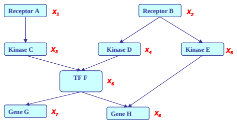
Computational Genomics

10-810/02-710, Spring 2009

Inferring gene regulatory

Eric Xing

Lecture 26, April 22, 2009



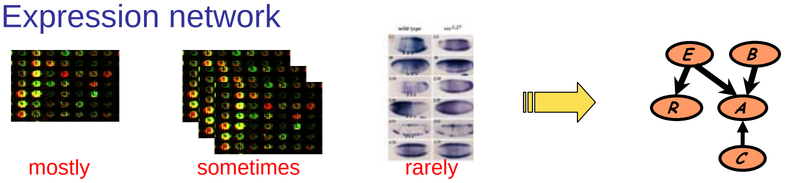
Reading: handouts

Gene regulatory networks

- Regulation of expression of genes is crucial:
  - Genes control cell behavior by controlling which proteins are made by a cell and when and where they are delivered
- Regulation occurs at many stages:
  - pre-transcriptional (chromatin structure)
  - transcription initiation
  - RNA editing (splicing) and transport
  - Translation initiation
  - Post-translation modification
  - RNA & Protein degradation
- Understanding regulatory processes is a central problem of biological research

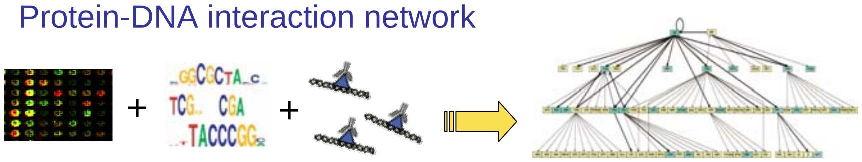
Inferring gene regulatory networks

- Expression network



  - gets most attentions so far, many algorithms
  - still algorithmically challenging

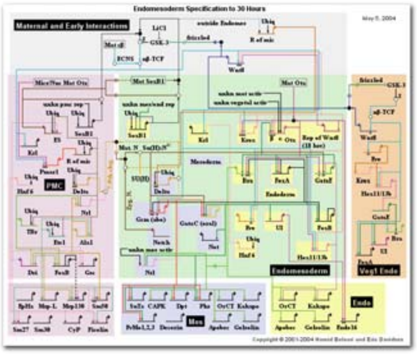
- Protein-DNA interaction network



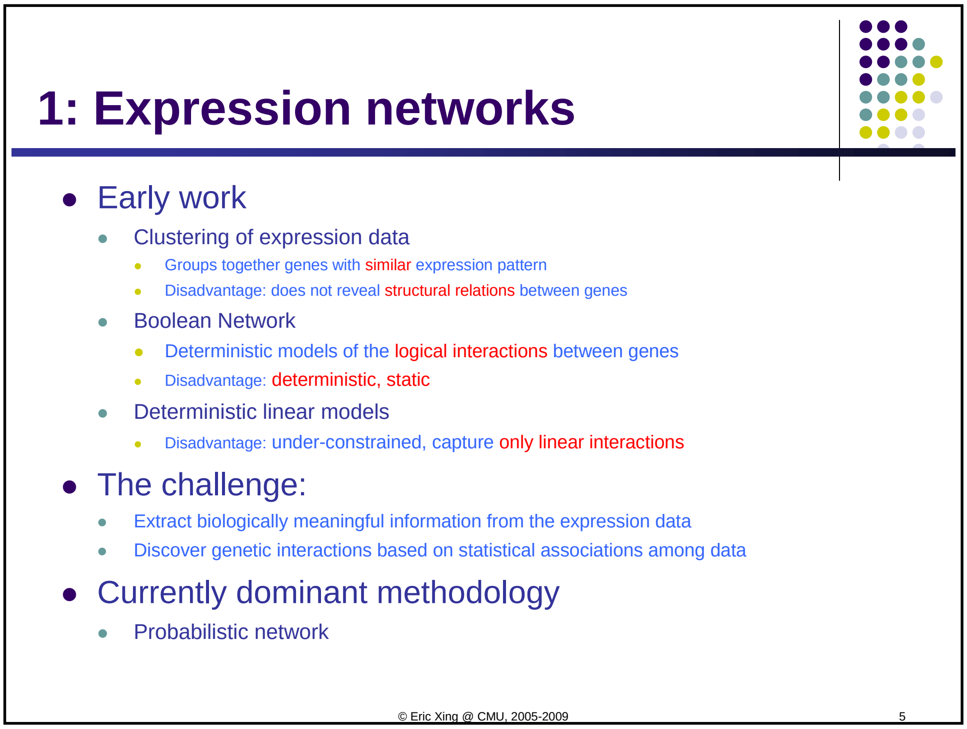
  - actively pursued recently, some interesting methods available
  - lots of room for algorithmic advance

Inferring gene regulatory networks

- Network of cis-regulatory pathways

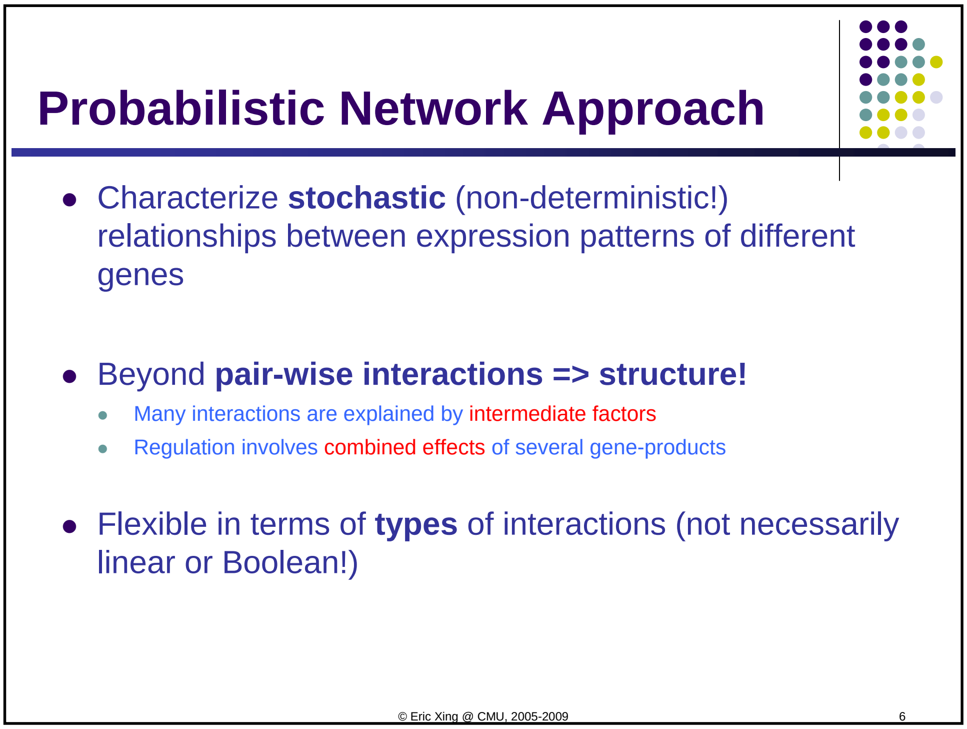

Endomesoderm Specification to 30 Hours
Maternal and Early Interactions
Endomesoderm
Endoderm
Mesoderm
PMC

  - Success stories in sea urchin, fruit fly, etc, from decades of experimental research
  - Statistical modeling and automated learning just started



1: Expression networks

- Early work
  - Clustering of expression data
    - Groups together genes with similar expression pattern
    - Disadvantage: does not reveal structural relations between genes
  - Boolean Network
    - Deterministic models of the logical interactions between genes
    - Disadvantage: deterministic, static
  - Deterministic linear models
    - Disadvantage: under-constrained, capture only linear interactions
- The challenge:
  - Extract biologically meaningful information from the expression data
  - Discover genetic interactions based on statistical associations among data
- Currently dominant methodology
  - Probabilistic network



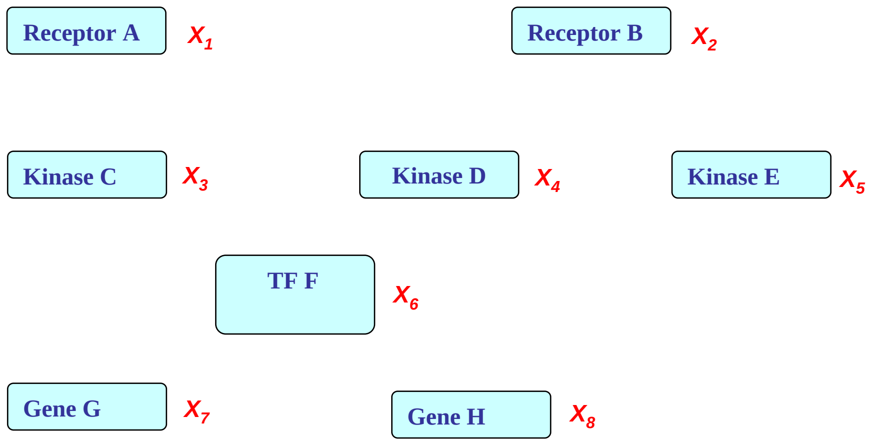
Probabilistic Network Approach

- Characterize **stochastic** (non-deterministic!) relationships between expression patterns of different genes
- Beyond **pair-wise interactions => structure!**
  - Many interactions are explained by intermediate factors
  - Regulation involves combined effects of several gene-products
- Flexible in terms of **types** of interactions (not necessarily linear or Boolean!)

What is a Graphical Model?

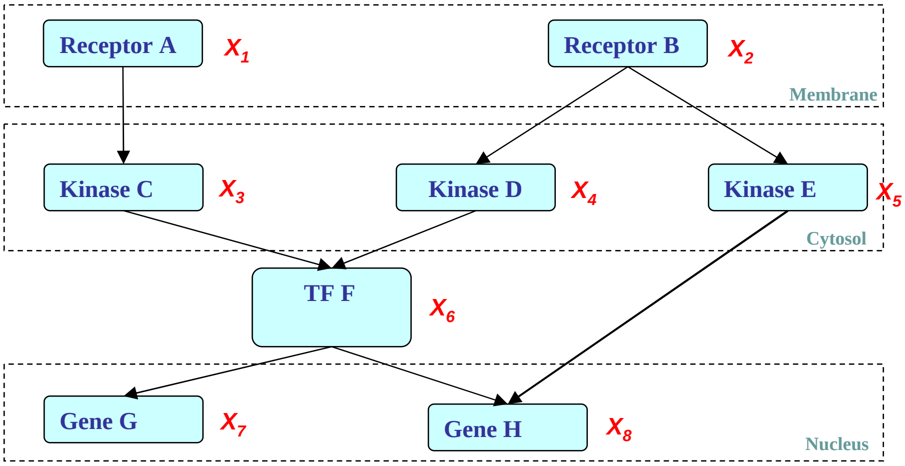
--- example from a signal transduction pathway

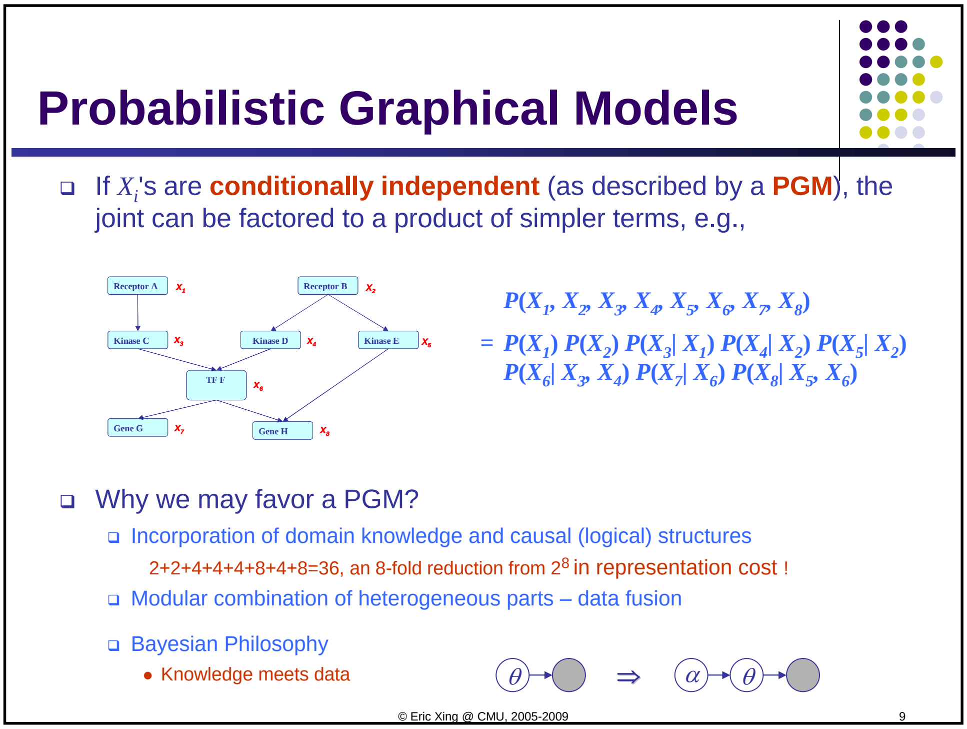
- A possible world for cellular signal transduction:



GM: Structure Simplifies Representation

- Dependencies among variables





Probabilistic Graphical Models

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,

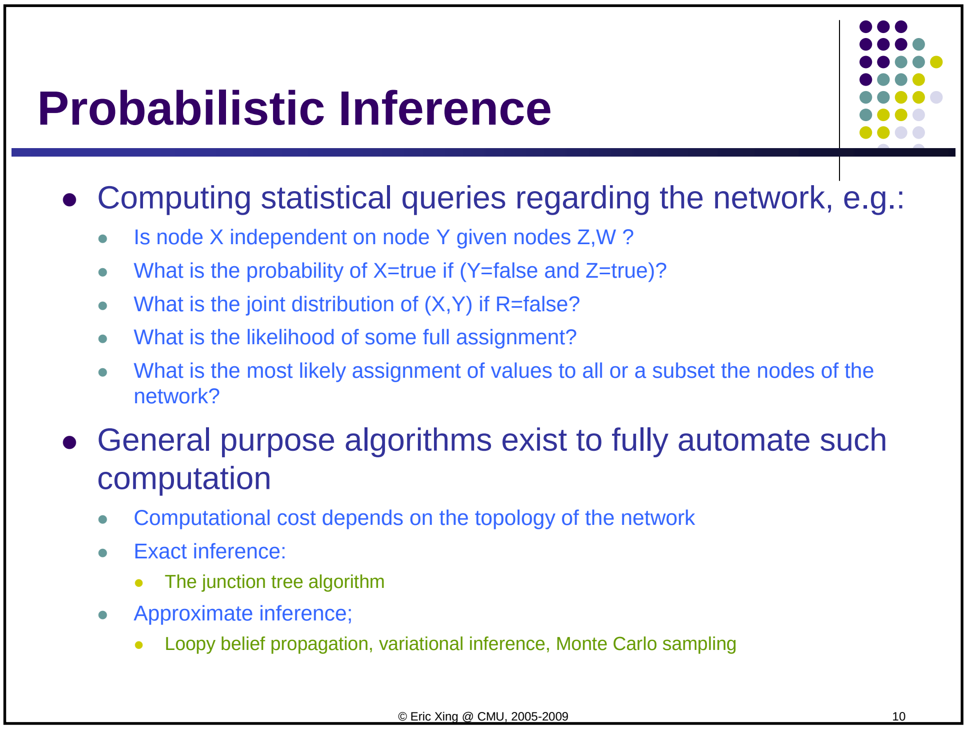
Receptor A
X_1
Receptor B
X_2
Kinase C
X_3
Kinase D
X_4
Kinase E
X_5
TF F
X_6
Gene G
X_7
Gene H
X_8

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\,P(X_2)\,P(X_3|\,X_1)\,P(X_4|\,X_2)\,P(X_5|\,X_2)\,P(X_6|\,X_3, X_4)\,P(X_7|\,X_6)\,P(X_8|\,X_5, X_6)$$

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures

    2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8 in representation cost !
  - Modular combination of heterogeneous parts – data fusion
  - Bayesian Philosophy
    - Knowledge meets data

$$\theta \rightarrow \bullet \quad \Rightarrow \quad \alpha \rightarrow \theta \rightarrow \bullet$$



Probabilistic Inference

- Computing statistical queries regarding the network, e.g.:
  - Is node X independent on node Y given nodes Z,W ?
  - What is the probability of X=true if (Y=false and Z=true)?
  - What is the joint distribution of (X,Y) if R=false?
  - What is the likelihood of some full assignment?
  - What is the most likely assignment of values to all or a subset the nodes of the network?
- General purpose algorithms exist to fully automate such computation
  - Computational cost depends on the topology of the network
  - Exact inference:
    - The junction tree algorithm
  - Approximate inference;
    - Loopy belief propagation, variational inference, Monte Carlo sampling

Two types of GMs

- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

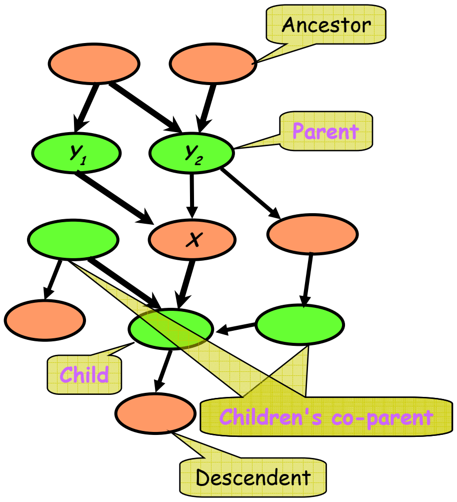
- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

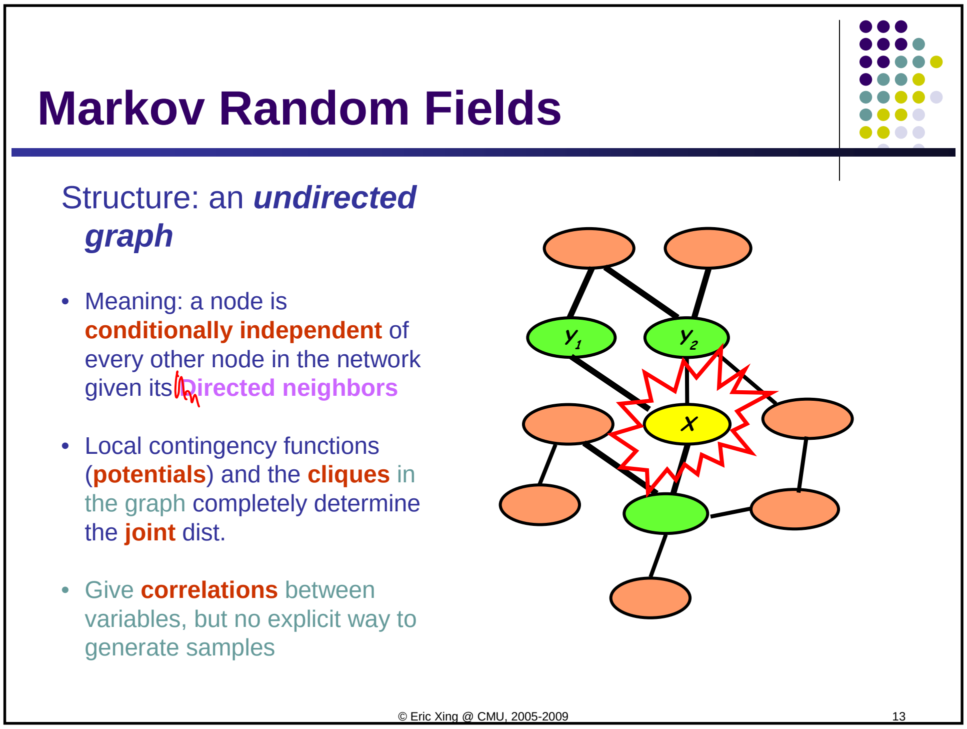
Bayesian Network

- Structure: ***DAG***

- Meaning: a node is **conditionally independent** of every other node in the network outside its **Markov blanket**

- Location conditional distributions (**CPD**) and the **DAG** completely determines the **joint** dist.



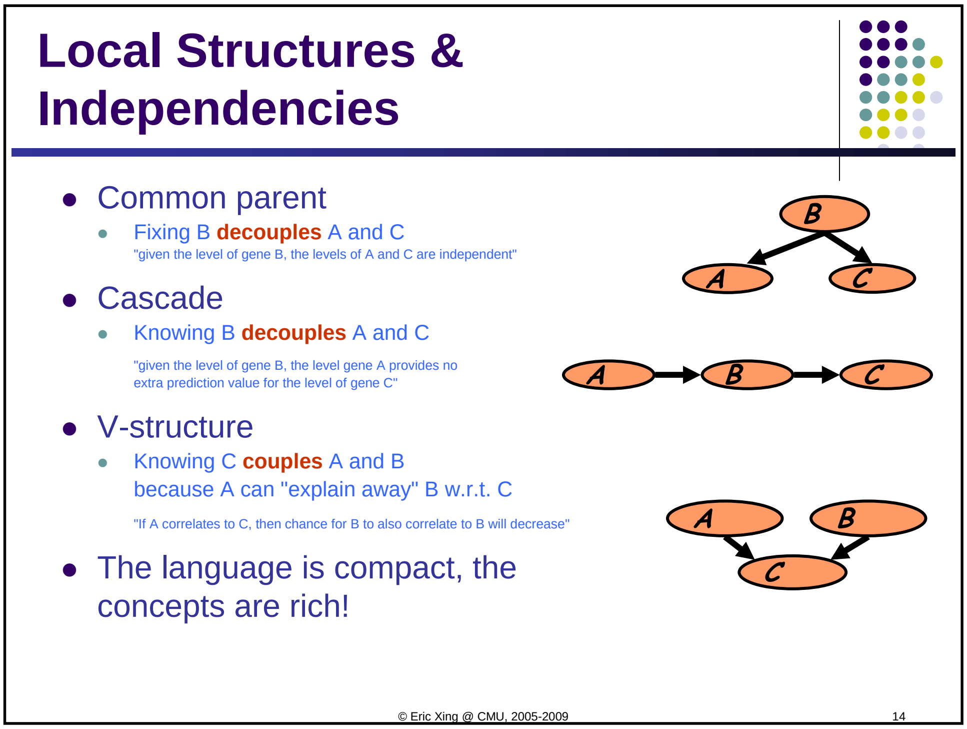


Markov Random Fields

Structure: an ***undirected graph***

- Meaning: a node is **conditionally independent** of every other node in the network given its UnDirected neighbors
- Local contingency functions (**potentials**) and the **cliques** in the graph completely determine the **joint** dist.
- Give **correlations** between variables, but no explicit way to generate samples

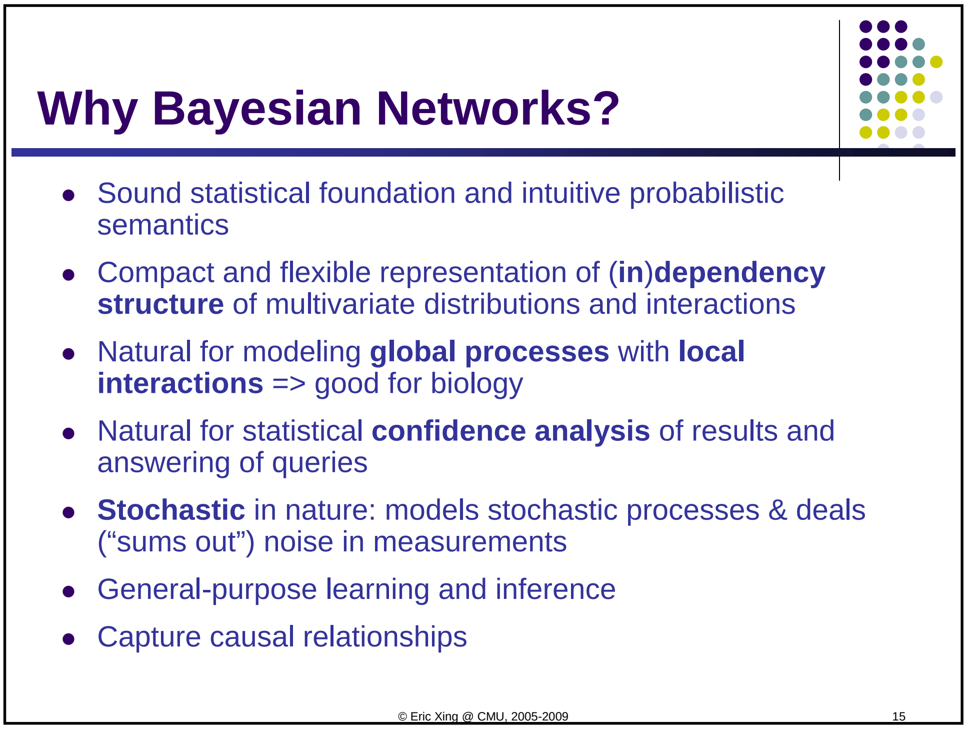
© Eric Xing @ CMU, 2005-2009 13



Local Structures & Independencies

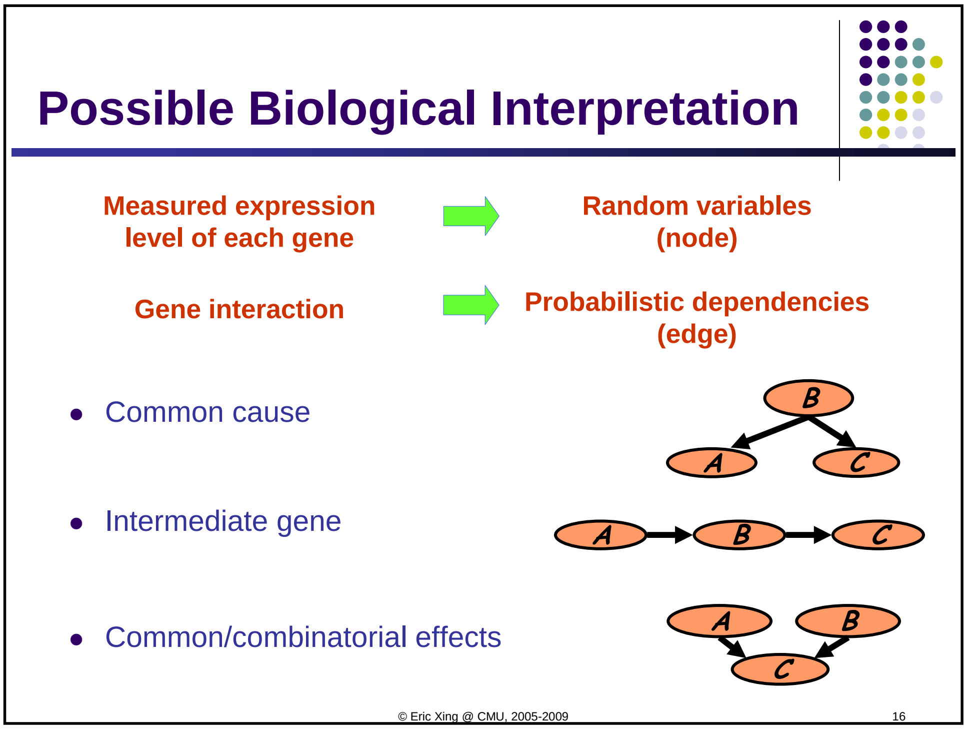
- Common parent
  - Fixing B **decouples** A and C
    "given the level of gene B, the levels of A and C are independent"
- Cascade
  - Knowing B **decouples** A and C
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"
- V-structure
  - Knowing C **couples** A and B because A can "explain away" B w.r.t. C
    "If A correlates to C, then chance for B to also correlate to B will decrease"
- The language is compact, the concepts are rich!

© Eric Xing @ CMU, 2005-2009 14



Why Bayesian Networks?

- Sound statistical foundation and intuitive probabilistic semantics
- Compact and flexible representation of (**in**)**dependency structure** of multivariate distributions and interactions
- Natural for modeling **global processes** with **local interactions** => good for biology
- Natural for statistical **confidence analysis** of results and answering of queries
- **Stochastic** in nature: models stochastic processes & deals ("sums out") noise in measurements
- General-purpose learning and inference
- Capture causal relationships

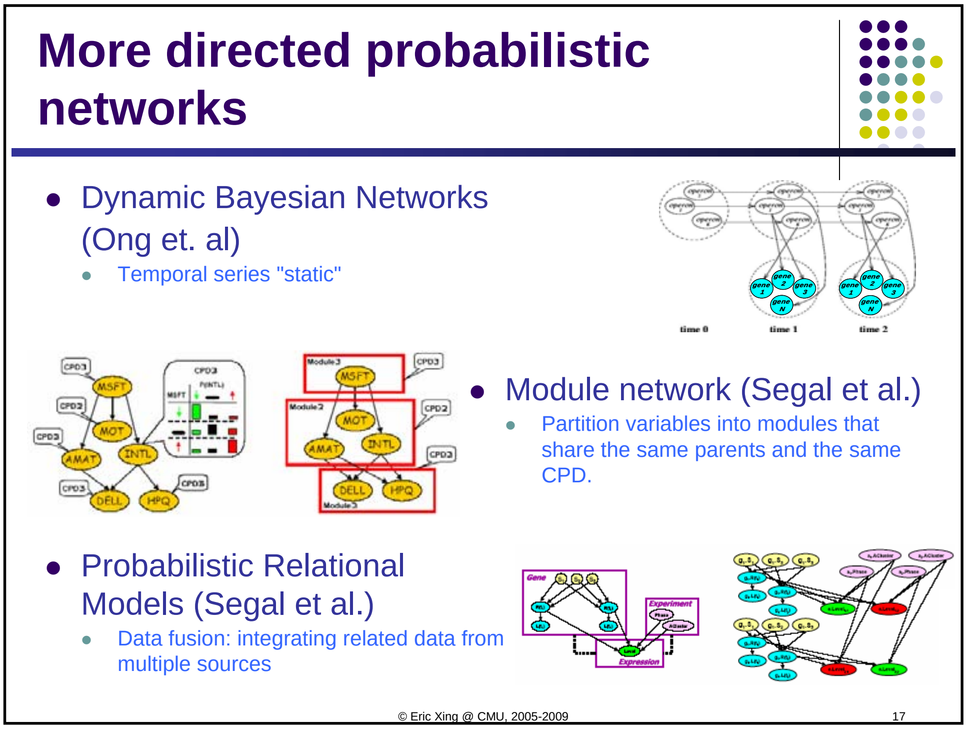


Possible Biological Interpretation

Measured expression level of each gene → **Random variables (node)**

Gene interaction → **Probabilistic dependencies (edge)**

- Common cause
- Intermediate gene
- Common/combinatorial effects

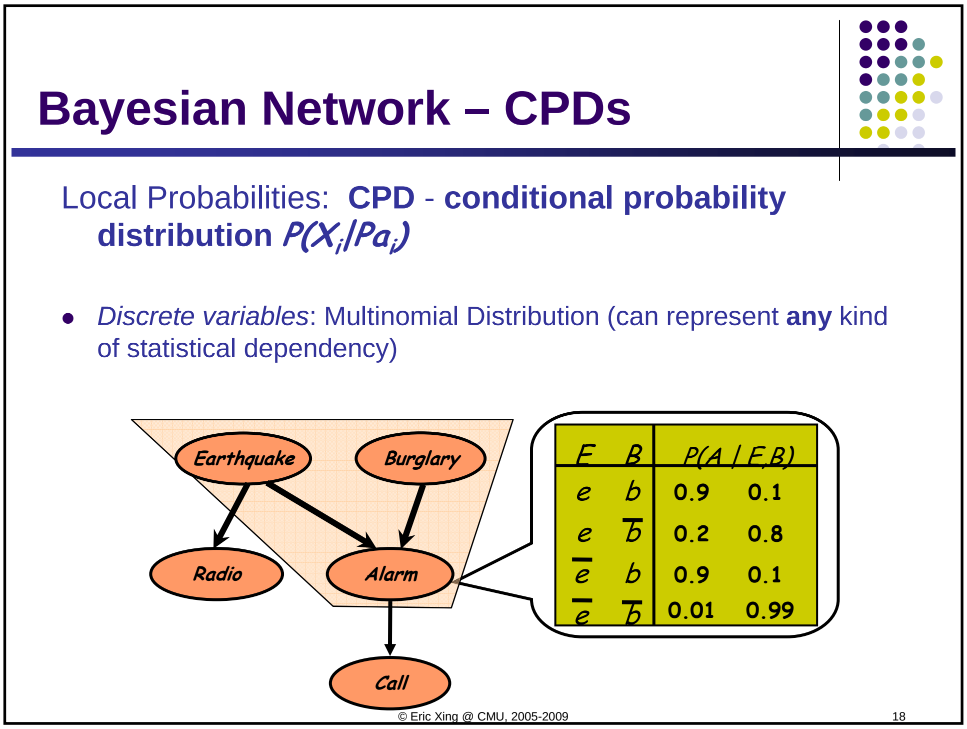


More directed probabilistic networks

- Dynamic Bayesian Networks (Ong et. al)
  - Temporal series "static"

- Module network (Segal et al.)
  - Partition variables into modules that share the same parents and the same CPD.
- Probabilistic Relational Models (Segal et al.)
  - Data fusion: integrating related data from multiple sources

© Eric Xing @ CMU, 2005-2009 17



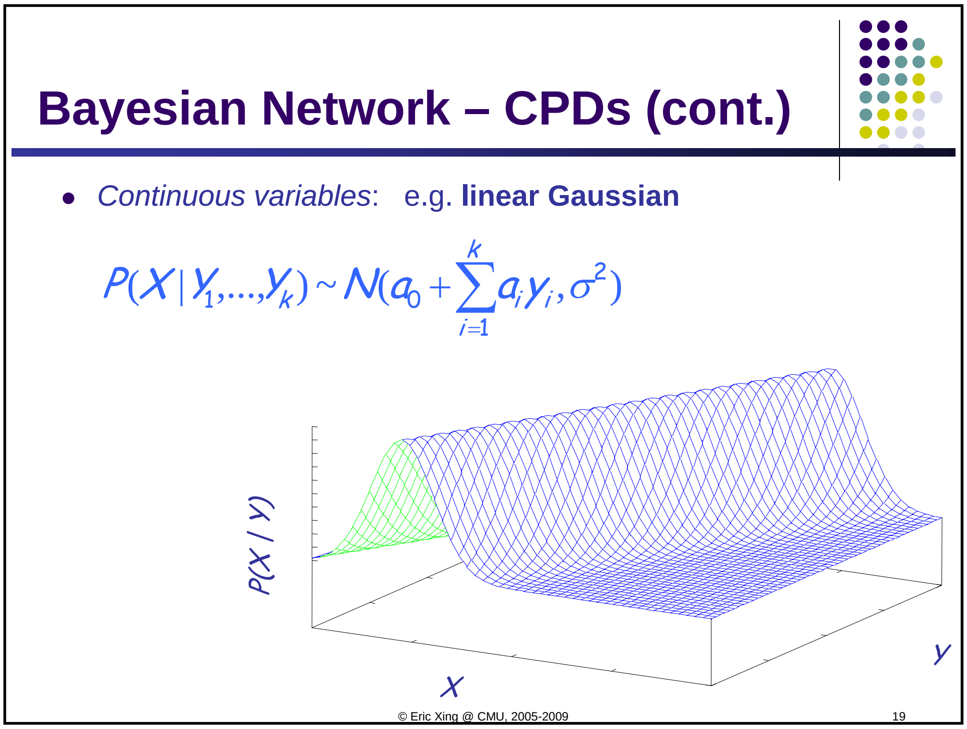
Bayesian Network – CPDs

Local Probabilities: **CPD - conditional probability distribution *$P(X_i|Pa_i)$***

- *Discrete variables*: Multinomial Distribution (can represent **any** kind of statistical dependency)

E	B	P(A | E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

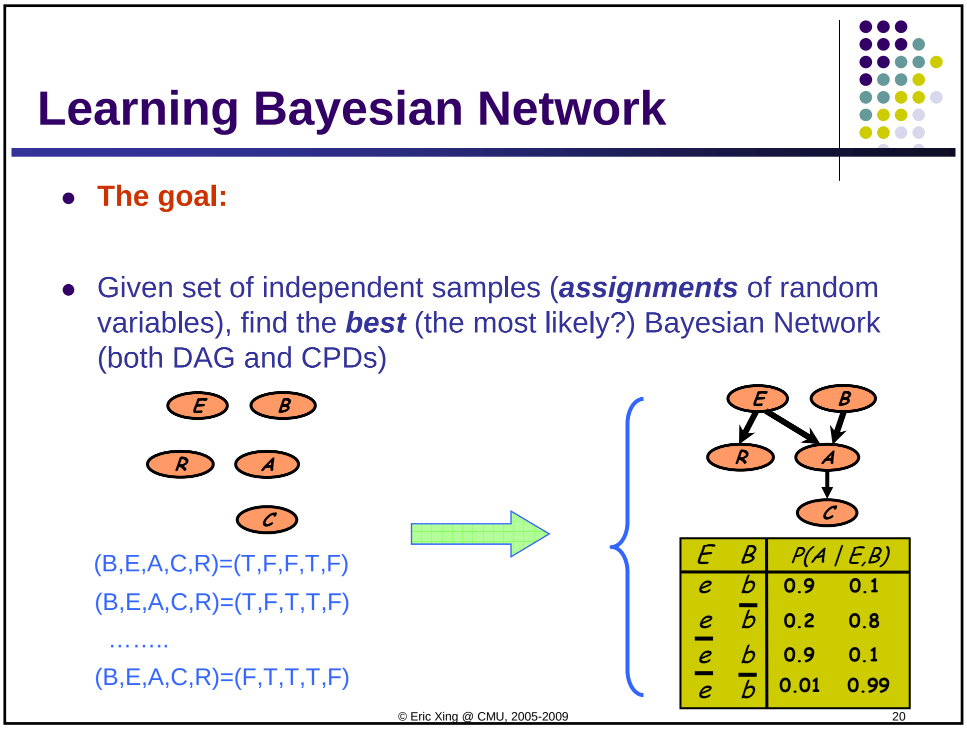
© Eric Xing @ CMU, 2005-2009 18



Bayesian Network – CPDs (cont.)

- *Continuous variables*: e.g. **linear Gaussian**

$$P(X \mid Y_1,\ldots,Y_k) \sim N(a_0 + \sum_{i=1}^{k} a_i y_i, \sigma^2)$$



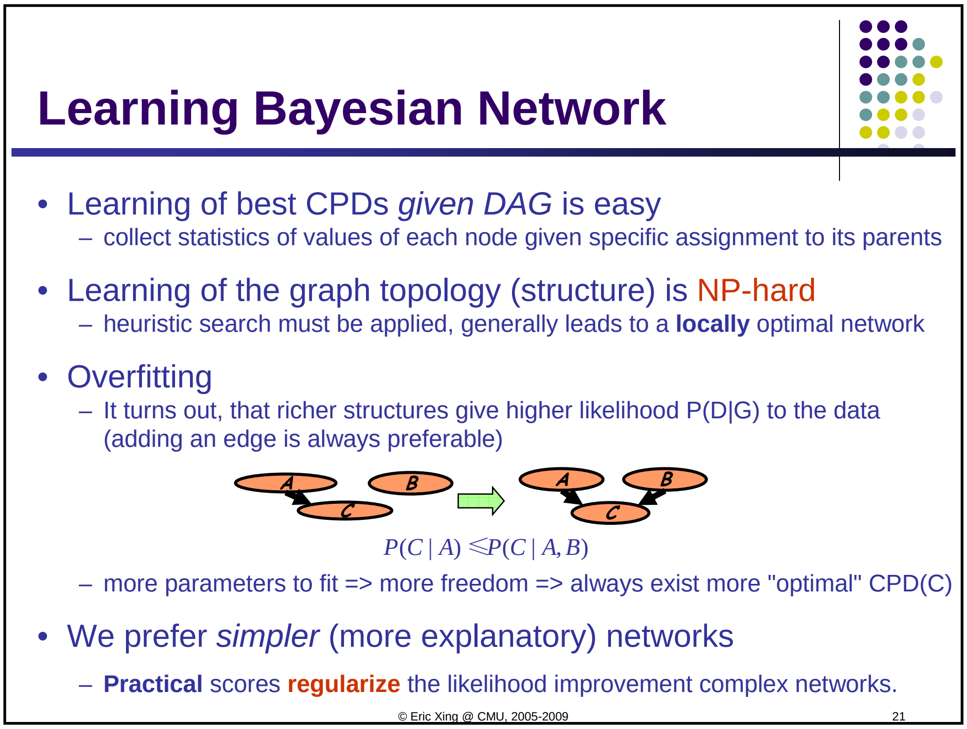
Learning Bayesian Network

- **The goal:**

- Given set of independent samples (***assignments*** of random variables), find the ***best*** (the most likely?) Bayesian Network (both DAG and CPDs)

(B,E,A,C,R)=(T,F,F,T,F)
(B,E,A,C,R)=(T,F,T,T,F)
……..
(B,E,A,C,R)=(F,T,T,T,F)

E	B	P(A | E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99



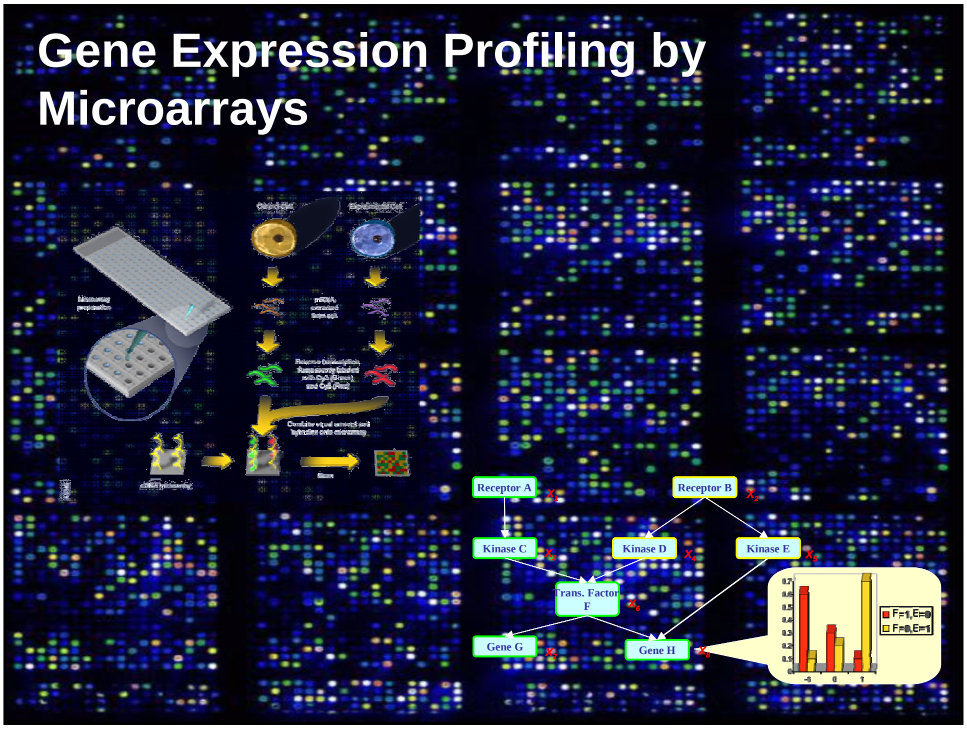
Learning Bayesian Network

- Learning of best CPDs *given DAG* is easy
  - collect statistics of values of each node given specific assignment to its parents
- Learning of the graph topology (structure) is NP-hard
  - heuristic search must be applied, generally leads to a **locally** optimal network
- Overfitting
  - It turns out, that richer structures give higher likelihood P(D|G) to the data (adding an edge is always preferable)

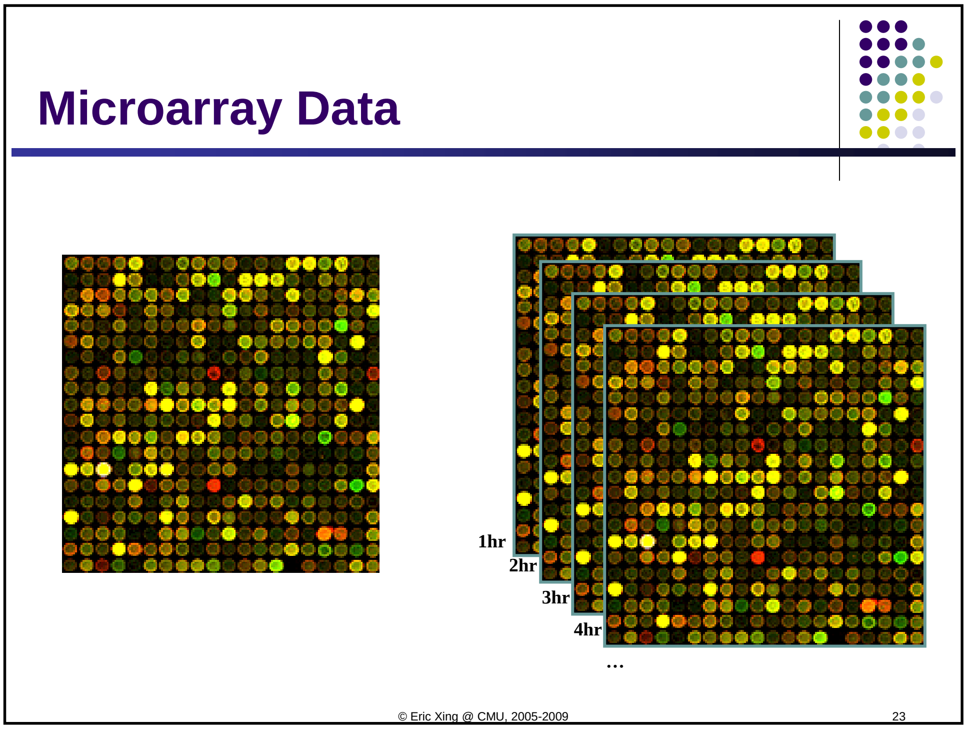
$$P(C \mid A) \leq P(C \mid A, B)$$

  - more parameters to fit => more freedom => always exist more "optimal" CPD(C)
- We prefer *simpler* (more explanatory) networks
  - **Practical** scores **regularize** the likelihood improvement complex networks.

© Eric Xing @ CMU, 2005-2009 21

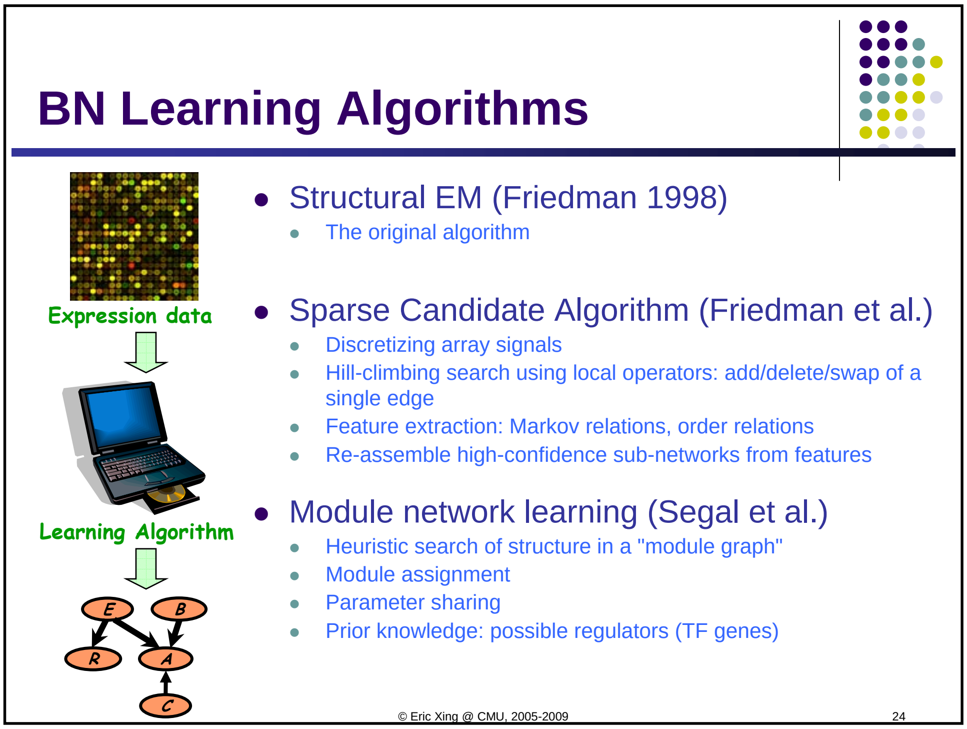


Gene Expression Profiling by Microarrays



Microarray Data

1hr
2hr
3hr
4hr
…



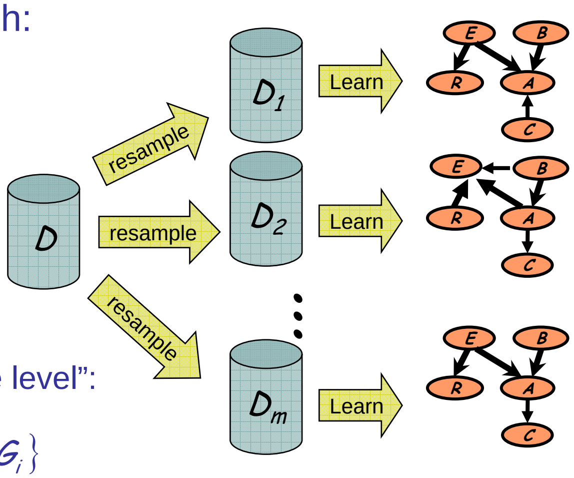
BN Learning Algorithms

Expression data
Learning Algorithm
E
B
R
A
C

- Structural EM (Friedman 1998)
  - The original algorithm
- Sparse Candidate Algorithm (Friedman et al.)
  - Discretizing array signals
  - Hill-climbing search using local operators: add/delete/swap of a single edge
  - Feature extraction: Markov relations, order relations
  - Re-assemble high-confidence sub-networks from features
- Module network learning (Segal et al.)
  - Heuristic search of structure in a "module graph"
  - Module assignment
  - Parameter sharing
  - Prior knowledge: possible regulators (TF genes)

Confidence Estimates

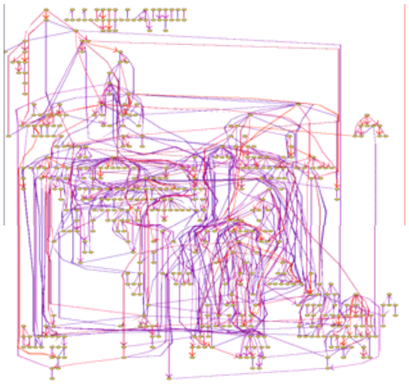
Bootstrap approach:



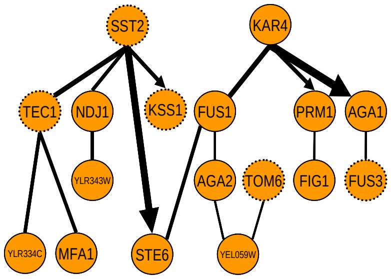
Estimate “Confidence level”:

$$C(f) = \frac{1}{m}\sum_{i=1}^{m} 1\{f \in G_i\}$$

Results from SCA + feature extraction (Friedman et al.)



The initially learned network of ~800 genes



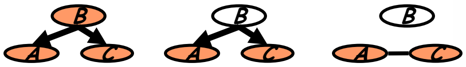
The “mating response” substructure

Gaussian Graphical Models

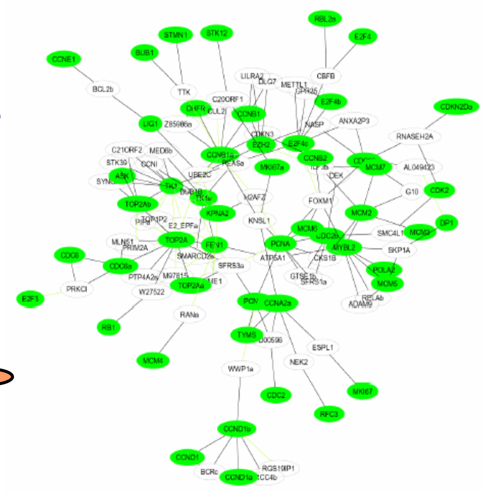
- **Why?**

  Sometimes an UNDIRECTED association graph makes more sense and/or is more informative

  - gene expressions may be influenced by unobserved factor that are **post-transcriptionally** regulated



  - The unavailability of the state of B results in a constrain over A and C



Covariance Selection

- Multivariate Gaussian over all continuous expressions

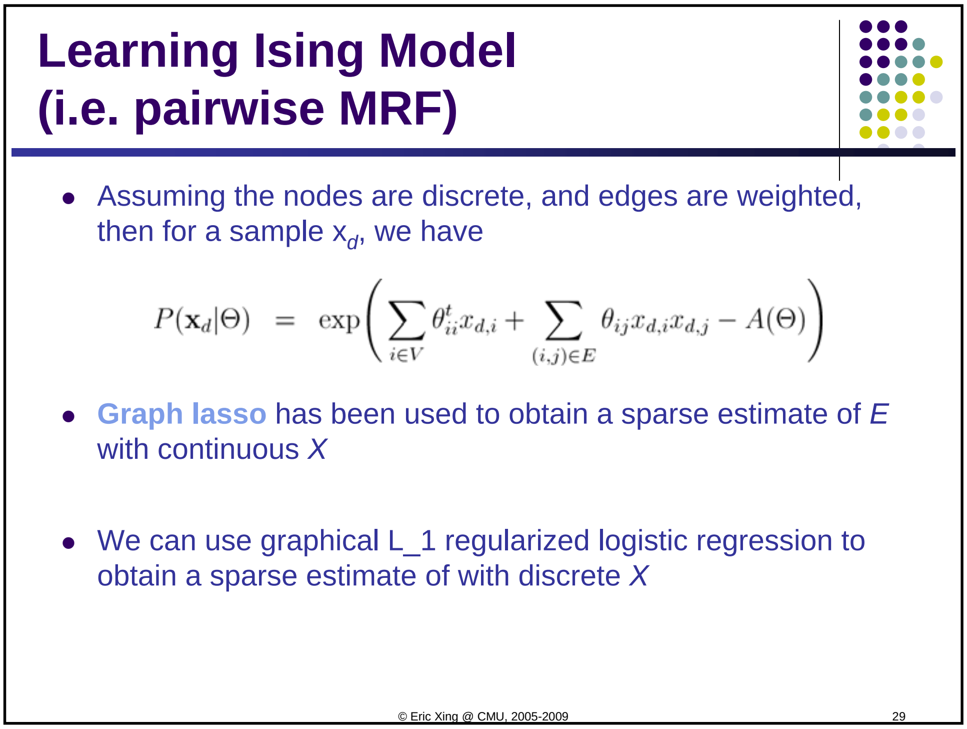
$$p([x_1,...,x_n]) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\tfrac{1}{2}(\vec{x}-\mu)^T \Sigma^{-1}(\vec{x}-\mu)\right\}$$

- The precision matrix $K=\Sigma^{-1}$ reveals the topology of the (undirected) network

$$E(x_i \mid x_{-i}) = \sum_j (K_{ij}/K_{ii}) x_j$$

  - Edge ~ $|K_{ij}| > 0$

- Learning Algorithm: Covariance selection
  - Want a sparse matrix
    - Regression for each node with degree constraint (Dobra et al.)
    - Regression for each node with hierarchical Bayesian prior (Li, et al)



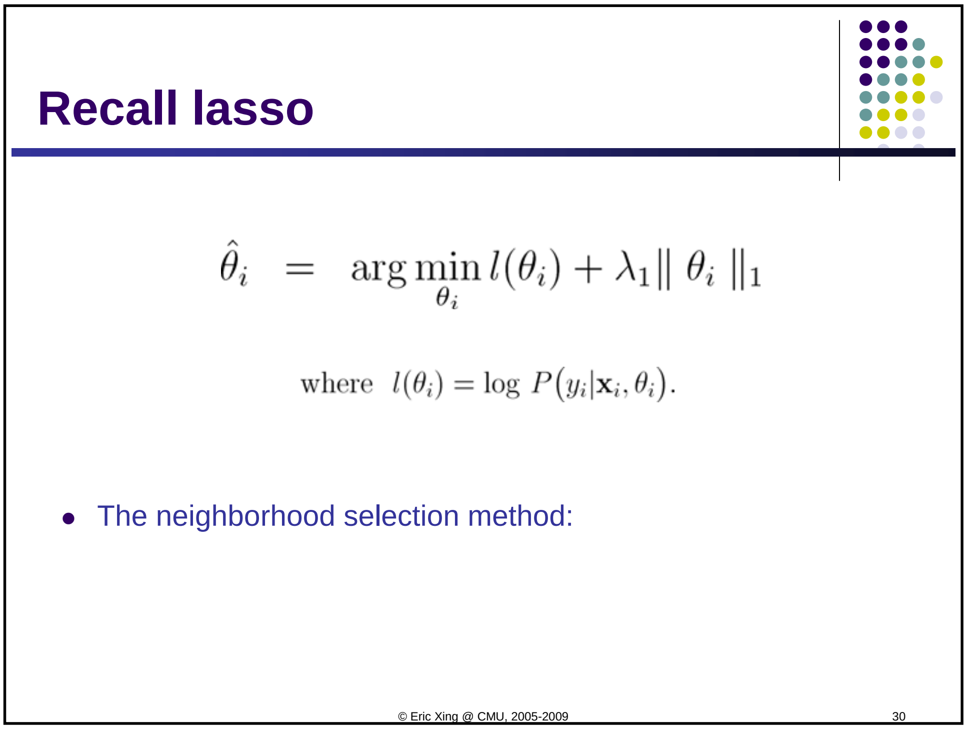
Learning Ising Model (i.e. pairwise MRF)

- Assuming the nodes are discrete, and edges are weighted, then for a sample x$_d$, we have

$$P(\mathbf{x}_d|\Theta) \quad = \quad \exp\left(\sum_{i\in V}\theta^t_{ii}x_{d,i} + \sum_{(i,j)\in E}\theta_{ij}x_{d,i}x_{d,j} - A(\Theta)\right)$$

- **Graph lasso** has been used to obtain a sparse estimate of *E* with continuous *X*

- We can use graphical L_1 regularized logistic regression to obtain a sparse estimate of with discrete *X*

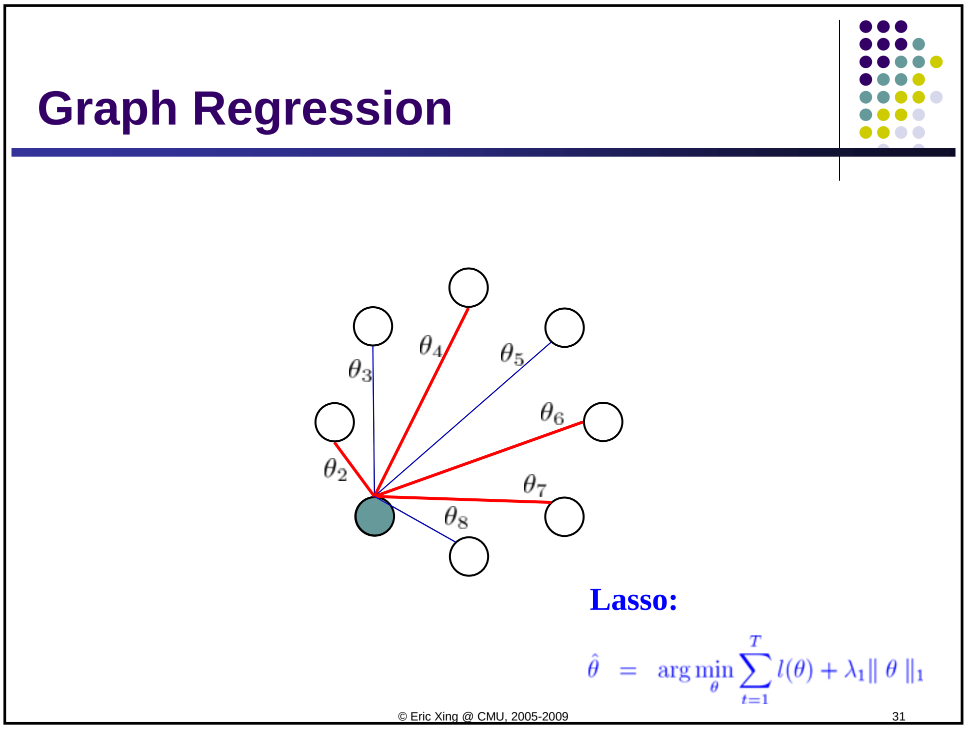


Recall lasso

$$\hat{\theta}_i \quad = \quad \arg\min_{\theta_i} l(\theta_i) + \lambda_1 \|\ \theta_i\ \|_1$$

$$\text{where} \quad l(\theta_i) = \log\ P\big(y_i|\mathbf{x}_i, \theta_i\big).$$

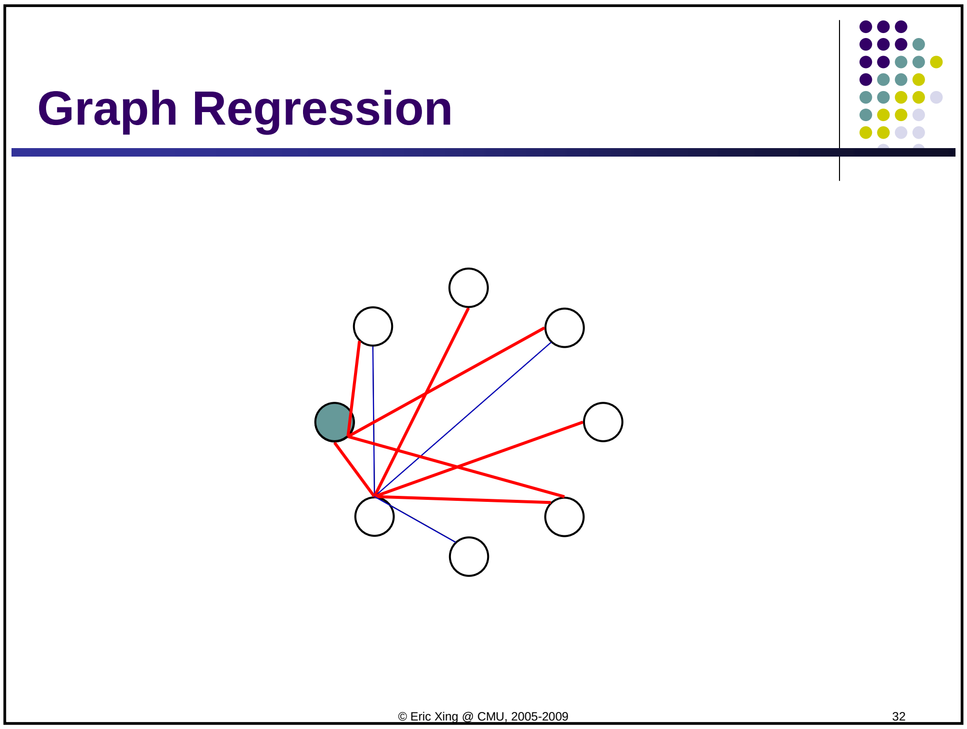
- The neighborhood selection method:



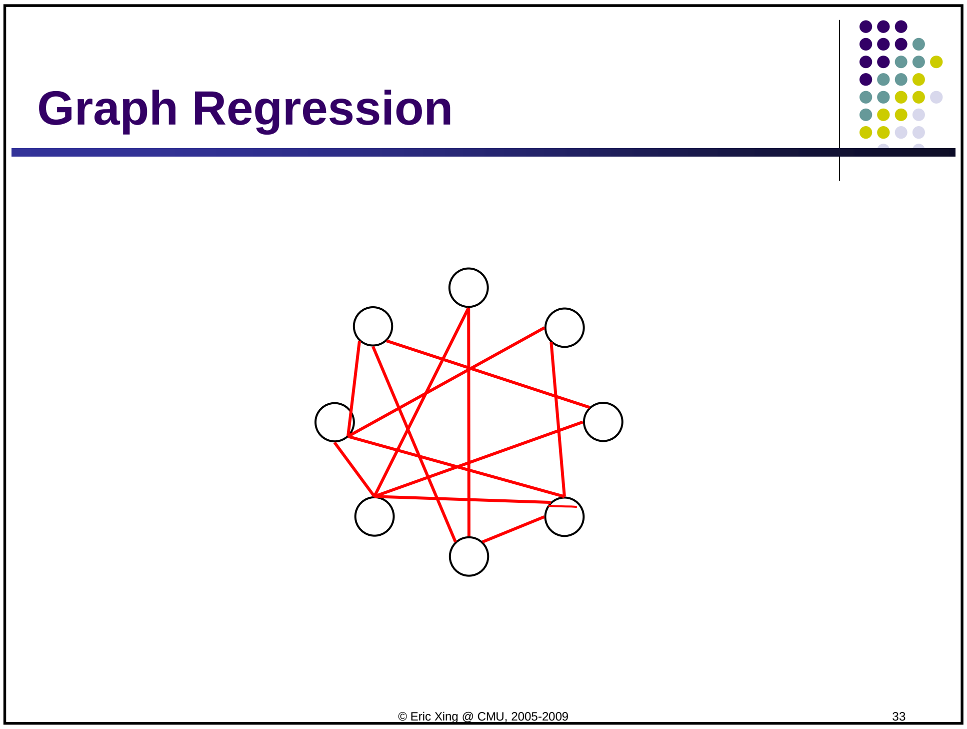
Graph Regression

Lasso:

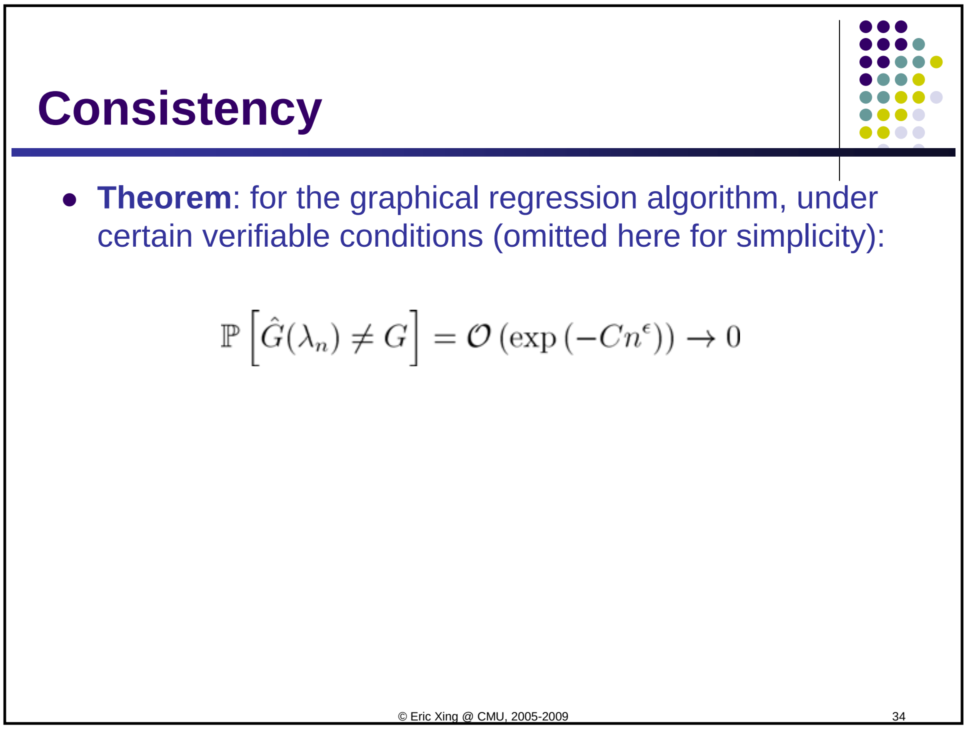
$$\hat{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} l(\theta) + \lambda_1 \| \theta \|_1$$



Graph Regression



Graph Regression



Consistency

- **Theorem**: for the graphical regression algorithm, under certain verifiable conditions (omitted here for simplicity):

$$\mathbb{P}\left[\hat{G}(\lambda_n) \neq G\right] = \mathcal{O}\left(\exp\left(-Cn^{\epsilon}\right)\right) \to 0$$

Summary: Learning GM

- Learning of best CPDs *given DAG* is easy
  - collect statistics of values of each node given specific assignment to its parents
- Learning of the graph topology (structure) is NP-hard
  - heuristic search must be applied, generally leads to a **locally** optimal network
- We prefer *simpler* (more explanatory) networks
  - **Regularized graph regression**

References

Using Bayesian Networks to Analyze Expression Data N. Friedman, M. Linial, I. Nachman, D. Pe'er. In Proc. Fourth Annual Inter. Conf. on Computational Molecular Biology RECOMB, 2000.

Genome-wide Discovery of Transcriptional Modules from DNA Sequence and Gene Expression, E. Segal, R. Yelensky, and D. Koller. Bioinformatics, 19(Suppl 1): 273--82, 2003.

A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules Joshua M. Stuart, Eran Segal, Daphne Koller, Stuart K. Kim, *Science*, 302 (5643):249-55, 2003.

Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. *Nature Genetics, 34(2): 166-76, 2003.*

Learning Module Networks E. Segal, N. FriedmanD. Pe'er, A. Regev, and D. Koller. In Proc. Nineteenth Conf. on Uncertainty in Artificial Intelligence (UAI), 2003

From Promoter Sequence to Expression: A Probabilistic Framework, E. Segal, *Y. Barash, I. Simon, N. Friedman, and D. Koller. In Proc. 6th Inter. Conf. on Research in Computational Molecular Biology (RECOMB), 2002*

Modelling Regulatory Pathways in *E.coli* from Time Series Expression Profiles, Ong, I. M., J. D. Glasner and D. Page, *Bioinformatics*, 18:241S-248S, 2002.

References

Sparse graphical models for exploring gene expression data, Adrian Dobra, Beatrix Jones, Chris Hans, Joseph R. Nevins and Mike West, *J. Mult. Analysis*, 90, 196-212, 2004.

Experiments in stochastic computation for high-dimensional graphical models, Beatrix Jones, Adrian Dobra, Carlos Carvalho, Chris Hans, Chris Carter and Mike West, Statistical Science, 2005.

Inferring regulatory network using a hierarchical Bayesian graphical gaussian model, Fan Li, Yiming Yang, Eric Xing (working paper) 2005.

Computational discovery of gene modules and regulatory networks. Z. Bar-Joseph*, G. Gerber*, T. Lee*, N. Rinaldi, J. Yoo, F. Robert, B. Gordon, E. Fraenkel, T. Jaakkola, R. Young, and D. Gifford. Nature Biotechnology, 21(11) pp. 1337-42, 2003.

Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data, A. Tanay, R. Sharan, M. Kupiec, R. Shamir *Proc. National Academy of Science USA* 101 (9) 2981--2986, 2004.

Informative Structure Priors: Joint Learning of Dynamic Regulatory Networks from Multiple Types of Data, Bernard, A. & Hartemink, A. In *Pacific Symposium on Biocomputing 2005 (PSB05)*,