Inferring Regulatory Networks from Gene Expression Data

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

April 2002

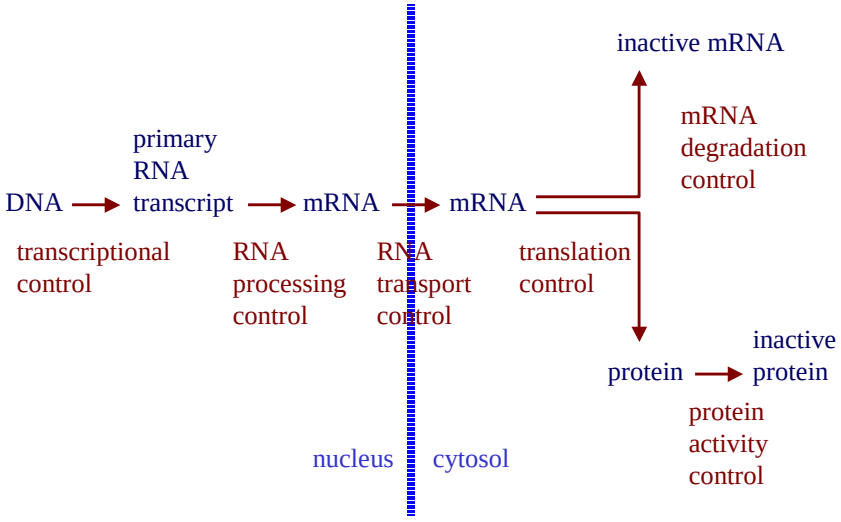
Announcements

- HW #2 due Monday
- project proposals due Monday
- reading for next week
  - Clustering chapter from *Foundations of Statistical Natural Language Processing*, Manning & Schütze

Regulatory Networks

- all cells in an organism have the same genomic data, but the proteins synthesized in each vary according to cell type, time, environmental factors
- there are networks of interactions among various biochemical entities in a cell (DNA, RNA, protein, small molecules).
- can we infer the networks of interactions among genes?

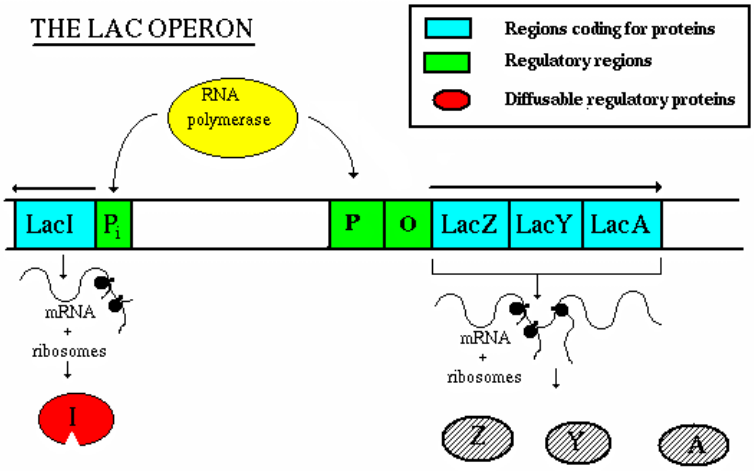
Eukaryotic Expression Regulation



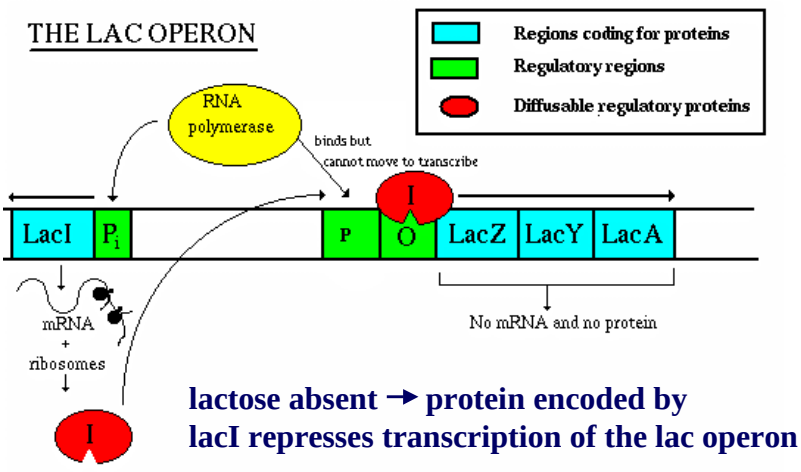
Regulatory Networks

- there are lots of regulatory interactions that occur after transcription, but we'll focus on *transcriptional regulation*:
  - it plays a major role in the regulation of protein synthesis
  - we have good technology for measuring mRNA levels

Transcriptional Regulation Example: the lac Operon

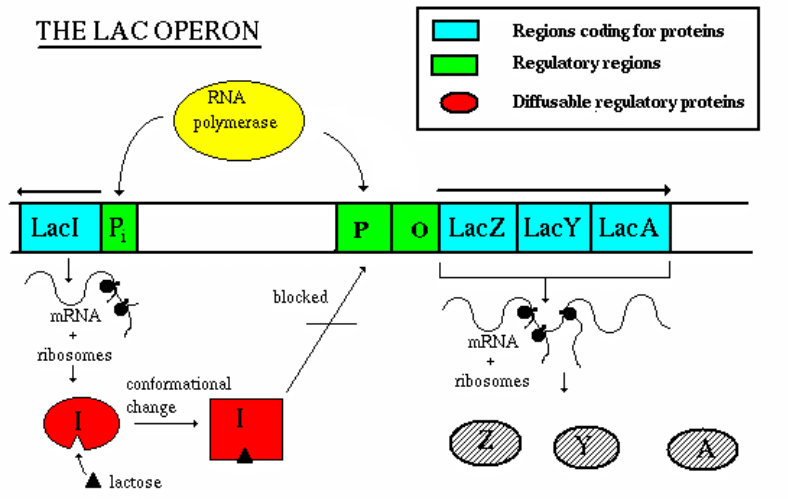


Transcriptional Regulation Example: the lac Operon



lactose absent → protein encoded by lacI represses transcription of the lac operon

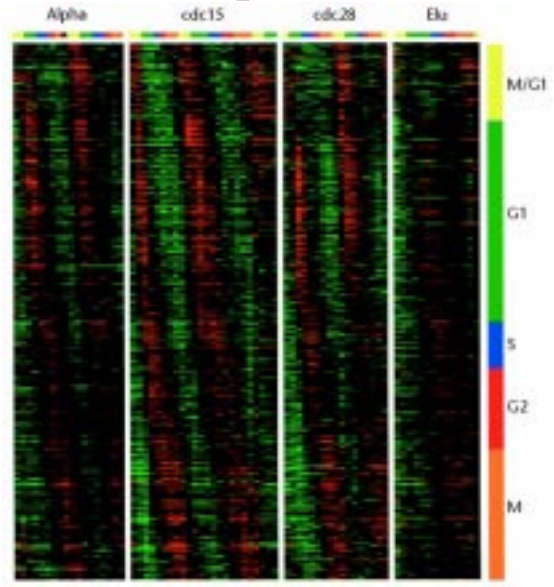
Transcriptional Regulation Example: the lac Operon



Inferrring Regulatory Networks

- given: expression data for a set of genes (data might be temporal)
- do: infer the network of regulatory relationships among the genes

A Gene Expression Profile



Regulatory Network Models

- there are various representations that have been applied to model regulatory networks, including
  - Boolean networks
    [Kaufmann, 1993; Liang, Fuhrman & Somogyi, 1998]
  - differential equations
    [Chen, He & Church, 1999]
  - weight matrices
    [Weaver, Workman & Stormo, 1999]
  - ✓Bayesian networks
    [Friedman et al., 2000]

Probabilistic Model of *lac* Operon

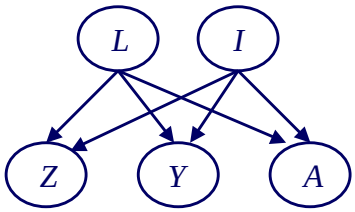
- each gene represented by a random variable in one of three states: under-expressed (-1), normal (0), over-expressed (1)
- lactose represented by a random variable with two states: absent (0), present (1)
- joint probability distribution

$$\Pr(L,I,Z,Y,A) = \Pr(L)\times\Pr(I\mid L)\times\Pr(Z\mid L,I)\times \Pr(Y\mid L,I,Z)\times\Pr(A\mid L,I,Z,Y)$$

- representing the distribution this way requires 162 (2×3^4) parameters

Bayesian Networks

- now consider the following Bayesian network for the *lac* operon

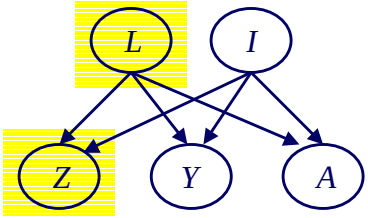


- nodes represent random variables
- edges represent dependencies

Bayesian Networks

- each node has a table representing conditional distribution given *parent* variables

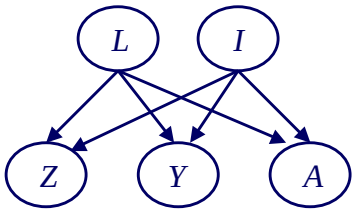
L	Pr(*L*)
0	0.8
1	0.2



L	*I*	Pr(*Z*=-1 | *L*, *I*)	Pr(*Z*=0|*L*, *I*)	Pr(*Z*=1|*L*, *I*)
0	-1	0.1	0.2	0.7
0	0	0.2	0.4	0.4
0	1	0.8	0.1	0.1
1	-1	0.1	0.1	0.8
1	0	0.1	0.2	0.7
1	1	0.1	0.2	0.7

Bayesian Networks

- a Bayesian network provides a *factored* representation of the joint probability distribution



$$\Pr(L, I, Z, Y, A) = \Pr(L) \times \Pr(I) \times \Pr(Z \mid L, I) \times \Pr(Y \mid L, I) \times \Pr(A \mid L, I)$$

- representing the joint distribution this way requires 59 ($2+3+18\times3$) parameters

Linear Gaussian Models

- we can also model the distribution of continuous variables in Bayesian networks
- one approach: linear Gaussian conditional densities

$$\Pr(X \mid u_1, ..., u_k) \sim N(a_0 + \sum_i a_i \times u_i,\ \sigma^2)$$

- X normally distributed around a mean that depends linearly on values of its parents u_i
- a_i parameters estimated from data during training

Learning Bayesian Networks

- given: training set *D* consisting of independent measurements for random variables
- do: find a Bayesian network that best "matches" *D*

- two parts to the approach
  - scoring function to evaluate a given network
  - search procedure to explore space of networks

Learning Bayesian Networks

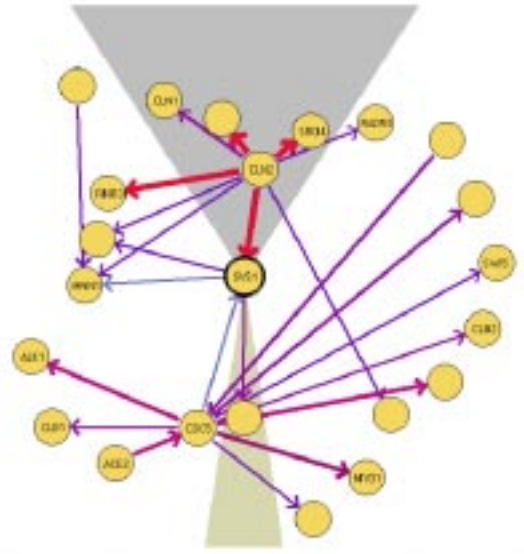


figure from Friedman et al., *Journal of Computational Biology*, 2000

Learning Bayesian Networks

- scoring function to evaluate a given network

$$\text{score}(G : D) = \log \Pr(G \mid D)$$
$$\propto \log \Pr(D \mid G) + \log \Pr(G)$$

log probability of data given graph *G*

log prior probability of graph *G*

- search procedure
  - operations: add, remove, reverse single arcs
  - search methods: hill climbing etc.

Representing Partial Models

- since there are many variables but data is sparse, focus on finding "features" common to lots of models that could explain the data
  - Markov relations: is Y in the *Markov blanket* of X?
    - X, given its Markov blanket, is independent of other variables in network
  - order relations: is X an ancestor of Y

Estimating Confidence in Features: The Bootstrap Method

- for $i = 1$ to m
  - sample (with replacement) N expression experiments
  - learn a Bayesian network from this sample

- the confidence in a feature is the fraction of the m models in which it was represented

Causaulity & Bayesian Networks

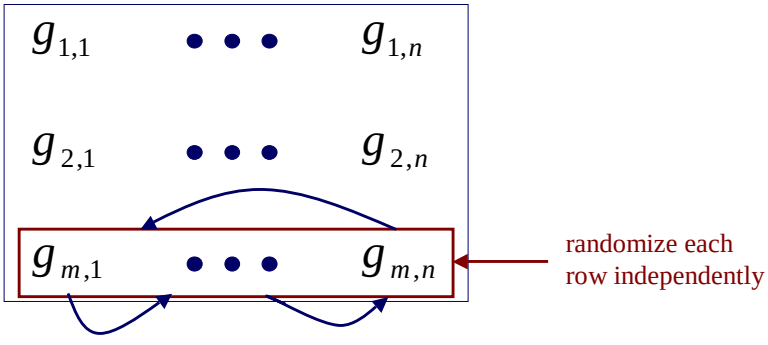
- more than one graph can represent the same set of independences
- from *observations* alone, we cannot distinguish causal relationships in general
- with *interventions* (e.g. gene knockouts) we can

Application to Yeast Cell Cycle Data

- learned Bayesian network models from Stanford yeast cell-cycle data
  - 76 measurements of 6177 genes
  - focused on 800 genes whose expression varied over the cell-cyle stages
- added variable representing cell cycle phase
- each measurement treated as an independent sample from a distribution

Confidence Levels of Features

- how can we tell if the confidence values for features are meaningful?
- compare against confidence values for *randomized* data – genes should then be independent and we shouldn't find "real" features



Confidence Levels of Features: Real vs. Randomized Data

Markov features order features

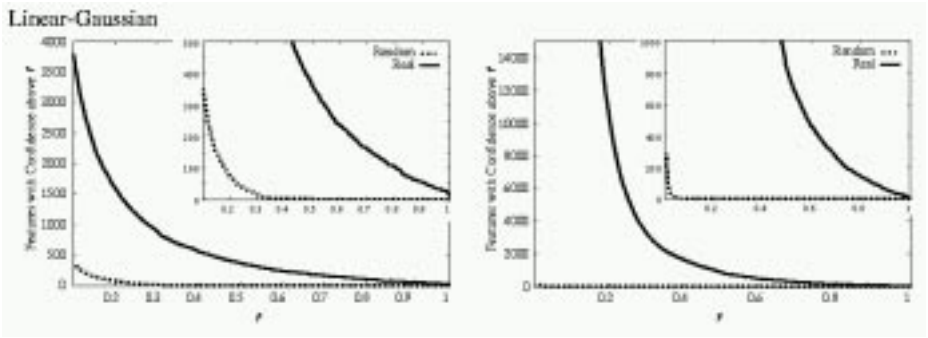


figure from Friedman et al., *Journal of Computational Biology*, 2000

Biological Analysis

- using confidence in order relations, identified *dominant genes*
  - several of these are known to be involved in cell-cycle control
  - several have inviable null mutants
  - many encode proteins involved in replication, sporulation, budding
- assessing confident Markov relations
  - most pairs are functionally related

Top Markov Relations

Table 2: List of top Markov relations, multinomial experiment.

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 mn 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

figure from Friedman et al., *Journal of Computational Biology*, 2000

Discussion

- extracts a richer structure from data than clustering methods
  - interactions among genes other than positive correlation
  - causal relationships (in some cases)
- compared to other approaches for extracting genetic networks
  - models have probabilistic (not deterministic) semantics
  - focus is on extracting “features” of networks, not complete networks themselves