

Inferring regulatory, signaling & metabolic networks from data

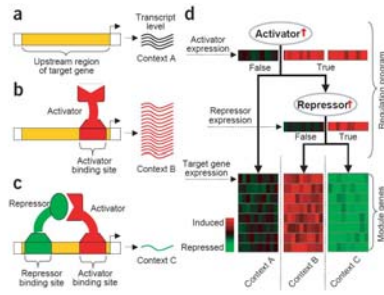
Devika Subramanian
Comp 470

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.

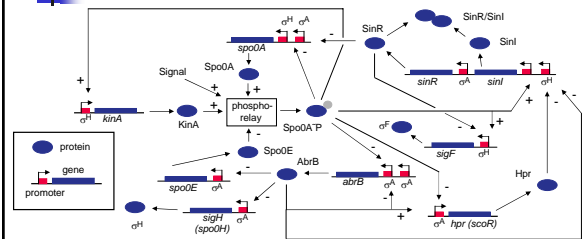
(c) Devika Subramanian, 2006

Regulators



(c) Devika Subramanian, 2006

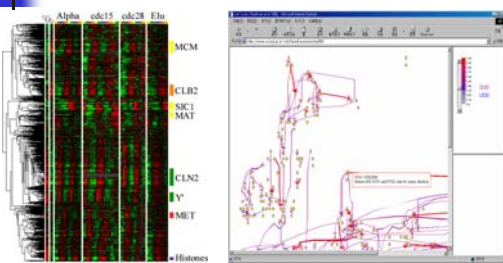
Genetic regulatory network of *B. subtilis*



Genetic regulatory network controlling the initiation of sporulation.

(c) Devika Subramanian, 2006

From expression data to gene regulatory networks

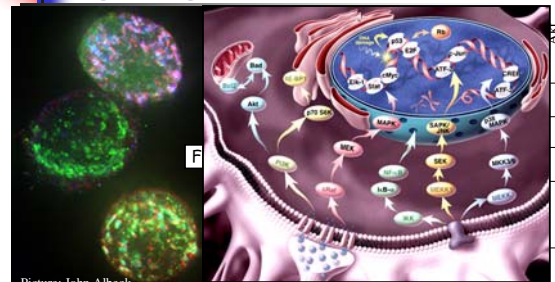


Microarray data

Yeast cell cycle

(c) Devika Subramanian, 2006

From flow cytometry data to signaling networks



K. Sachs, 2005

High throughput data signaling Pathways

(c) Devika Subramanian, 2006

Some initial approaches

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

(c) Devika Subramanian, 2006

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!

(c) Devika Subramanian, 2006

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
 - $P(\text{disease/problem}|\text{set of symptoms})$
 - $P(\text{email is spam}|\text{email text+header})$
 - $P(\text{hurricane will hit place X}|\text{movement history})$
 - $P(\text{sentence}|\text{acoustic signal})$
 - $P(\text{regulatory network}|\text{gene exp data})$

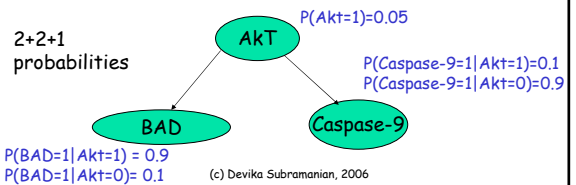
(c) Devika Subramanian, 2006

Example: Akt pathway

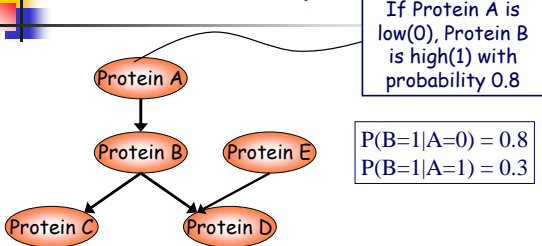
Random variables: Akt, BAD, caspase-9

Conditional independencies:

$$P(\text{BAD and caspase-9}|\text{AKT}) = P(\text{BAD}|\text{AKT})P(\text{Caspase-9}|\text{AKT})$$



Another example



Adapted from Sachs, 2005

(c) Devika Subramanian, 2006

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 $P(X|\text{Parents}(X))$ associated with it.

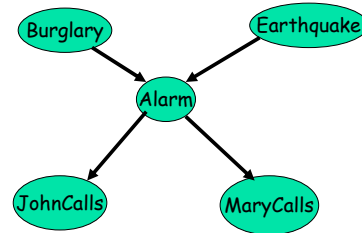
(c) Devika Subramanian, 2006

Segue

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

(c) Devika Subramanian, 2006

A simple Bayesian network



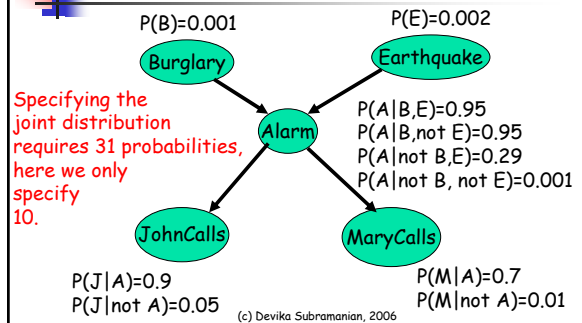
(c) Devika Subramanian, 2006

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.

(c) Devika Subramanian, 2006

Bayesian network with CPTs



(c) Devika Subramanian, 2006

Computing joint probability distributions

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

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

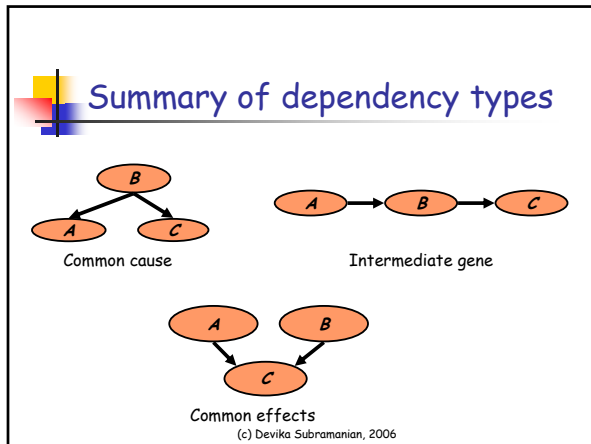
(c) Devika Subramanian, 2006

Computing joint probabilities

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

$$\begin{aligned}
 P(\text{Burglary} | \text{Alarm}) &= 0.376 \\
 P(\text{Burglary} | \text{Alarm}, \text{Earthquake}) &= 0.003
 \end{aligned}$$

(c) Devika Subramanian, 2006



feedforward loop

single input module (SIM)

dense overlapping regulations (DOR)

Little is known about the design principles¹⁻¹⁰ of transcriptional regulation networks that control gene expression in cells. Recent advances in data collection and analysis^{11,12}, however, are generating unprecedented amounts of information about gene regulation networks. To understand these complex wiring diagrams¹³⁻¹⁵, we sought to break down such networks into basic building blocks. We generated the notion of motifs, widely used for sequence analysis, to the level of networks. We define 'network motifs' as patterns of interactions that occur 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 *E. coli* (see ref. 16). 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 spectrum 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 428 operons (involving 116 transcription factors); it was formed on the basis of an existing database (RegulonDB^{17,18}). We enhanced RegulonDB by an extensive literature search, adding 33 new transcription factors, including alternative σ -factor (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 regulator site.

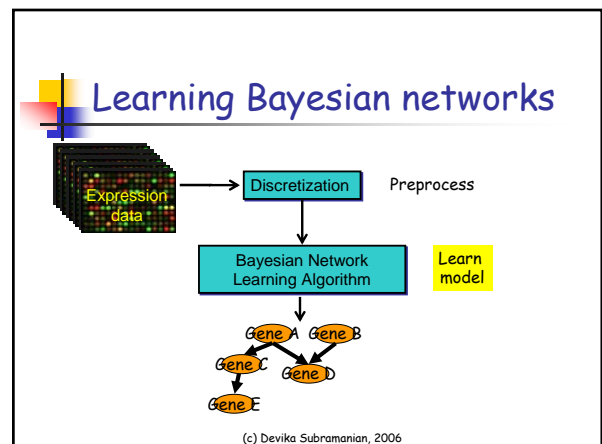
The transcriptional network can be represented as a directed graph, in which each node represents an operon and edges represent direct transcriptional interactions. Each edge is directed

Fig. 1 Network motifs found in the *E. coli* transcriptional regulation network. Operons representing the motifs are also shown. a, Feedforward loop: transcription factor A regulates a second transcription factor B and both jointly regulate the operon Z. b, Example of a feedforward loop in *Salmonella enteritidis*. c, SIM motif: a single transcription factor, X, regulates a set of operons Z₁, Z₂, Z₃, Z₄, ..., Z_m. d, Example of a SIM motif in *E. coli*. e, A densely overlapping set of operons Z₁, Z₂, Z₃, Z₄, ..., Z_m is regulated by a single transcription factor X.

- ## Conditional probability distributions
- Multinomial model
 - Discrete values
 - Linear Gaussian model
 - $P(X|u_1, u_2, \dots, u_k) = N(a_0 + \sum_i a_i u_i, \sigma^2)$
- (c) Devika Subramanian, 2006

- ## 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
- (c) Devika Subramanian, 2006

- ## 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.
- (c) Devika Subramanian, 2006



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.

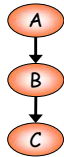
(c) Devika Subramanian, 2006

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.

(c) Devika Subramanian, 2006

Learning CPTs

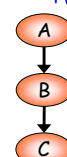


Known structure!

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
Off	Off	Off

(c) Devika Subramanian, 2006

Learning CPTs



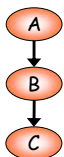
$$P(B='On'|A='On') = 0.83$$

$$5/6 = 0.83$$

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
Off	Off	Off

(c) Devika Subramanian, 2006

Learning CPTs



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

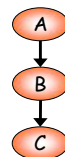
$$P(B='Off'|A='Off') = 0.8$$

$$4/5 = 0.8$$

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
Off	Off	Off

(c) Devika Subramanian, 2006

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
Off	Off	Off

(c) Devika Subramanian, 2006

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
Off	Off	Off
Off	On	On
Off	Off	Off
Off	Off	Off
Off	Off	Off

(c) Devika Subramanian, 2006

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.

(c) Devika Subramanian, 2006

Modeling cellular processes: topology of glutathione network

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.

(c) Devika Subramanian, 2006

Modeling cellular processes: the quantitative parameters

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

(c) Devika Sut

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.
 - $P(Y|X) = \text{count}(X \& Y) / \text{count}(Y)$
- 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.

(c) Devika Subramanian, 2006

Component network learning

A portion of the GSH network

- 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.

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.

(c) Devika Subramanian, 2006

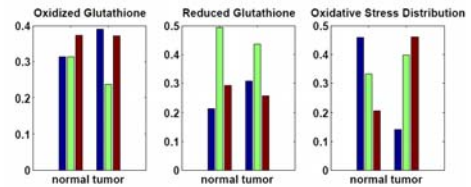
Robustness of EM learning

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

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

(c) Devika Subramanian, 2006

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!

(c) Devika Subramanian, 2006

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!

(c) Devika Subramanian, 2006

Learning network structure

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

(c) Devika Subramanian, 2006

The Bayesian approach

Network Posterior

Marginal Likelihood

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

Prior over Networks

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

(c) Devika Subramanian, 2006

Learning network structure

- The structure (G) learning problem is NP-hard \Rightarrow 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).

(c) Devika Subramanian, 2006

Learning structure

- If we add B to Parents(C), we have more parameters to fit \rightarrow more freedom \rightarrow
- 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.

(c) Devika Subramanian, 2006

Local search

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

(c) Devika Subramanian, 2006

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

(c) Devika Subramanian, 2006

Difficulty #1

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

(c) Devika Subramanian, 2006

Evaluating networks

Look for features **common to many models**

(c) Devika Subramanian, 2006

Difficulty #2

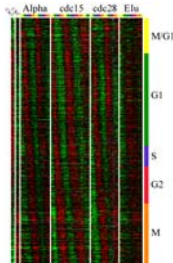
- What space of graph perturbations to consider?
- Solution: sparse candidate algorithm (Friedman 1999)
 - Limit potential parents to k most correlated variables.

(c) Devika Subramanian, 2006

Experiment

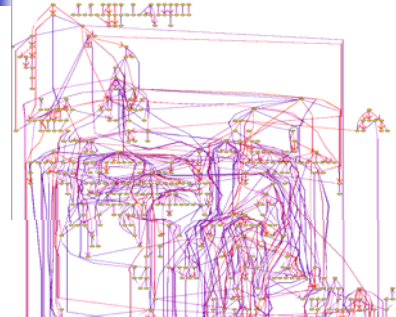
Data from *Spellman et al.* (Mol. Bio. of the Cell 1998).

- 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.



(c) Devika Subramanian, 2006

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).

(c) Devika Subramanian, 2006

Representing partial models

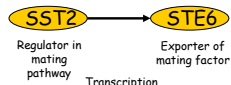
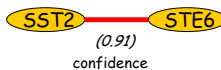
- Analyze the set of plausible networks and attempt to characterize features that are common to most of these networks.
- Features
 - Markov relations: Is Y in the Markov blanket of X ?
 - Order relations: Is X an ancestor of Y in all the networks of a given equivalence class?

(c) Devika Subramanian, 2006

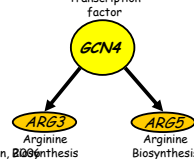
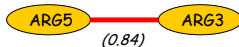
Overview of features

- Question: Do X and Y directly interact?

- Parent-child



- Hidden parent



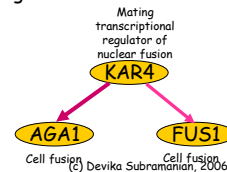
(c) Devika Subramanian, 2006

Features contd.

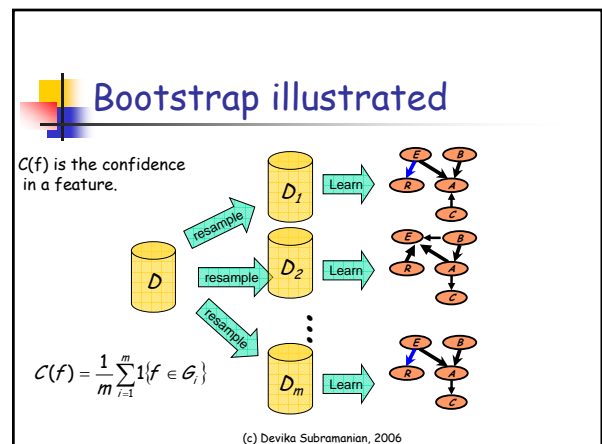
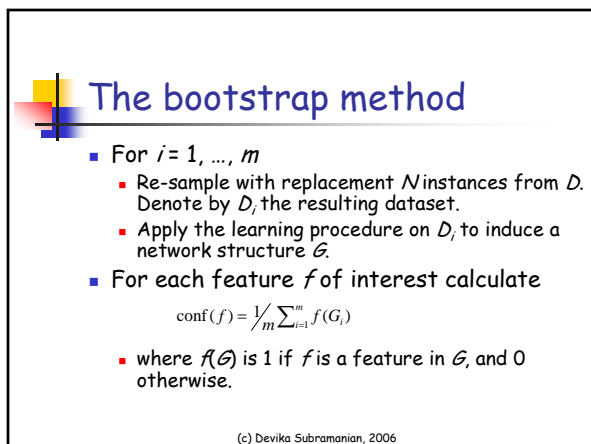
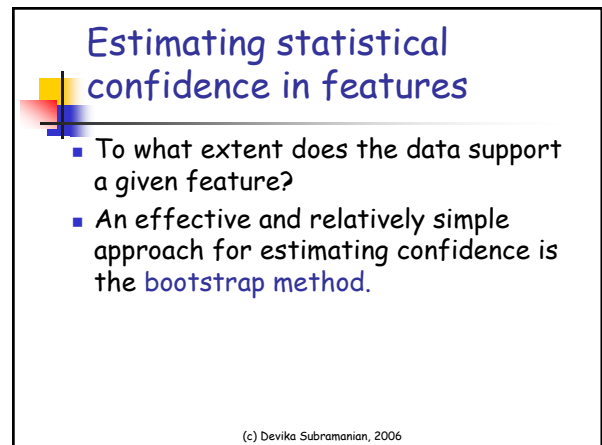
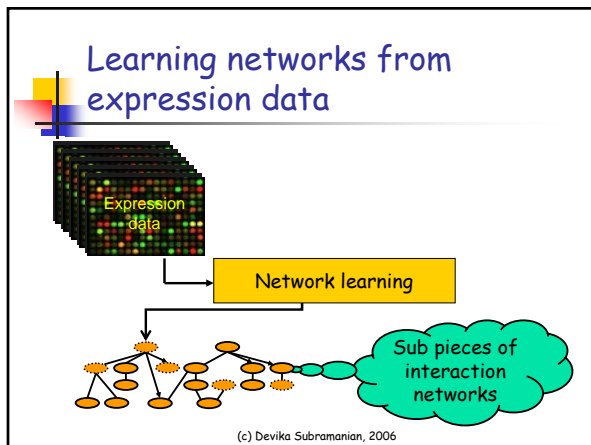
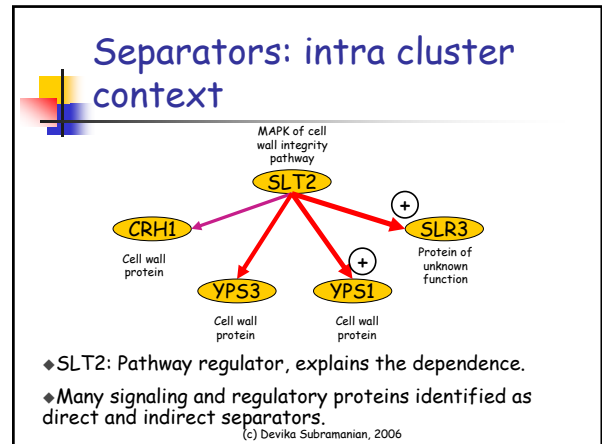
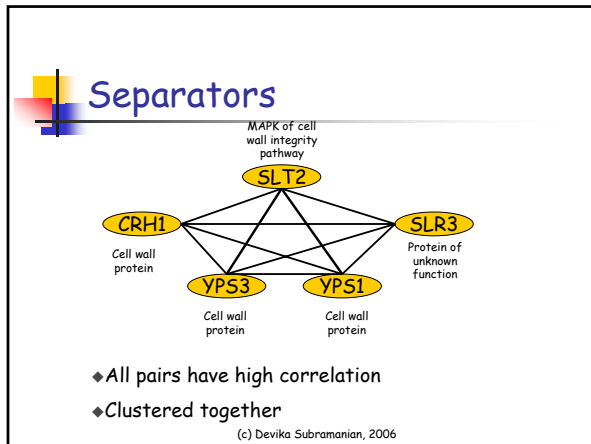
- Question: Given that X and Y are indirectly dependent, who mediates this dependence?

- Separator relation:

- X affects Z who in turn affects Y
- Z regulates both X and Y



(c) Devika Subramanian, 2006



Improving statistical significance

Sparse Data

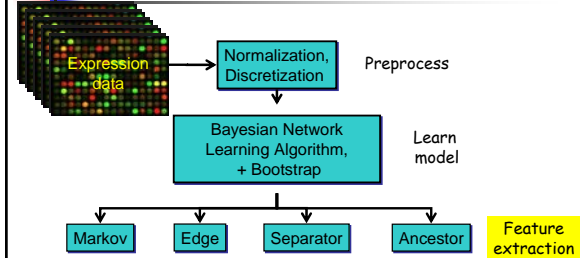
- Small number of samples
- "Flat posterior" -- many networks fit the data.

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

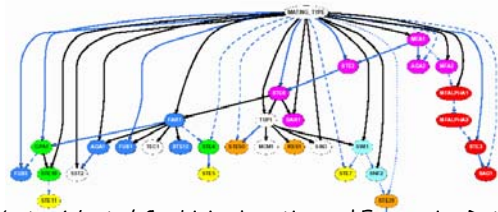
(c) Devika Subramanian, 2006

Summary of method



(c) Devika Subramanian, 2006

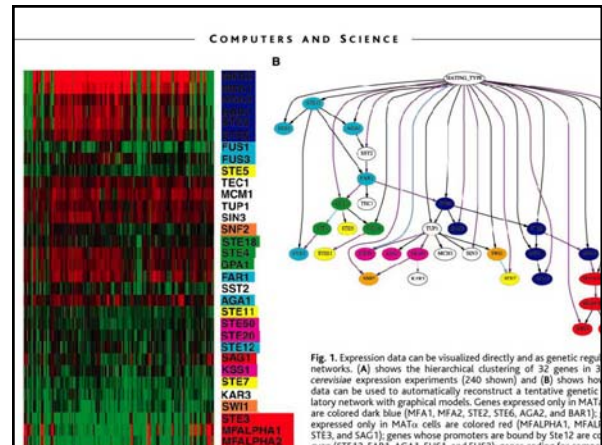
Bayesian network learned for yeast



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

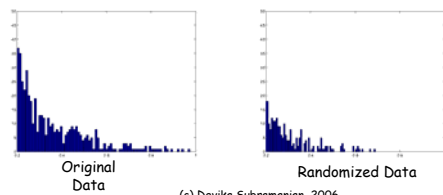
PSB 2002 psb.stanford.edu/psb-online

(c) Devika Subramanian, 2006



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



(c) Devika Subramanian, 2006

Biological Analysis of order relations

Gene/ORF	Score in Experiment		Notes
	Multinomial	Gaussian	
MCD1	550	525	Mitotic Chromosome Determinant, null mutant is inviable
MSH6	292	508	Required for mismatch repair in mitosis and meiosis
CS2	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
MFA2	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	Homologous domain protein
SR04	463	239	Involved in cellular polarization during budding
CLN1	324	-	Role in cell cycle START, null mutant exhibits G1 arrest
YBR089W	298	-	

(c) Devika Subramanian, 2006

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	HBT1	HTR1	Both are Histones
0.97	HTR2	HTR1	Both are Histones
0.94	YNL057W	YNL058C	Close locality on chromosome
0.94	YBR143W	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	HBF1	HBF2	Both are Histones
0.85	MET10	ECM17	Both are sulfite reductases
0.85	CDC9	RAD27	Both participate in Okazaki fragment processing

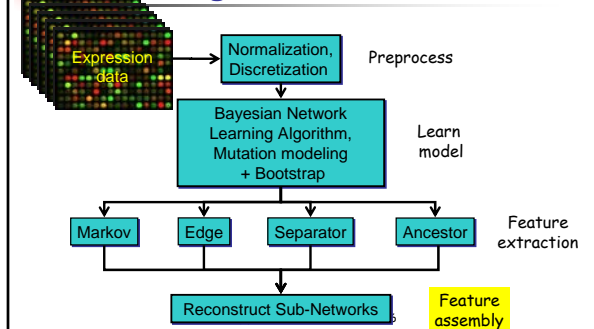
(c) Devika Subramanian, 2006

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

(c) Devika Subramanian, 2006

Learning subnetworks

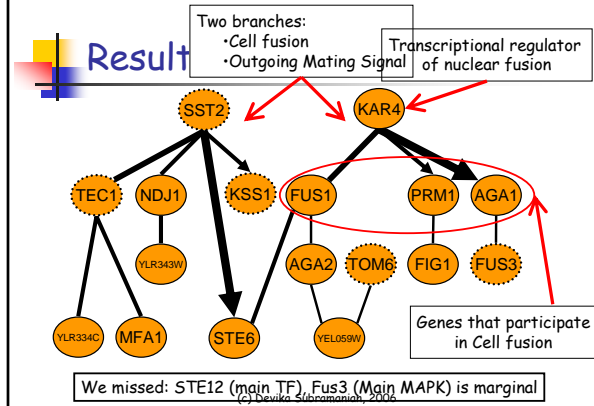


Results

- 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

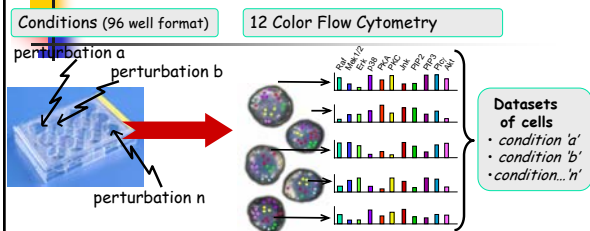
(c) Devika Subramanian, 2006

Result



(c) Devika Subramanian, 2006

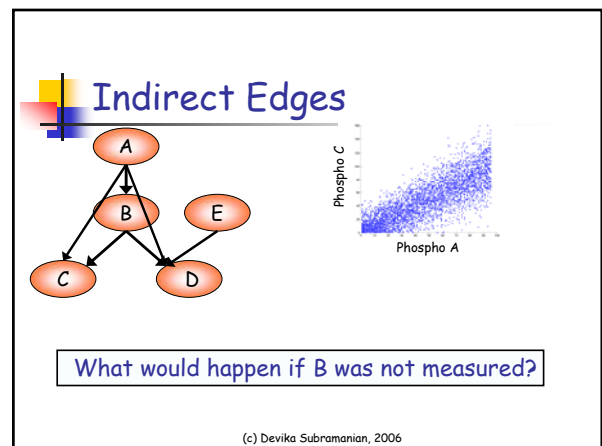
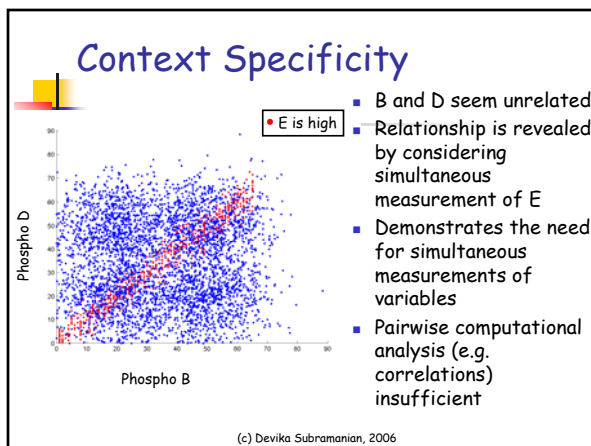
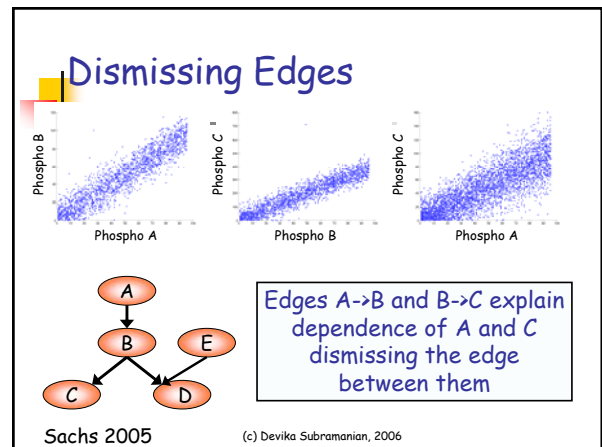
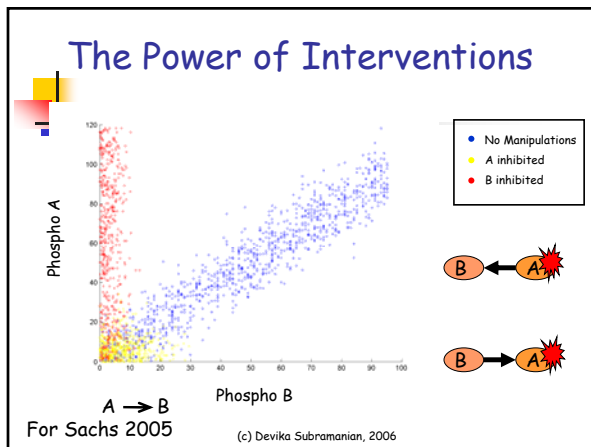
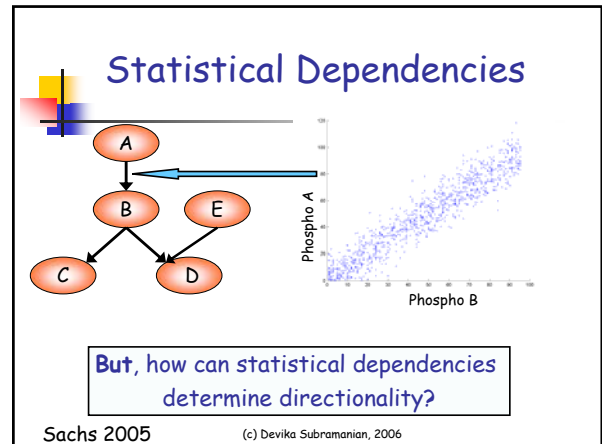
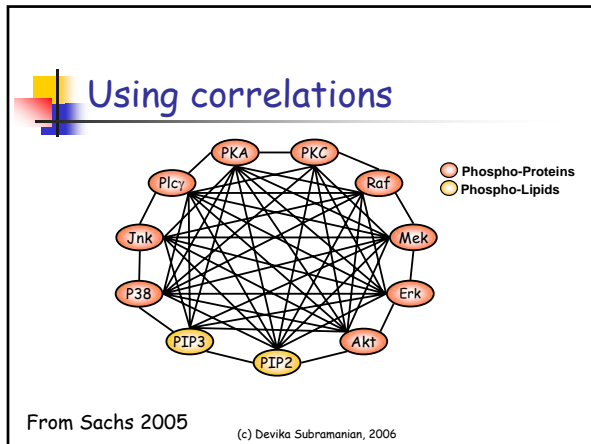
T-Lymphocyte Data (Sachs 2005)

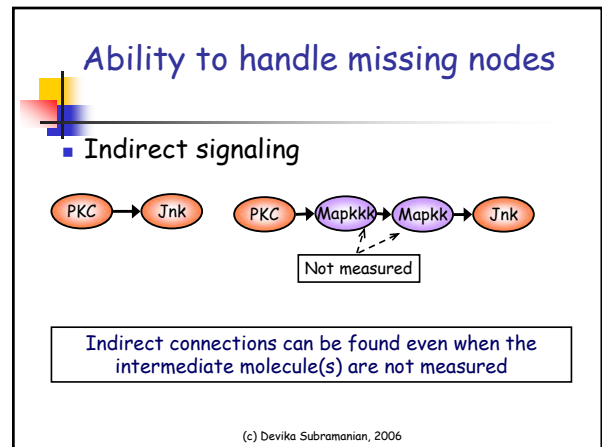
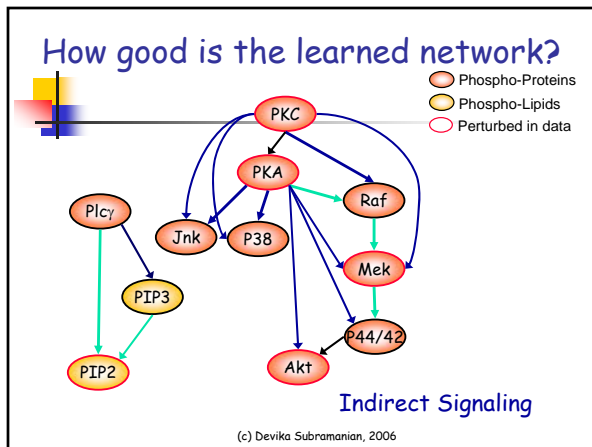
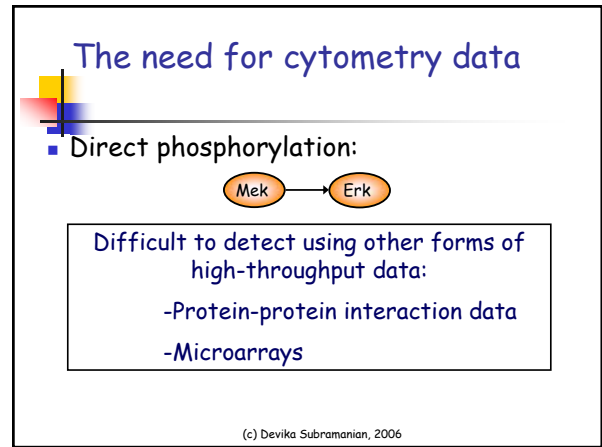
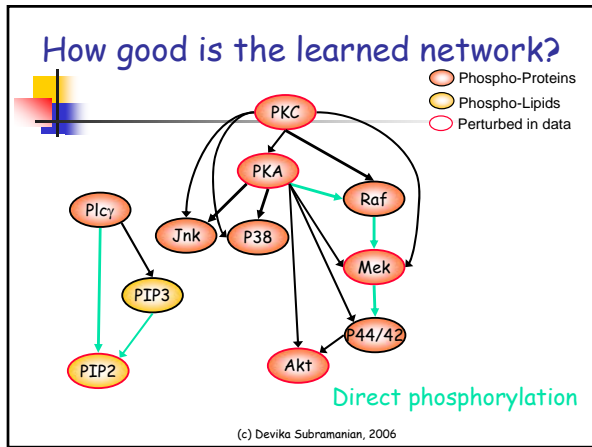
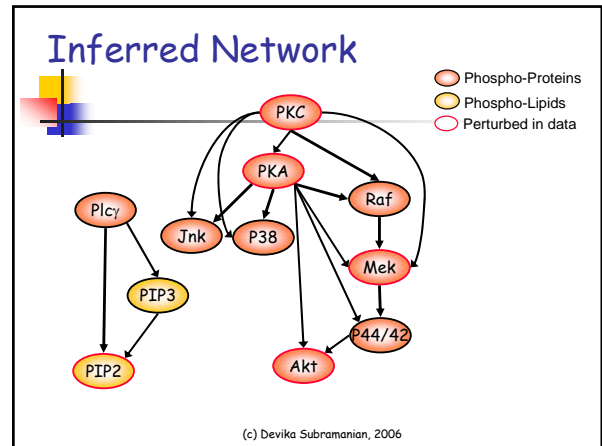
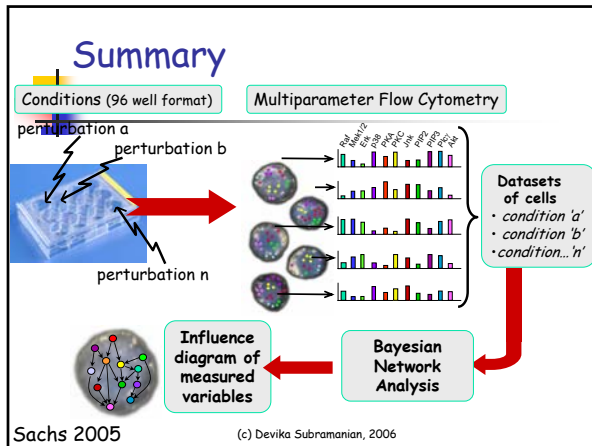


- Primary human T-Cells
- 9 conditions
 - (6 Specific interventions)
- 9 phosphoproteins, 2 phospholipids
- 600 cells per condition
 - 5400 data-points

From Sachs 2005

(c) Devika Subramanian, 2006





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?

- Phospho-Proteins
- Phospho-Lipids
- Perturbed in data
- Expected Pathway

15/17 Classic

(c) Devika Subramanian, 2006

How good is the learned network?

- Phospho-Proteins
- Phospho-Lipids
- Perturbed in data
- Expected Pathway
- Reported
- Reversed
- Missed

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

— control, stimulated
— Erk1 siRNA, stimulated

P-Akt P-PKA

(c) Devika Subramanian, 2006

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

