Bayesian Networks Applied to Modeling Cellular Networks

BMI/CS 576
www.biostat.wisc.edu/bmi576/
Mark Craven
craven@biostat.wisc.edu
Fall 2011

Bayesian networks

- a BN is a Directed Acyclic Graph (DAG) in which
  - the nodes denote random variables
  - each node *X* has a *conditional probability distribution* (CPD) representing P(*X* | *Parents*(*X*))
- the intuitive meaning of an arc from *X* to *Y* is that *X directly influences Y*
- formally: each variable *X* is independent of its non-descendants given its parents

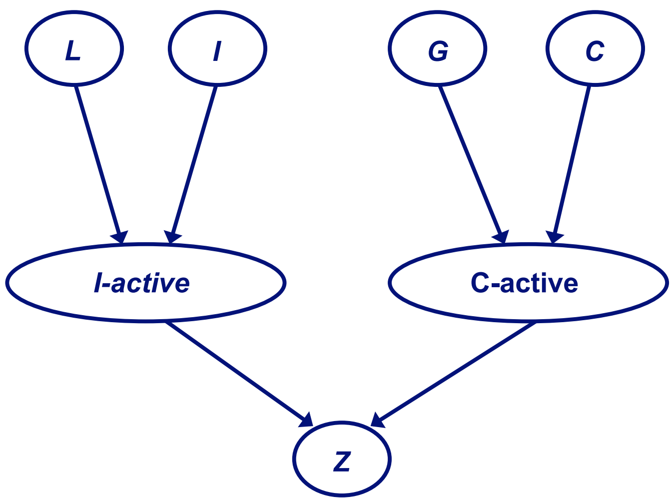
Probabilistic model of *lac* operon

- suppose we represent the system by the following discrete variables

L (lactose)	present, absent
G (glucose)	present, absent
I (lacI)	present, absent
C (CAP)	present, absent
I-active (lacI unbound)	present, absent
C-active (CAP bound)	present, absent
Z (lacZ)	high, low, absent

- suppose (realistically) the system is not completely deterministic
- the joint distribution of the variables could be specified by $2^6 \times 3 = 192$ parameters

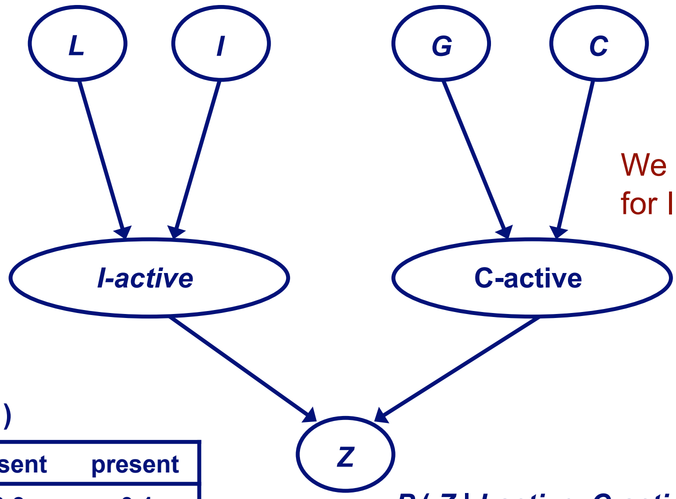
A Bayesian network for the lac system



A Bayesian network for the lac system

***P*(*L*)**

absent	present
0.9	0.1



We also have CPDs for I, G, C, C-active

***P*(I-active | *L, I*)**

L	*I*	absent	present
absent	absent	0.9	0.1
absent	present	0.1	0.9
present	absent	0.9	0.1
present	present	0.9	0.1

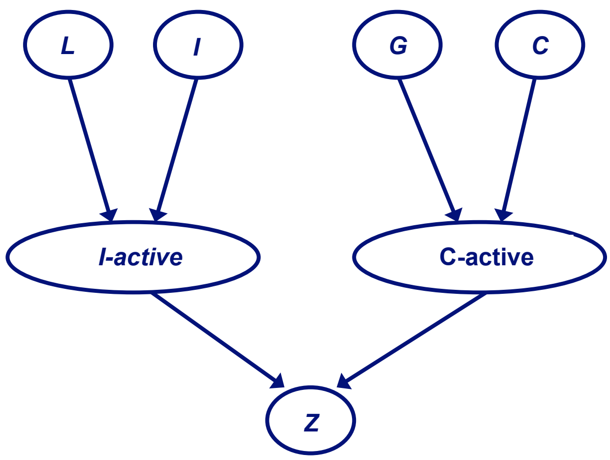
***P*(*Z* | *I-active, C-active*)**

I-active	C-active	absent	low	high
absent	absent	0.1	0.8	0.1
absent	present	0.1	0.1	0.8
present	absent	0.8	0.1	0.1
present	present	0.8	0.1	0.1

Bayesian networks

- a BN provides a *factored* representation of the joint probability distribution

$$P(L,I,Ia,G,Ca,Z) = P(L) \times P(I) \times P(Ia \mid L,I) \times P(G) \times P(C) \times P(Ca \mid G,C) \times P(Z \mid Ia,Ca)$$



- this representation of the joint distribution can be specified with 36 parameters (vs. 192 for the unfactored representation)

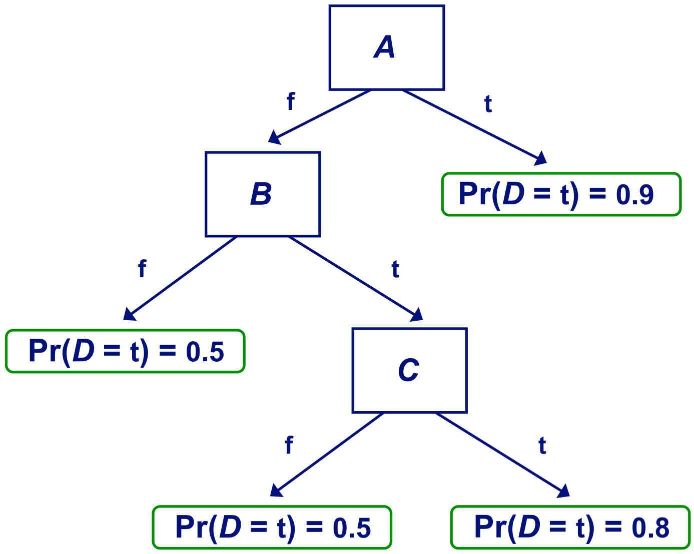
Representing CPDs for discrete variables

- CPDs can be represented using tables or trees
- consider the following case with Boolean variables *A, B, C, D*

P(*D* | *A, B, C*)

A	*B*	*C*	t	f
t	t	t	0.9	0.1
t	t	f	0.9	0.1
t	f	t	0.9	0.1
t	f	f	0.9	0.1
f	t	t	0.8	0.2
f	t	f	0.5	0.5
f	f	t	0.5	0.5
f	f	f	0.5	0.5

P(*D* | *A, B, C*)



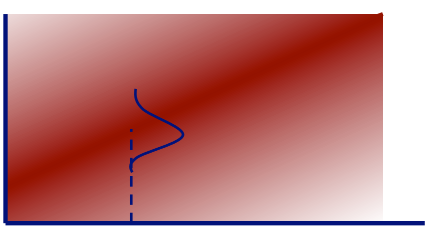
Representing CPDs for continuous variables

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

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

- X normally distributed around a mean that depends linearly on values of its parents u_i

$P(X \mid u_1)$

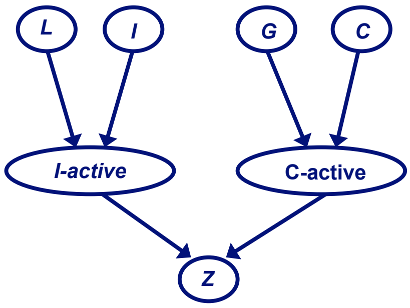


u_1

The parameter learning task

- Given: a set of training instances, the graph structure of a BN

L	G	I	C	I-active	C-active	Z
present	present	present	present	absent	absent	low
present	present	present	present	absent	absent	absent
absent	present	present	present	present	absent	high
			...			



- Do: infer the parameters of the CPDs

- this is straightforward when there aren't missing values, hidden variables

The structure learning task

- Given: a set of training instances

L	G	I	C	I-active	C-active	Z
present	present	present	present	absent	absent	low
present	present	present	present	absent	absent	absent
absent	present	present	present	present	absent	high
			...			

- Do: infer the graph structure (and perhaps the parameters of the CPDs too)