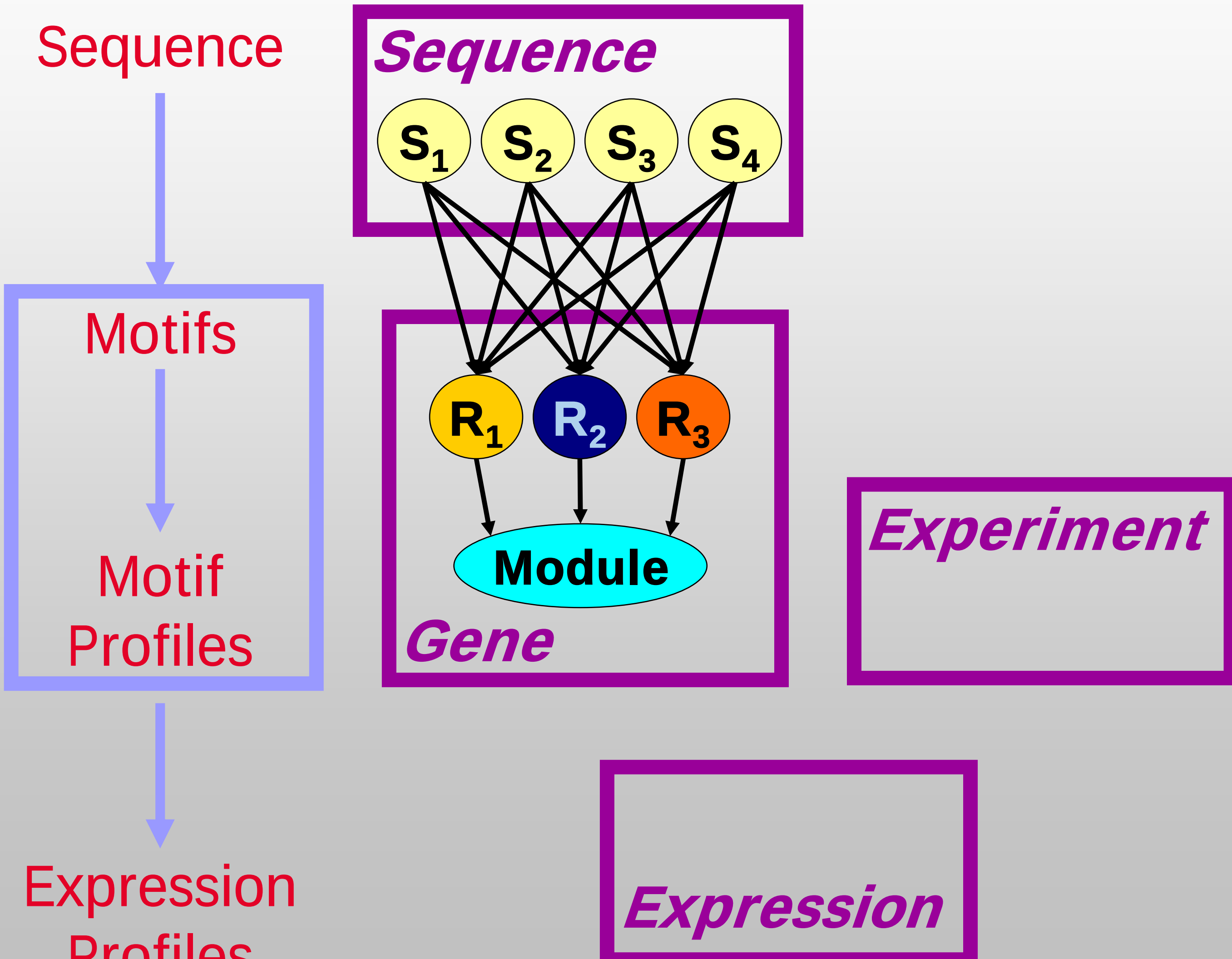


Unified Probabilistic Model



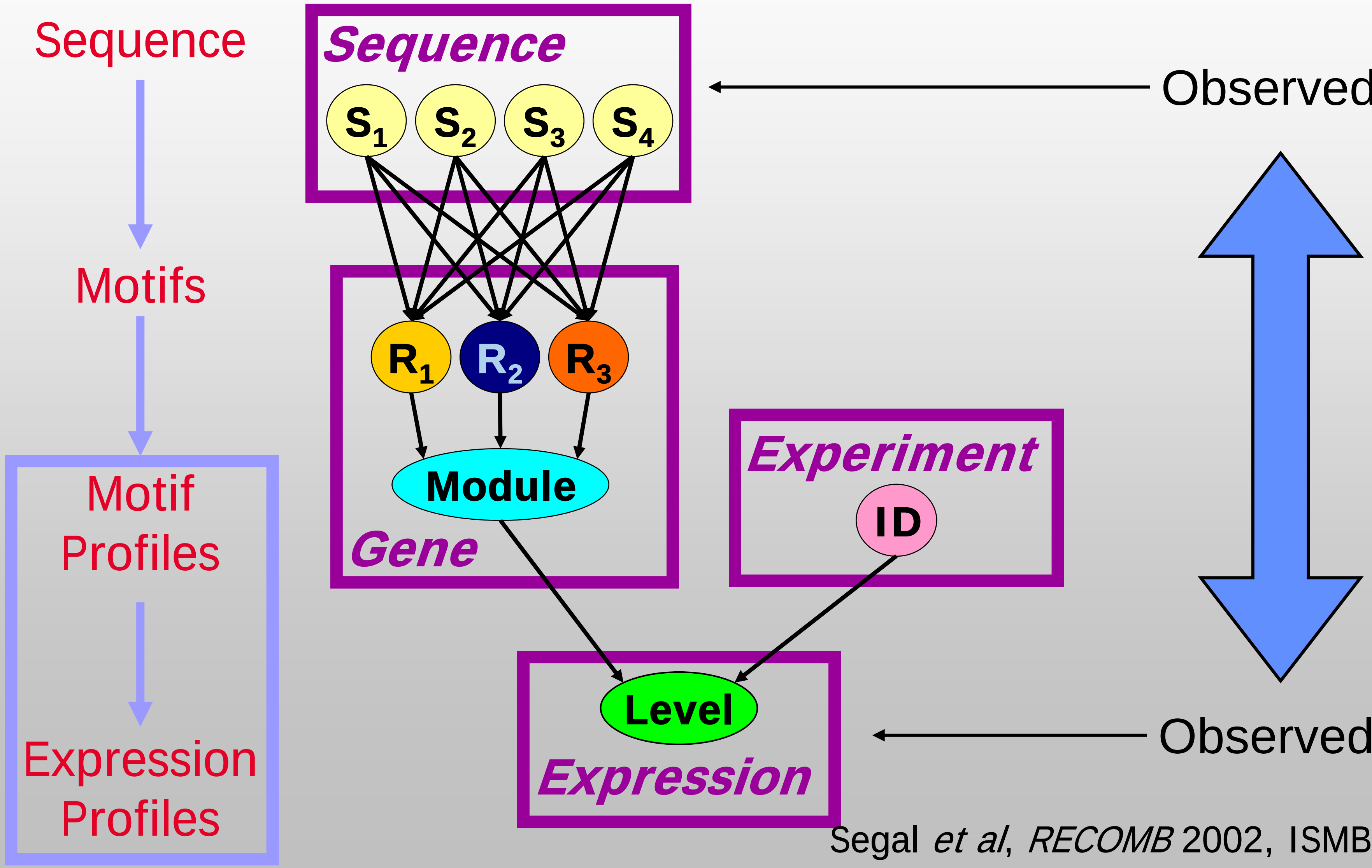
Segal *et al*, *RECOMB* 2002, ISMB 2003

Unified Probabilistic Model



Segal *et al*, *RECOMB* 2002, ISMB 2003

Unified Probabilistic Model



Segal *et al*, *RECOMB* 2002, ISMB 2003

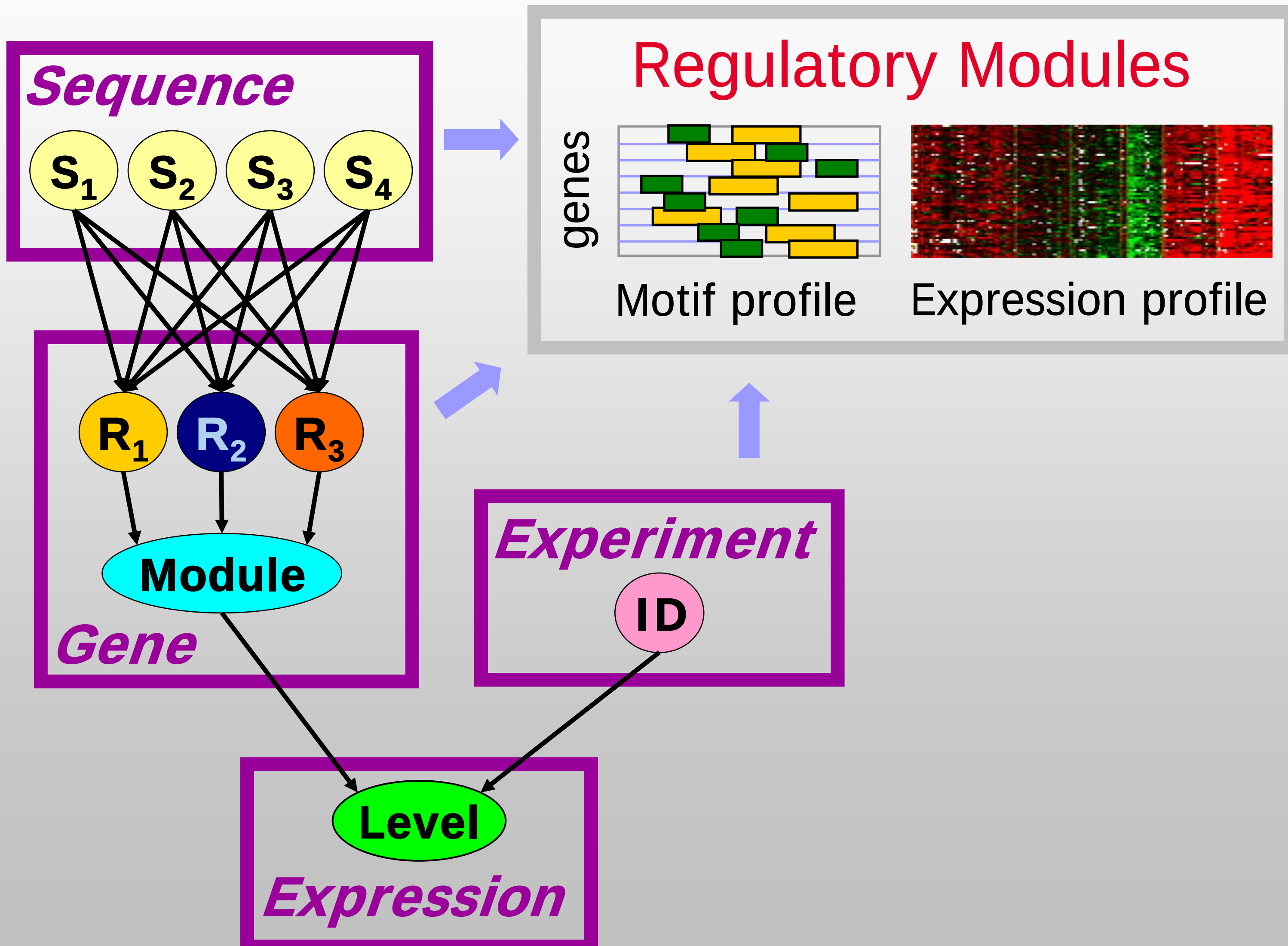
Probabilistic Model

Sequence

Motifs

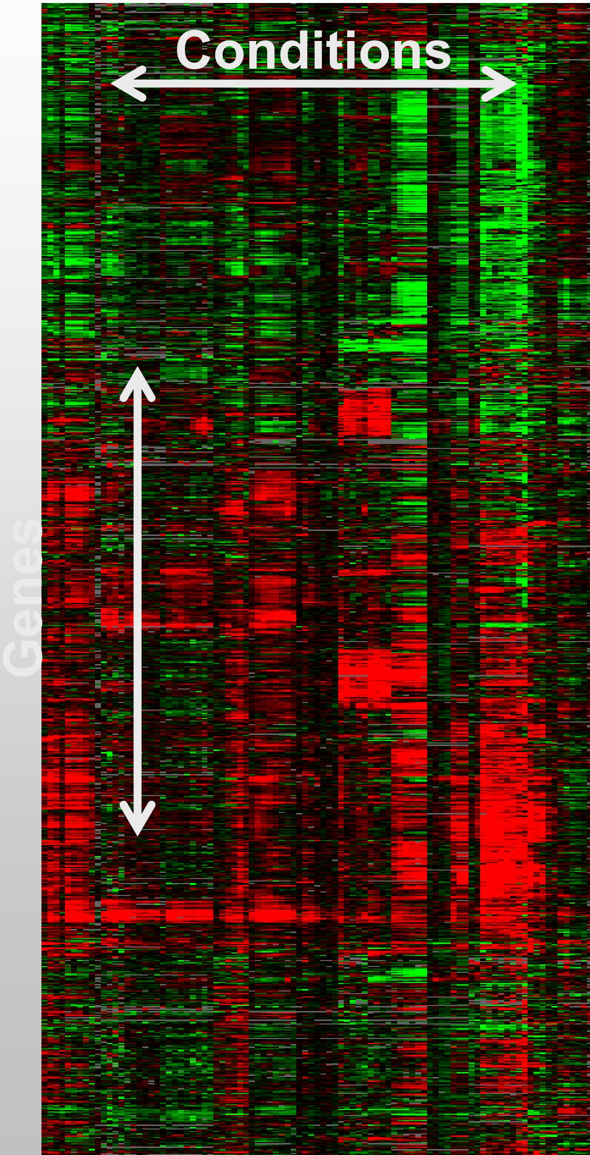
Motif Profiles

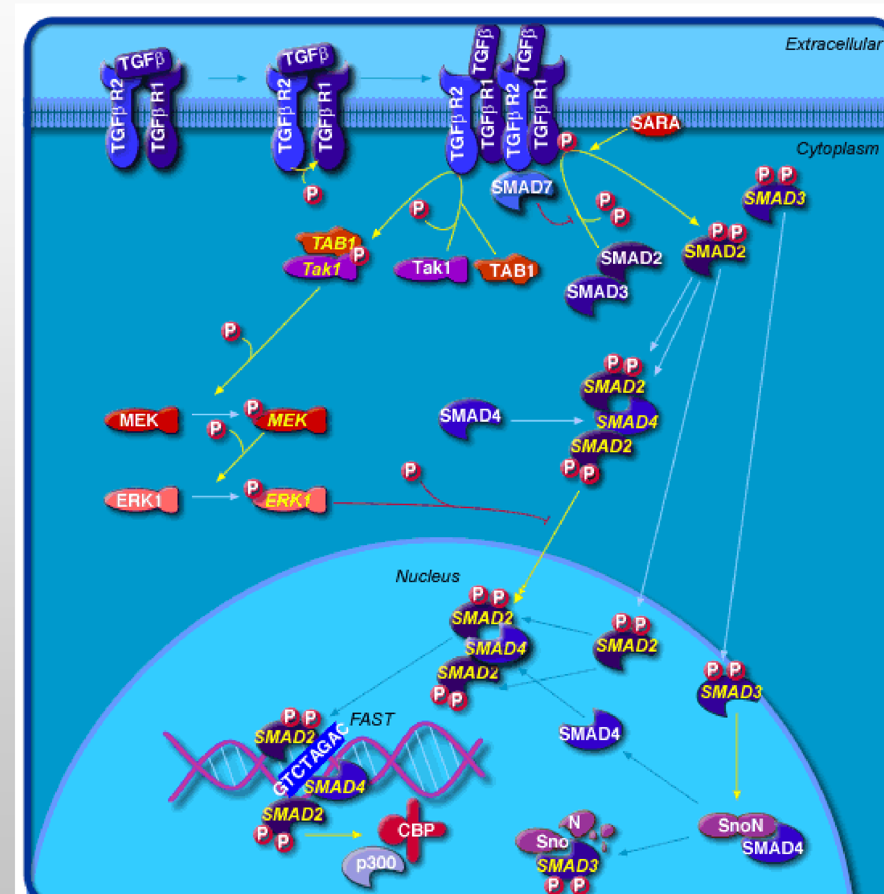
Expression Profiles



Segal *et al*, *RECOMB* 2002, ISMB 2003

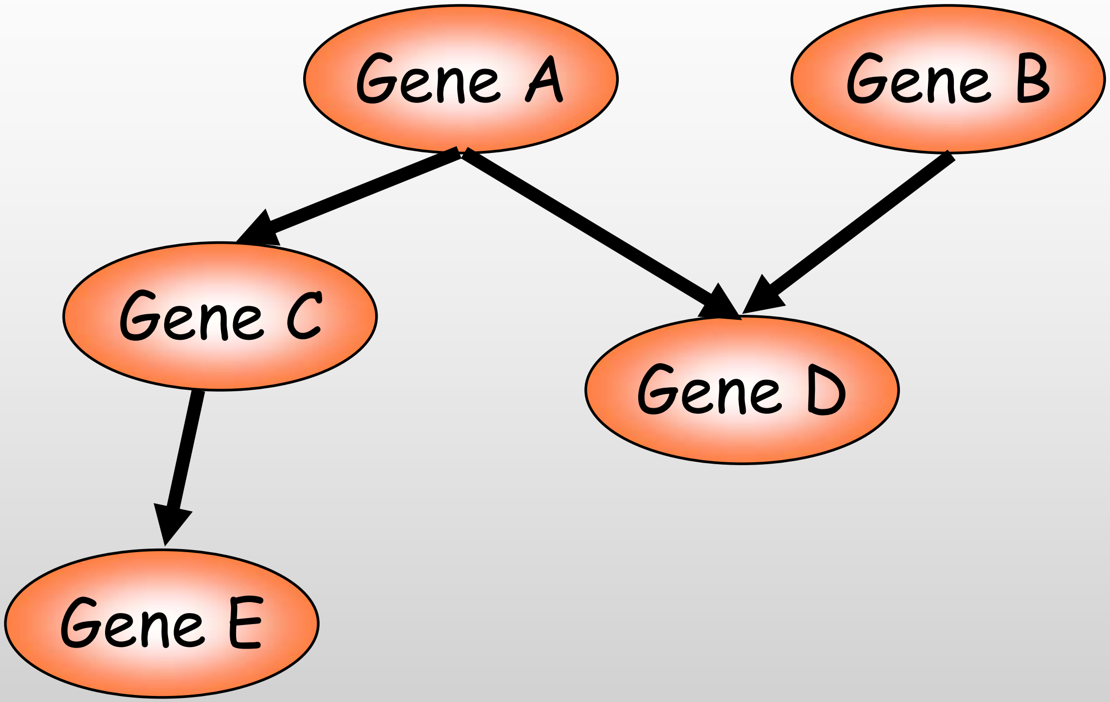
Goal: Reconstruct Cellular Networks





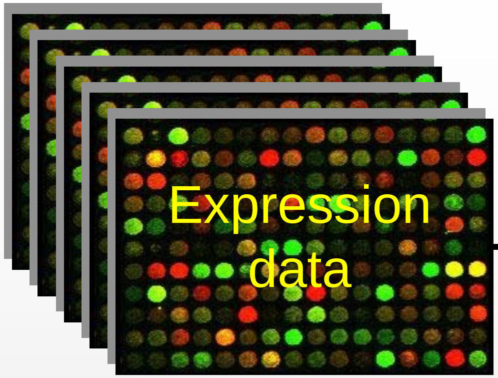
Biocarta. http://www.biocarta.com/

Causal Reconstruction for Gene Expression



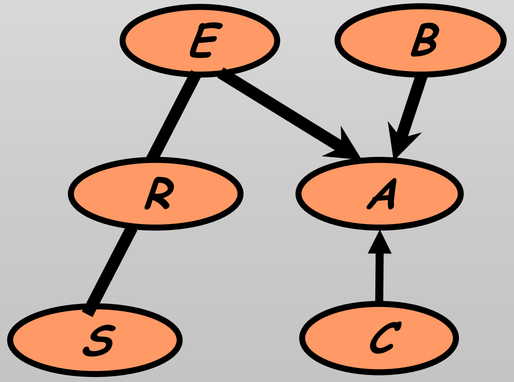
- Use language of Bayesian networks to reconstruct causal connections

Friedman et al, JCB 2000



Guided K-means Discretization

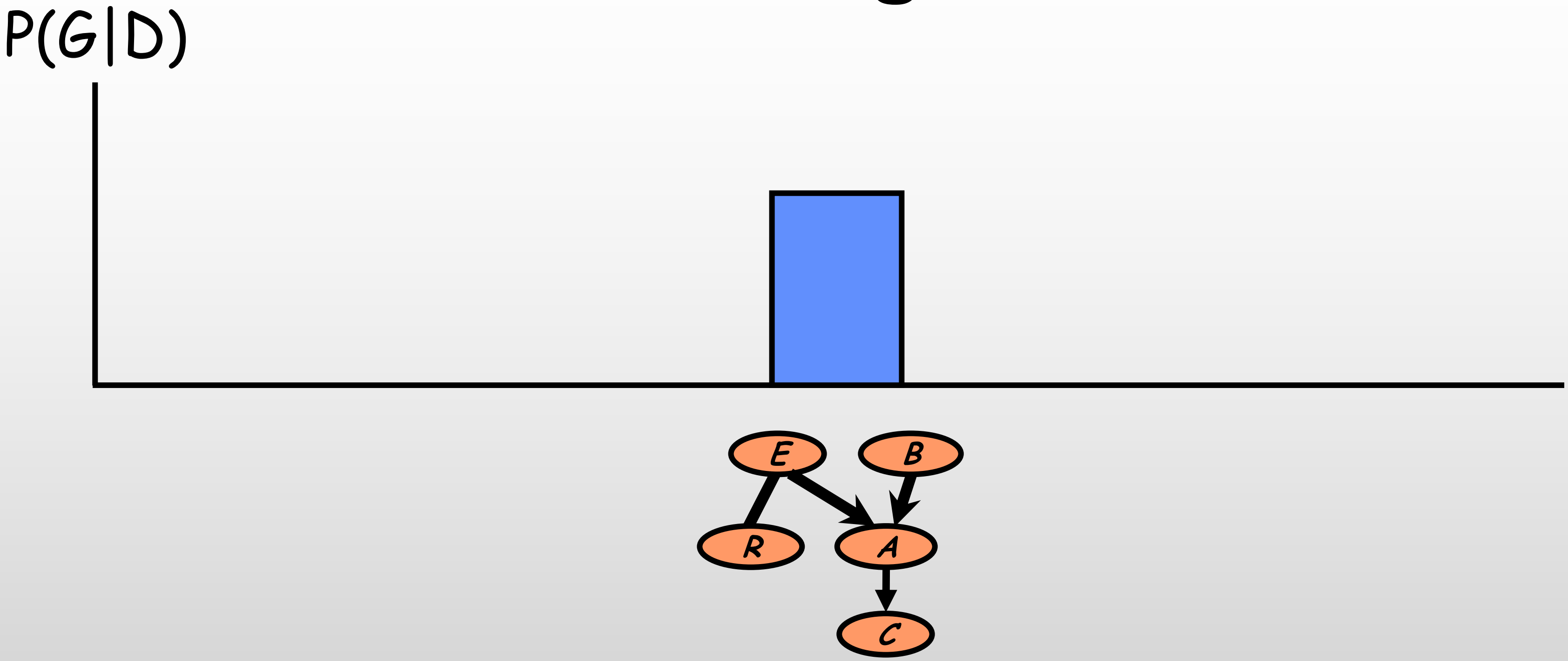
Bayesian Network Learning Algorithm



Properties of interactions among genes

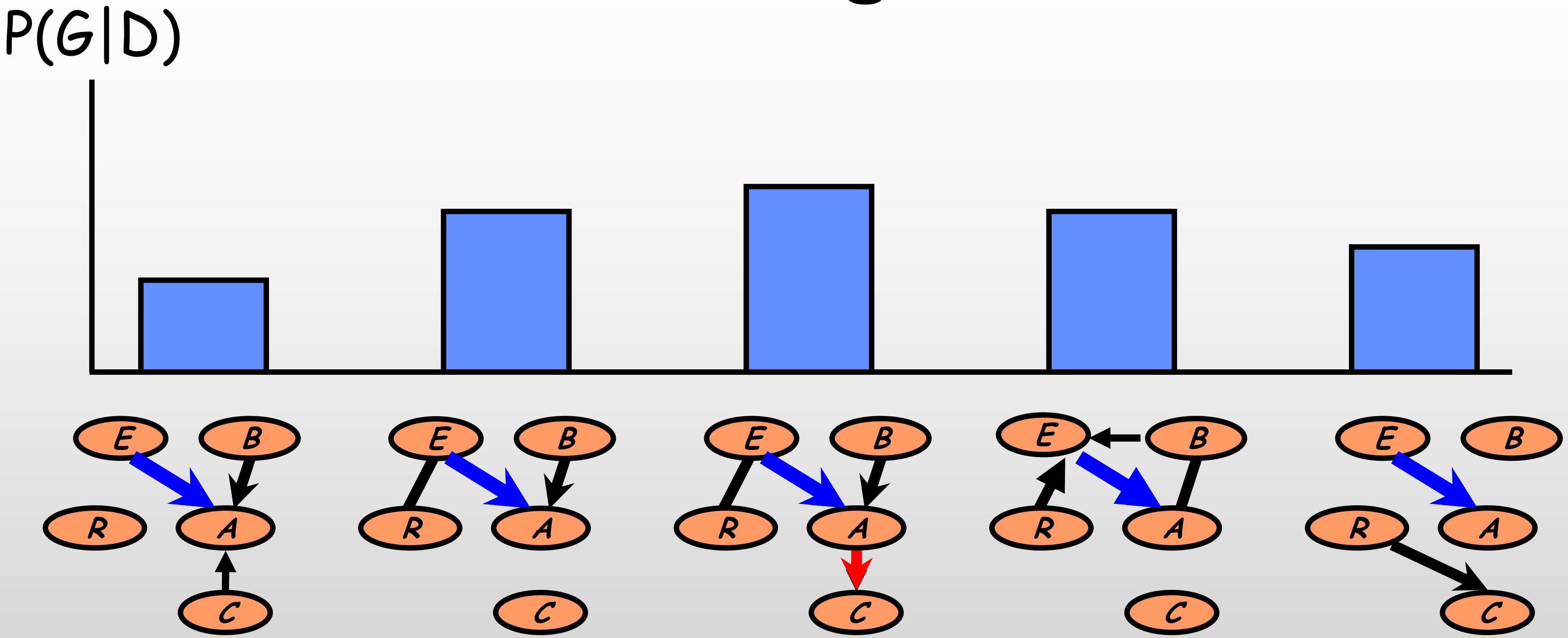
Critical question: do we believe the structure?

Discovering Structure



- Model selection
  - Pick a single high-scoring model
  - Use that model to infer domain structure

Discovering Structure



Problem

- Small sample size $\Rightarrow$ many high scoring models
- Answer based on one model often useless
- Want features common to many models

Bayesian Approach

- Posterior distribution over structures
- Estimate probability of **features**
  - Edge X→Y
  - Path X→… → Y
  - …

$$P(f \mid D) = \sum_G f(G) P(G \mid D)$$

Bayesian score for G

Feature of G, e.g., $X \rightarrow Y$

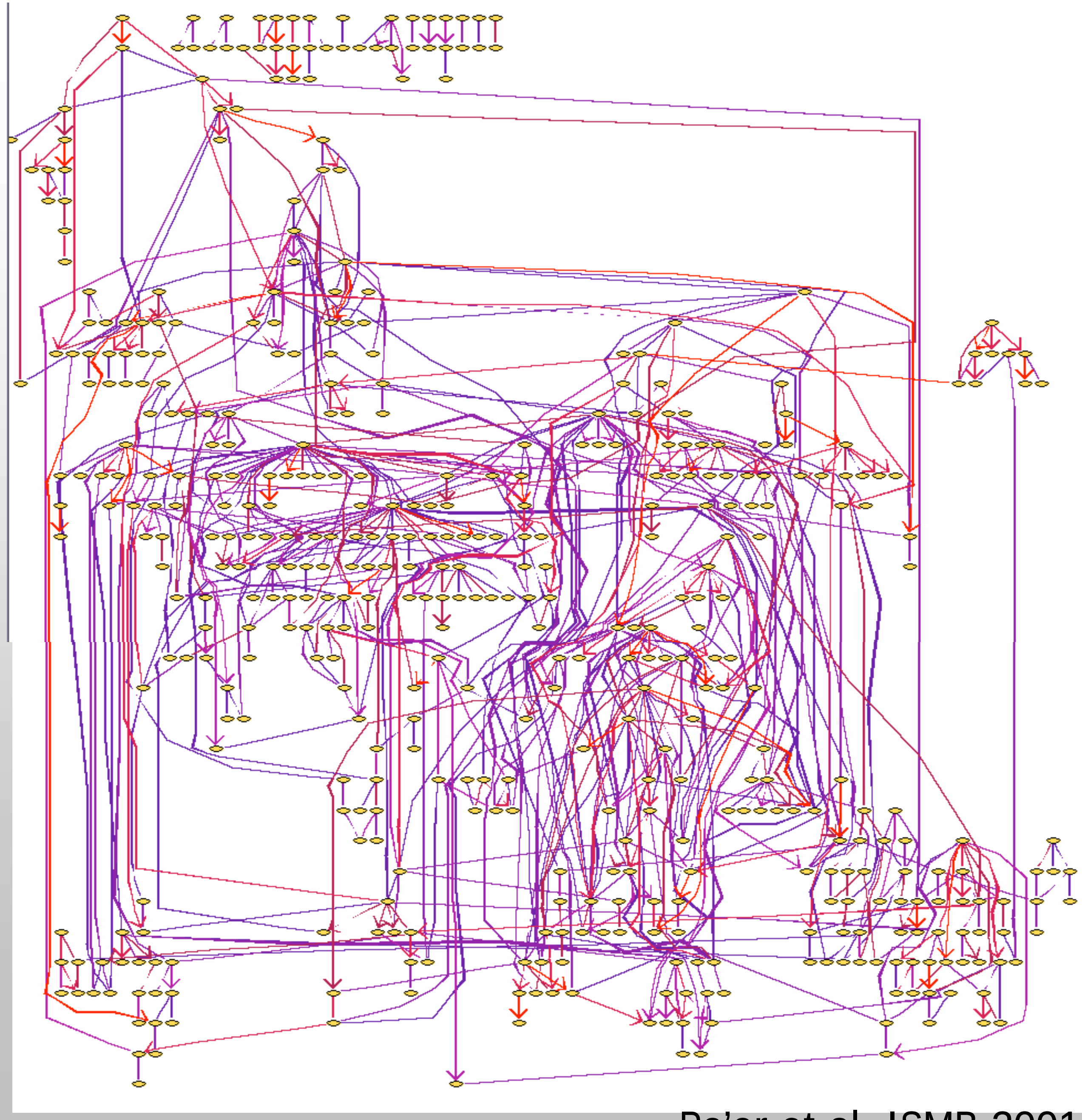
Indicator function for feature f

Two estimation methods:

- Bootstrap
- Markov Chain Monte Carlo

Experiment

- 300 deletion knockout in yeast [Hughes et al 2000]
- 600 genes
- Color code showing confidence on edges

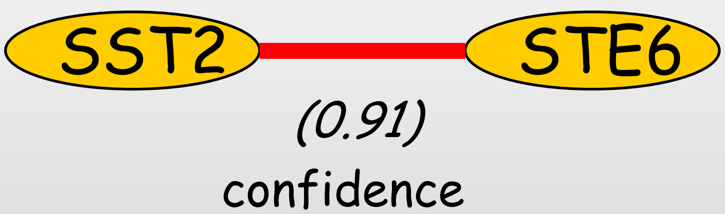


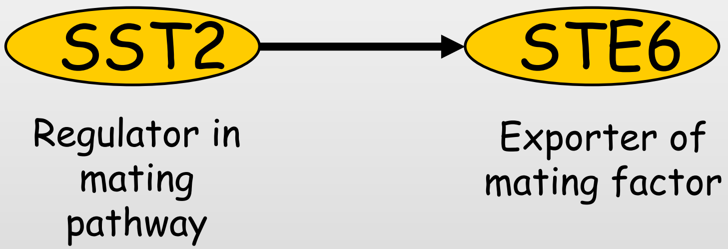
Pe’er et al, ISMB 2001

Markov Relations

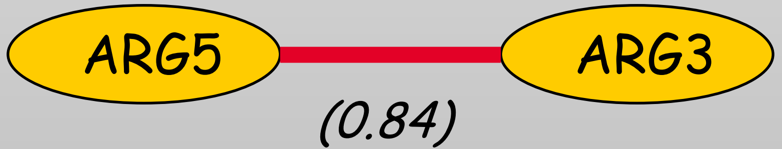
Question: Do X and Y directly interact?

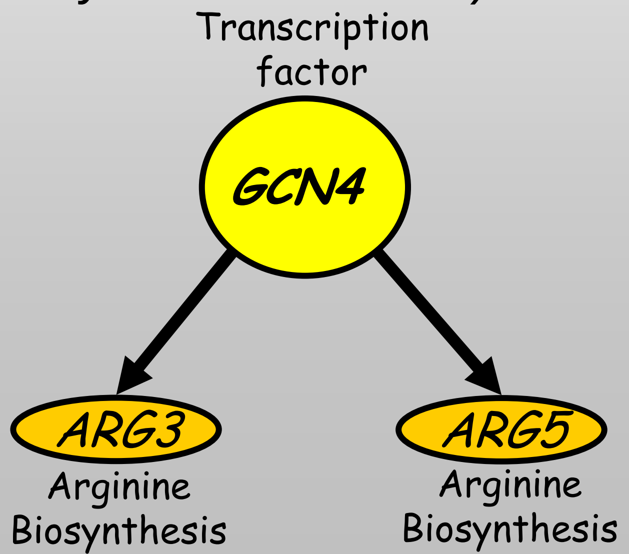
Parent-Child (one gene regulating the other)



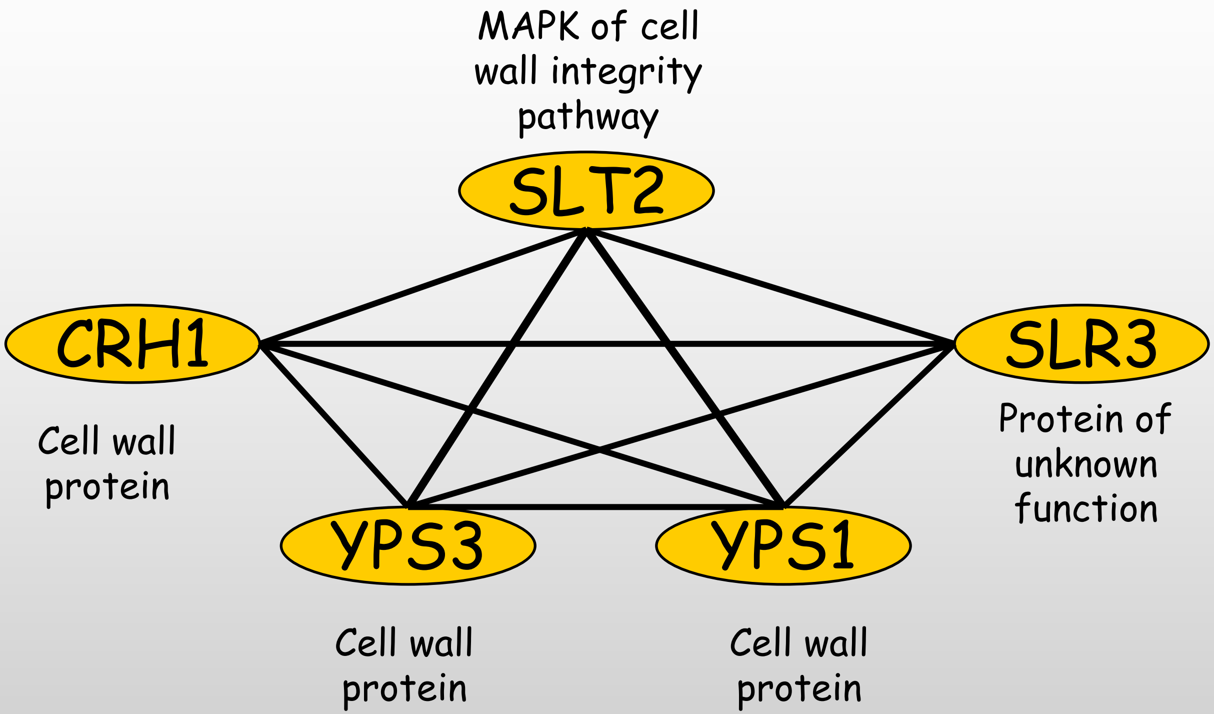


Hidden Parent (two genes co-regulated by a hidden factor)



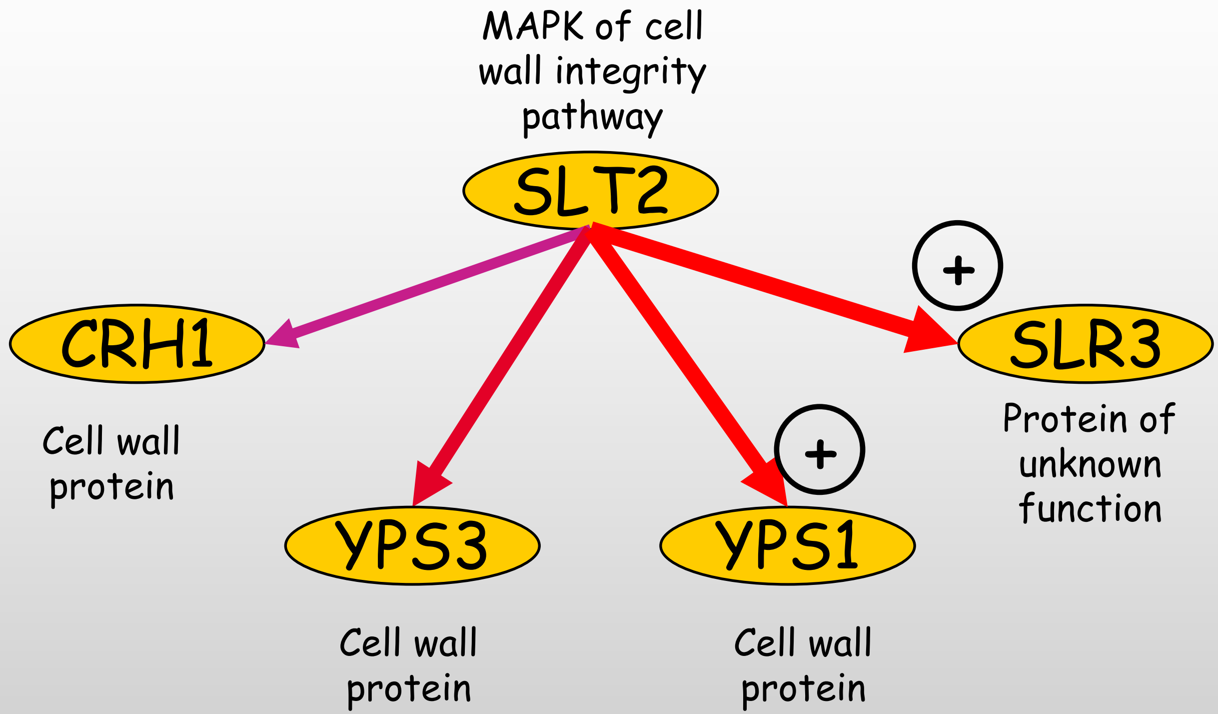


Separators: Intra-cluster Context

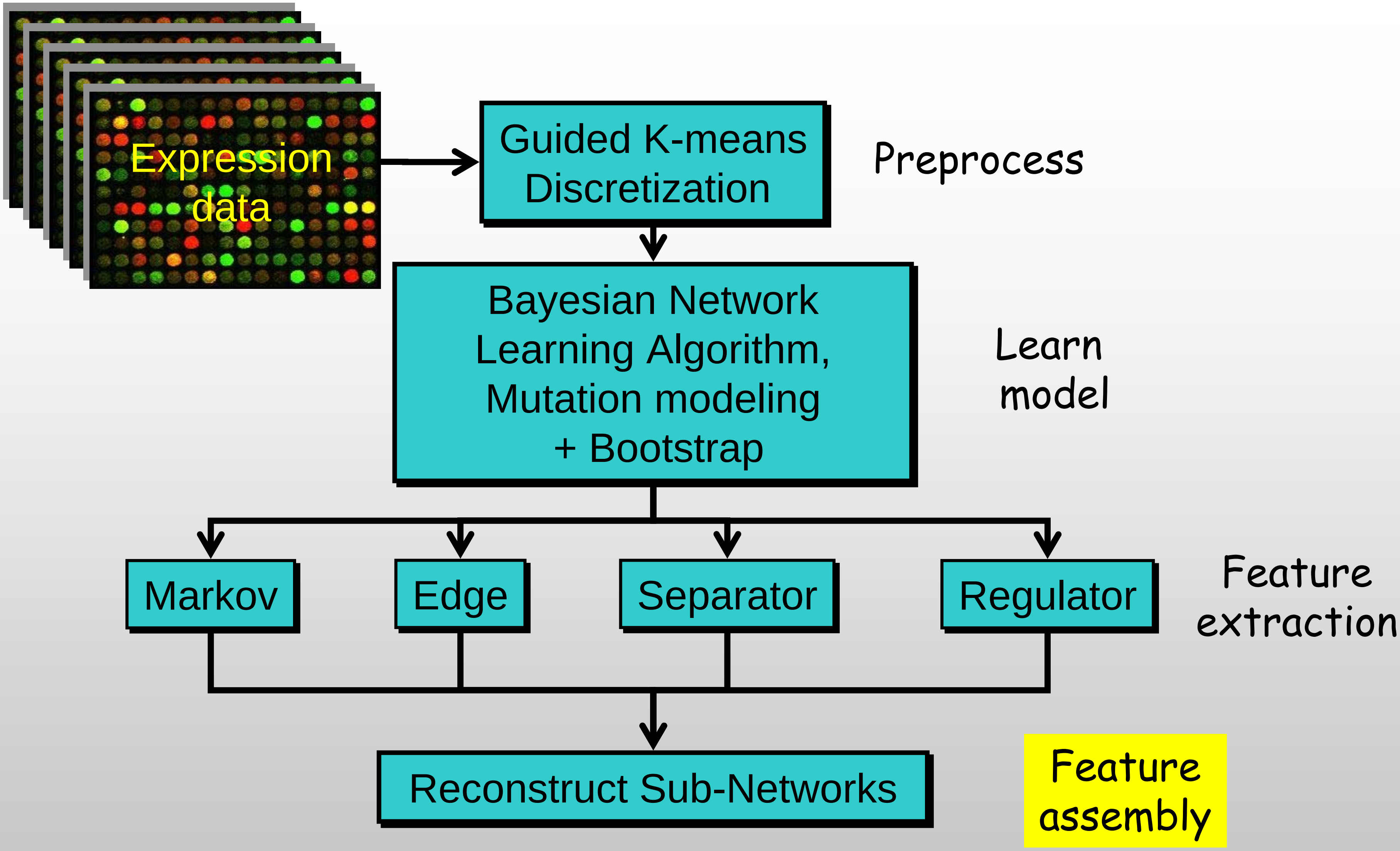


- All pairs have high correlation
- Clustered together

Separators: Intra-cluster Context



- SLT2: Pathway regulator, explains the dependence
- Many signaling and regulatory proteins identified as direct and indirect separators

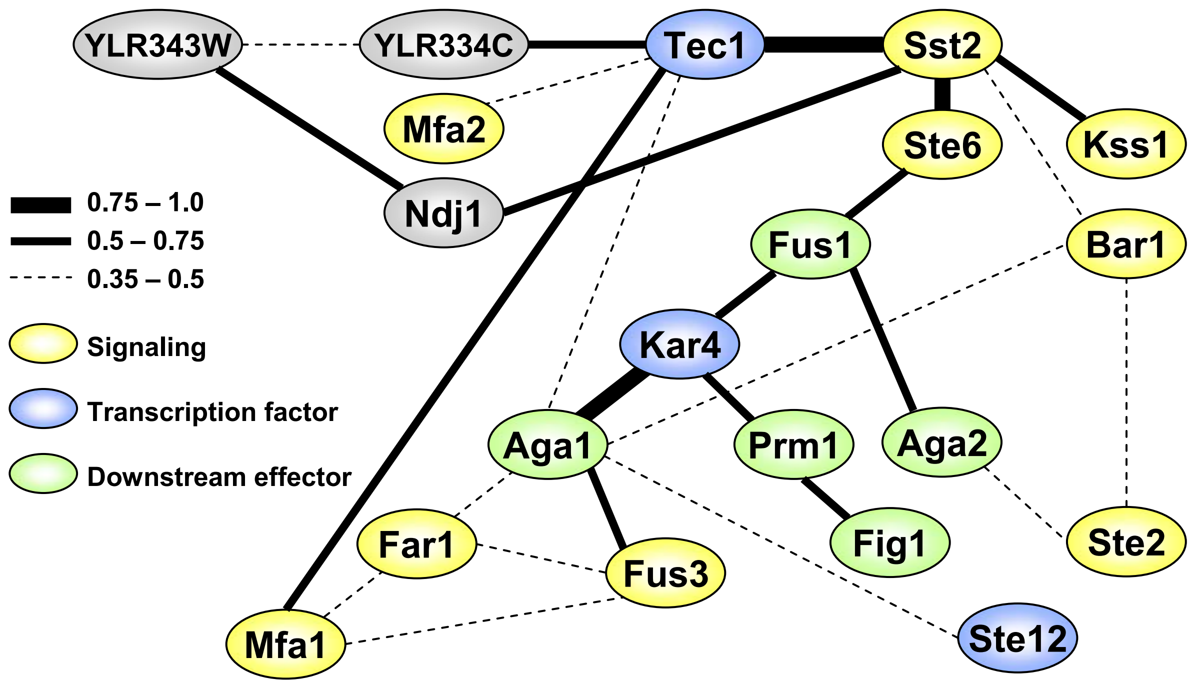


Global network → Local features → Sub-network

Subnetworks in Compendium Dataset

◆6 … ng
co

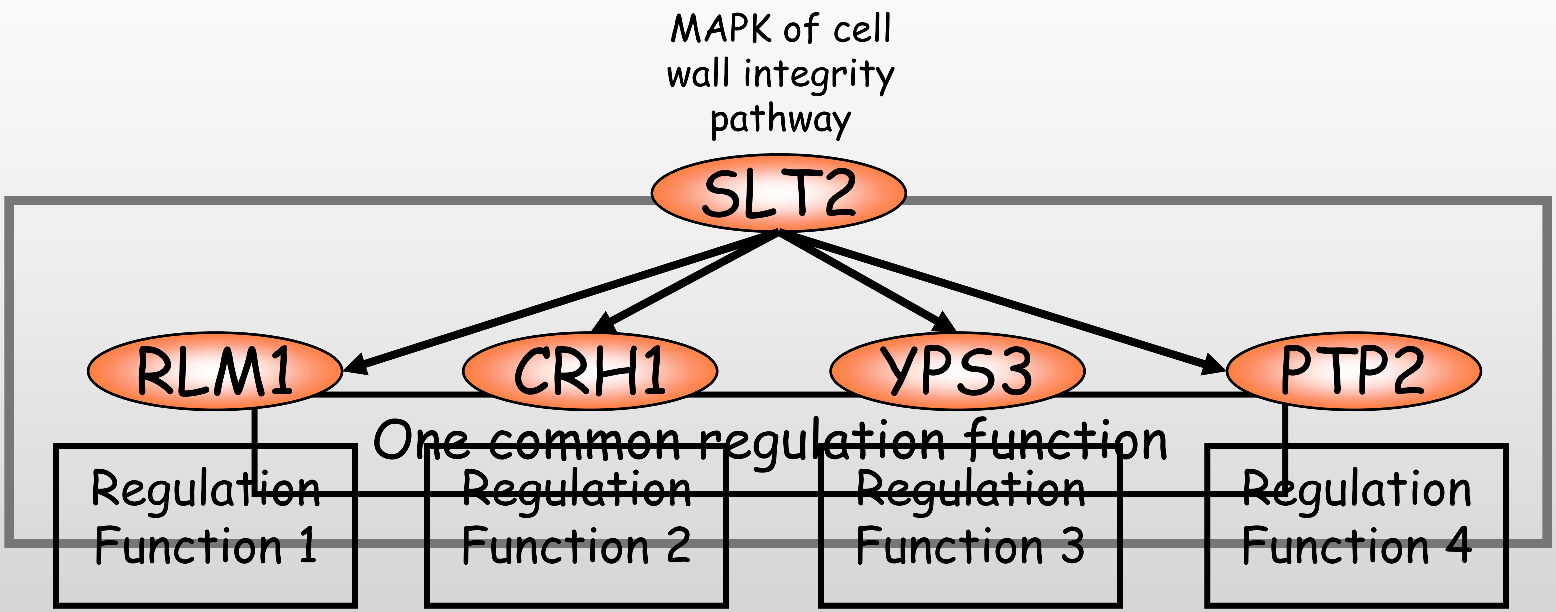
… on



"Mating response" Substructure

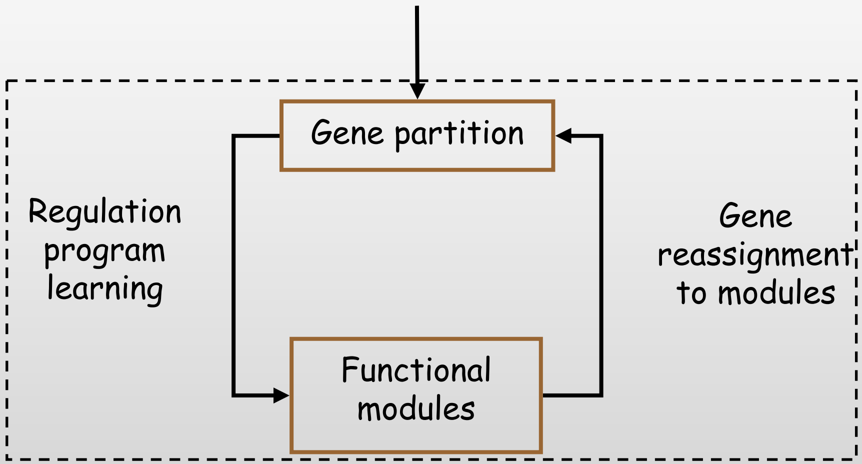
STE12 (main TF), Fus3 (Main MAPK) are marginal

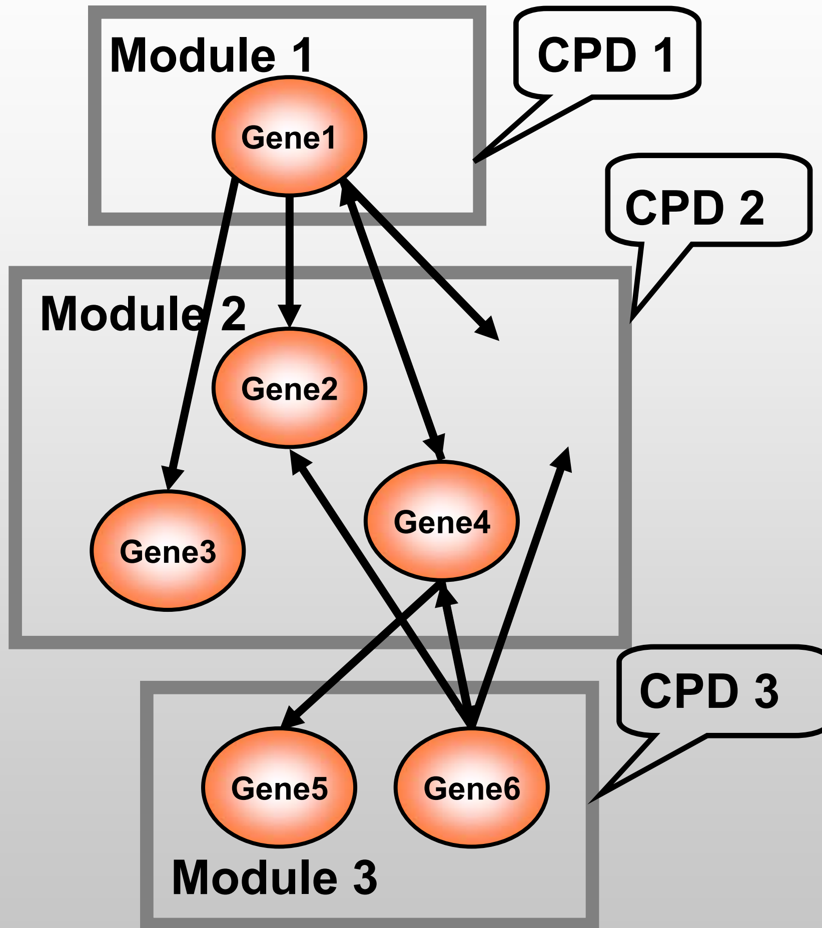
From Networks to Modules



Idea: enforce common regulatory program

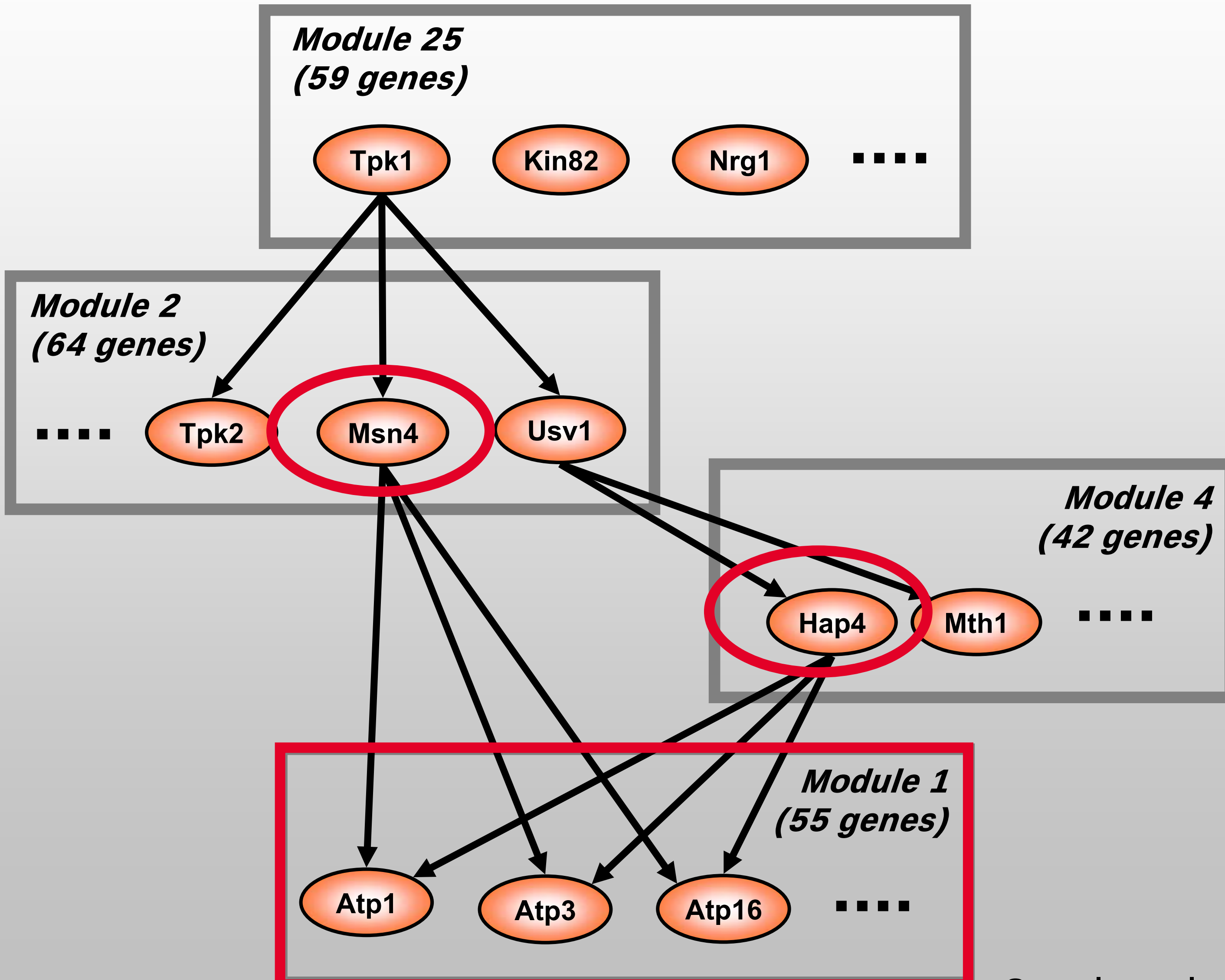
- Statistical robustness: Regulation programs are estimated from $m*k$ samples
- Organization of genes into regulatory modules: Concise biological description



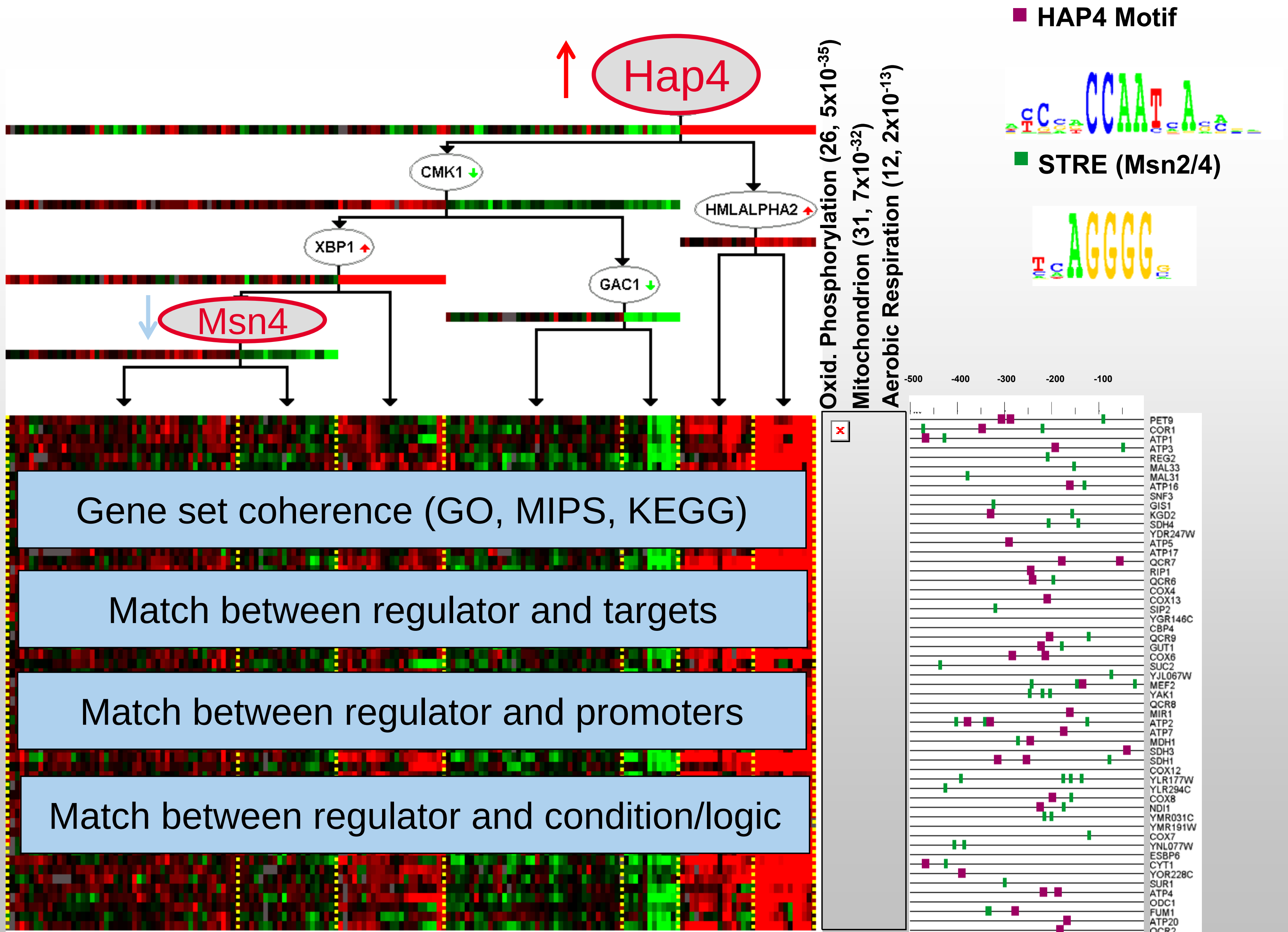


Segal et al, Nat Gen 2003

Learned Network (fragment)

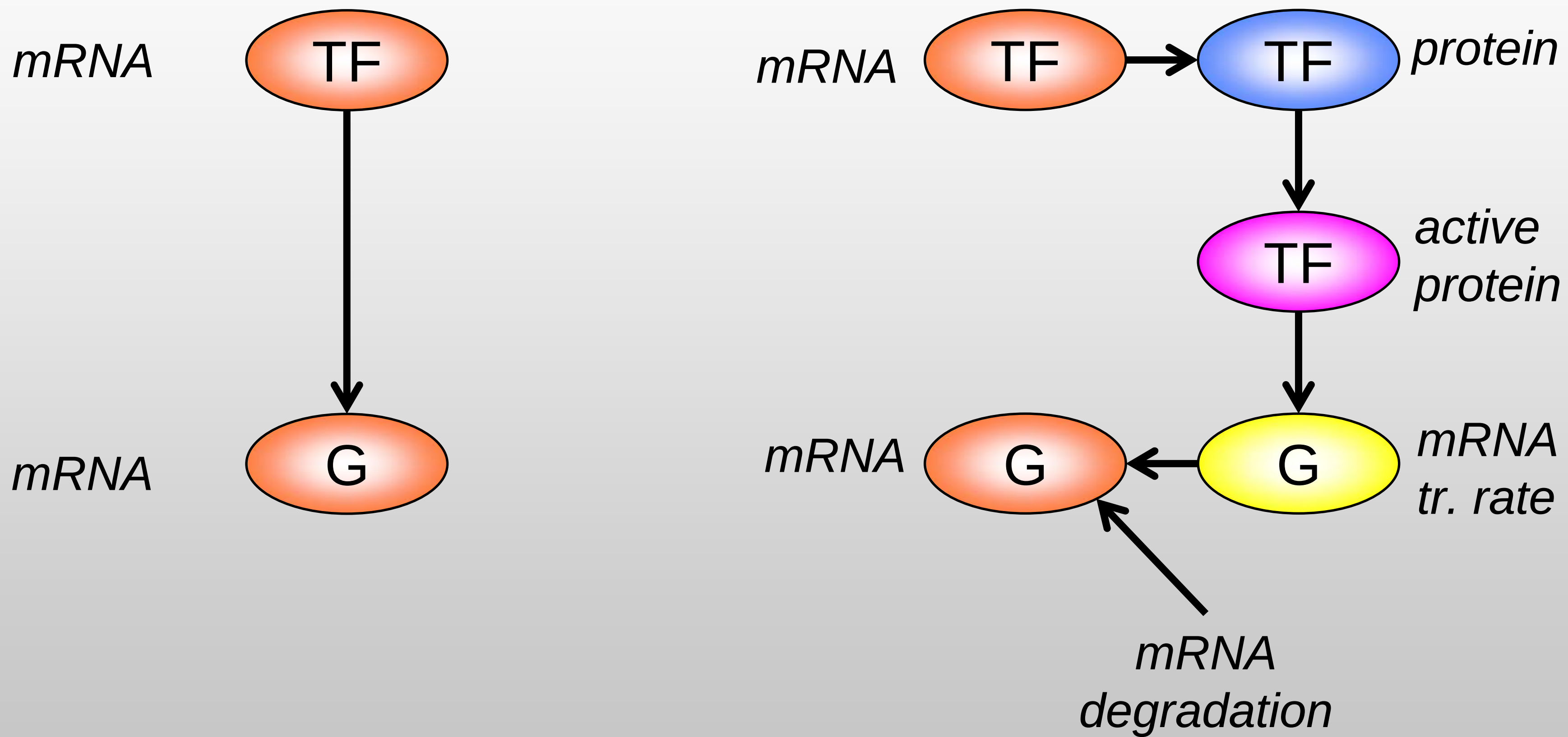


Segal et al, Nat Gen 2003



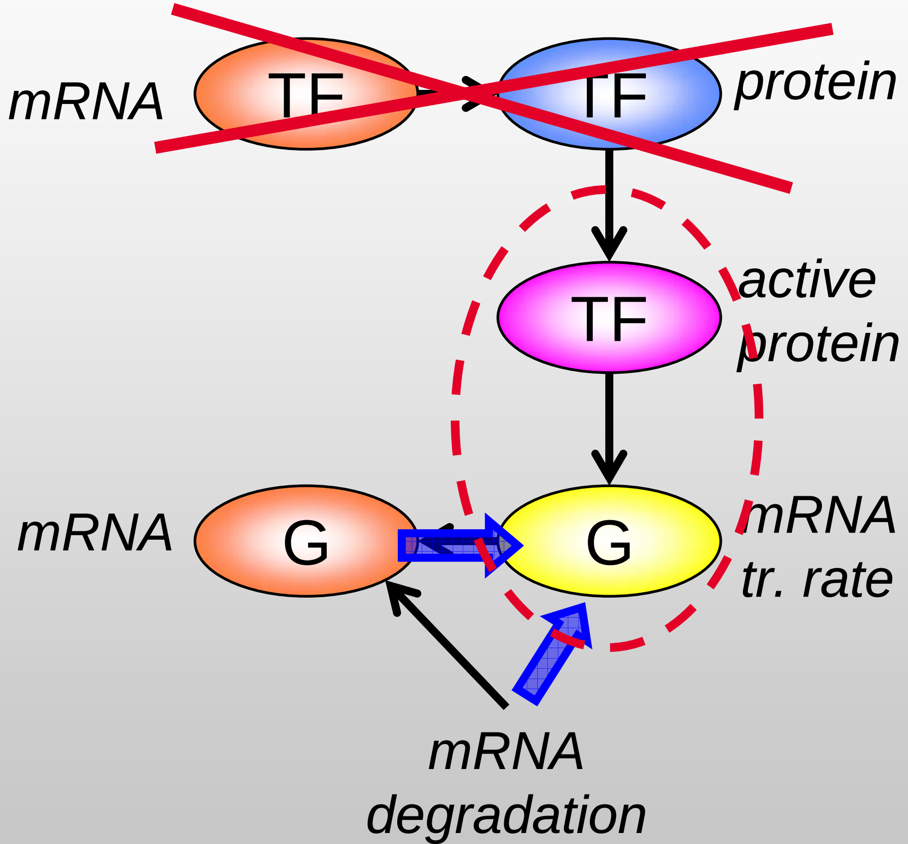
Segal et al, Nat Gen 2003

A Major Assumption



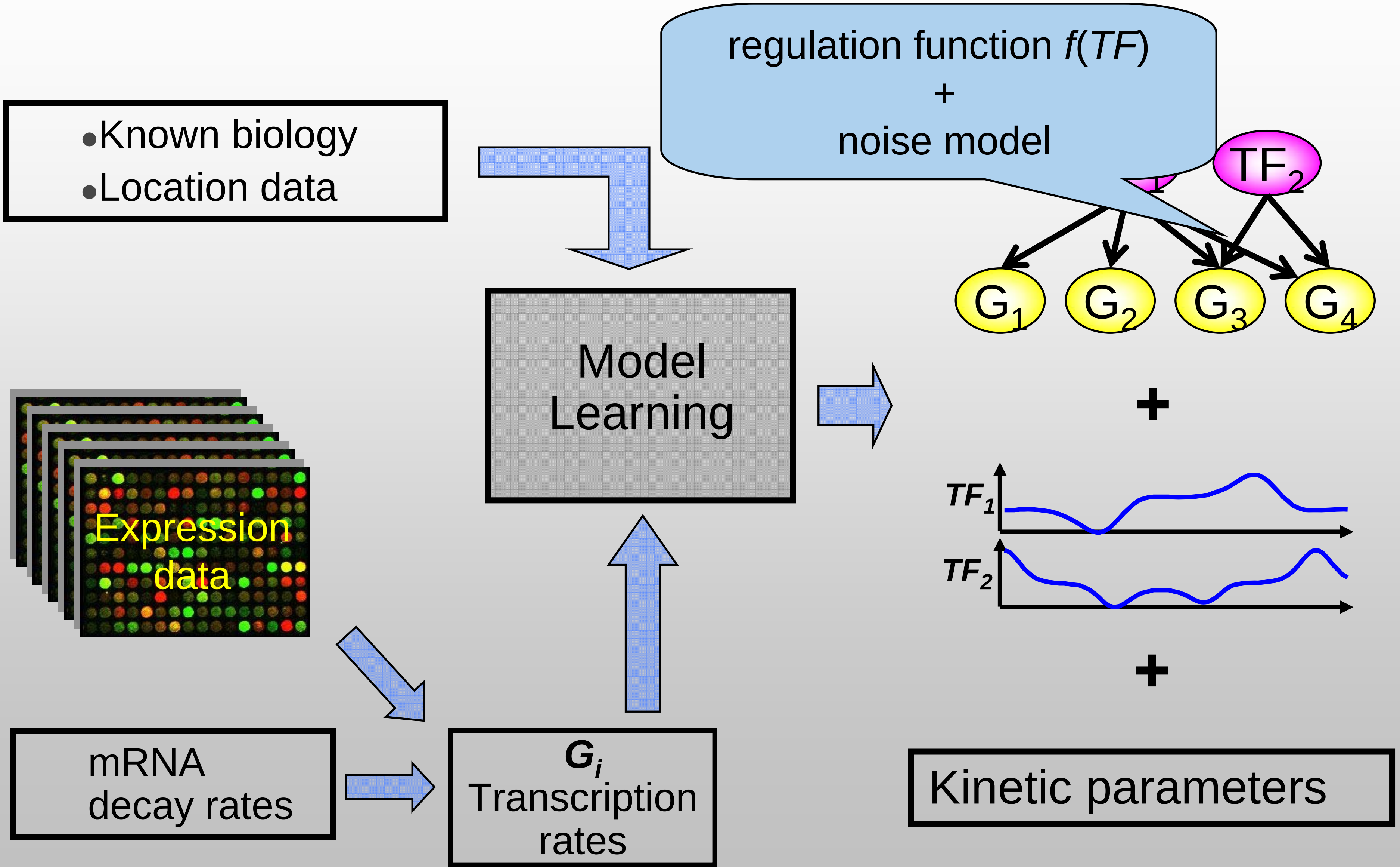
Realistic Regulation Modeling

- Model the closest connection



- Active protein levels are not measured
- Transcript rates are computed from expression data and mRNA decay rates

New Proposed Scheme



Nachman et al, ISMB 2004

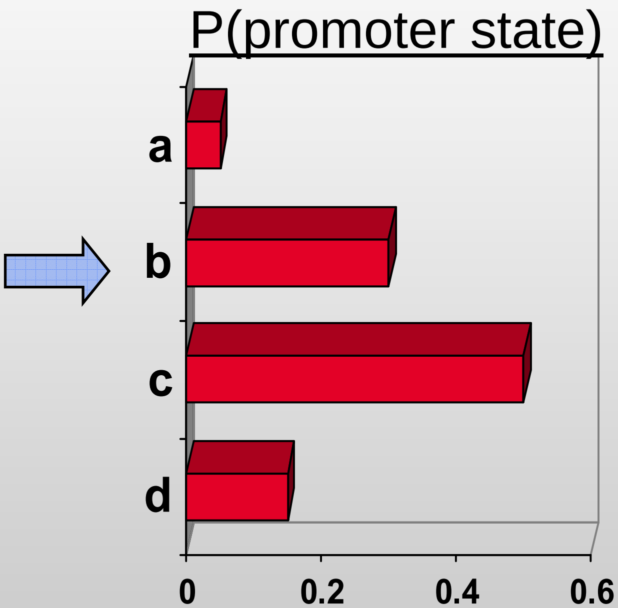
General Two Regulator Function

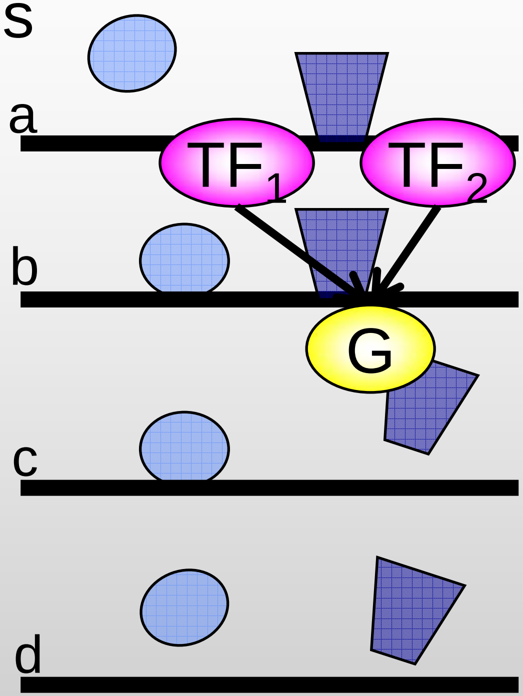
I. Compute distribution of promoter states

<u>State Equations</u>

$$S_1 + tf_1 \overset{\gamma_1}{\longleftrightarrow} S_1tf_1$$

$$S_2 + tf_2 \overset{\gamma_2}{\longleftrightarrow} S_2tf_2$$





f(TF₁, TF₂) should describe mean transcription rate of *G*

Nachman et al, ISMB 2004

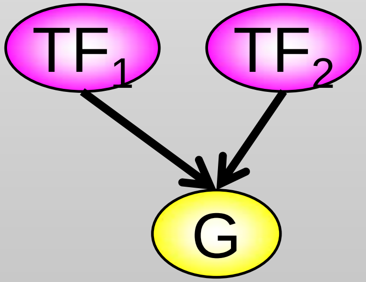
General Two Regulator Function

II. Assign activation level to each state

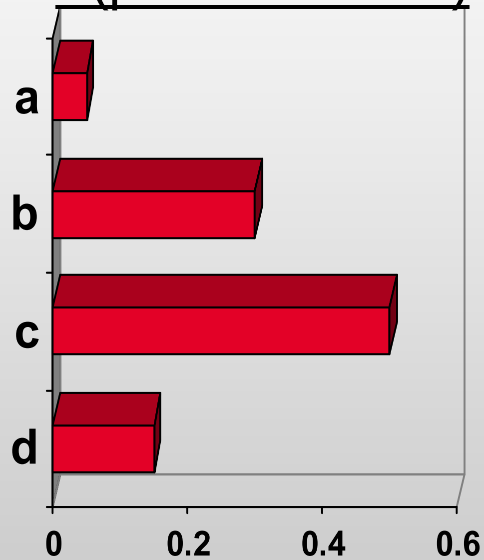
State Equations

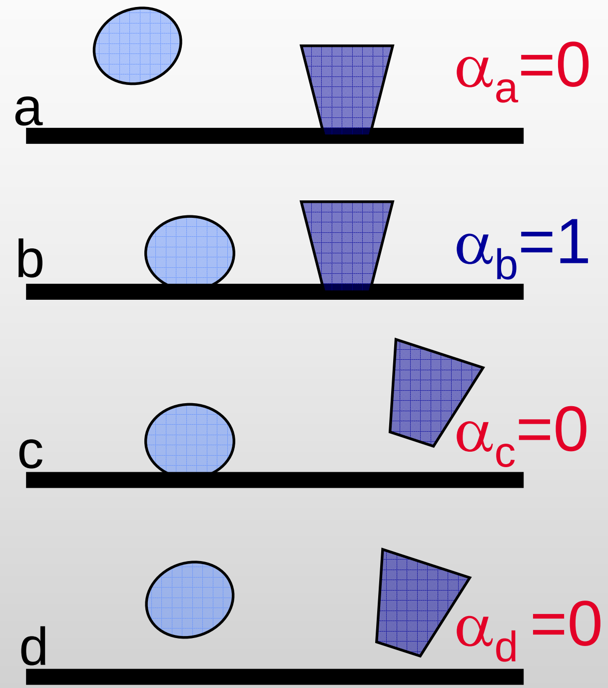
$$S_1 + tf_1 \overset{\gamma_1}{\longleftrightarrow} S_1 tf_1$$

$$S_2 + tf_2 \overset{\gamma_2}{\longleftrightarrow} S_2 tf_2$$



P(promoter state)





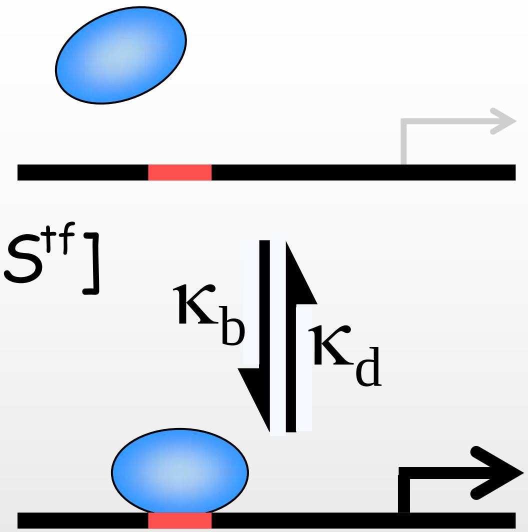
$$f = (\alpha_a \text{ X } \quad + \alpha_b \text{ X } \quad + \alpha_c \text{ X } \quad + \alpha_d \text{ X } \quad) \text{ X } \beta$$

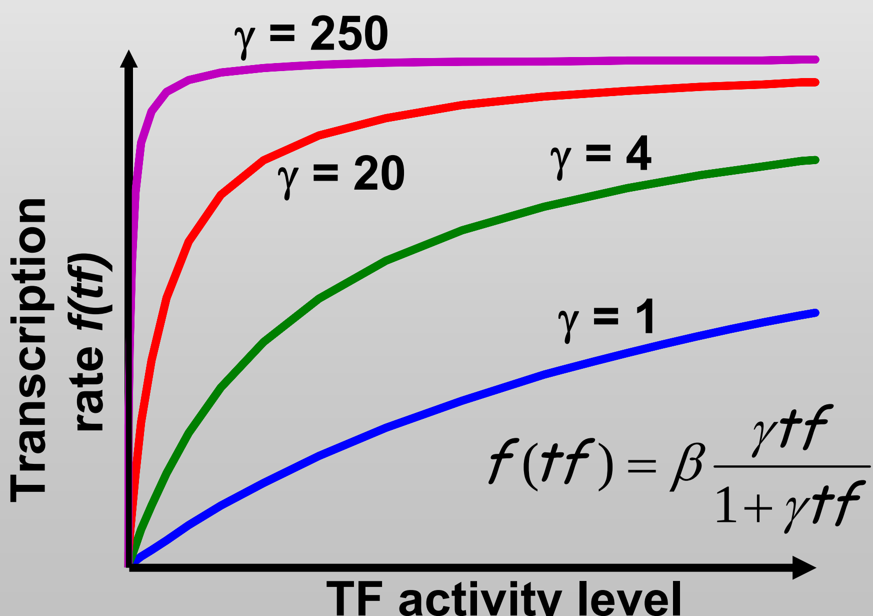
Nachman et al, ISMB 2004

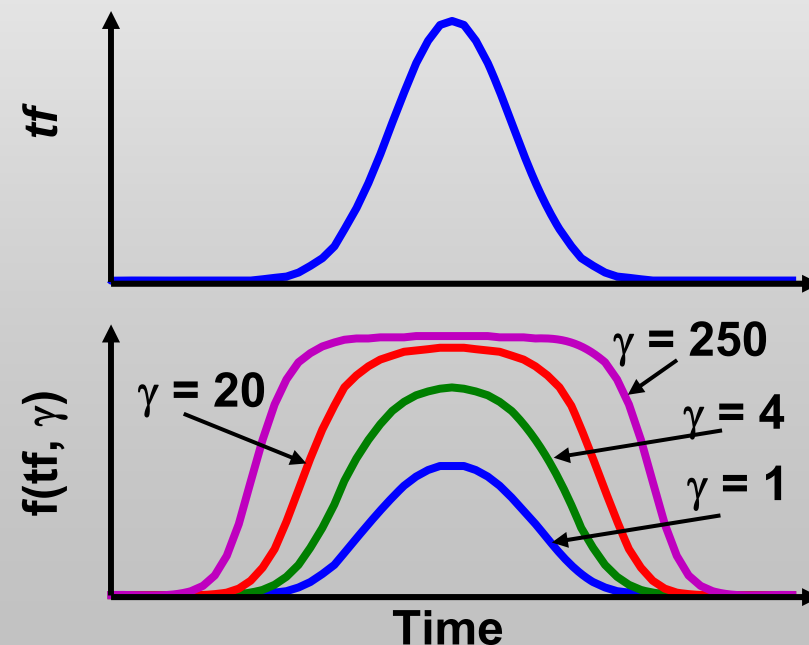
Example: One Activator Function

$$\kappa_b[S^-][tf] = \kappa_d[S^{tf}]$$

$$[S^-] + [S^{tf}] = 1$$



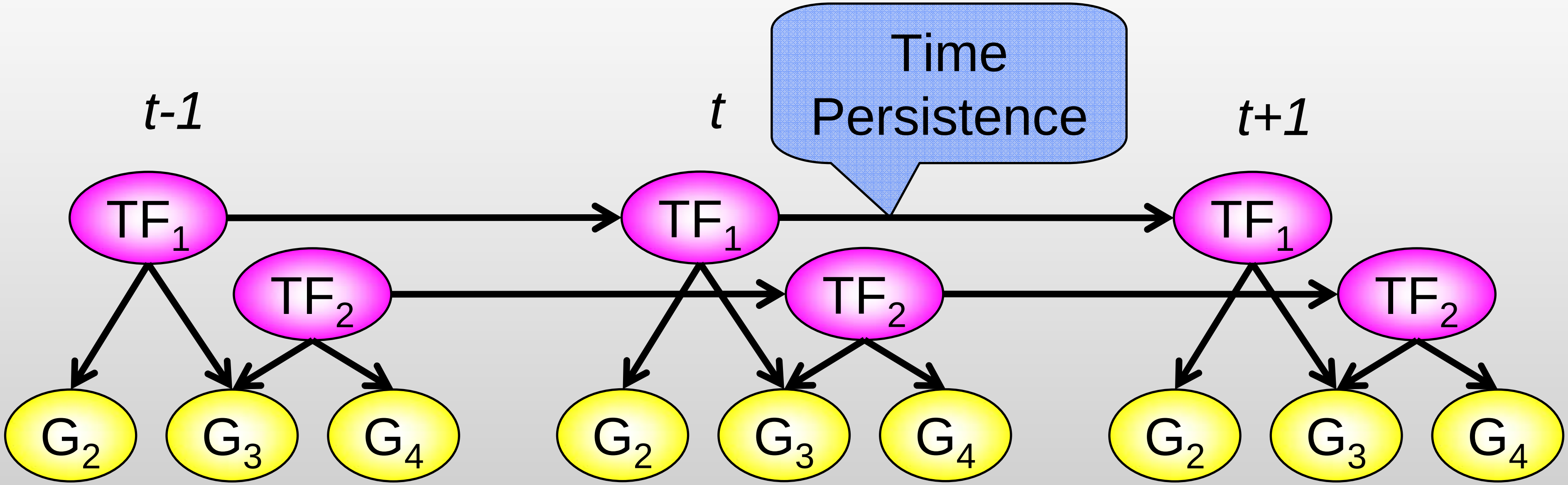




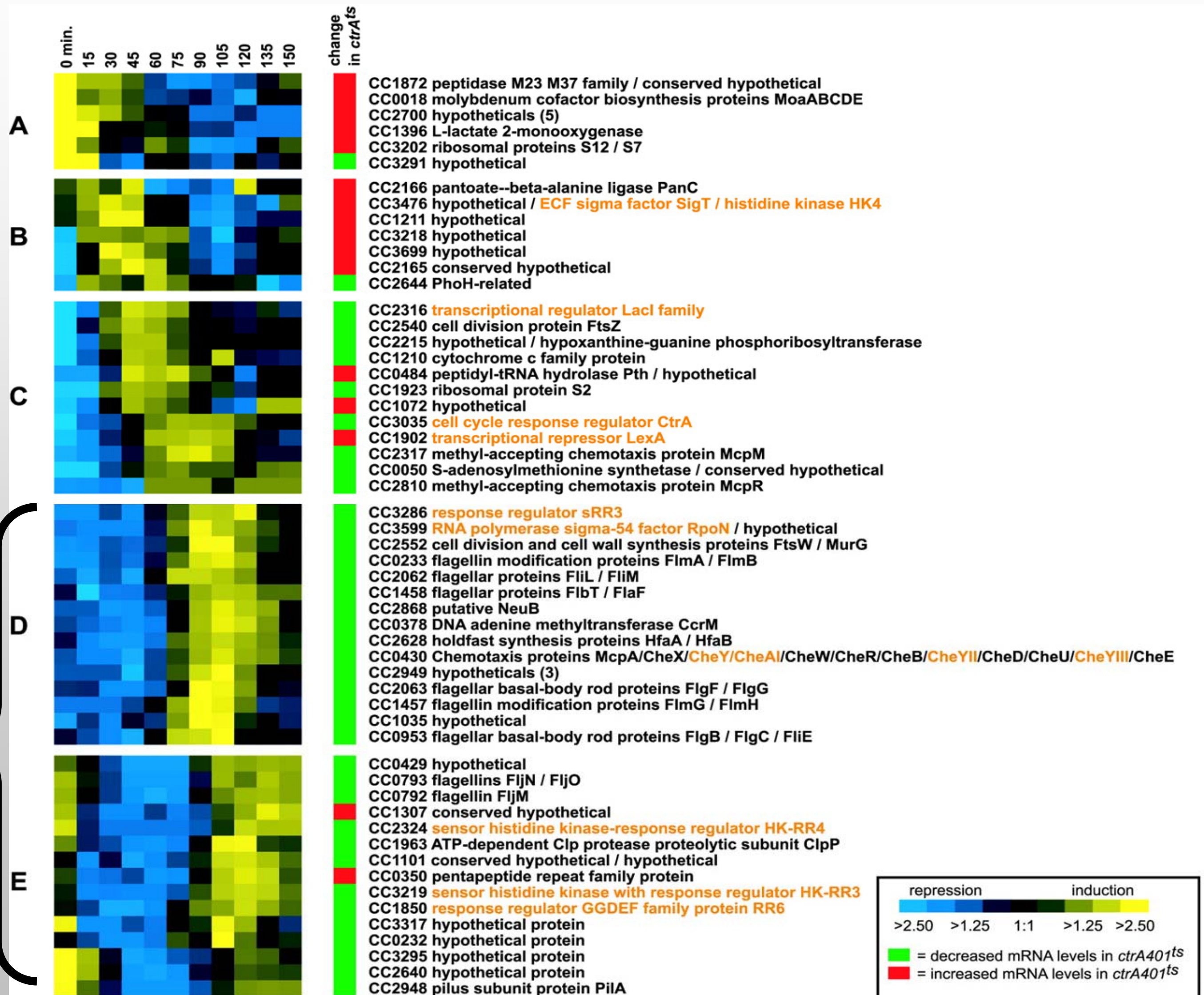
Nachman et al, ISMB 2004

Adding a Temporal Aspect

For time series – add explicit time modeling



Caulobacter CtrA regulon

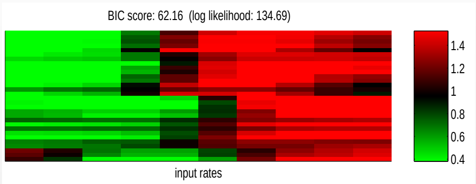


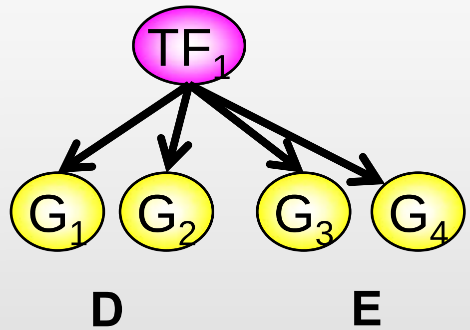
Is There a Second Regulator?

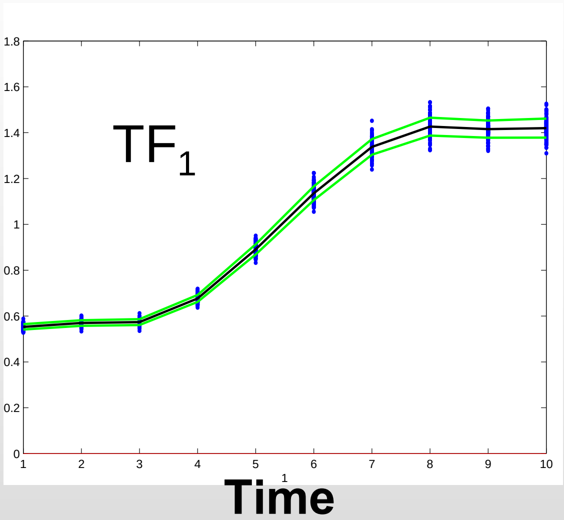
i1

Caulobacter CtrA regulon

Input r



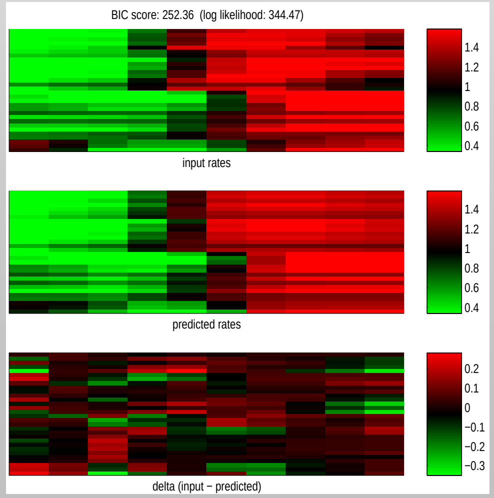


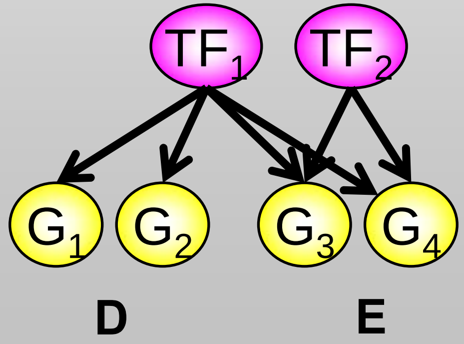


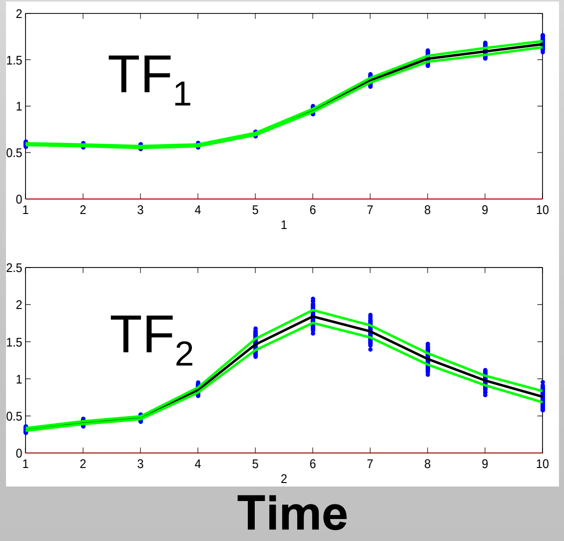
Input r

Pred. r

error







i1

show in animation:
input r -> model -> predicted h -> pred r -> error

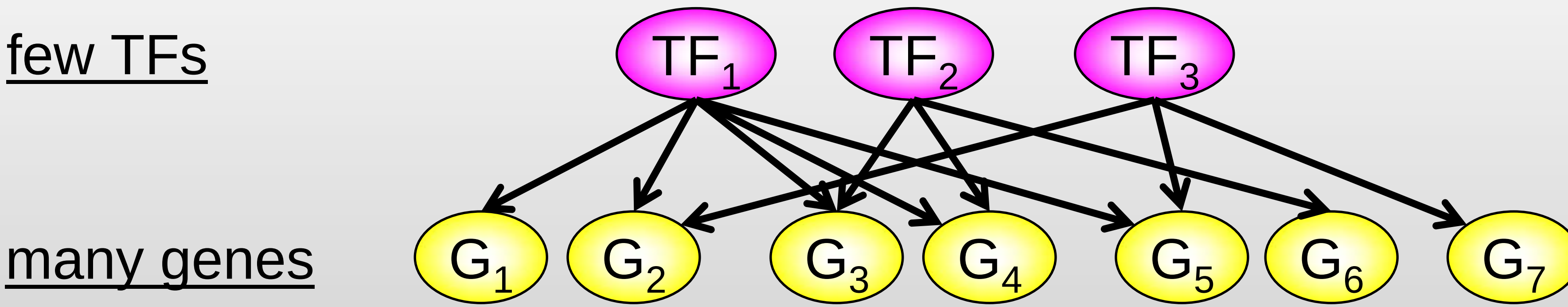
remove p-val and say in words

stress "realistic"

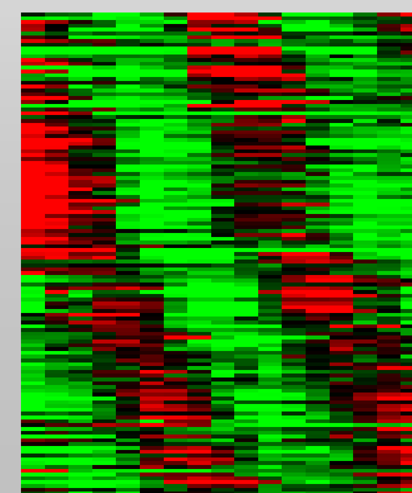
iftach, 3/31/2004

Multiple Regulon Experiments

Can we describe the cell transcriptome using a small number of hidden regulators?



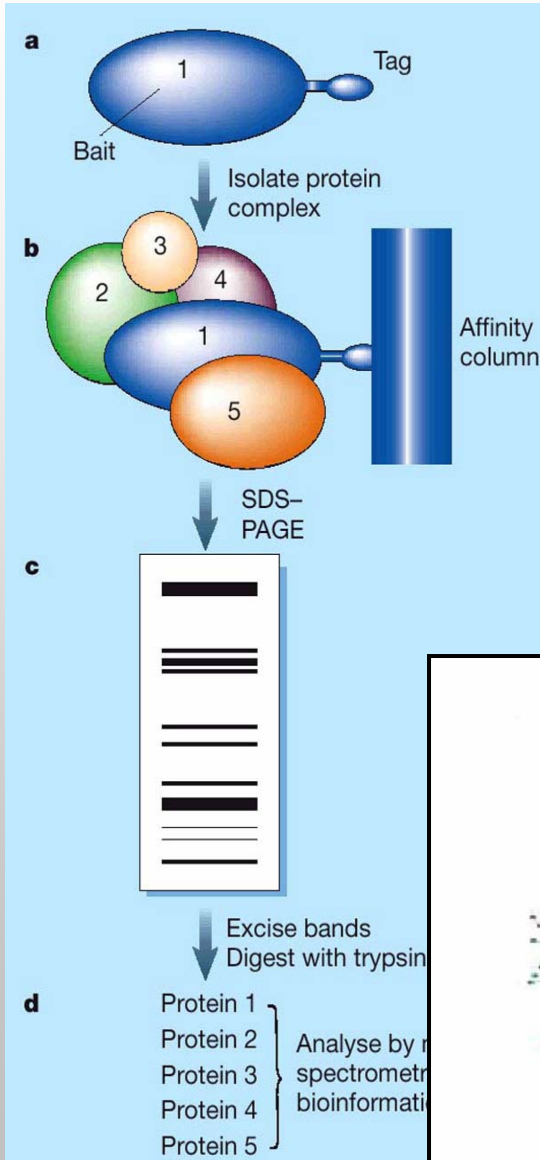
- "Realistic" dimensionality reduction
- Allows prediction of target gene dynamics

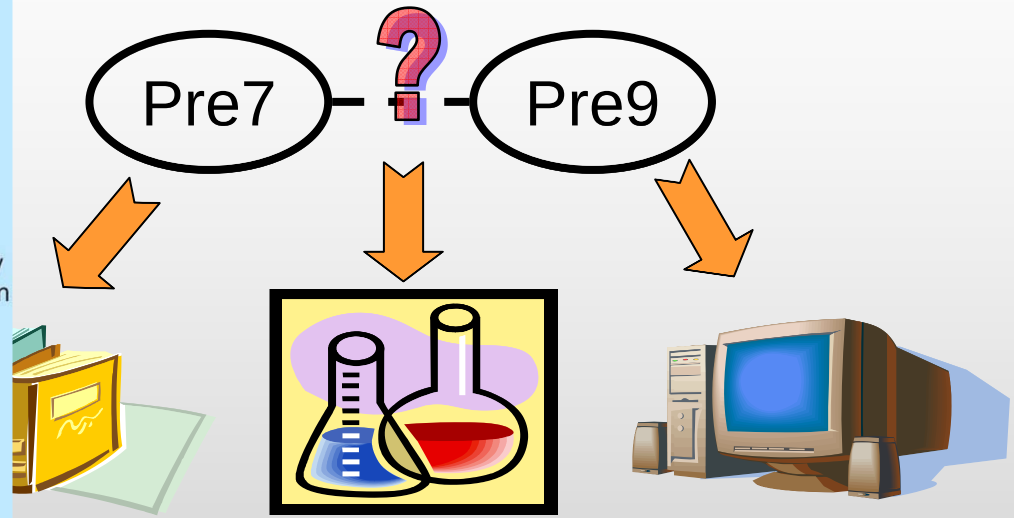


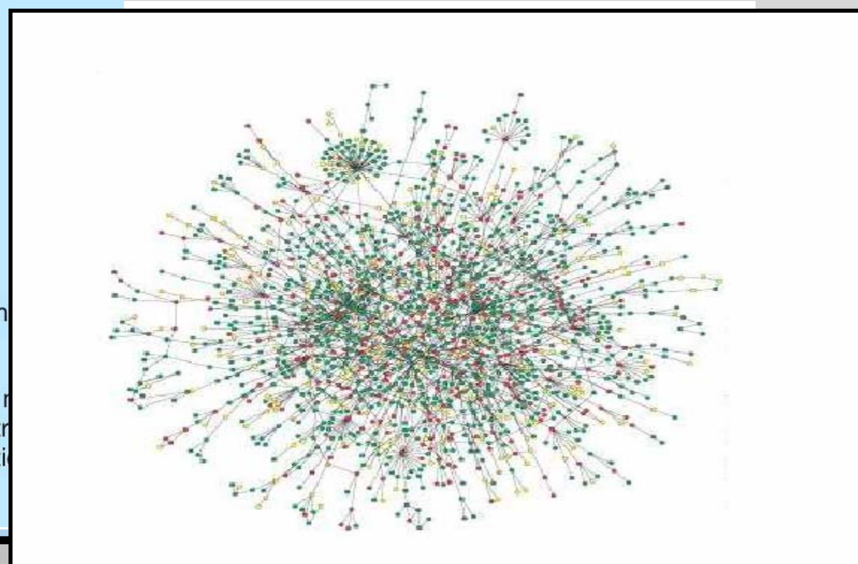
Outline

- Introduction
- Bayesian Networks
- Learning Bayesian Networks
- Transcriptional regulation
- Gene expression
- Markov Networks
- Protein-Protein Interactions
- Discussion

Protein-Protein Interactions







t two hybrid
et al. 2000
ogenetic profiles
2000

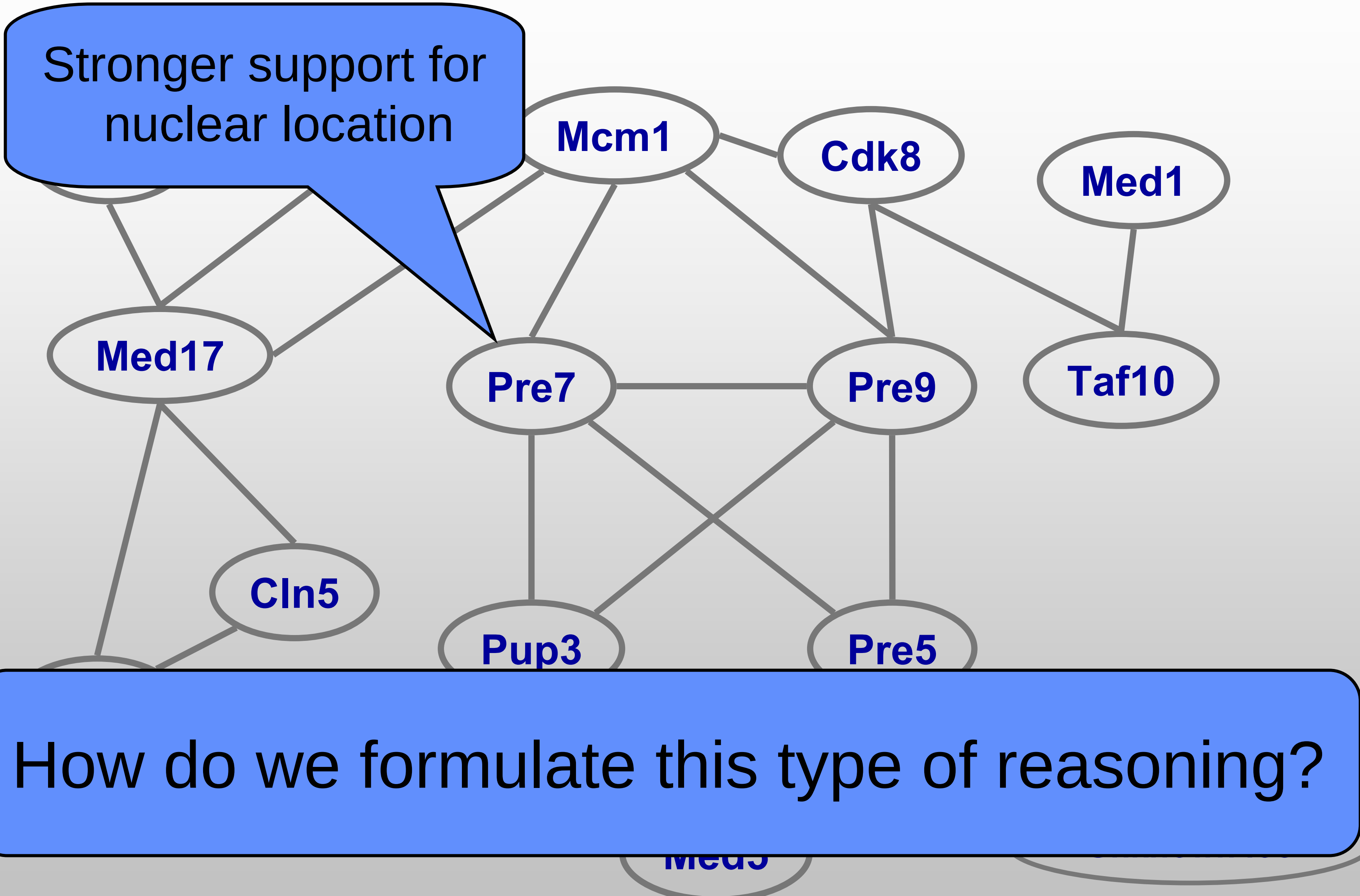
Using Protein-Protein Interactions

- Can we use interactions to better understand protein attributes?

Intuition: Interacting proteins tend to be similar

- In the same cellular compartment
- Involved in the same function
- Have similar expression patterns
- ...

Motivation

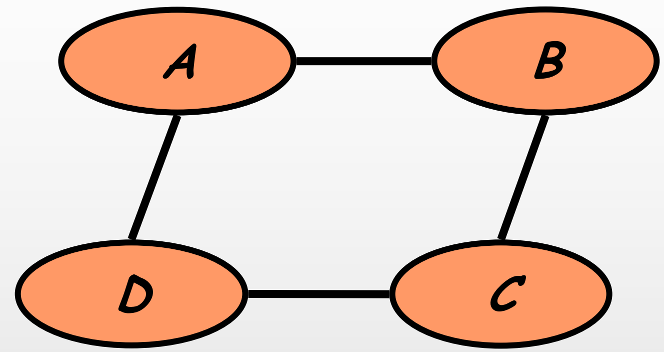


Stronger support for nuclear location

How do we formulate this type of reasoning?

Markov Networks

A	B	f_1
0	0	-1
0	1	0
1	0	1
1	1	1



Define jo … n

$$P(A,B,C,D) = \frac{1}{Z} \mathrm{Exp}\left(f_1(A,B) + f_2(B,C) + f_3(C,D) + f_4(D,E)\right)$$

Normalization constant

Potential function

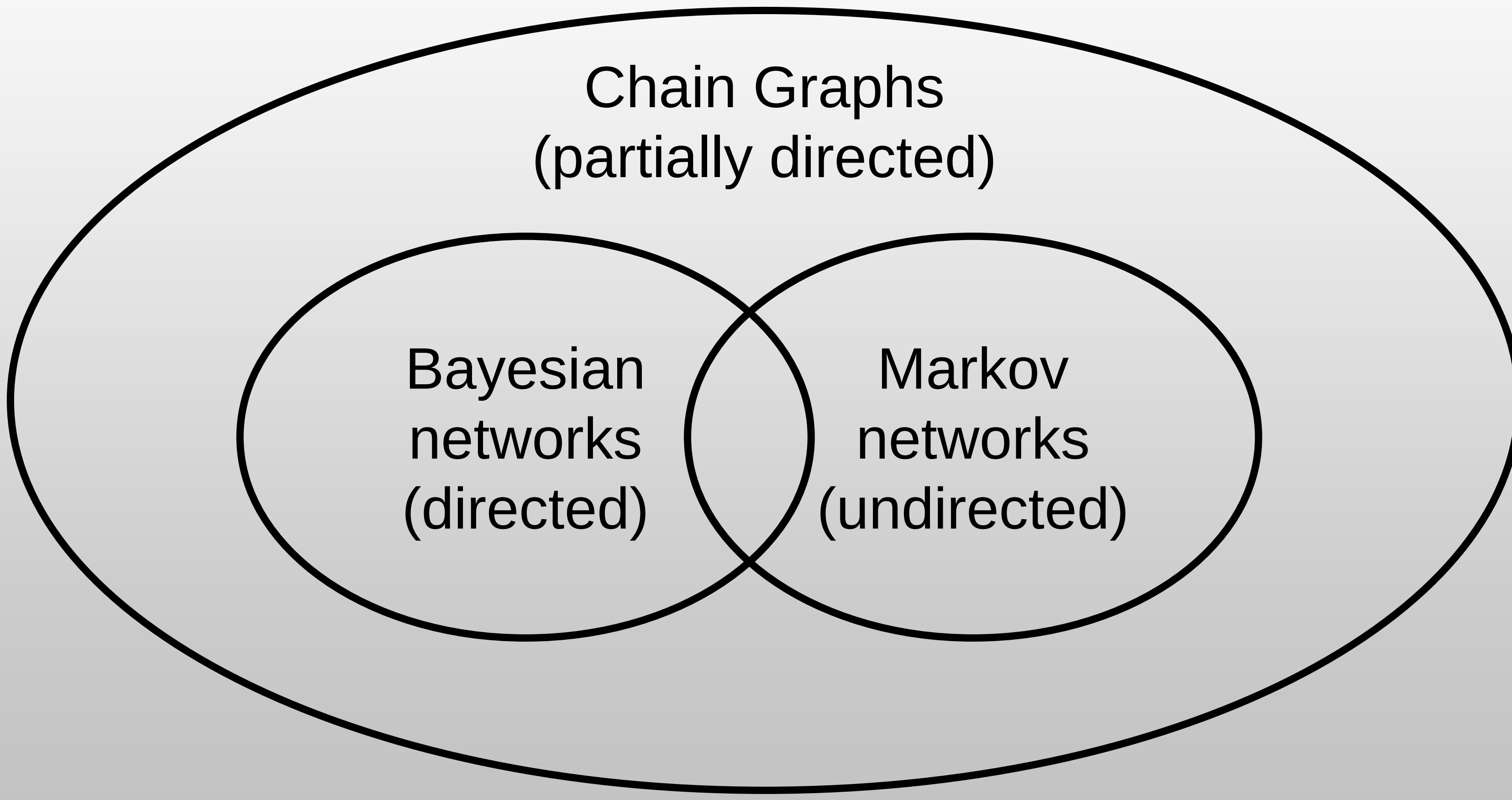
Undirected graph:

- Edge X – Y if there is a factor that includes both X and Y in the same scope

Markov Networks vs Bayesian Network

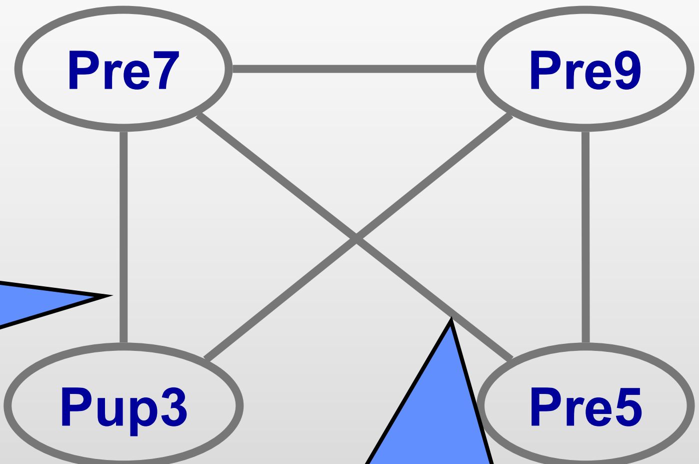
- Undirected graph
  - ➔no acyclicity constraints
- Potential functions
  - ➔less natural and interpretable than conditional distributions
- Inference is similar to that of Bayesian networks
- Learning is computationally harder

Relationship between Directed & Undirected Models



Relational Markov Networks

- Similar to Relational Bayesian Networks
- Duplicate potentials



Pre7	Pup3	f
0	0	0
0	1	-1
1	0	-1
1	1	4

Pre7	Pre5	f
0	0	0
0	1	-1
1	0	-1
1	1	4

Outline

- Introduction
- Bayesian Networks
- Learning Bayesian Networks
- Transcriptional regulation
- Gene expression
- Markov Networks
- Protein-Protein Interactions
- Discussion

Relational Markov Networks for Protein-Protein Interaction

- Random variable for each attribute of protein
  - *Pre7.nucleus*
  - *Pre7.mitochondria*
  - *Pre7.cytoplasam*
  - *…*
  - *Pre7.ribosomal*
  - *Pre7.DNA-binding*
  - *…*

Cellular compartment

Functional category (GO)

- Introduce potential between interacting pairs

$$\prod_{p \text{ interacts with } q} f_{\text{nucleus}}(p.nucleus, q.nucleus)$$

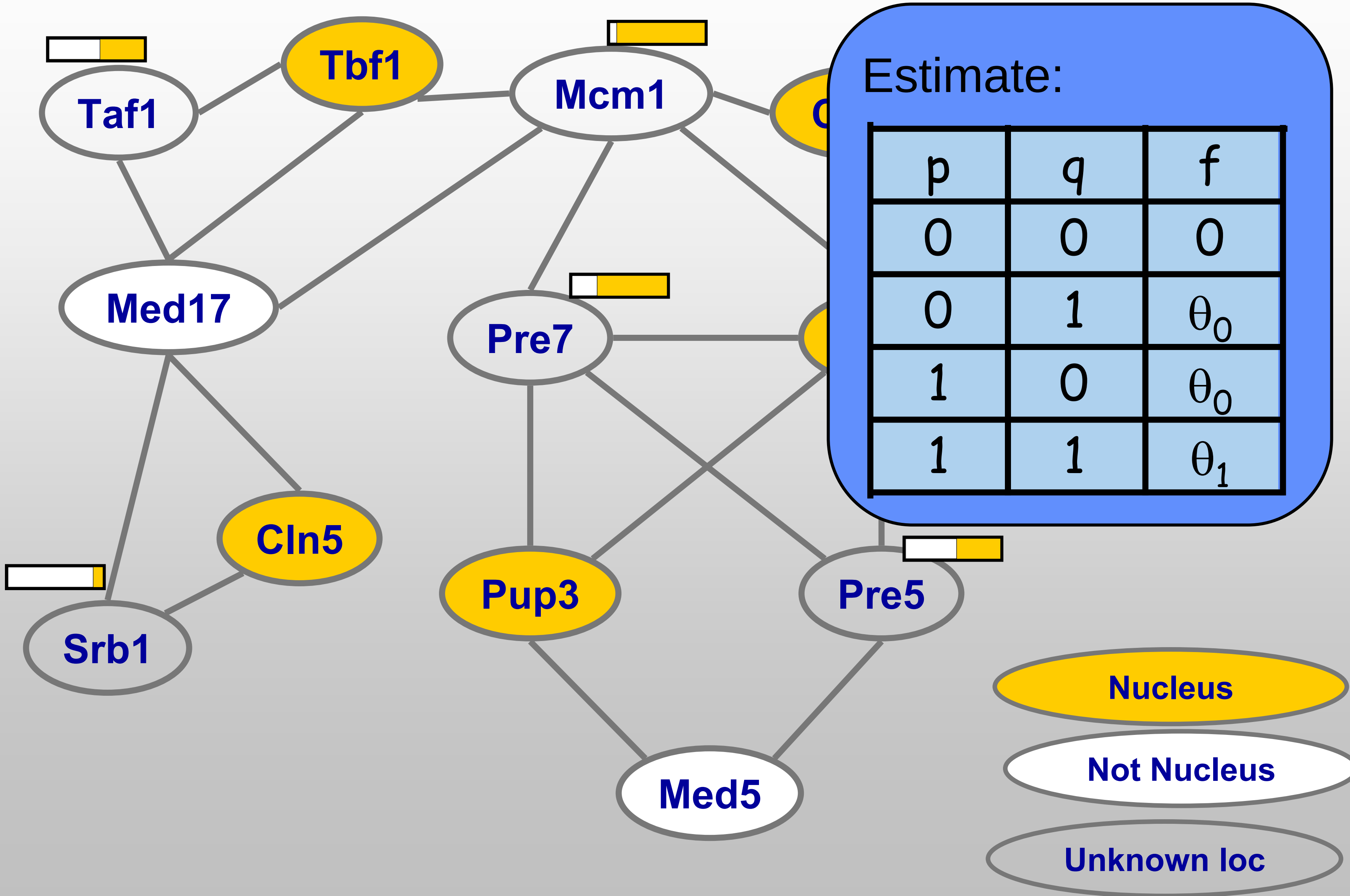
Deng et al, RECOMB 2003

Relational Markov Networks for Protein-Protein Interaction

Three phase process

- Model construction
  - Use interaction network to construct model
- Learning phase
  - Use know proteins attributes to estimate potentials for each type of attribute
- Prediction phase
  - Use inference to predict attributes for all proteins given evidence
  - Simultaneous predictions for all the proteins in the network

Relational Markov Networks for Protein-Protein Interaction

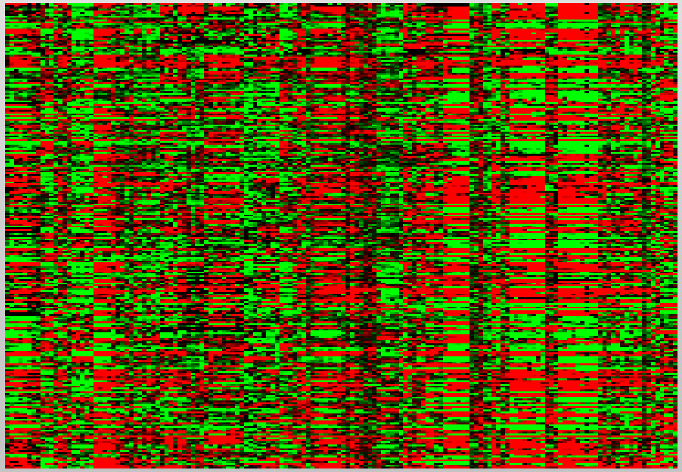


Estimate:

p	q	f
0	0	0
0	1	θ_0
1	0	θ_0
1	1	θ_1

Inferring "Pathways"

- **Assumption**: pathways exhibit two properties
  - Have similar expression profiles
  - Protein products more likely to interact
- Use both types of data to find pathways



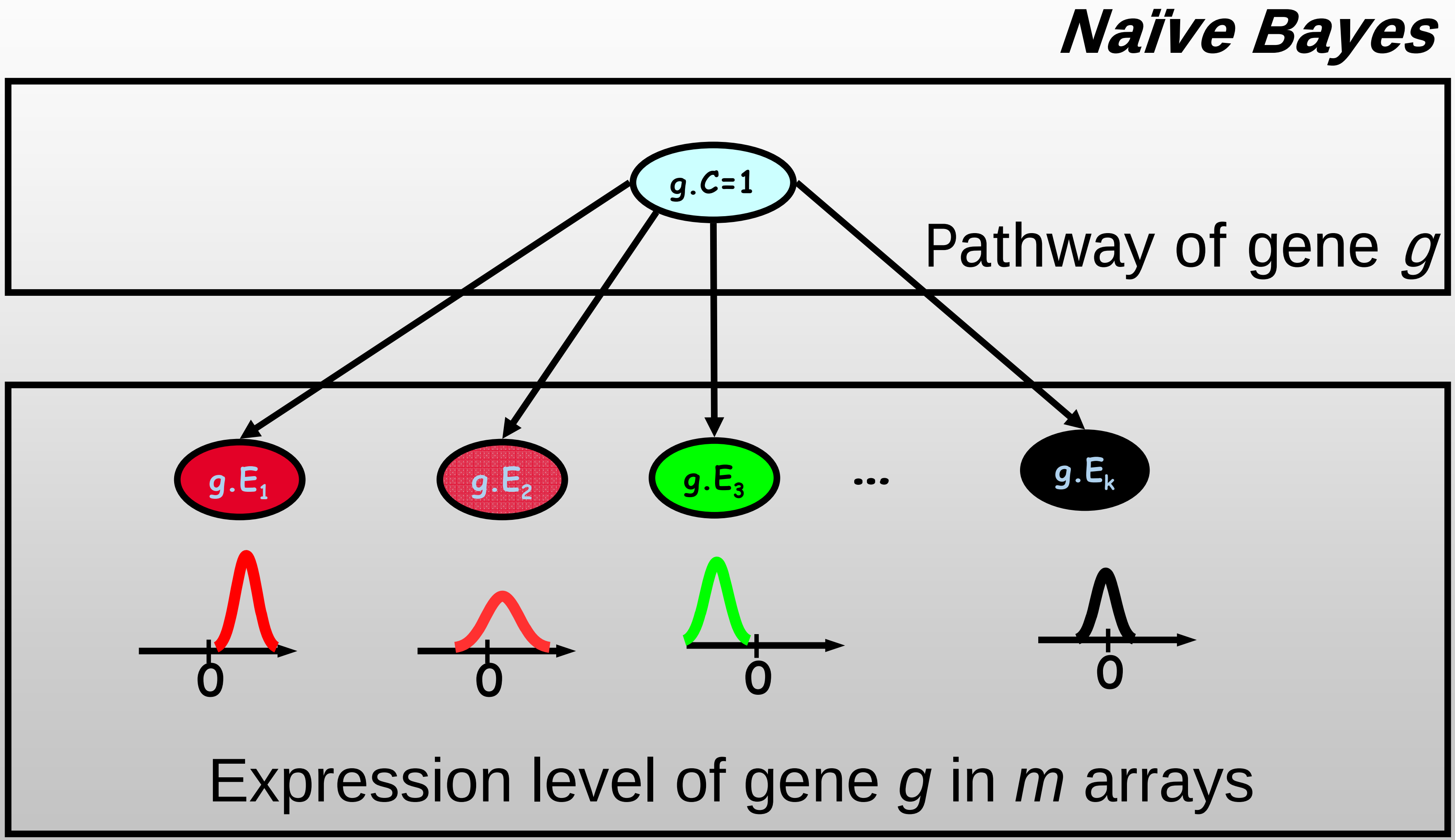
Segal et al, ISMB 2003

Probabilistic Model

- Genes are partitioned into “pathways”:
  - Every gene is assigned to one of ‘k’ pathways
  - Random variable for each gene with domain {1,…,k}
- Expression component:
  - Model likelihood is higher when genes in the same pathway have similar expression profiles
- Interaction component:
  - Model likelihood is higher when genes in the same pathway interact

Segal et al, ISMB 2003

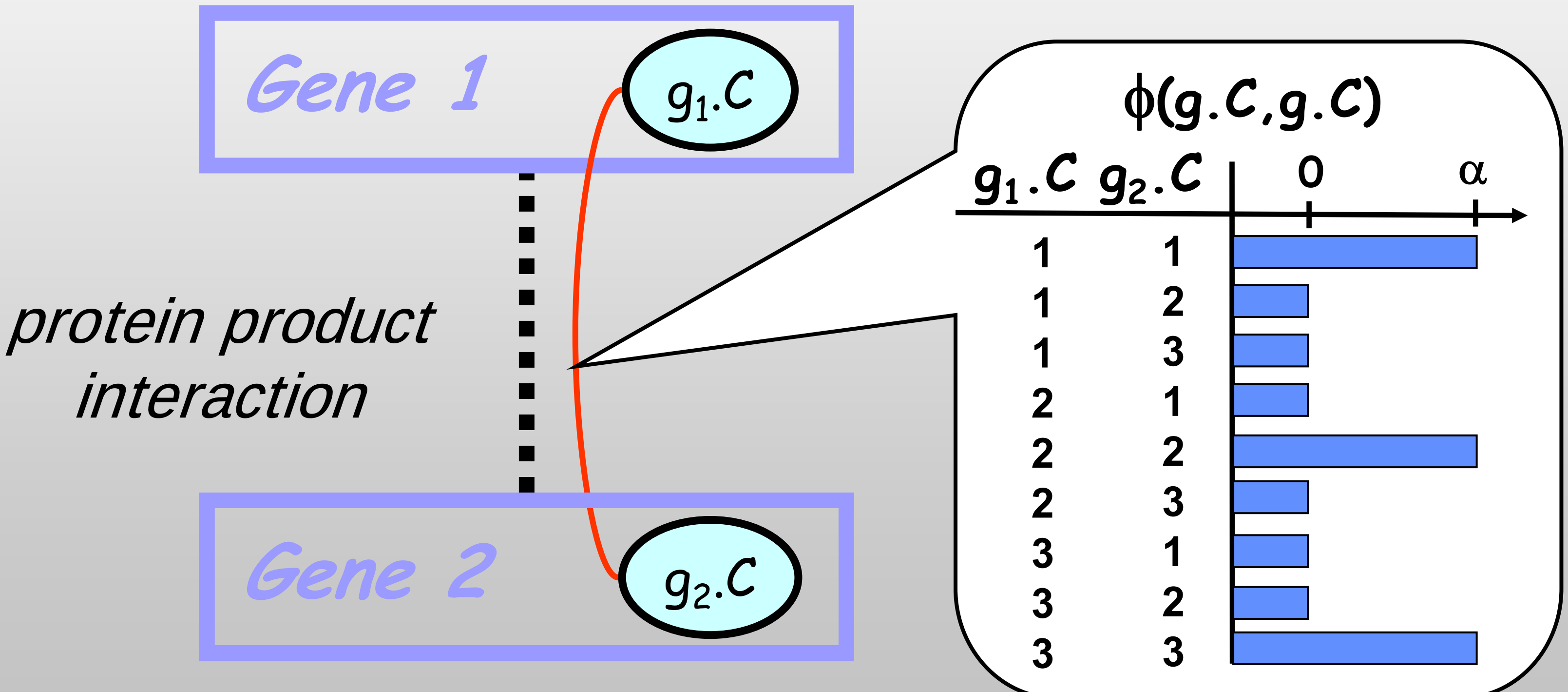
Expression Component



Segal et al, ISMB 2003

Protein Interaction Component

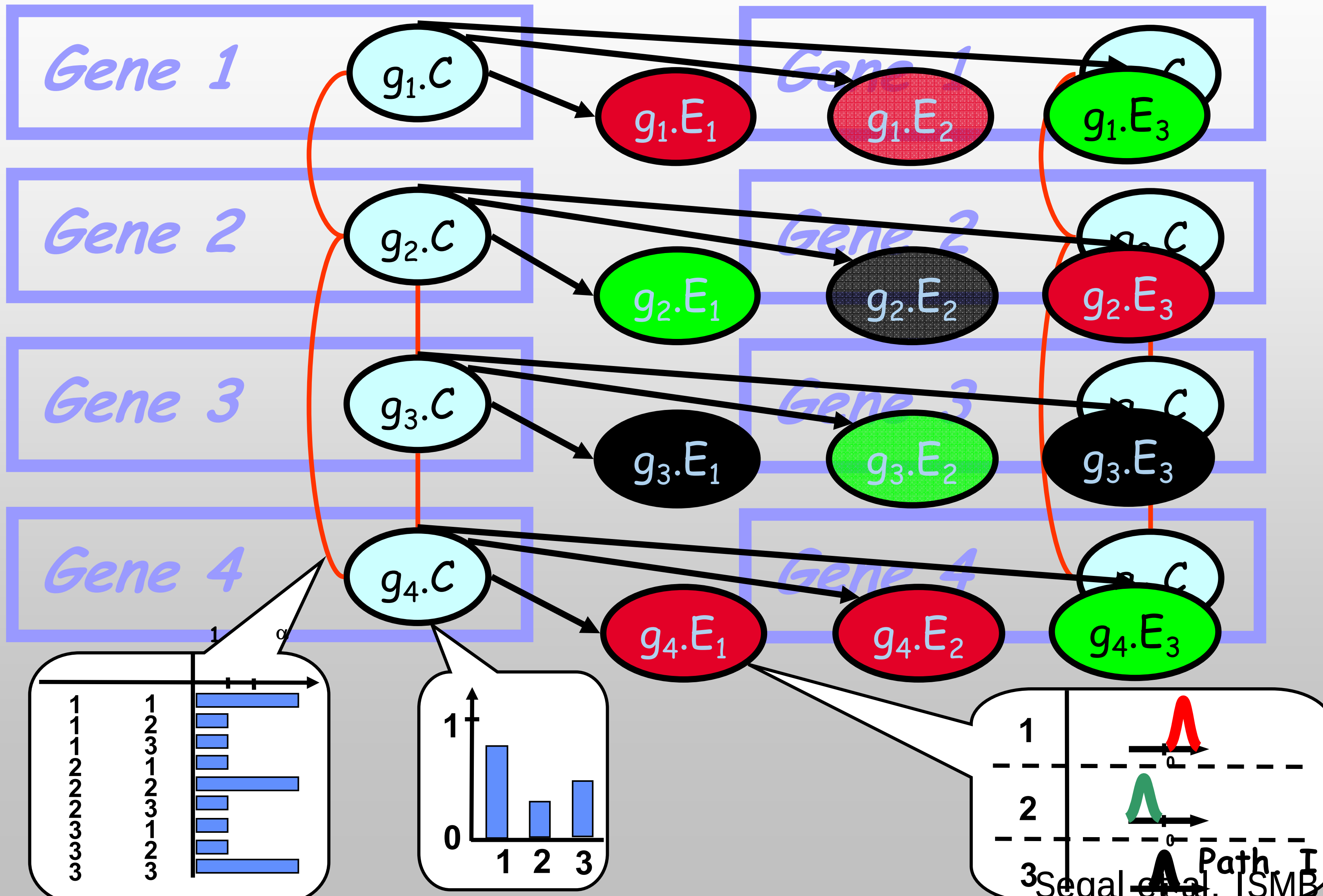
- Interacting genes are more likely to be in the same pathway



Compatibility potential

Segal et al, ISMB 2003

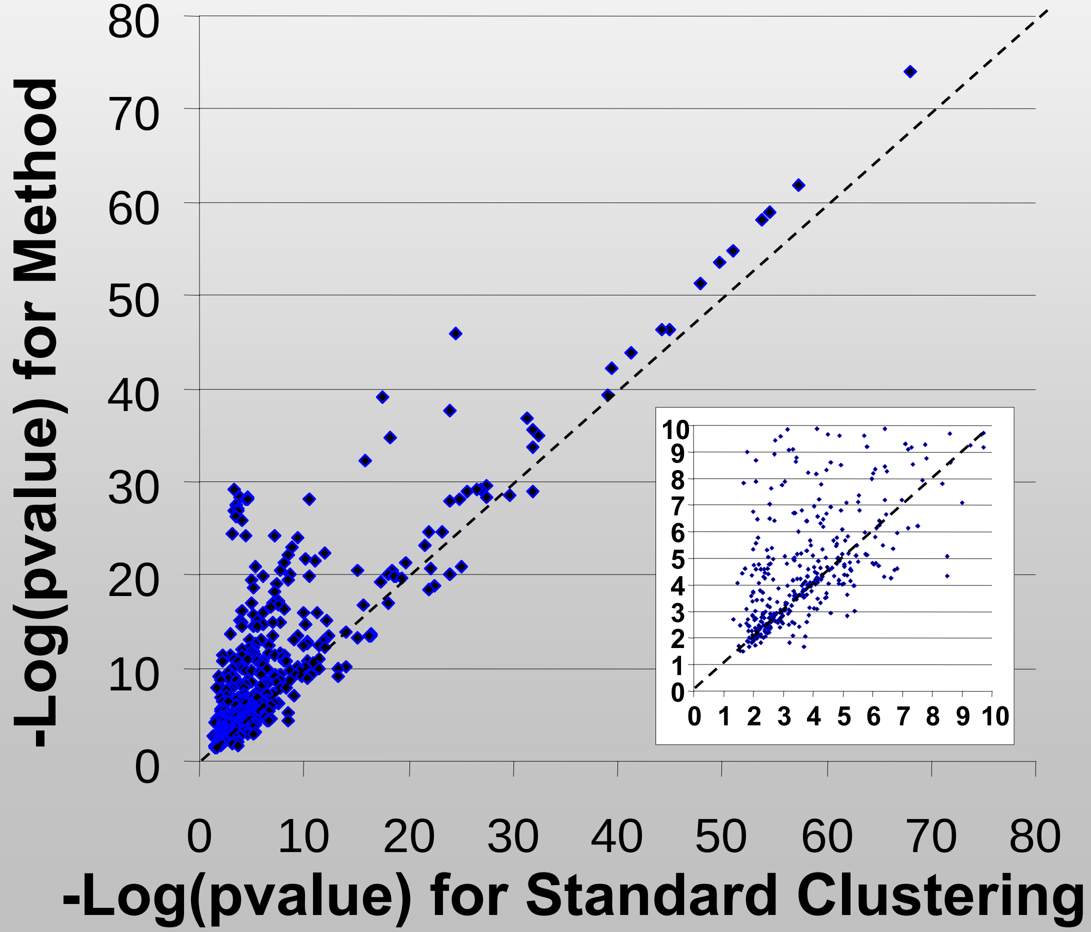
Joint Probabilistic Model



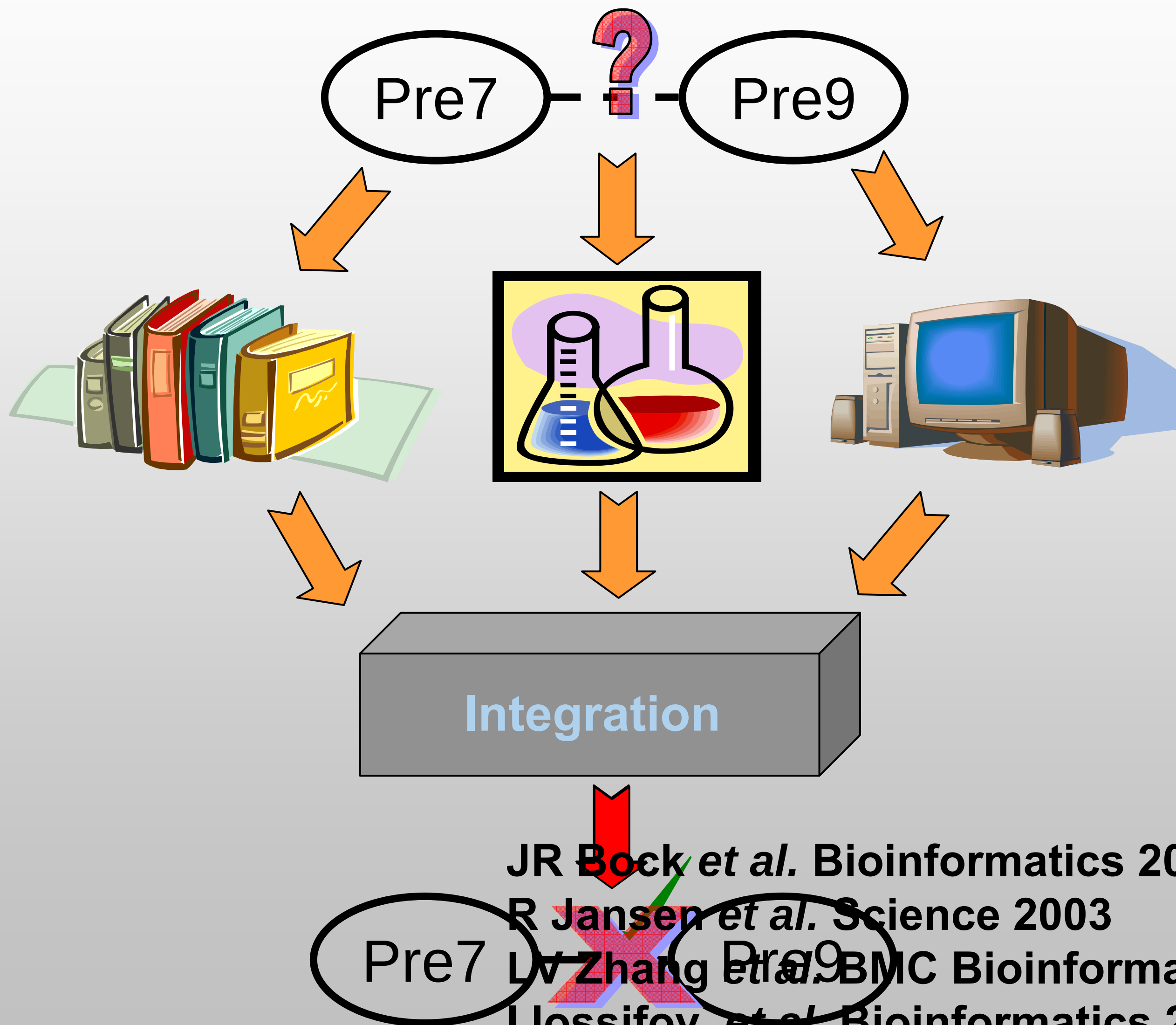
Segal et al, ISMB 2003

Comparison to Clustering

- Check enrichment of known gene annotations in pathways
- Calculate significance (negative log p-value)

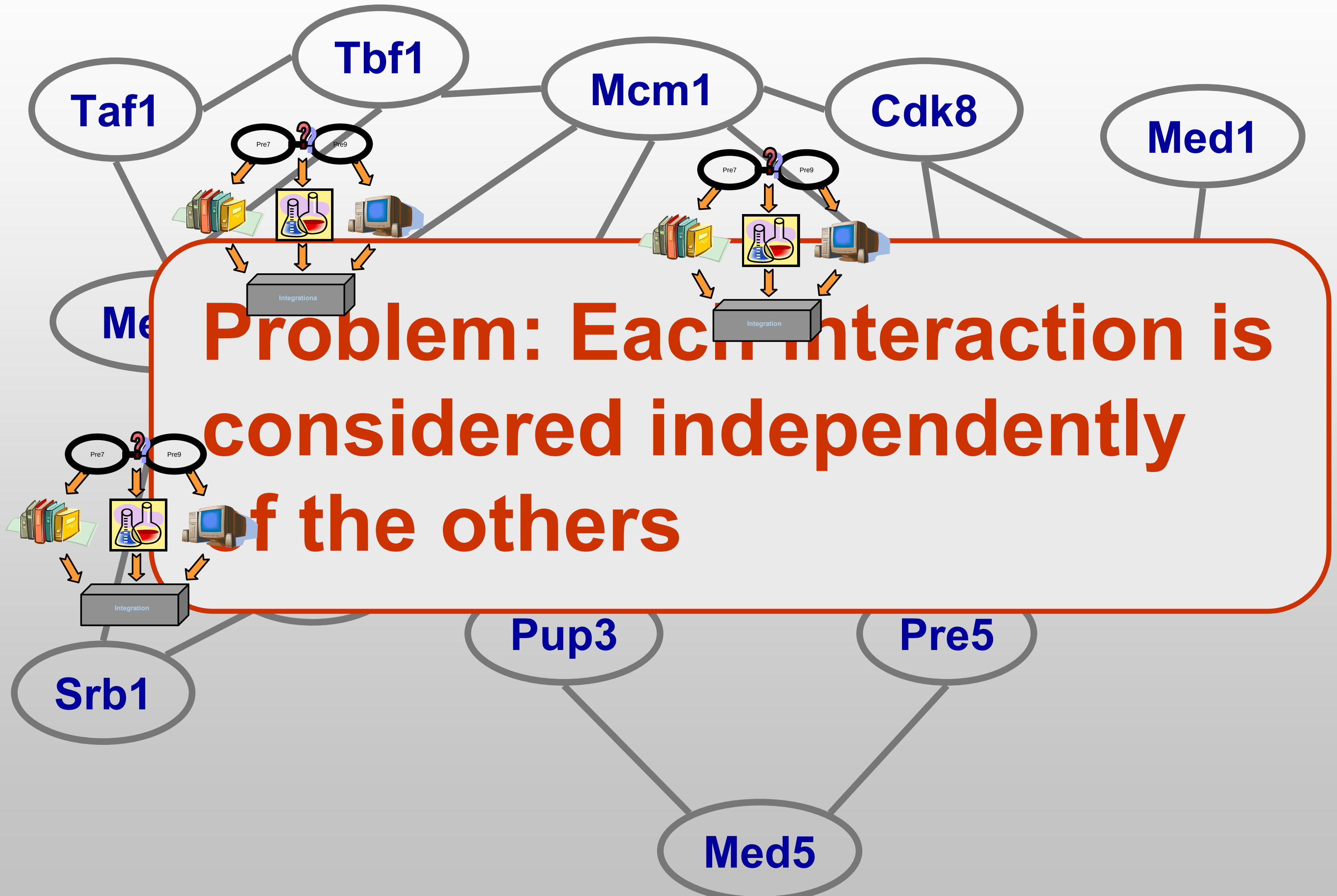


Predicting Protein-Protein Interactions

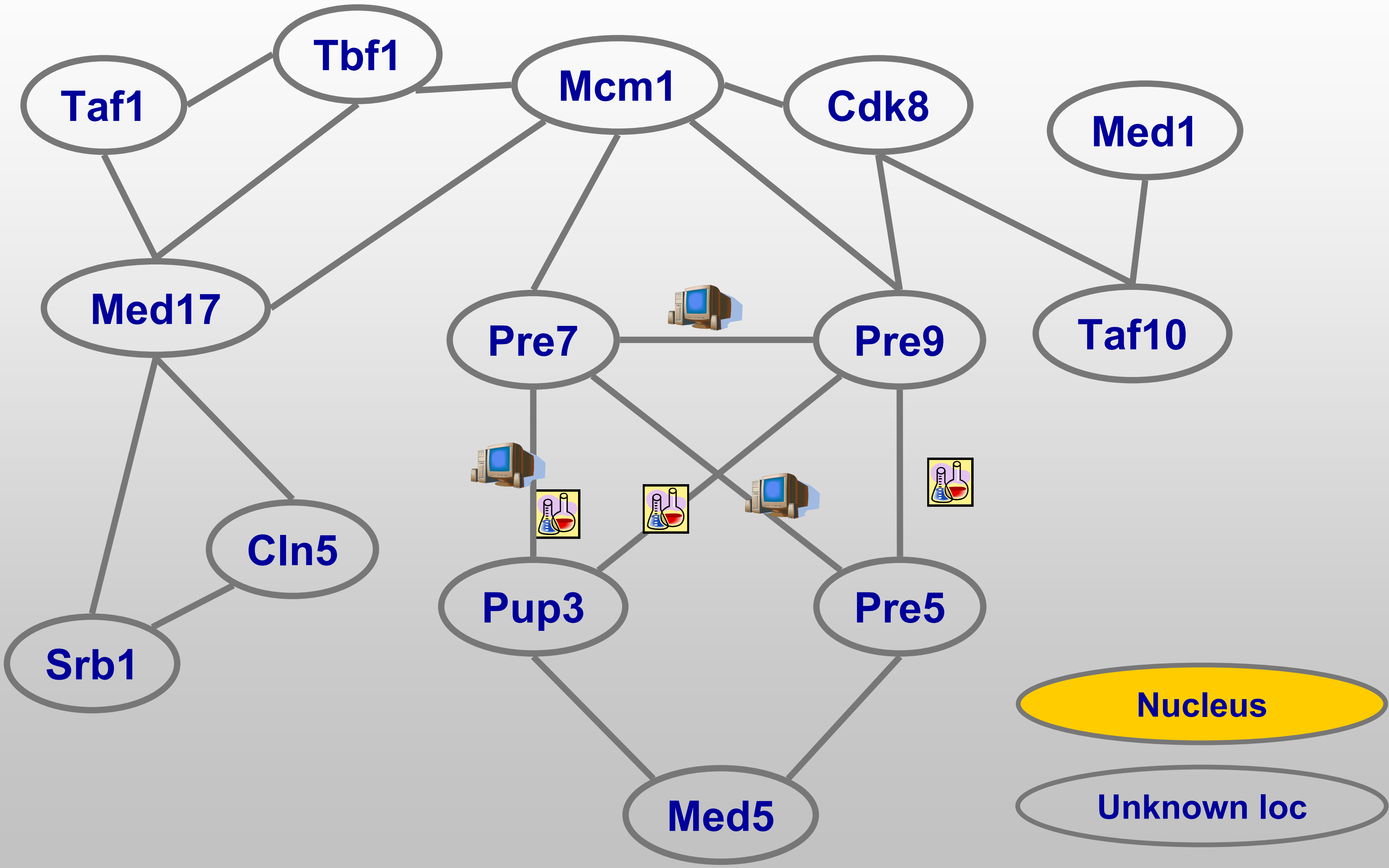


JR Bock *et al.* Bioinformatics 2001
R Jansen *et al.* Science 2003
LV Zhang *et al.* BMC Bioinformatics 2003
I Iossifov *et al.* Bioinformatics 2004 ...

Predicting Interactions

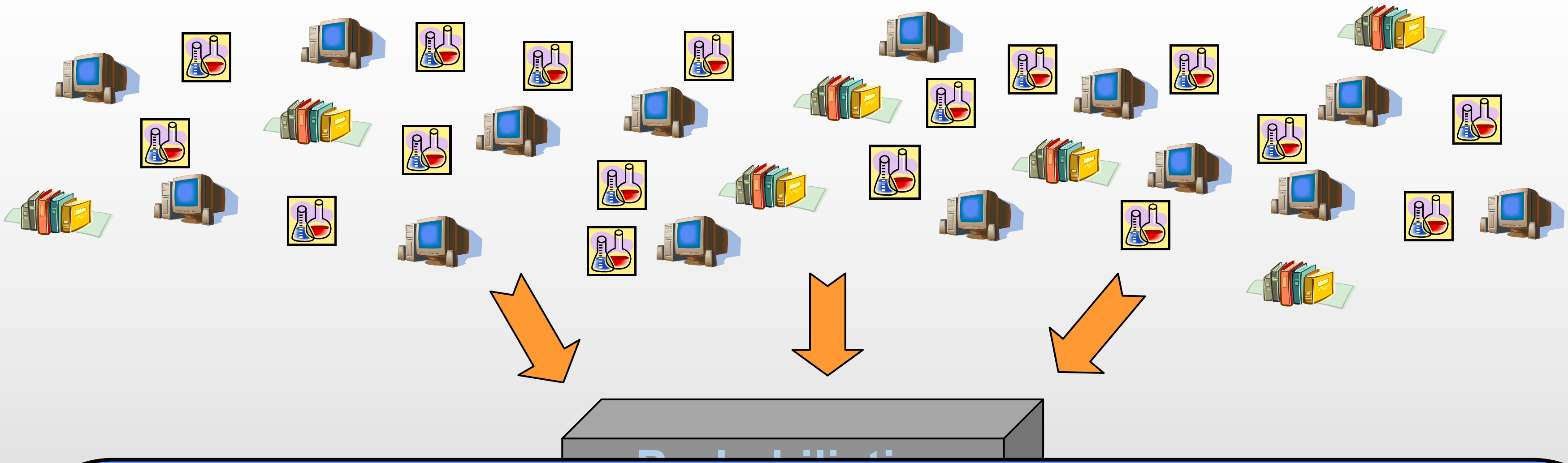


Motivation



Tbf1
Taf1
Mcm1
Cdk8
Med1
Med17
Pre7
Pre9
Taf10
Cln5
Pup3
Pre5
Srb1
Med5
Nucleus
Unknown loc

Design Plan



New variables denoting

- Whether two proteins interacts
- Experimental observations about each interaction

Main difficulty:

High connectivity between these variables

Building the Model

$IA(p_1,p_2)$ \ $I(p_1,p_2)$	0
0	0.9
1	0.1

L_1	L_2	I	Score
0	0	0	1
0	0	1	1
0	1	0	1
	1	1	0.4
1	0	0	1
1	0	1	0.4
1	1	0	1
L(Pre7,nuc)			1.2

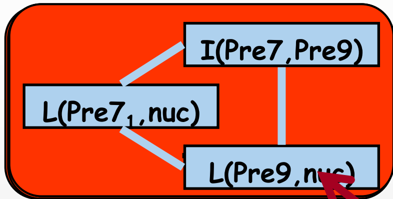
Med1

Taf10

Srb1

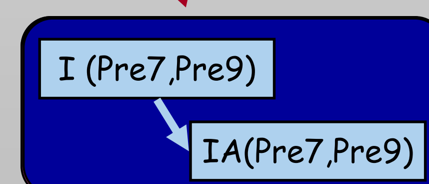
L(Pre9,nuc)

Using a Relational Model

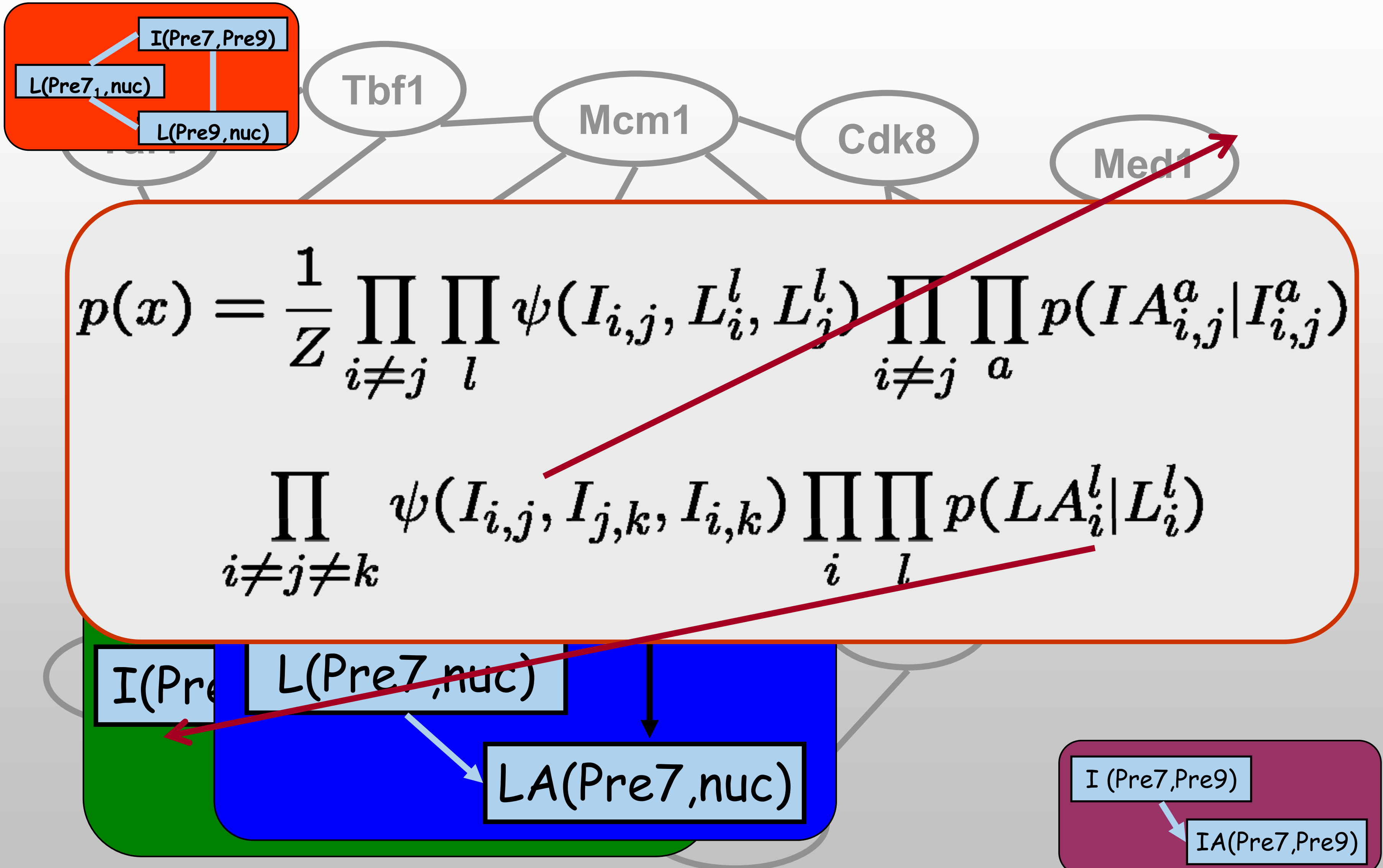


So far, equivalent to integrated prediction of each interaction independently

$$p(x) = \frac{1}{Z} \prod_{i \neq j} \left(\prod_{l} \psi(I_{i,j}, L_i^l, L_j^l) \prod_{a} p(IA_{i,j}^a \mid I_{i,j}) \right)$$



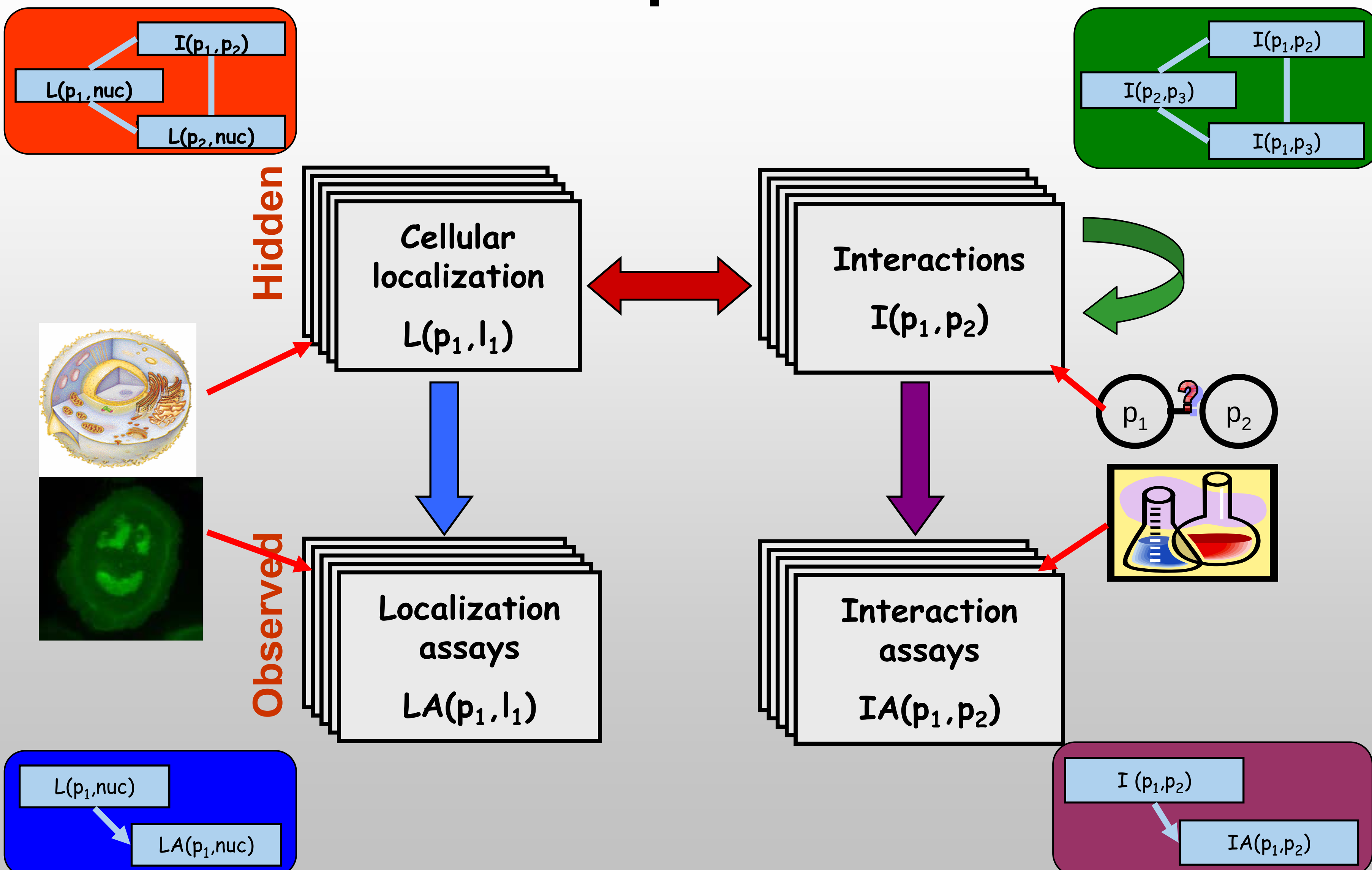
Building the Model



$$p(x) = \frac{1}{Z} \prod_{i \neq j} \prod_{l} \psi(I_{i,j}, L_i^l, L_j^l) \prod_{i \neq j} \prod_{a} p(IA_{i,j}^a | I_{i,j}^a)$$

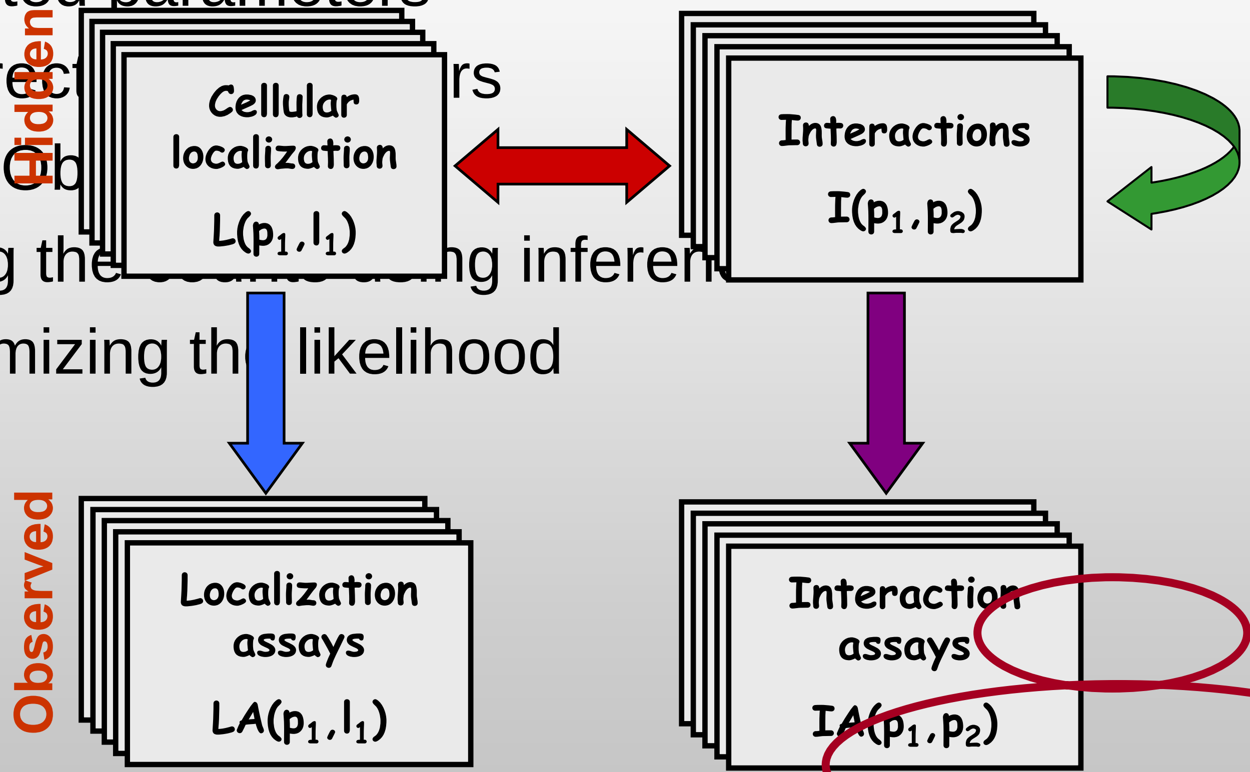
$$\prod_{i \neq j \neq k} \psi(I_{i,j}, I_{j,k}, I_{i,k}) \prod_{i} \prod_{l} p(LA_i^l | L_i^l)$$

The Complete Model



Learning the Parameters

- Maximizing the likelihood (fully observed case)
  - Directed parameters
  - Undirec… rs
- Partially Ob…
  - Filling the … g inferen…
  - Maximizing th likelihood



$$p(x) = \frac{1}{Z_x}\prod_c \psi_c(x_c)\prod_i p(y_i|Pa_i)$$

Model Evaluation: *S.cerevisiae*

Large scale data:

- **Yeast two hybrid (Ito *et al.* + Uetz *et al.*)**
- **Complexes (MIPS)**
- **Correlated domain signatures (Sprinzak *et al.*)**
- **Protein localization (Huh *et al.*)**

38,000 potentials

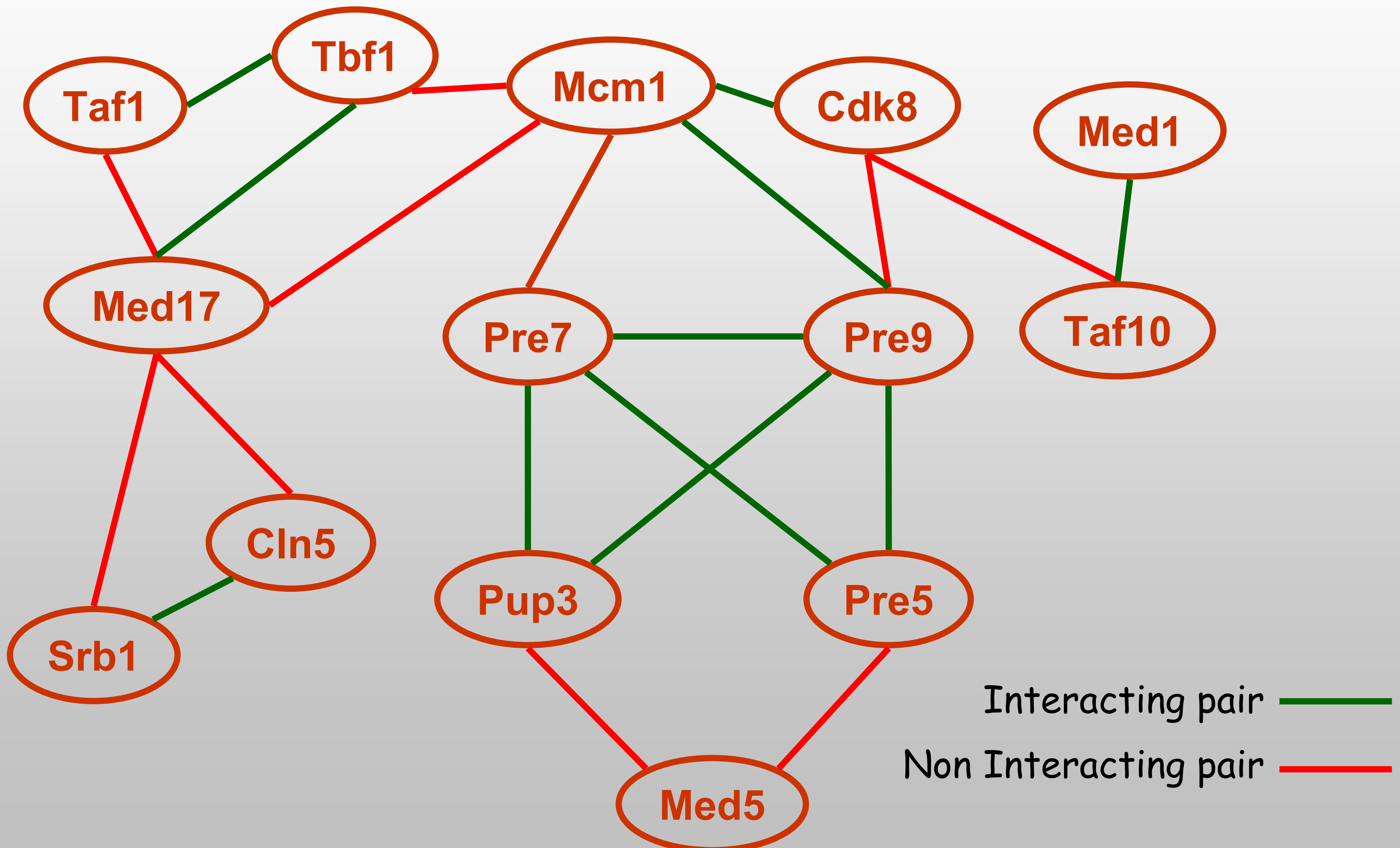
37 free parameters

Med5

pair

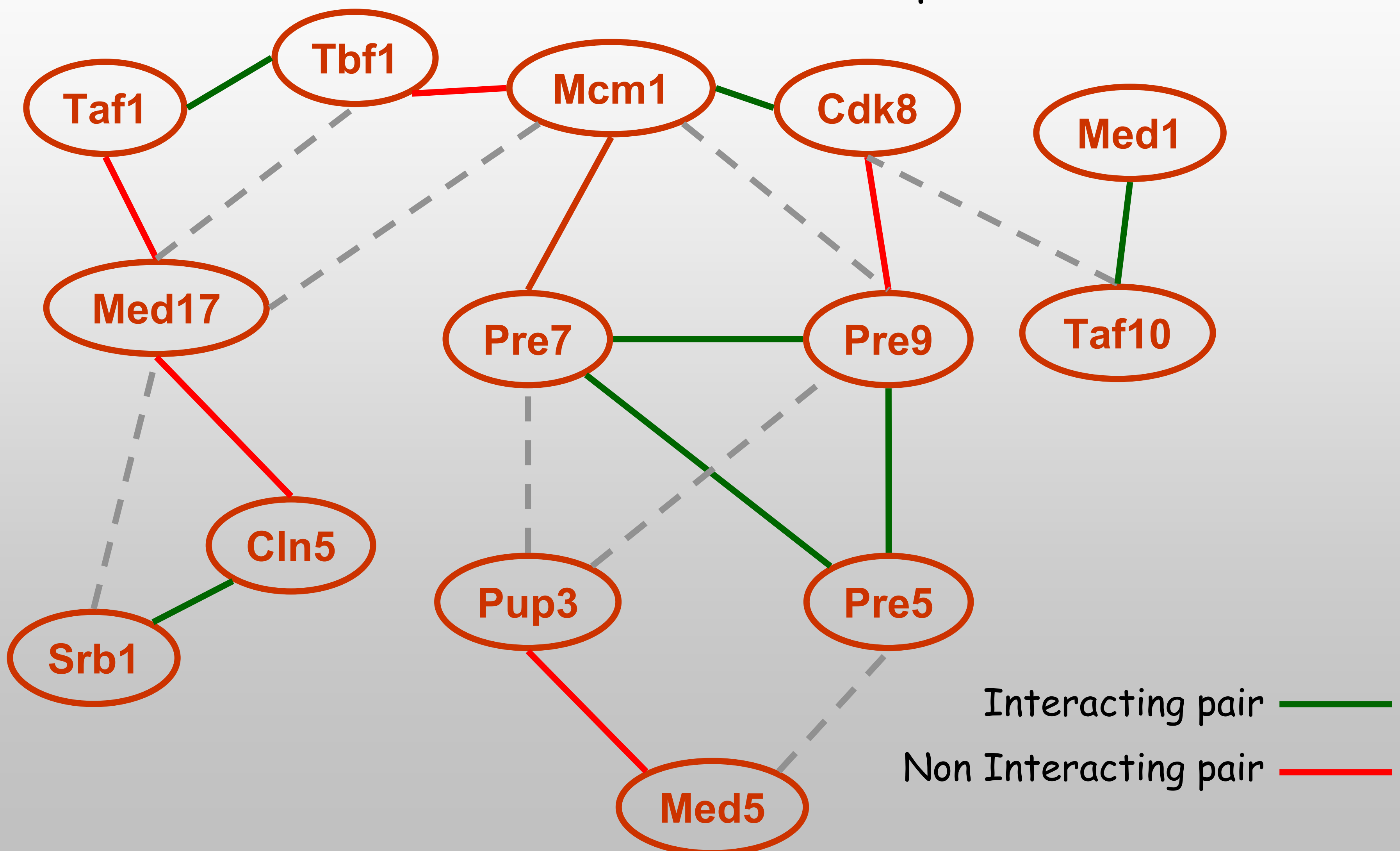
von Mering *et al.* Nature 2002

Evaluation: Cross Validation



Evaluation: Parameter Estimation

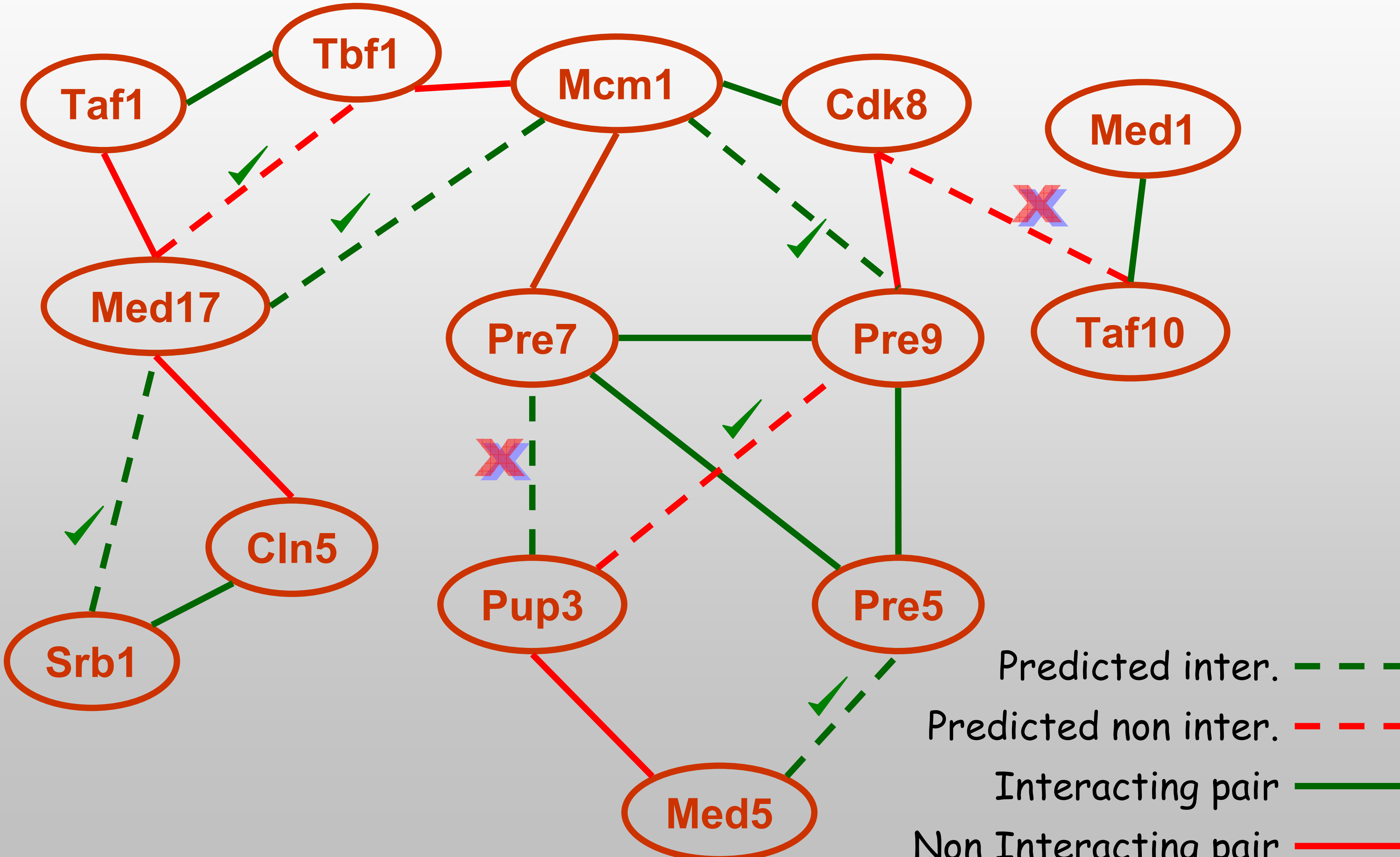
Hide a set of test interactions and learn parameters



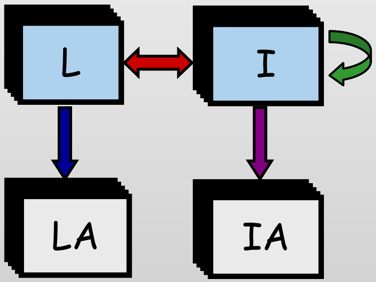
Evaluation: Validate Predictions

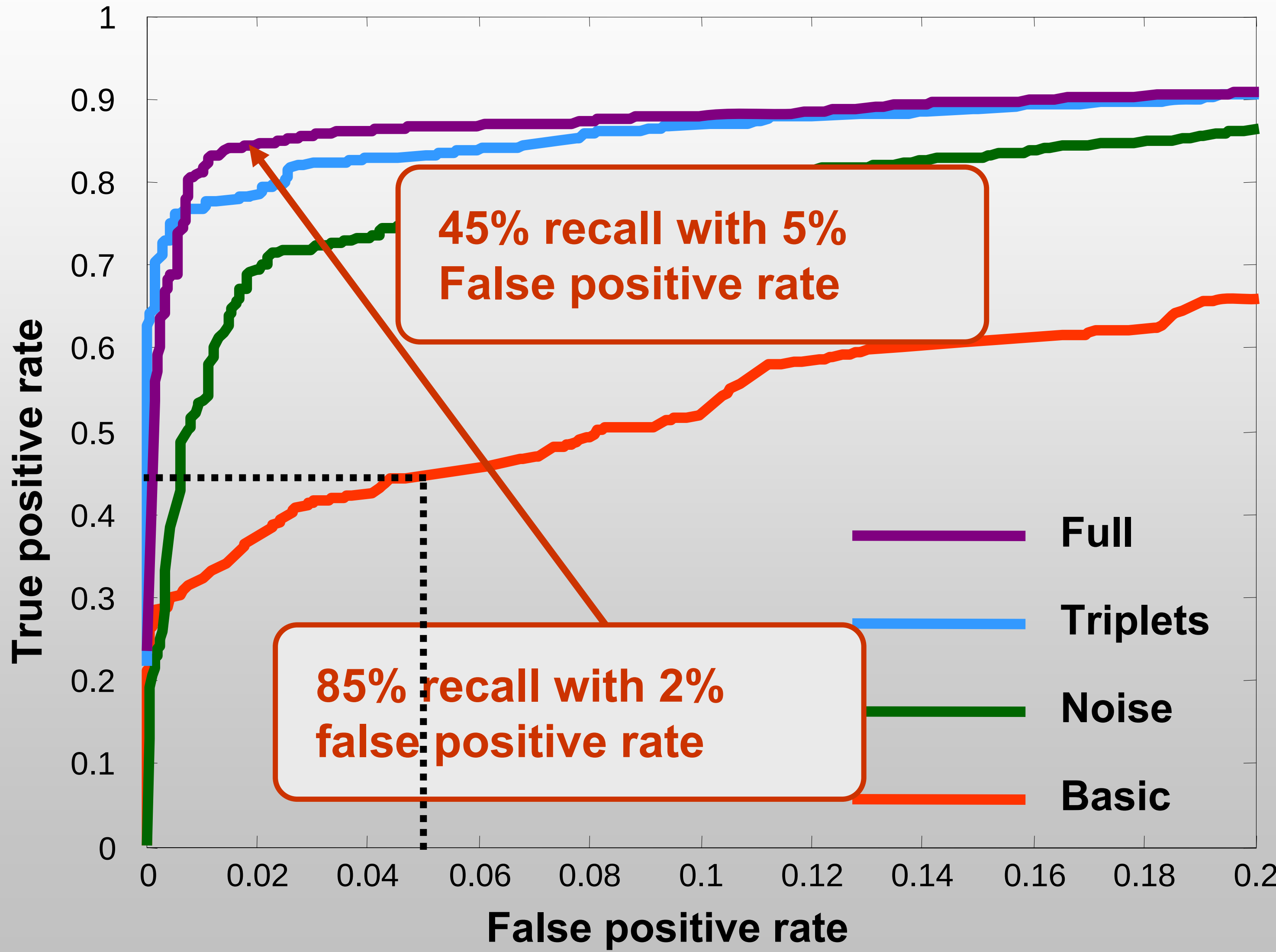
Use Learned parameters to predict hidden interactions

Evaluate prediction



Evaluation: ROC curve

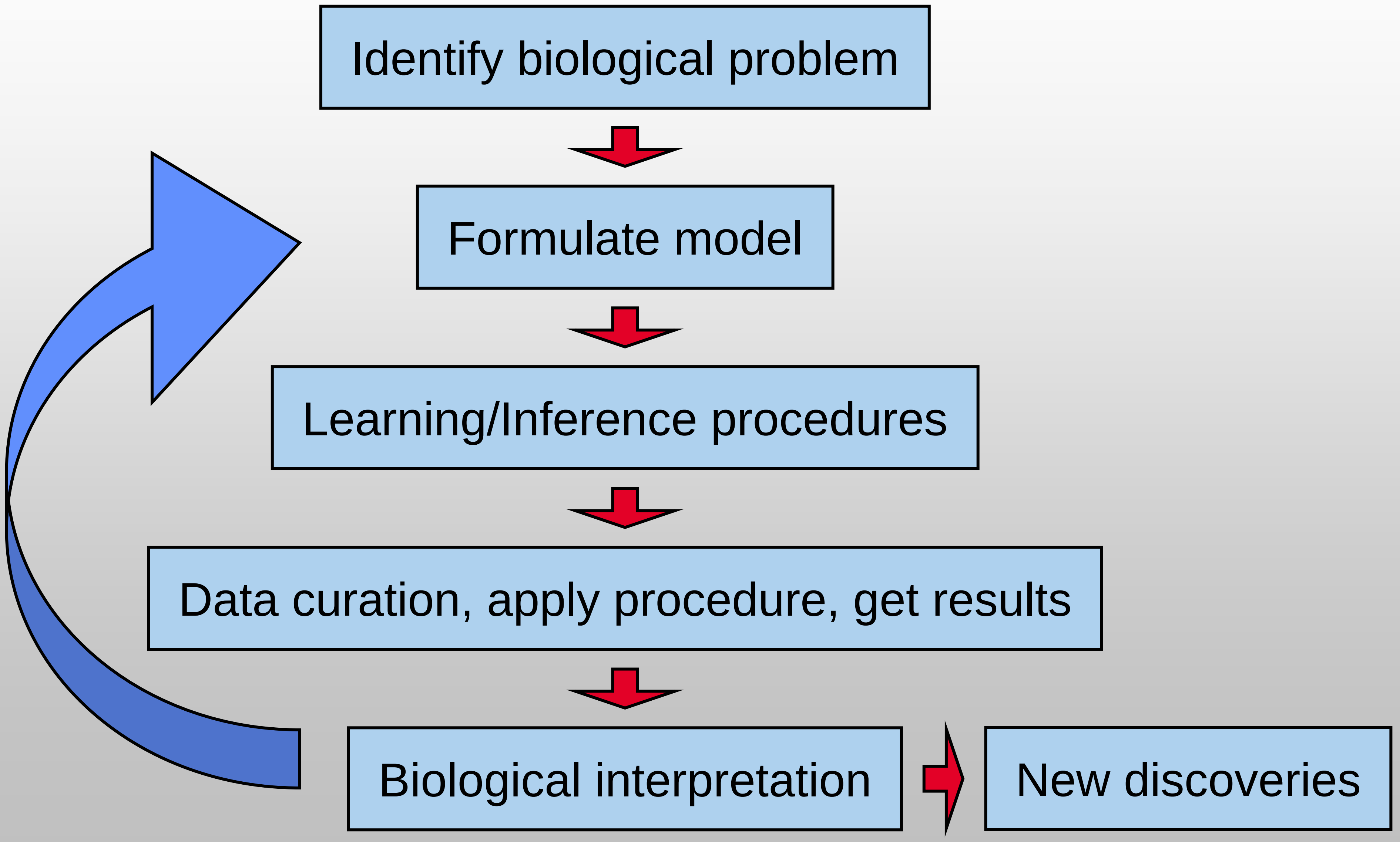




Outline

- Introduction
- Bayesian Networks
- Learning Bayesian Networks
- Transcriptional regulation
- Gene expression
- Markov Networks
- Protein-Protein Interactions
- Discussion

Philosophy


Identify biological problem
Formulate model
Learning/Inference procedures
Data curation, apply procedure, get results
Biological interpretation
New discoveries

Recap

- Models of evolution
  - Pedigree analysis
  - Sequence evolution
- Transcription Factors
  - Binding sites
- Gene Expression
  - Clustering, interaction networks
- Protein-Protein interaction networks
- Combination of subsets of these

Additional Areas

- Gene finding
  - Extended HMMs + evolutionary models
- Analysis of genetic variation
  - SNPs, haplotypes, and recombination
- Protein structure
  - 2^{nd}-ary and 3^{rd}-ary structure, molecular recognition

Take Home Message

- **Graphical models as a methodology**
  - Modeling language
  - Foundations & algorithms for learning
  - Allows to incorporate prior knowledge about biological mechanisms
  - Learning can reveal "structure" in data
- **Exploring unified system models**
  - Learning from heterogeneous data
    - Not simply combining conclusions
  - Combine weak evidence from multiple sources
    $\Rightarrow$ detect subtle signals
  - Get closer to **mechanistic** understanding of the signal

The END