

# Using Bayesian Networks to Analyze Expression Data

Friedman, Linial, Nachman and Pe'er

Presented by Andreas Kägi

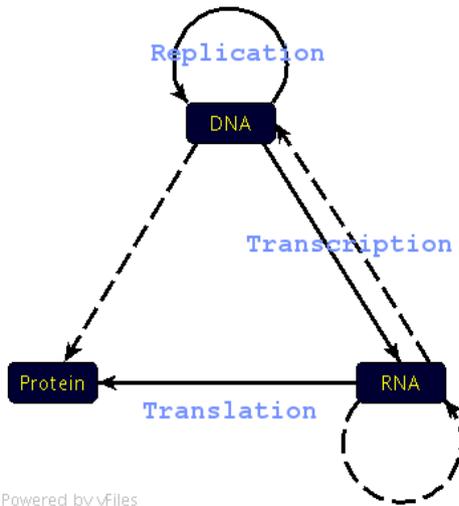
13. June 2006

# Overview

- Introduction of the solution domain
- Overview of the approach by Friedman et al.
- Interesting details
- Experimental results

# Molecular Biology

- Study of structure and function of DNA, RNA and proteins
- Central dogma of molecular biology
  - Standard information flow  
DNA → RNA → Protein
  - No information flow from protein to protein or protein to nucleic acid  
(this is the actual dogma)

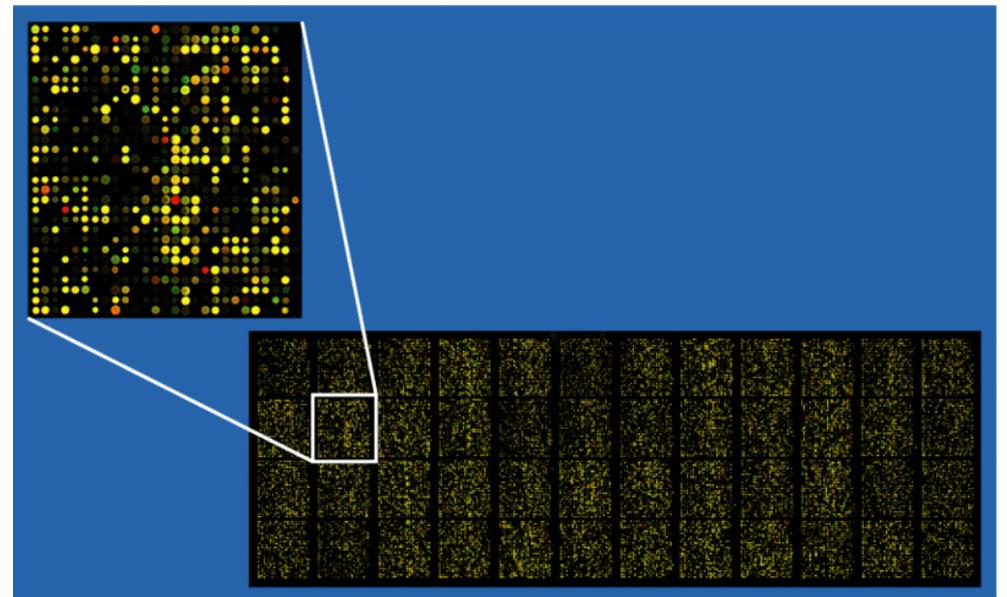


- DNA
  - carrier of genetic information
  - resides in the cell nucleus
- Protein
  - most important “building part” of organisms
  - functions: enzymes, structure, regulation, signalling, defence
- mRNA
  - transports information from the nucleus to the location of protein synthesis

- Gene expression
  - Process of converting a gene's DNA sequence into structure and function of a cell. (proteins)
  - One says a cell expresses a certain amount (level) of proteins
- Regulation
  - Cellular control of the amount and timing of appearance of the functional product (proteins) of a gene
  - Regulation in different stages of expression
  - Often done by proteins itself

# “DNA” microarrays

- Measure amount of gene expression by measuring cellular concentration of mRNA  
→ thus we consider only transcriptional regulation
- Thousands of spots on the chip “bind” pieces of cDNA from the sample



# CS Task

- Analyse the microarray results
- Find properties of the transcriptional program  
“What factors (genes) influence the expression of a certain gene?”
- Problems:
  - Noisy data
  - mRNA concentration gives only partial image
  - sample sets relatively small

## Approach by Friedman et al.

- Probabilistic approach with random variables
- Try to obtain joint probability distribution
- Modelling based on Bayesian Networks
  - only locally interacting components
- RVs = expression level of individual genes (one can include other factors)

# Steps

- (1) Local probability models
- (2) Learn network structure
- (3) Learn causal patterns
- (4) Interesting features
- (5) Confidence in features

# Local probability models

- Multinomial model
  - Only discrete variables
  - Gene expression level is discretised: under-expressed, normal, over-expressed
  - Loss of information by discretisation
- Linear Gaussian model
  - $P(X|u_1..u_k) \sim N(a_0 + \sum_i a_i u_i, \sigma^2)$
  - Only dependencies close to linear can be detected

# Learn network structure

- Find network  $B = (G, \Theta)$  that best matches the data
- Optimisation problem: Maximise score
- $S(G : D) = \log P(G|D) = \log P(D|G) + \log P(G) + C$   
 $P(D|G) = \int P(D|G, \Theta) P(\Theta|G) d\Theta$
- This score can be computed locally
- Search space of DAGs is too large  
→ restrict search space

- Consider for each node a relatively small set of candidate parents
- Candidate parents are chosen according to their contribution to the score
- In the restricted search space of DAGs find the ones with best scores by some local search procedure.
  - They propose a greedy hill-climbing algorithm

# Causal networks

- Ultimate goal is a causal network
  - a link  $X \rightarrow Y$  should indicate that  $X$  influences the expression of  $Y$
- Same as a BN (DAG, local prob. model)
- But link  $X \rightarrow Y$  means:  
 $X$  is a direct cause of  $Y$
- Causal Markov assumption
  - Variable independent of earlier causes given its immediate causes
  - then a causal network can be interpreted as a BN

# Learn causal patterns

- Concept of equivalent DAGs (Pearl 1.2.8)
  - same underlying undirected Graph
  - same v-structures (converging directed edges into the same node, such as  $a \rightarrow b \leftarrow c$ )
  - Representation of the equivalence class as PDAG
    - $X \rightarrow Y$  iff all BNs in the class share this edge
    - $X - Y$  iff only some BNs in the class have it

- If all equivalent networks are present in a class, one of them is the true causal network
- Thus if we have  $X \rightarrow Y$  in the PDAG we can say that  $X$  is (probably\*) a cause of  $Y$
- \*probably because either
  - not all members of the equivalence class can be learned
  - or the Causal Markov assumption does not fully hold

# Interesting features on variable pairs

- Markov relation
  - Y is in the Markov blanket of X
  - Indicator that X and Y are related in some joint biological process
- Order relation
  - X is ancestor of Y in the PDAG
  - Thus X is a cause of Y

# Confidence in features

- We want statistical confidence in inferred features (Markov and order relations)
- We apply the bootstrap method
  - regard data set as empirical distribution and sample from it (with replacement)
  - calculate statistic on it
- $\text{conf}(f) = \frac{1}{m} \sum_{i=1}^m f(G_i)$ 
  - $G_i$  are the resampled data sets
  - $f(G_i) = 1$  iff the feature is in  $G_i$

# Experimental results

- Data from Spellman et al.  
(<http://cellcycle-www.stanford.edu/>)
- Data from yeast cell cycle measurements
- 76 measurements of 6177 ORFs  
(mRNA locations) (small data sets!)
- 6 time series
- Spellman found 800 regulated genes in these 6177 ORFs

- Robustness analysis by comparing the data sets with random data sets
  - difference of confidence higher for LG model
  - multinomial model is susceptible to over-fitting
  - order relation more robust than Markov relation
  - analysis sensitive to local model and discretisation method
- Biological analysis
  - Order relations reveal dominant genes (genes that appear before many genes in the network) that are plausible. (Involved in Cell-cycle initiation, etc.)
  - Also the Markov relations reveal biologically plausible genes

# Questions

- Can one trust the results?  
How much would a different local probability model change them?
- How could one better make use of the (available) temporal data?
- How could one integrate interventions?
- Why don't they consider Functional Causal models? (Pearl)