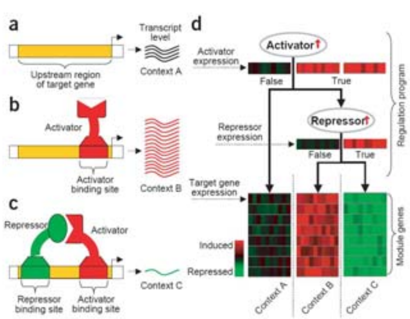
Inferring regulatory, signaling & metabolic networks from data

Devika Subramanian
Comp 470

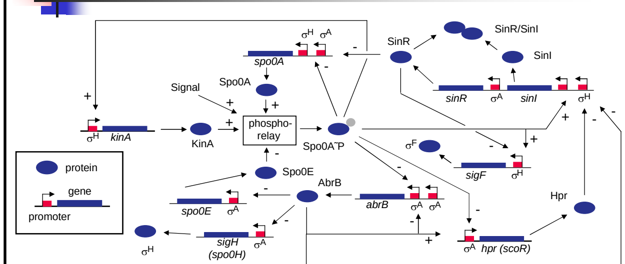
Networks

- 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

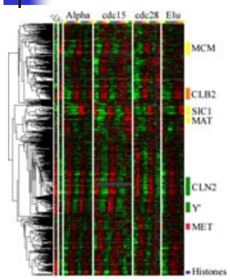


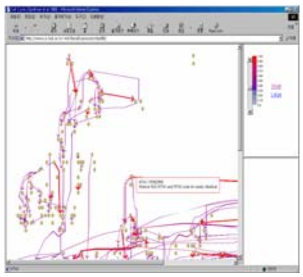
Genetic regulatory network of *B. subtilis*



Genetic regulatory network controlling the initiation of sporulation.

From expression data to gene regulatory networks

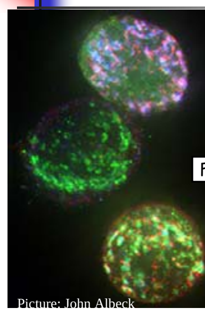




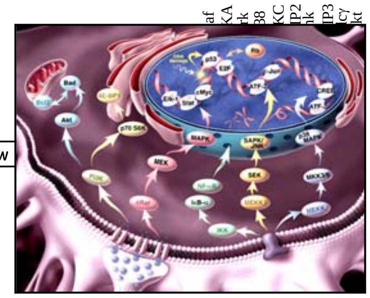
Microarray data

Yeast cell cycle

From flow cytometry data to signaling networks



Flow



K. Sachs, 2005

High throughput data

Signaling Pathways

Outline

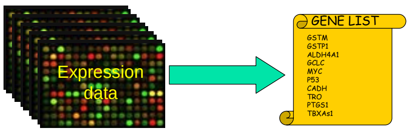
- The problem of learning regulatory, signaling and metabolic networks from data
- A quick intro to Bayesian networks
- Algorithms for learning Bayesian networks from data
- Examples
  - Glutathione metabolism from humans (expression data)
  - Regulatory network from yeast cell cycle (expression data)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

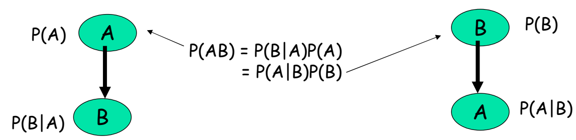
- The important science/technology to come out of AI in the last 15 years.
- Underlies all important applications today.
- Frames every question as the estimation of a conditional probability
  - P(disease/problem|set of symptoms)
  - P(email is spam|email text+header)
  - P(hurricane will hit place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

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.

An example

- A Bayesian network is a compact representation of the joint distribution over a set of random variables.
  - $P(X_1,X_2,\ldots,X_n)$

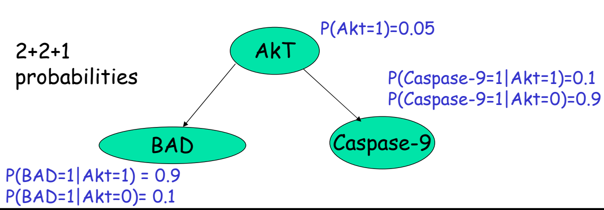


Example: Akt pathway

Random variables: Akt, BAD, caspase-9

Conditional independencies:

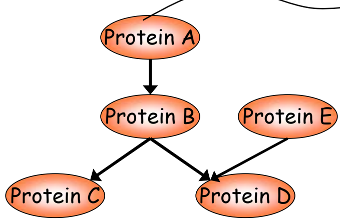
P(BAD and caspase-9|AKT) =P(BAD|Akt)P(Caspase-9|AkT)



Akt pathway

- To specify full distribution, assuming that the three variables are discretized into high and low, we need 2^3-1=7 probabilities.
- The Bayesian netwok representation needs 5 probabilities.
- In general, for an n variable problem, reduction of parameters from 2^n to $n*2^k$, if every node has k parents (k<<n).

Another example

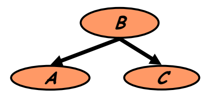


If Protein A is low(0), Protein B is high(1) with probability 0.8

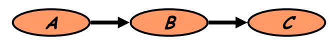
$P(B=1|A=0) = 0.8$
$P(B=1|A=1) = 0.3$

Adapted from Sachs, 2005

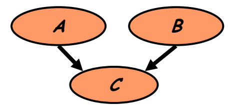
Summary of dependency types



Common cause

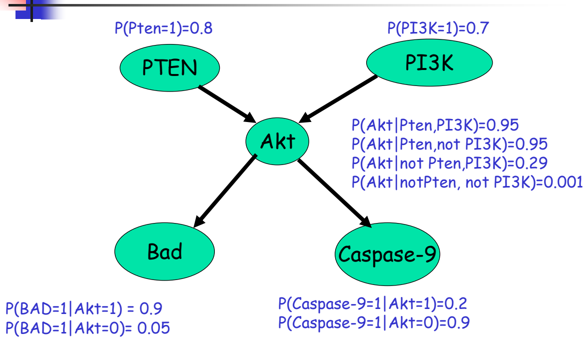


Intermediate gene



Common effects

A simple Bayesian network



Conditional independence

- The topology of the network reflects a set of conditional independence statements.
  - PTEN and PI3K directly affect the probability of the Akt levels being high, but whether or not Bad or Caspase-9 is high depends on the Akt levels alone. Bad and Caspase-9 do not directly respond to PTEN and PI3K levels, the interaction is mediated only through Akt.
  - Bad level is conditionally independent of Caspase-9 level given Akt level.

Computing joint probability distributions

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

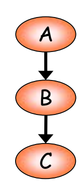
Computing joint probabilities

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

Learning Bayesian Models

- Using data D, find the Bayesian network G that is most likely given the data, i.e. G that maximizes P(G|D).
- 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



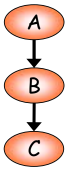
Known structure!

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

From Sachs 2005

Learning CPTs

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



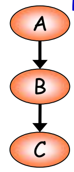
5/6 = 0.83

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

From Sachs 2005

Learning CPTs

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



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

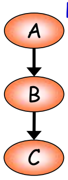
4/5 = 0.8

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

From Sachs 2005

Learning CPTs

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



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

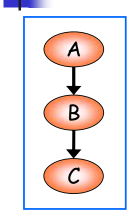
P(C='On'|A='On') = 0.66

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

From Sachs 2005

Learning CPTs



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

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

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

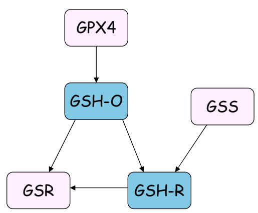
P(C='On'|B='On') = 0.8

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

From Sachs 2005

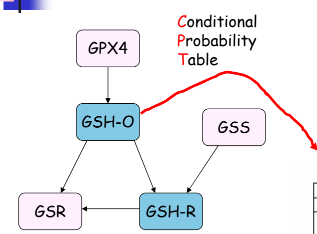
Modeling cellular processes: topology of glutathione network



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.

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

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.

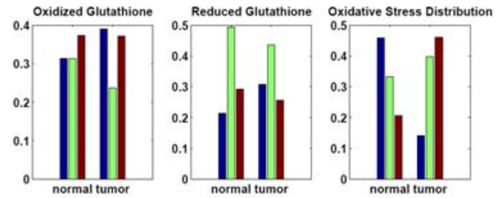
- We learn **separate network** parameters for normal cells and diseased cells for each metabolic process we model.
- Differences in parameters indicate differences in the underlying process.

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!

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

Network Posterior

Marginal Likelihood

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

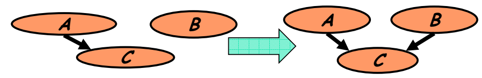
Prior over Networks

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

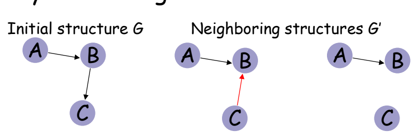
Learning structure



- 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

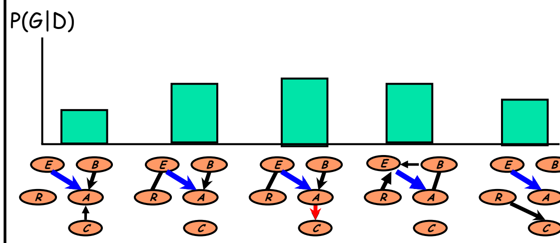
- Start with a random graph G. Evaluate its likelihood wrt D, P(G|D).
- Until little improvement in likelihood
  - Perturb structure G by adding, deleting or reversing edge
  - Accept change if likelihood improves.
- End

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 networks 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.

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.

The bootstrap method

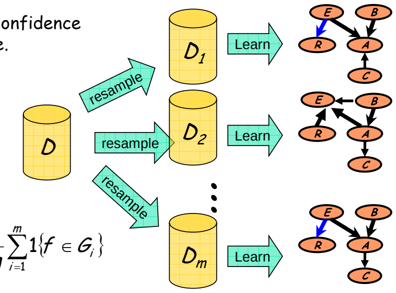
- For $i = 1, \ldots, m$
  - Re-sample with replacement N instances from D. Denote by D_i the resulting dataset.
  - Apply the learning procedure on D_i to induce a network structure G_i
- For each feature f of interest calculate

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

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

Bootstrap illustrated

C(f) is the confidence in a feature.



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

Improving statistical significance

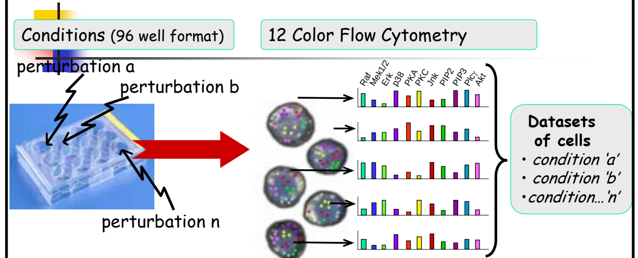
Sparse Data

- 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

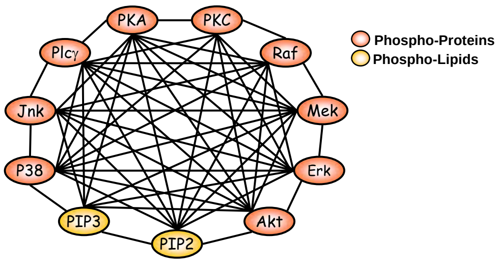
T-Lymphocyte Data (Sachs 2005)



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

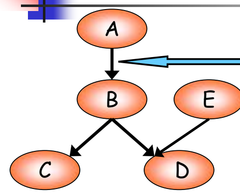
From Sachs 2005

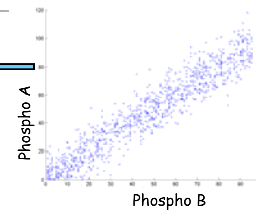
Using correlations



From Sachs 2005

Statistical Dependencies

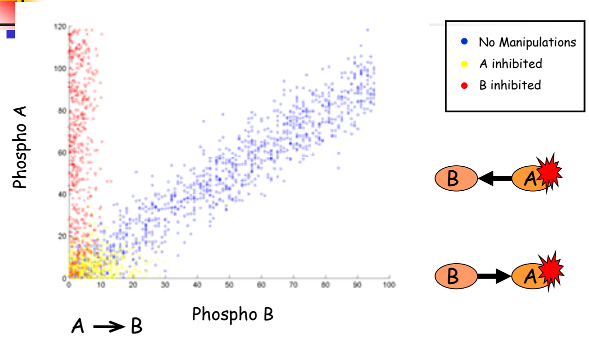




But, how can statistical dependencies determine directionality?

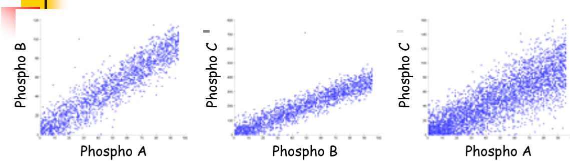
Sachs 2005

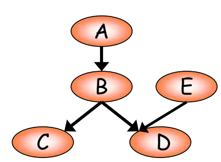
The Power of Interventions



From Sachs 2005

Dismissing Edges

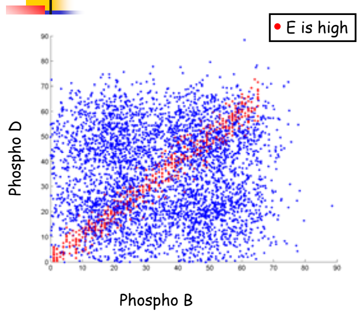




Edges A->B and B->C explain dependence of A and C dismissing the edge between them

Sachs 2005

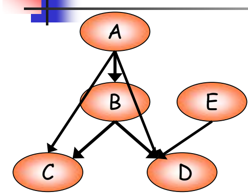
Context Specificity

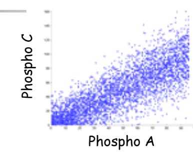


- B and D seem unrelated
- Relationship is revealed by considering simultaneous measurement of E
- Demonstrates the need for simultaneous measurements of variables
- Pairwise computational analysis (e.g. correlations) insufficient

Sachs 2005

Indirect Edges

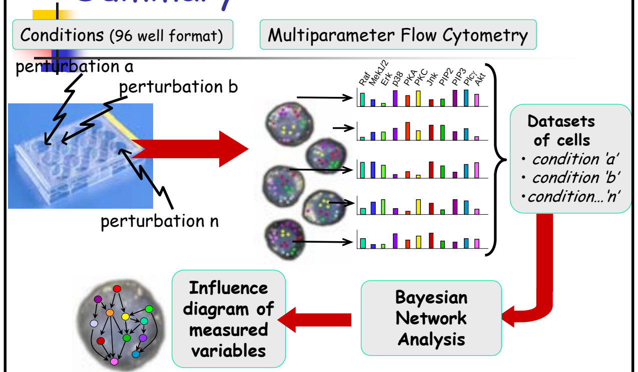




What would happen if B was not measured?

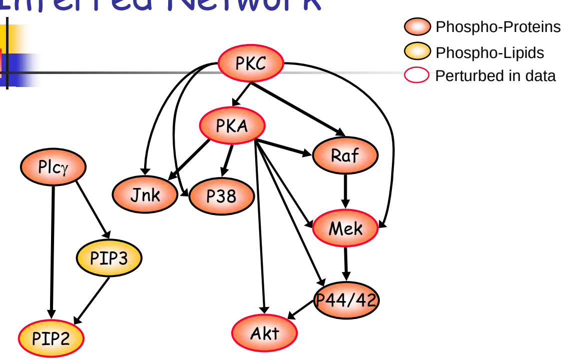
Sachs 2005

Summary



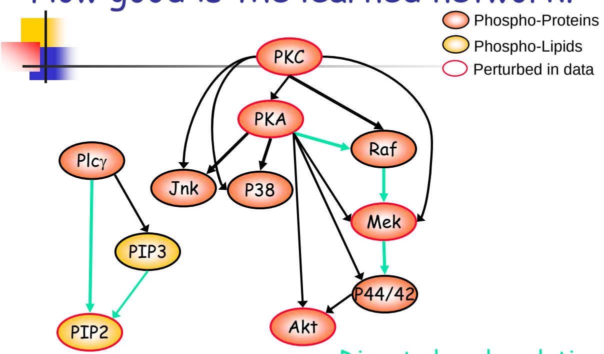
Sachs 2005

Inferred Network



Sachs 2005

How good is the learned network?

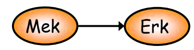


Direct phosphorylation

Sachs 2005

The need for cytometry data

- Direct phosphorylation:



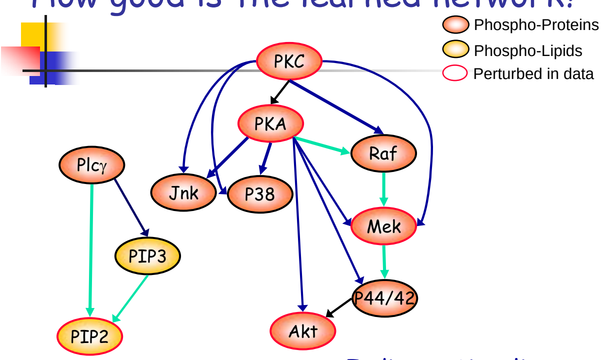
Difficult to detect using other forms of high-throughput data:

-Protein-protein interaction data

-Microarrays

Sachs 2005

How good is the learned network?

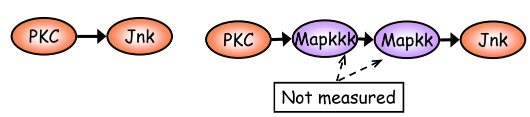


Indirect Signaling

Sachs 2005

Ability to handle missing nodes

- Indirect signaling

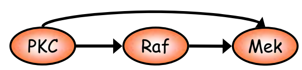


Indirect connections can be found even when the intermediate molecule(s) are not measured

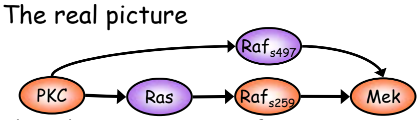
Sachs 2005

Indirect signaling

- Is this a mistake?

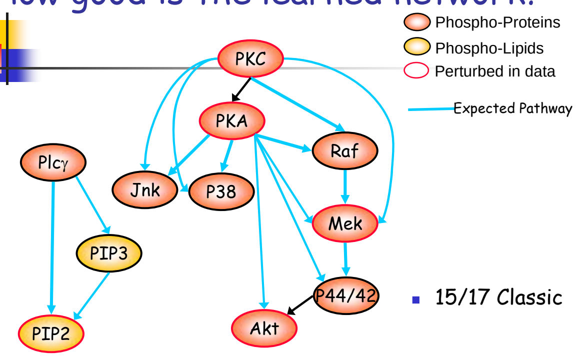


- The real picture



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

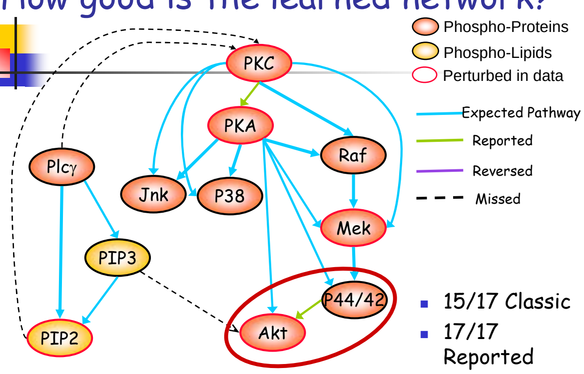
How good is the learned network?



- 15/17 Classic

Sachs 2005

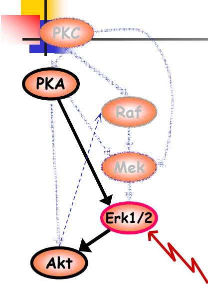
How good is the learned network?



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

Sachs 2005

Prediction



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

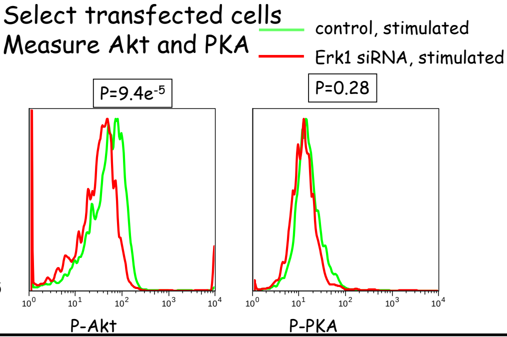
Predictions:

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

Sachs 2005

Validation

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



Sachs 2005

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

Spectrum of modeling tools in systems biology