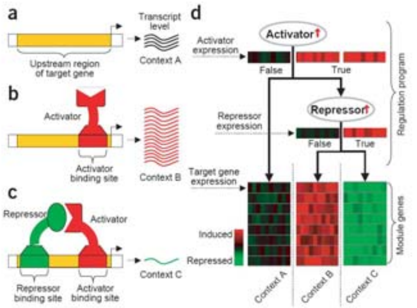
Inferring regulatory, signaling & metabolic networks from data

Devika Subramanian
Comp 470

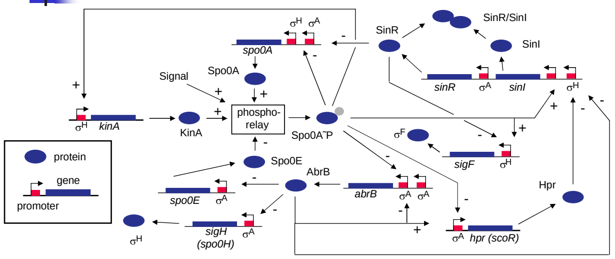
Networks

- Regulatory network: network of control decisions used to turn genes on/off.
- Signaling network: interactions among genes, gene products and small molecules that activate cellular processes.
- Metabolic network: network of proteins that synthesize and breakdown cellular molecules.

Regulators

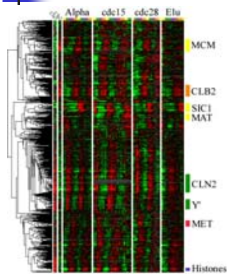


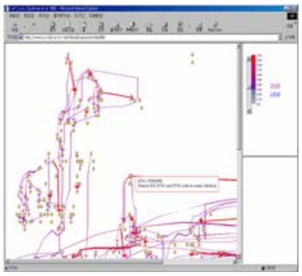
Genetic regulatory network of *B. subtilis*



Genetic regulatory network controlling the initiation of sporulation.

From expression data to gene regulatory networks



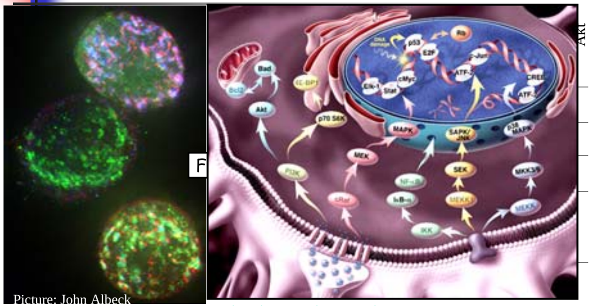


Microarray data

Yeast cell cycle

(c) Devika Subramanian, 2006

From flow cytometry data to signaling networks



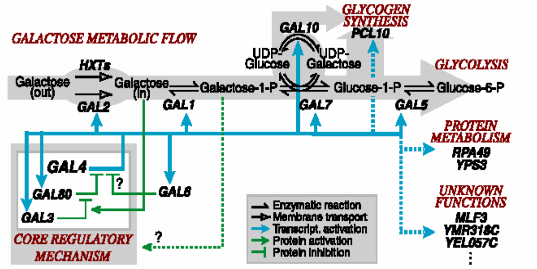
High throughput data

Signaling Pathways

K. Sachs, 2005

(c) Devika Subramanian, 2006

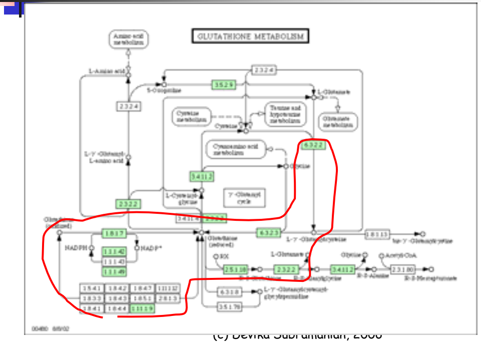
The galactose pathway (manually discovered)



T. Ideker, et al., Science 292 (May 4, 2001) 929-934.

(c) Devika Subramanian, 2006

The glutathione metabolism



An important metabolic process; detoxification in cells. Known to be disrupted in several cancers.

Kegg pathway

(c) Devika Subramanian, 2006

Outline

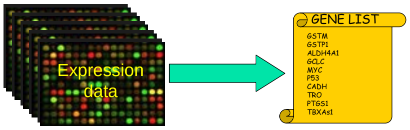
- The problem of learning regulatory, signaling and metabolic networks from data
- A quick intro to Bayesian networks
- Algorithms for learning Bayesian networks from data
- Examples
  - Glutathione metabolism from humans (expression data)
  - Regulatory network from yeast cell cycle (expression data)
  - T-cell signaling from humans (flow cytometry data)

Challenges

- The cell is a complex stochastic domain: signal transduction, metabolic and regulatory pathways all interconnected.
- Pathways are controlled by combination of many mechanisms.
- We only observe mRNA levels and/or phospho-lipid levels.
- Many interactions are not directly observed at the mRNA level
- Measurements are noisy.

Some initial approaches

- Classification of expression data
  - Reveals genes that are differentially expressed.
  - Disadvantage: does not reveal structural relationships between genes.



Some initial approaches

- Clustering techniques
  - Many interesting clusters of co-regulated genes
  - No system-level insight.

Some initial approaches

- Boolean networks
  - Deterministic models of interactions between genes.
  - Disadvantage: deterministic. We need stochastic models for representing interactions.

Why probabilistic models?

Gene regulation occurs at many stages:

- pre-transcriptional (chromatin structure)
- transcription initiation
- RNA editing (splicing) and transport
- Translation initiation
- Post-translation modification
- RNA & Protein degradation

All these processes are stochastic!

Why Bayesian networks?

- The important science/technology to come out of AI in the last 15 years.
- Underlies all important applications today.
- Frames every question as the estimation of a conditional probability
  - P(disease/problem|set of symptoms)
  - P(email is spam|email text+header)
  - P(hurricane will hit place X|movement history)
  - P(sentence|acoustic signal)
  - P(regulatory network|gene exp data)

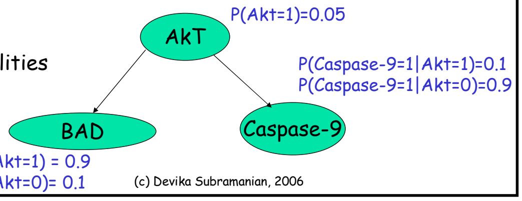
Example: Akt pathway

Random variables: Akt, BAD, caspase-9

Conditional independencies:

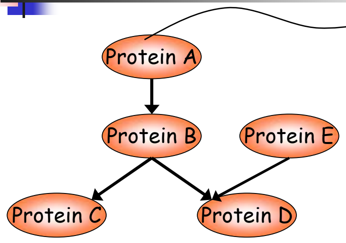
P(BAD and caspase-9|AKT) =P(BAD|Akt)P(Caspase-9|AkT)

2+2+1
probabilities



P(BAD=1|Akt=1) = 0.9
P(BAD=1|Akt=0)= 0.1

Another example



If Protein A is low(0), Protein B is high(1) with probability 0.8

$P(B=1|A=0) = 0.8$
$P(B=1|A=1) = 0.3$

Adapted from Sachs, 2005

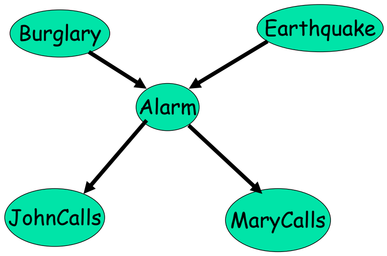
Bayesian networks: the model

- A Bayesian network B = (V,E) is a directed acyclic graph in which each node in V is annotated with quantitative probability information.
  - A set V of random variables are the nodes of the network. They can be continuous or discrete.
  - If there is an edge from node X to node Y in E, then X is said to be the parent of Y.
  - Each node X in V has a conditional probability distribution P(X|Parents(X)) associated with it.

Segue

- … to an old example from Pearl 1986.
- Illustrates the major kinds of stochastic dependencies that can be modeled using Bayesian networks

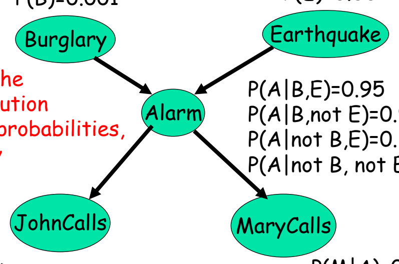
A simple Bayesian network



Semantics of Bayesian networks

- The topology of the network reflects a set of conditional independence statements.
  - Burglary and Earthquake directly affect the probability of the alarm going off, but whether or not John or Mary calls depends on the alarm. John and Mary do not directly perceive burglary or minor earthquakes.
  - JohnCalls is conditionally independent of MaryCalls given Alarm.

Bayesian network with CPTs



P(B)=0.001

P(E)=0.002

P(A|B,E)=0.95
P(A|B,not E)=0.95
P(A|not B,E)=0.29
P(A|not B, not E)=0.001

Specifying the joint distribution requires 31 probabilities, here we only specify 10.

P(J|A)=0.9
P(J|not A)=0.05

P(M|A)=0.7
P(M|not A)=0.01

Computing joint probability distributions

- Any entry in the joint probability distribution can be calculated from the Bayesian network.

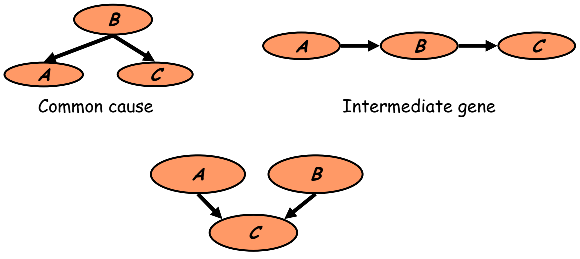
$$
\begin{aligned}
P(J,M,A,\neg B,\neg E) &= P(J \mid M,A,\neg B,\neg E)P(M,A,\neg B,\neg E) \\
&= P(J \mid A)P(M \mid A,\neg B,\neg E)P(A,\neg B,\neg E) \\
&= P(J \mid A)P(M \mid A)P(A \mid \neg B,\neg E)P(\neg B,\neg E) \\
&= P(J \mid A)P(M \mid A)P(A \mid \neg B,\neg E)P(\neg B)P(\neg E)
\end{aligned}
$$

Computing joint probabilities

$$P(X_1 = x_1,\ldots,X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid Parents(X_i))$$

P(Burglary|Alarm) = 0.376
P(Burglary|Alarm,Earthquake) = 0.003

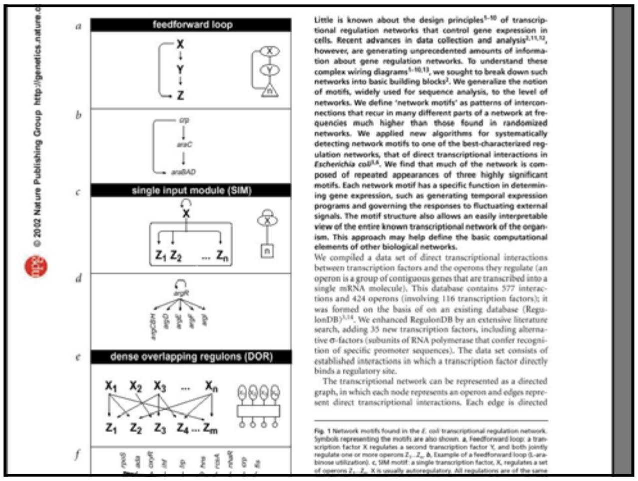
Summary of dependency types



Common cause

Intermediate gene

Common effects


© 2002 Nature Publishing Group http://genetics.nature.c
a
feedforward loop
X
Y
Z
X
Y
n
b
crp
araC
araBAD
c
single input module (SIM)
X
Z1 Z2 ... Zn
X
n
d
argR
argCBH
argD
argE
argF
argI
e
dense overlapping regulons (DOR)
X1 X2 X3 ... Xn
Z1 Z2 Z3 Z4 ... Zm
f
rpoS
ada
oxyR
lrp
hns
rcsA
nhaR
crp

Little is known about the design principles[1-10] of transcriptional regulation networks that control gene expression in cells. Recent advances in data collection and analysis[2,11,12], however, are generating unprecedented amounts of information about gene regulation networks. To understand these complex wiring diagrams[1-10,13], we sought to break down such networks into basic building blocks[2]. We generalize the notion of motifs, widely used for sequence analysis, to the level of networks. We define 'network motifs' as patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks. We applied new algorithms for systematically detecting network motifs to one of the best-characterized regulation networks, that of direct transcriptional interactions in *Escherichia coli*[3,6]. We find that much of the network is composed of repeated appearances of three highly significant motifs. Each network motif has a specific function in determining gene expression, such as generating temporal expression programs and governing the responses to fluctuating external signals. The motif structure also allows an easily interpretable view of the entire known transcriptional network of the organism. This approach may help define the basic computational elements of other biological networks.

We compiled a data set of direct transcriptional interactions between transcription factors and the operons they regulate (an operon is a group of contiguous genes that are transcribed into a single mRNA molecule). This database contains 577 interactions and 424 operons (involving 116 transcription factors); it was formed on the basis of on an existing database (RegulonDB)[3,14]. We enhanced RegulonDB by an extensive literature search, adding 35 new transcription factors, including alternative σ-factors (subunits of RNA polymerase that confer recognition of specific promoter sequences). The data set consists of established interactions in which a transcription factor directly binds a regulatory site.

The transcriptional network can be represented as a directed graph, in which each node represents an operon and edges represent direct transcriptional interactions. Each edge is directed

Fig. 1 Network motifs found in the *E. coli* transcriptional regulation network. Symbols representing the motifs are also shown. ***a***, Feedforward loop: a transcription factor X regulates a second transcription factor Y, and both jointly regulate one or more operons $Z_1...Z_n$. ***b***, Example of a feedforward loop (L-arabinose utilization). ***c***, SIM motif: a single transcription factor, X, regulates a set of operons $Z_1...Z_n$. X is usually autoregulatory. All regulations are of the same

Conditional probability distributions

- Multinomial model
  - Discrete values
- Linear Gaussian model
  - $P(X|u_1, u_2, \ldots, u_k) = N(a_0 + \Sigma_i a_i u_i, \sigma^2)$

Modeling genetic networks

Variables of interest:

- Expression levels of genes
- Concentration levels of proteins
- Exogenous variables: Nutrient levels, Metabolite Levels, Temperature
- Phenotype information
- …

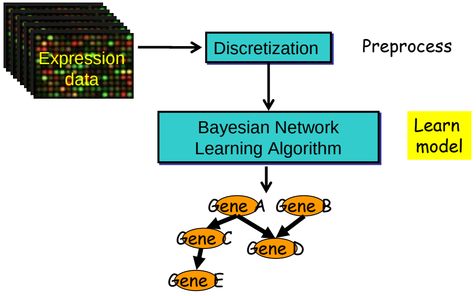
Bayesian Network Structure:

- Capture dependencies among these variables

Advantages of Bayesian networks

- Flexible representation of (**in**)**dependency structure** of multivariate distributions and interactions.
- Natural for modeling **global processes** with **local interactions**.
- Clear probabilistic semantics.
- Natural for statistical **confidence analysis** of results and answering of queries.
- **Stochastic** in nature: models stochastic processes & deals well with noise in measurements.

Learning Bayesian networks

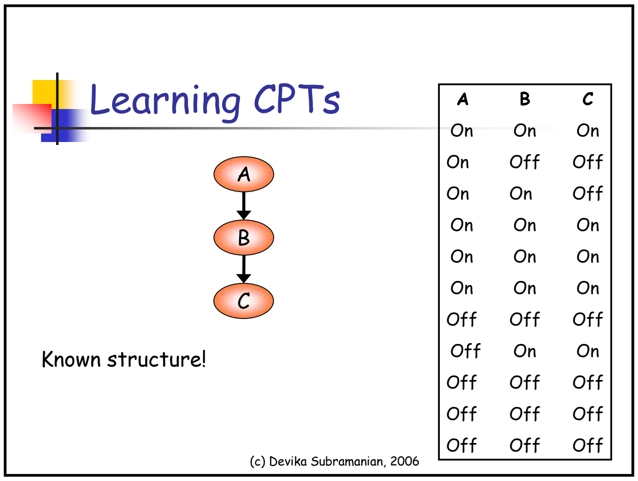


Need for discretization

- The expression measurements are **real numbers**.
  - We need to discretize them in order to learn general conditional probability distributions. This step entails a loss of information.
  - If we don't discretize, we must assume some specific type of conditional probability distribution (like "linear Gaussian"), and this assumption causes loss of modeling fidelity.

Learning Bayesian Models

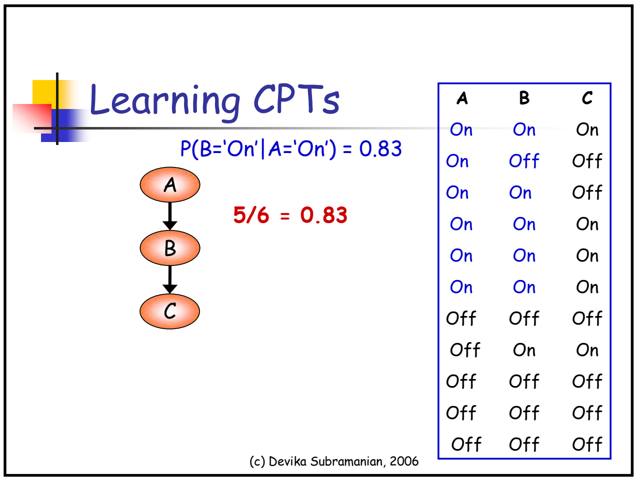
- Using gene expression data D, find the Bayesian network G that is most likely given the data, i.e. G that maximizes P(G|D).
- Two cases
  - Graph structure is known; the conditional probability distributions are unknown.
    - Recovering optimal conditional probability distributions when the graph is known is "easy".
  - Graph structure and the conditional probability distributions are unknown.
    - Recovering optimal graph structure is NP-hard.



Learning CPTs

Known structure!

A	B	C
On	On	On
On	Off	Off
On	On	Off
On	On	On
On	On	On
On	On	On
Off	Off	Off
Off	On	On
Off	Off	Off
Off	Off	Off
Off	Off	Off

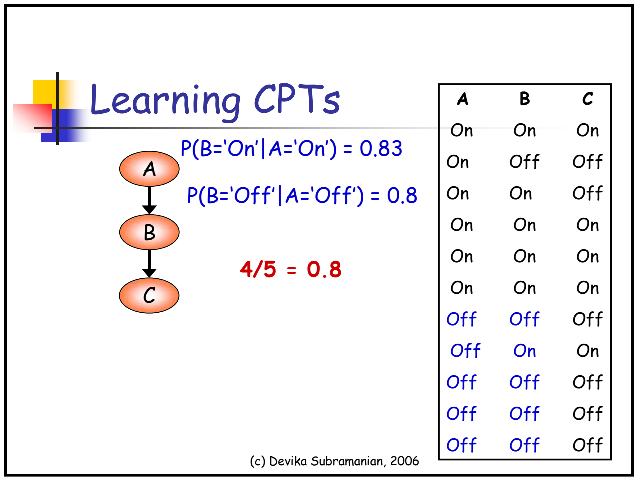


Learning CPTs

P(B='On'|A='On') = 0.83

5/6 = 0.83

A	B	C
On	On	On
On	Off	Off
On	On	Off
On	On	On
On	On	On
On	On	On
Off	Off	Off
Off	On	On
Off	Off	Off
Off	Off	Off
Off	Off	Off



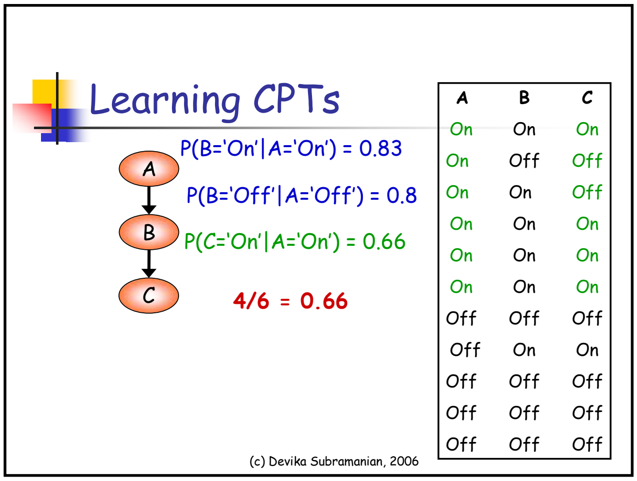
Learning CPTs

P(B='On'|A='On') = 0.83

P(B='Off'|A='Off') = 0.8

4/5 = 0.8

A	B	C
On	On	On
On	Off	Off
On	On	Off
On	On	On
On	On	On
On	On	On
Off	Off	Off
Off	On	On
Off	Off	Off
Off	Off	Off
Off	Off	Off



Learning CPTs

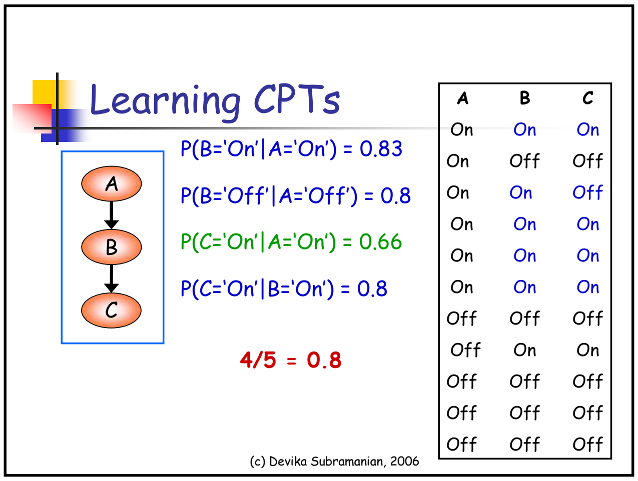
P(B='On'|A='On') = 0.83

P(B='Off'|A='Off') = 0.8

P(C='On'|A='On') = 0.66

4/6 = 0.66

A	B	C
On	On	On
On	Off	Off
On	On	Off
On	On	On
On	On	On
On	On	On
Off	Off	Off
Off	On	On
Off	Off	Off
Off	Off	Off
Off	Off	Off



Learning CPTs

P(B='On'|A='On') = 0.83

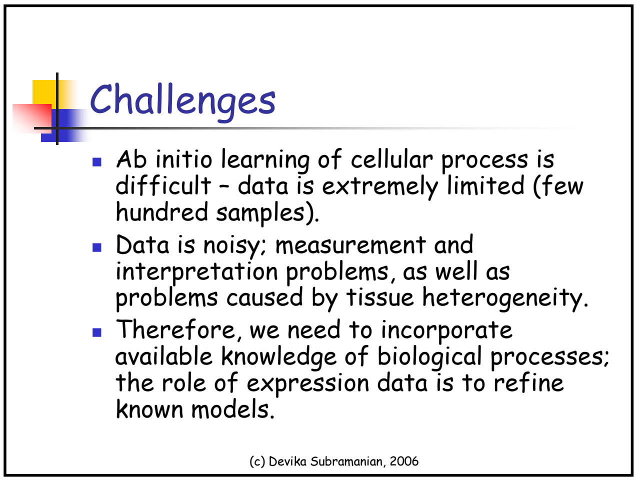
P(B='Off'|A='Off') = 0.8

P(C='On'|A='On') = 0.66

P(C='On'|B='On') = 0.8

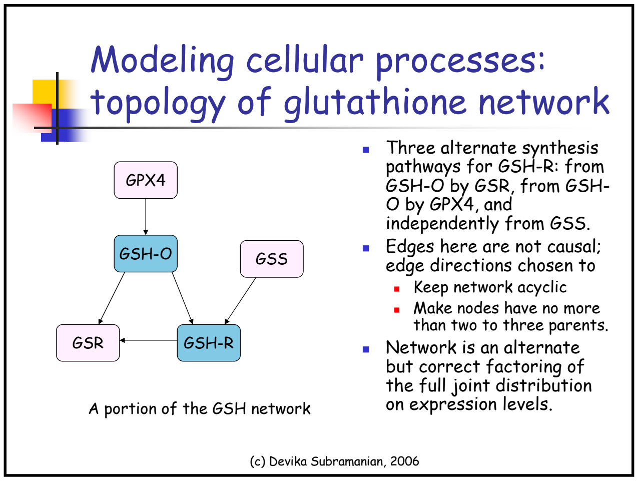
4/5 = 0.8

A	B	C
On	On	On
On	Off	Off
On	On	Off
On	On	On
On	On	On
On	On	On
Off	Off	Off
Off	On	On
Off	Off	Off
Off	Off	Off
Off	Off	Off



Challenges

- Ab initio learning of cellular process is difficult – data is extremely limited (few hundred samples).
- Data is noisy; measurement and interpretation problems, as well as problems caused by tissue heterogeneity.
- Therefore, we need to incorporate available knowledge of biological processes; the role of expression data is to refine known models.

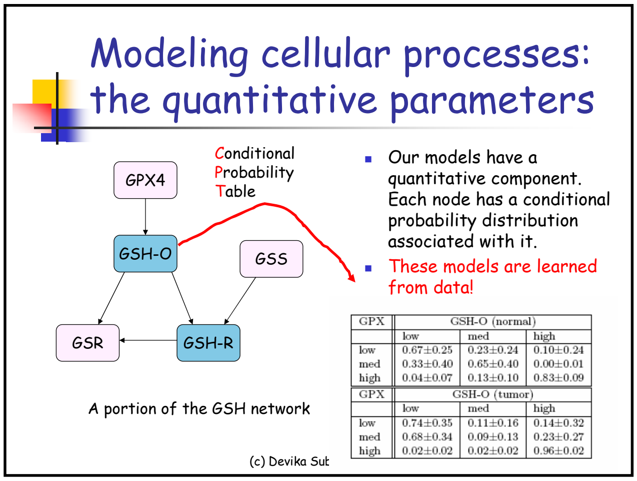


Modeling cellular processes: topology of glutathione network

GPX4
GSH-O
GSS
GSR
GSH-R

A portion of the GSH network

- Three alternate synthesis pathways for GSH-R: from GSH-O by GSR, from GSH-O by GPX4, and independently from GSS.
- Edges here are not causal; edge directions chosen to
  - Keep network acyclic
  - Make nodes have no more than two to three parents.
- Network is an alternate but correct factoring of the full joint distribution on expression levels.



Modeling cellular processes: the quantitative parameters

GPX4
GSH-O
GSS
GSR
GSH-R

Conditional Probability Table

A portion of the GSH network

- Our models have a quantitative component. Each node has a conditional probability distribution associated with it.
- These models are learned from data!

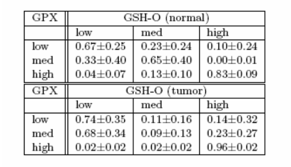
GPX	GSH-O (normal)		
	low	med	high
low	0.67±0.25	0.23±0.24	0.10±0.24
med	0.33±0.40	0.65±0.40	0.00±0.01
high	0.04±0.07	0.13±0.10	0.83±0.09

GPX	GSH-O (tumor)		
	low	med	high
low	0.74±0.35	0.11±0.16	0.14±0.32
med	0.68±0.34	0.09±0.13	0.23±0.27
high	0.02±0.02	0.02±0.02	0.96±0.02

Learning CPTs from data

- To learn a CPT of the form P(Y|X), where Y and X are both observed, we can use maximum likelihood estimation.
  - P(Y|X)=count(X&Y)/count(Y)
- When there are unobserved variables, we use the expectation maximization (EM) procedure to make the best guess for the values of the unobserved variables given the observed ones, and readjust the parameters of the network based on the guesses. We find the most likely network parameters given the observed data.

Component network learning



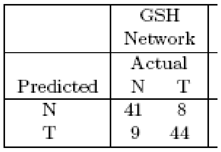
GPX	GSH-O (normal)		
	low	med	high
low	0.67±0.25	0.23±0.24	0.10±0.24
med	0.33±0.40	0.65±0.40	0.00±0.01
high	0.04±0.07	0.13±0.10	0.83±0.09
GPX	**GSH-O (tumor)**		
	low	med	high
low	0.74±0.35	0.11±0.16	0.14±0.32
med	0.68±0.34	0.09±0.13	0.23±0.27
high	0.02±0.02	0.02±0.02	0.96±0.02

Note that tumor cells produce lower than normal amounts of GSH-O when GPX levels are medium.

- We learn **separate network** parameters for normal cells and diseased cells for each metabolic process we model.
- Differences in parameters indicate differences in the underlying process.

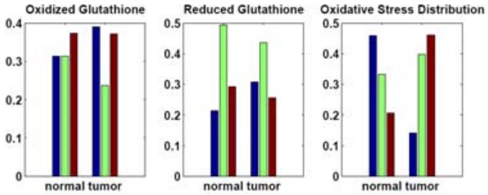
Robustness of EM learning

Leave-one-out Cross validation results for the GSH network



	GSH Network	
	Actual	
Predicted	N	T
N	41	8
T	9	44

Predictions from GSH network



We can make predictions about metabolite levels from the two learned networks. It is remarkable that we can predict that the level of oxidative stress in tumor cells is much higher in tumor cells using networks learned from the gene expression data alone!

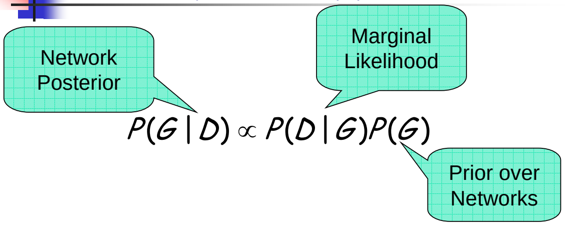
Bayesian network learning

- Computationally intensive.
- Require lots of data.
- Dynamical Bayesian networks can represent feedback loops and deal with temporal data.
- Dynamical Bayesian networks are generalizations of Hidden Markov Models!

Learning network structure

- Find the network structure that has maximum likelihood with respect to the data
  - Find G that maximizes P(G|D).

The Bayesian approach



$$P(G \mid D) \propto P(D \mid G)P(G)$$

Network Posterior

Marginal Likelihood

Prior over Networks

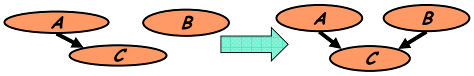
Key idea: Use *P(G|D)* to evaluate a network given a particular microarray data set.

Learning network structure

- The structure (G) learning problem is NP-hard => heuristic search for best model must be applied, generally bring out a **locally** optimal network.

- It turns out, that richer structures give higher likelihood P(D|G) to the data (adding an edge to the graph is always preferable).

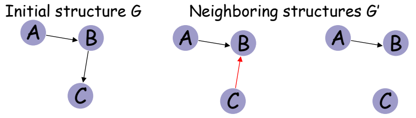
Learning structure



- If we add B to Parents(C) , we have more parameters to fit → more freedom →
- But we prefer *simpler* (more explanatory) networks (Occam's razor!)
- Therefore, **practical** scores of Bayesian Networks compensate for the likelihood improvement by imposing a penalty on complex networks.

Local search

We change one edge and evaluate the gains made by this change



Search algorithm recipe

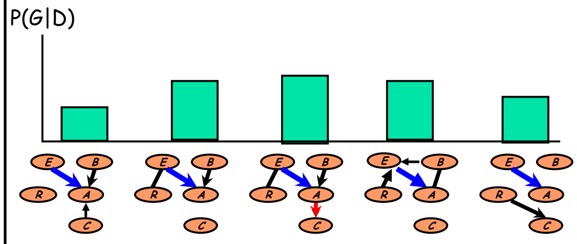
- Start with a random graph G. Evaluate its likelihood wrt D, P(G|D).
- Until little improvement in likelihood
  - Perturb structure G by adding, deleting or reversing edge
  - Accept change if likelihood improves.
- End

Randomized restarts

Difficulty #1

- We do not have enough data to uniquely identify a high-scoring network.
  - Exponentially many networks with the same P(G|data) score!
- Solution: generate many high-scoring network and extract common features.

Evaluating networks



Look for features **common to many models**

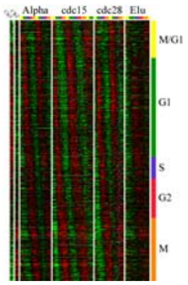
Difficulty #2

- What space of graph perturbations to consider?
- Solution: sparse candidate algorithm (Friedman 1999)
  - Limit potential parents to k most correlated variables.

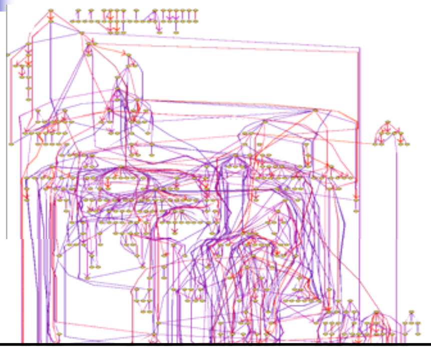
Experiment

Data from *Spellman et al.* (Mol.Bio. of the Cell 1998).

- Contains 76 samples of all the yeast genome:
  - Different methods for synchronizing cell-cycle in yeast.
  - Time series at few minutes (5-20min) intervals.
- *Spellman et al.* identified 800 cell-cycle regulated genes.



Learned network



The sparse data problem: summary

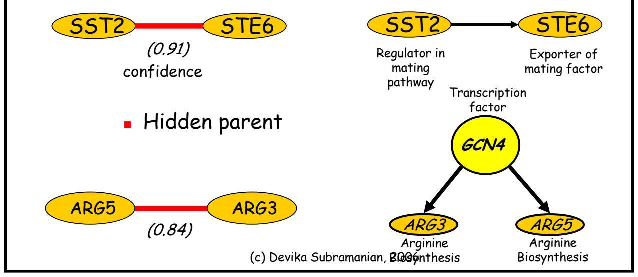
- **There are many more genes than experiments** Therefore, many different networks suit the data well.
- **Shrink the network search space**. E.g., in biological systems each gene is regulated directly by only a few regulators.
- Don't believe the learned networks, but use them to find reliable links between genes. (i.e., edges that are present in all learned networks).

Representing partial models

- Analyze the set of plausible networks and attempt to characterize features that are common to most of these networks.
- Features
  - Markov relations: Is *Y* in the Markov blanket of *X*?
  - Order relations: Is *X* an ancestor of *Y* in all the networks of a given equivalence class?

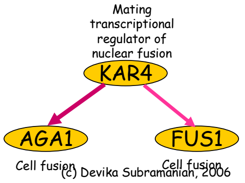
Overview of features

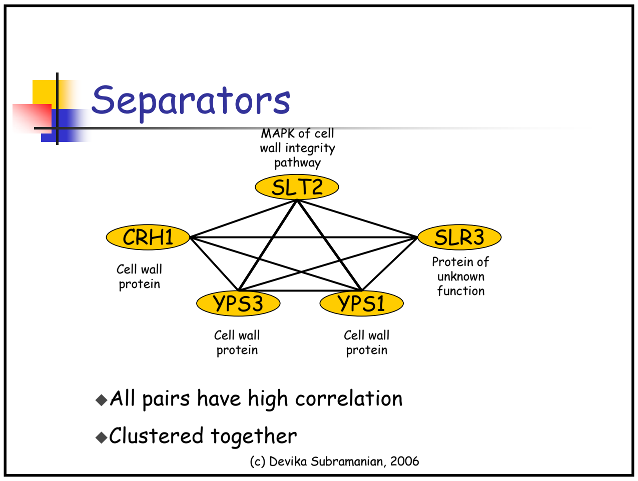
- **Question:** Do *X* and *Y* directly interact?
  - Parent-child
  - Hidden parent



Features contd.

- **Question:** Given that *X* and *Y* are indirectly dependent, who **mediates** this dependence?
- **Separator** relation:
  - *X* affects *Z* who in turn affects *Y*
  - *Z* regulates both *X* and *Y*



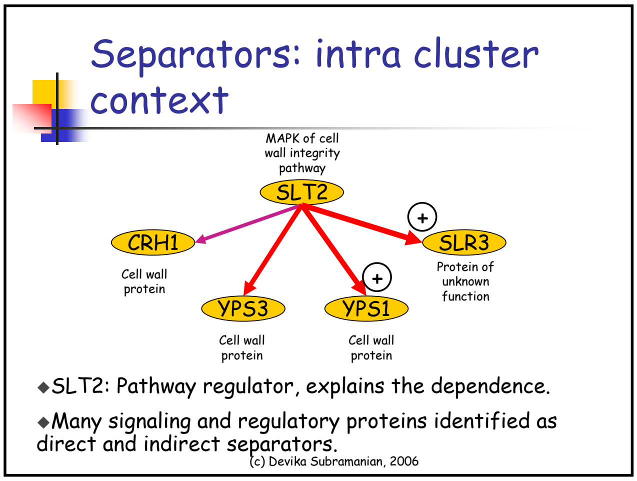


Separators

◆All pairs have high correlation

◆Clustered together

(c) Devika Subramanian, 2006

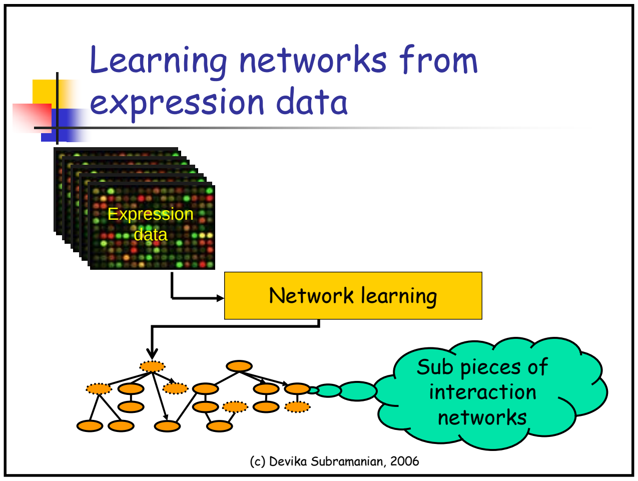


Separators: intra cluster context

◆SLT2: Pathway regulator, explains the dependence.

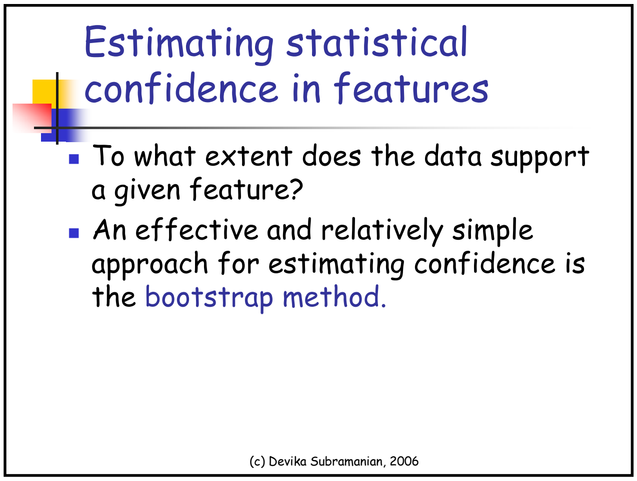
◆Many signaling and regulatory proteins identified as direct and indirect separators.

(c) Devika Subramanian, 2006



Learning networks from expression data

(c) Devika Subramanian, 2006



Estimating statistical confidence in features

- To what extent does the data support a given feature?
- An effective and relatively simple approach for estimating confidence is the bootstrap method.

(c) Devika Subramanian, 2006

The bootstrap method

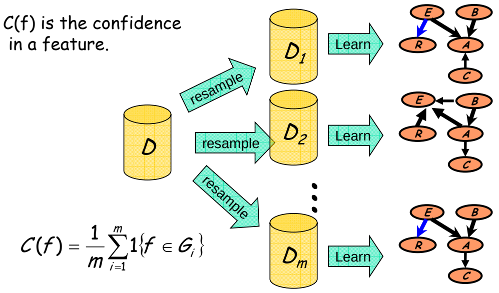
- For *i* = 1, …, *m*
  - Re-sample with replacement *N* instances from *D*. Denote by D_i the resulting dataset.
  - Apply the learning procedure on D_i to induce a network structure *G*.
- For each feature *f* of interest calculate

$$\mathrm{conf}(f) = \frac{1}{m}\sum_{i=1}^{m} f(G_i)$$

  - where *f*(*G*) is 1 if *f* is a feature in *G*, and 0 otherwise.

(c) Devika Subramanian, 2006

Bootstrap illustrated



C(f) is the confidence in a feature.

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

(c) Devika Subramanian, 2006

Improving statistical significance

Sparse Data

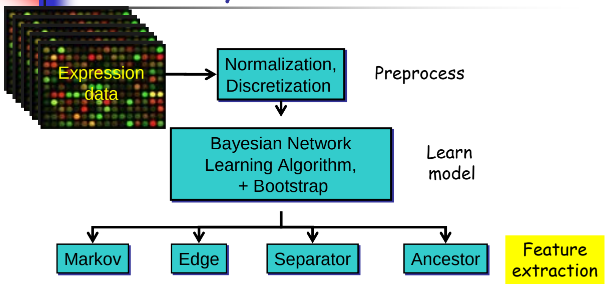
- Small number of samples
- "Flat posterior" -- many networks fit the data.

Solution

- estimate confidence in network **features**
- E.g., two types of features
  - **Markov** neighbors: *X* **directly** interacts with *Y* *(have mutual edge or a mutual child)*
  - **Order** relations: *X* is an **ancestor** of *Y*

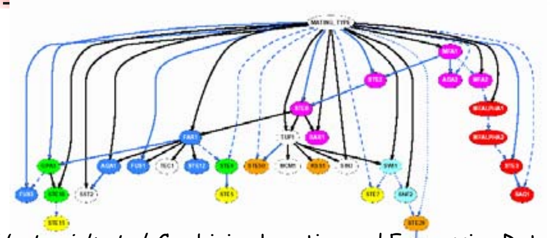
(c) Devika Subramanian, 2006

Summary of method



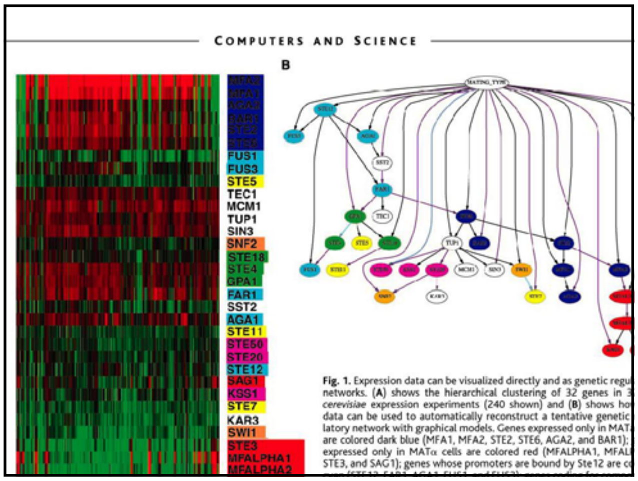
(c) Devika Subramanian, 2006

Bayesian network learned for yeast



Hartemink et al, Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models,

PSB 2002 psb.stanford.edu/psb-online



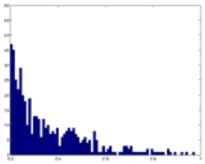
COMPUTERS AND SCIENCE

B
MATING_TYPE
FUS1
FUS3
STE5
TEC1
MCM1
TUP1
SIN3
SNF2
STE18
STE4
GPA1
FAR1
SST2
AGA1
STE11
STE50
STE20
STE12
SAG1
KSS1
STE7
KAR3
SWI1
STE3
MFALPHA1
MFALPHA2

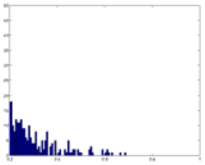
Fig. 1. Expression data can be visualized directly and as genetic regu networks. (**A**) shows the hierarchical clustering of 32 genes in 3 *cerevisiae* expression experiments (240 shown) and (**B**) shows ho data can be used to automatically reconstruct a tentative genetic latory network with graphical models. Genes expressed only in MAT are colored dark blue (MFA1, MFA2, STE2, STE6, AGA2, and BAR1); expressed only in MATα cells are colored red (MFALPHA1, MFAL STE3, and SAG1); genes whose promoters are bound by Ste12 are co

Permutation testing

- ◆Running the procedure on randomized data where the order of values for each gene is reshuffled.
- ◆Histograms of number of Markov features at each confidence level



Original Data

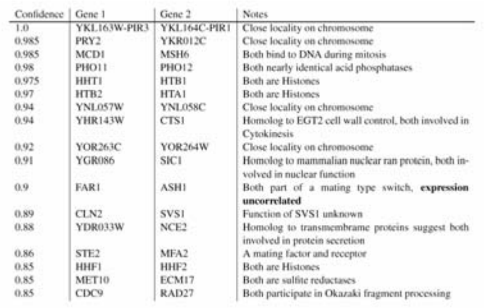


Randomized Data

Biological Analysis of order relations

Gene/ORF	Score in Experiment Multinomial	 Gaussian	Notes
MCD1	550	525	Mitotic Chromosome Determinant,null mutant is inviable
MSH6	292	508	Required for mismatch repair in mitosis and meiosis
CSI2	444	497	cell wall maintenance, chitin synthesis
CLN2	497	454	Role in cell cycle START, null mutant exhibits G1 arrest
YLR183C	551	448	Contains forkheaded associated domain, thus possibly nuclear
RFA2	456	423	Involved in nucleotide excision repair, null mutant is inviable
RSR1	352	395	GTP-binding protein of the RAS family involved in bud site selection
CDC45	-	394	Required for initiation of chromosomal replication, null mutant lethal
RAD53	60	383	Cell cycle control, checkpoint function, null mutant lethal
CDC5	209	353	Cell cycle control, required for exit from mitosis, null mutant lethal
POL30	376	321	Required for DNA replication and repair, null mutant is inviable
YOX1	400	291	Homeodomain protein
SRO4	463	239	Involved in cellular polarization during budding
CLN1	324	-	Role in cell cycle START, null mutant exhibits G1 arrest
YBR089W	298	-	

Biological Analysis of Markov relations

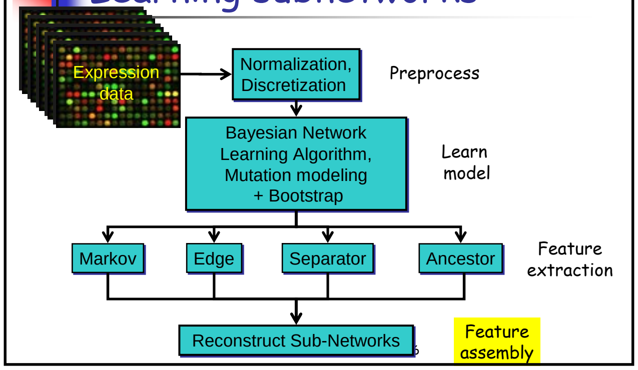


Confidence	Gene 1	Gene 2	Notes
1.0	YKL163W-PIR3	YKL164C-PIR1	Close locality on chromosome
0.985	PRY2	YKR012C	Close locality on chromosome
0.985	MCD1	MSH6	Both bind to DNA during mitosis
0.98	PHO11	PHO12	Both nearly identical acid phosphatases
0.975	HHT1	HTB1	Both are Histones
0.97	HTB2	HTA1	Both are Histones
0.94	YNL057W	YNL058C	Close locality on chromosome
0.94	YHR143W	CTS1	Homolog to EGT2 cell wall control, both involved in Cytokinesis
0.92	YOR263C	YOR264W	Close locality on chromosome
0.91	YGR086	SIC1	Homolog to mammalian nuclear ran protein, both involved in nuclear function
0.9	FAR1	ASH1	Both part of a mating type switch, **expression uncorrelated**
0.89	CLN2	SVS1	Function of SVS1 unknown
0.88	YDR033W	NCE2	Homolog to transmembrane proteins suggest both involved in protein secretion
0.86	STE2	MFA2	A mating factor and receptor
0.85	HHF1	HHF2	Both are Histones
0.85	MET10	ECM17	Both are sulfite reductases
0.85	CDC9	RAD27	Both participate in Okazaki fragment processing

Assembling subnetworks

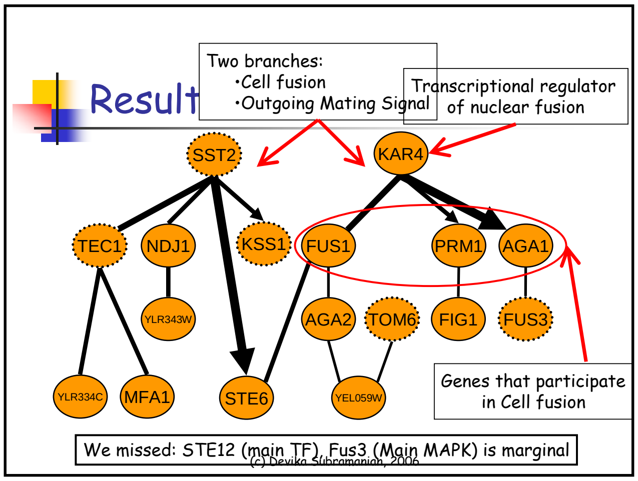
- Automatic reconstruction
  - **Goal:** Dense sub-network with highly confident pair-wise features
  - **Score**: Statistical significance
  - **Search**: High scoring sub-networks
- Advantages
  - Global picture
  - Structured context for interactions
  - Incorporate mid-confidence features

Learning subnetworks



Results

- 6 well structured sub-networks representing coherent molecular responses
  - Mating
  - Iron metabolism
  - Low osmolarity cell wall integrity pathway
  - Stationary phase and stress response
  - Amino acid metabolism, mitochondrial function and sulfate assimilation
  - Citrate metabolism
- Uncovered regulatory, signaling and metabolic interactions



Result

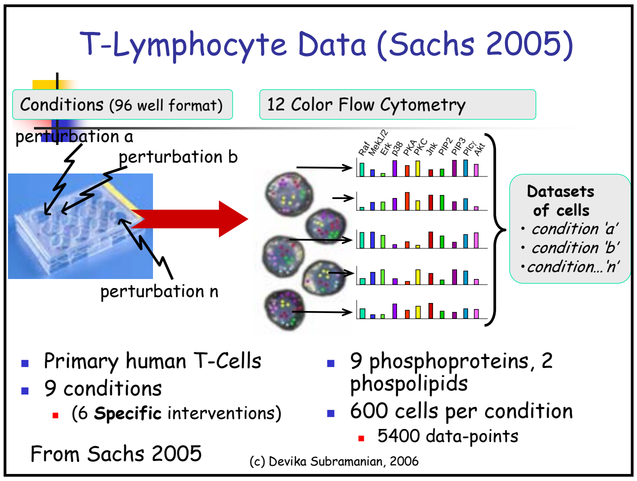
Two branches:
- Cell fusion
- Outgoing Mating Signal

Transcriptional regulator of nuclear fusion

Genes that participate in Cell fusion

We missed: STE12 (main TF), Fus3 (Main MAPK) is marginal

(c) Devika Subramanian, 2006



T-Lymphocyte Data (Sachs 2005)

Conditions (96 well format)

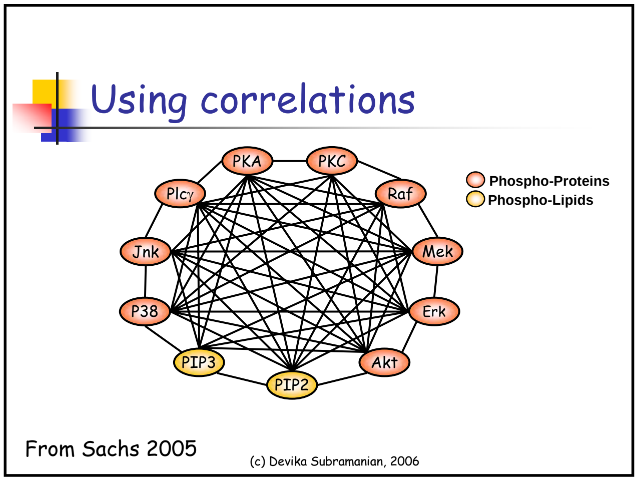
12 Color Flow Cytometry

Datasets of cells
- *condition 'a'*
- *condition 'b'*
- *condition…'n'*

- Primary human T-Cells
- 9 conditions
  - (6 **Specific** interventions)
- 9 phosphoproteins, 2 phospholipids
- 600 cells per condition
  - 5400 data-points

From Sachs 2005

(c) Devika Subramanian, 2006

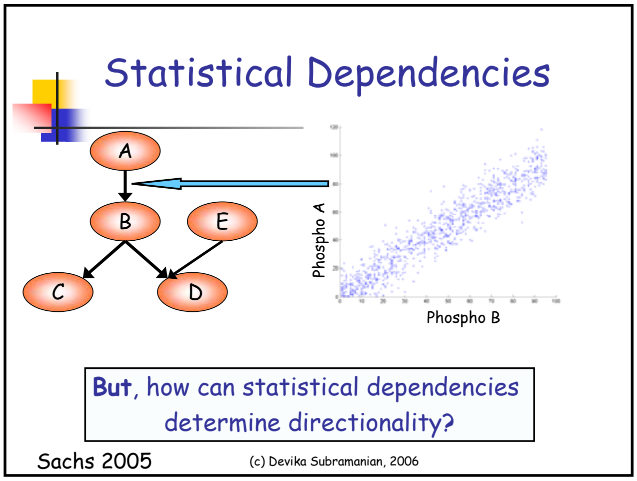


Using correlations

● Phospho-Proteins
○ Phospho-Lipids

From Sachs 2005

(c) Devika Subramanian, 2006

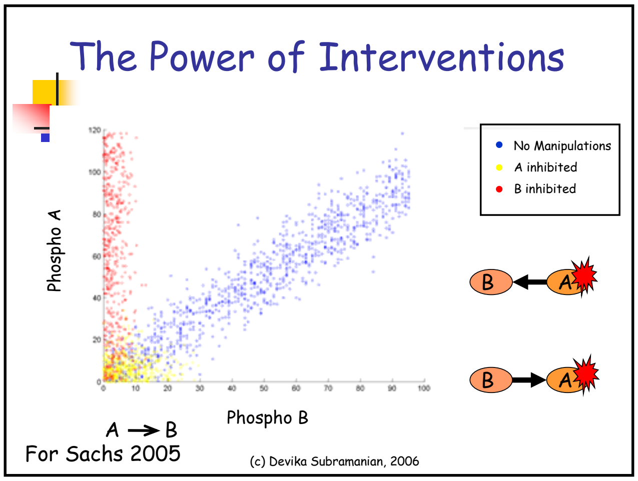


Statistical Dependencies

But, how can statistical dependencies determine directionality?

Sachs 2005

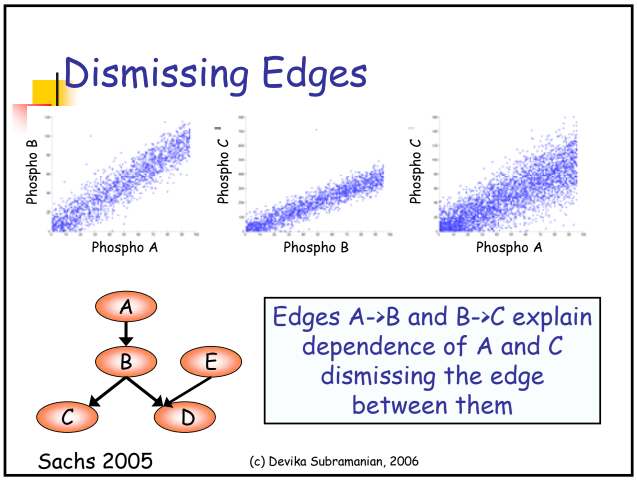
(c) Devika Subramanian, 2006



The Power of Interventions

A → B
For Sachs 2005

(c) Devika Subramanian, 2006

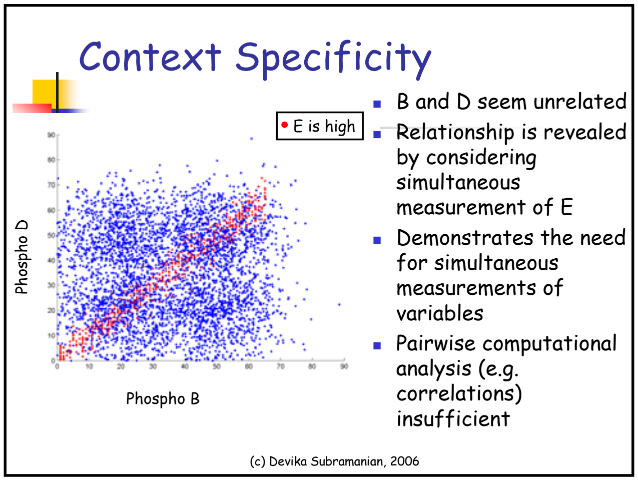


Dismissing Edges

Edges A->B and B->C explain dependence of A and C dismissing the edge between them

Sachs 2005

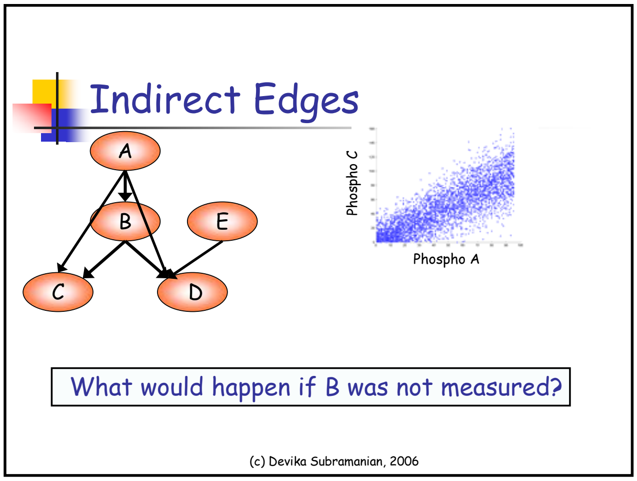
(c) Devika Subramanian, 2006



Context Specificity

- B and D seem unrelated
- Relationship is revealed by considering simultaneous measurement of E
- Demonstrates the need for simultaneous measurements of variables
- Pairwise computational analysis (e.g. correlations) insufficient

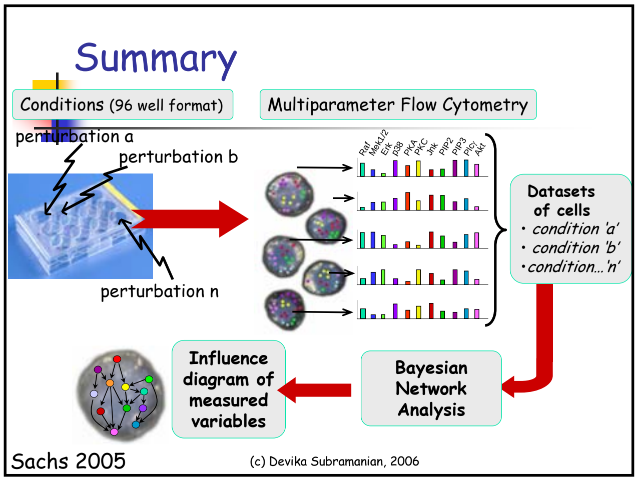
(c) Devika Subramanian, 2006



Indirect Edges

What would happen if B was not measured?

(c) Devika Subramanian, 2006



Summary

Conditions (96 well format)

Multiparameter Flow Cytometry

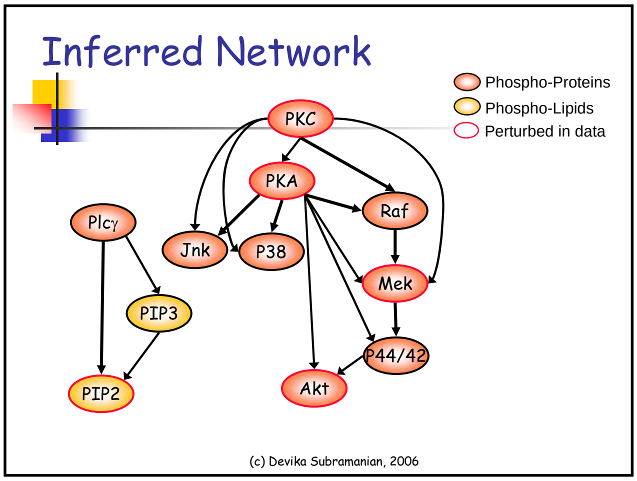
Datasets of cells
• *condition 'a'*
• *condition 'b'*
• *condition…'n'*

Bayesian Network Analysis

Influence diagram of measured variables

Sachs 2005

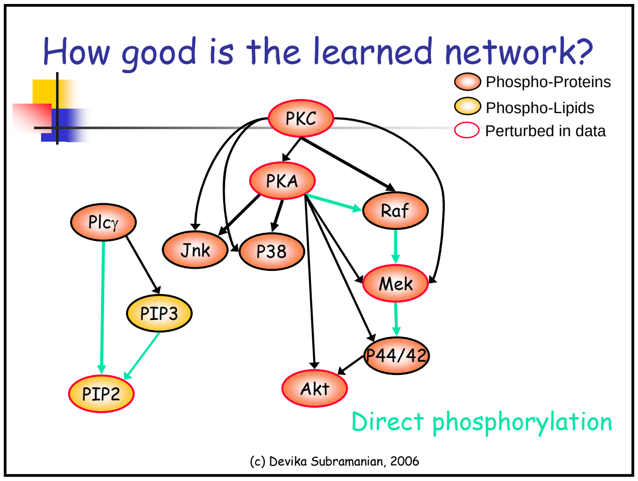
(c) Devika Subramanian, 2006



Inferred Network

○ Phospho-Proteins
○ Phospho-Lipids
○ Perturbed in data

(c) Devika Subramanian, 2006

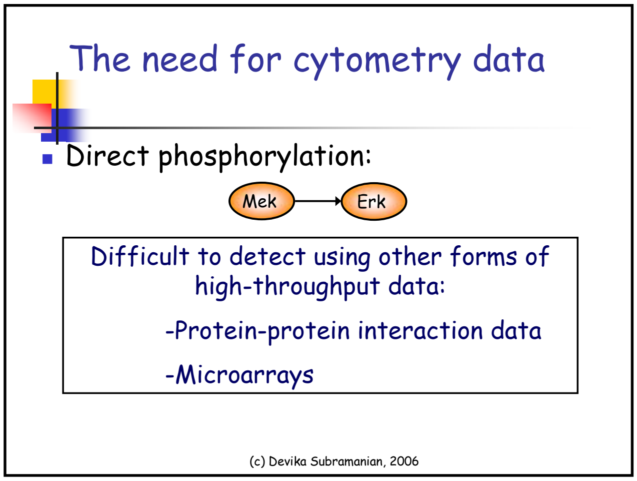


How good is the learned network?

○ Phospho-Proteins
○ Phospho-Lipids
○ Perturbed in data

Direct phosphorylation

(c) Devika Subramanian, 2006



The need for cytometry data

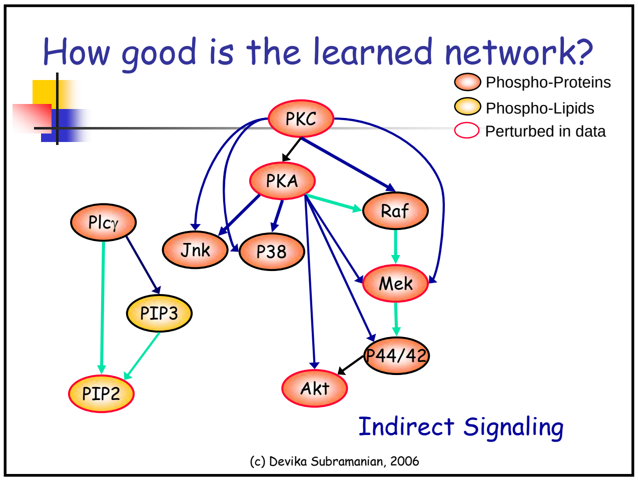
- Direct phosphorylation:

Difficult to detect using other forms of high-throughput data:

-Protein-protein interaction data

-Microarrays

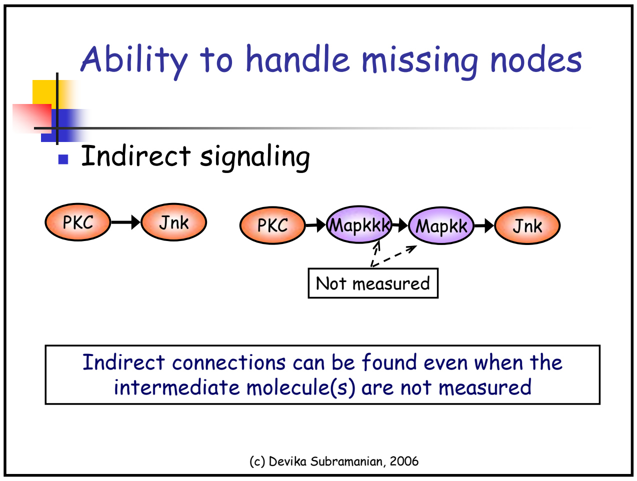
(c) Devika Subramanian, 2006



How good is the learned network?

Indirect Signaling

(c) Devika Subramanian, 2006

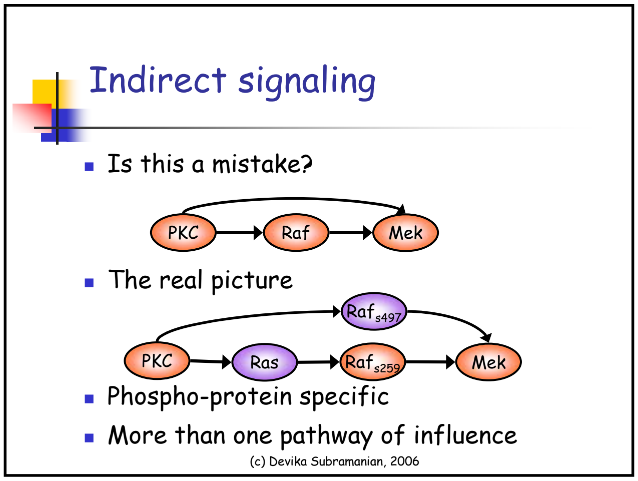


Ability to handle missing nodes

- Indirect signaling

Indirect connections can be found even when the intermediate molecule(s) are not measured

(c) Devika Subramanian, 2006



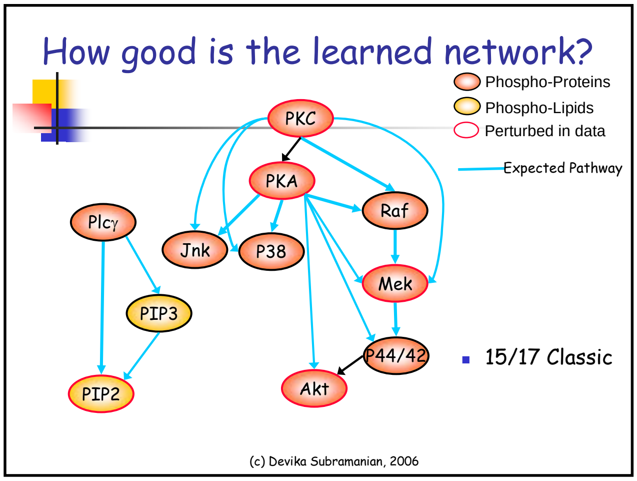
Indirect signaling

- Is this a mistake?

- The real picture

- Phospho-protein specific
- More than one pathway of influence

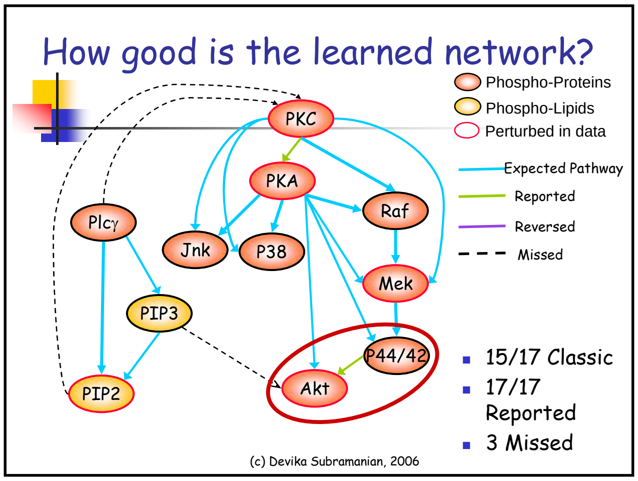
(c) Devika Subramanian, 2006



How good is the learned network?

- 15/17 Classic

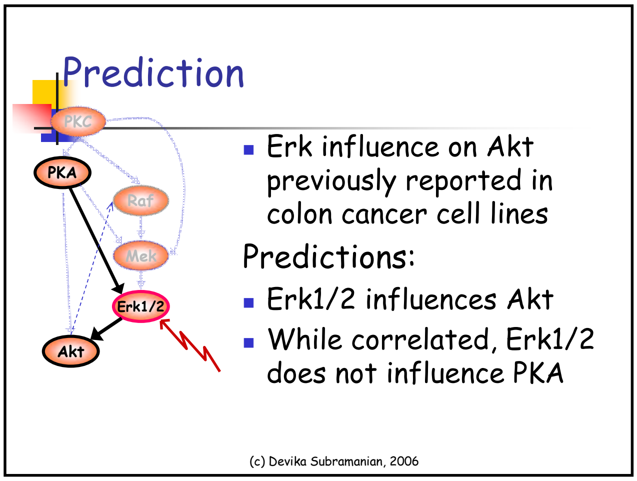
(c) Devika Subramanian, 2006



How good is the learned network?

- 15/17 Classic
- 17/17 Reported
- 3 Missed

(c) Devika Subramanian, 2006



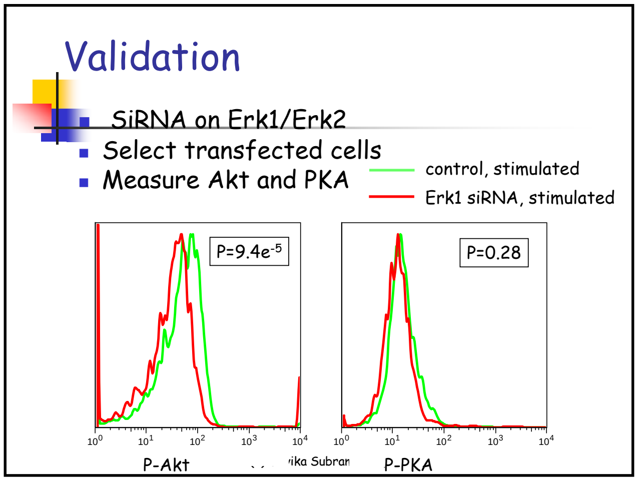
Prediction

- Erk influence on Akt previously reported in colon cancer cell lines

Predictions:

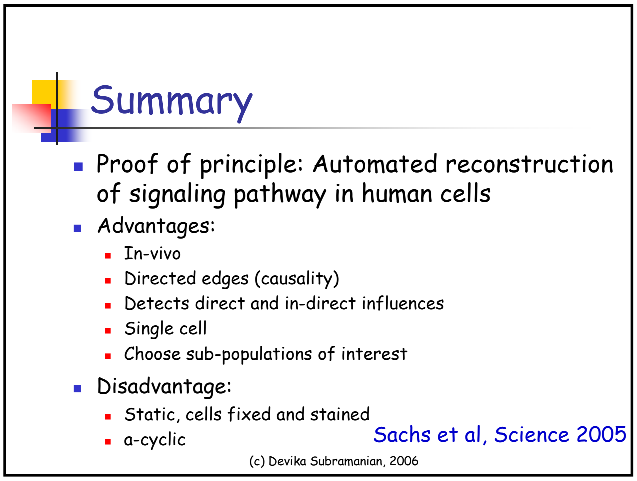
- Erk1/2 influences Akt
- While correlated, Erk1/2 does not influence PKA

(c) Devika Subramanian, 2006



Validation

- SiRNA on Erk1/Erk2
- Select transfected cells
- Measure Akt and PKA



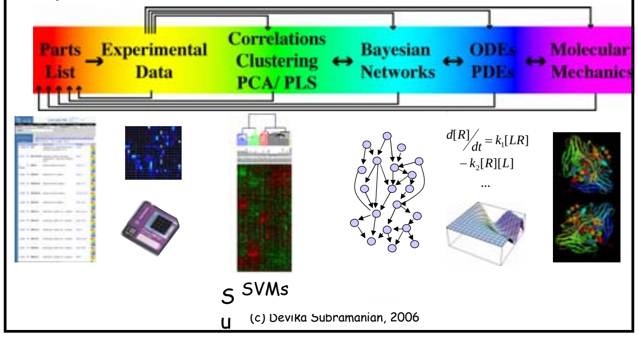
Summary

- Proof of principle: Automated reconstruction of signaling pathway in human cells
- Advantages:
  - In-vivo
  - Directed edges (causality)
  - Detects direct and in-direct influences
  - Single cell
  - Choose sub-populations of interest
- Disadvantage:
  - Static, cells fixed and stained
  - a-cyclic

Sachs et al, Science 2005

(c) Devika Subramanian, 2006

Spectrum of modeling tools in systems biology



$$\frac{d[R]}{dt} = k_1[LR] - k_2[R][L] \ldots$$

S SVMs

u