ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Bayesian networks for temporal progression

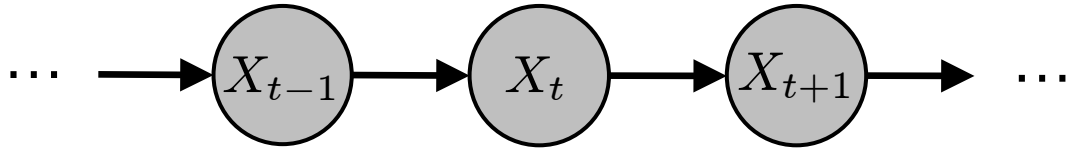
Niko Beerenwinkel

Outline

- Part 1: Dynamic Bayesian networks
  - Time series data
  - Example: Cell cycle gene expression data
- Part 2: Conjunctive Bayesian networks
  - Cross-sectional data
  - Example: Genetic progression of cancer

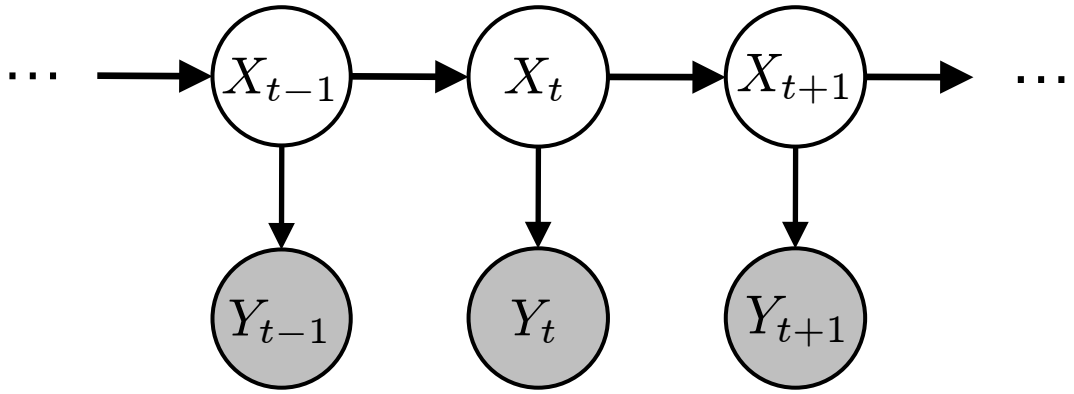
Dynamic Bayesian networks

Dynamic Bayesian network (DBN)

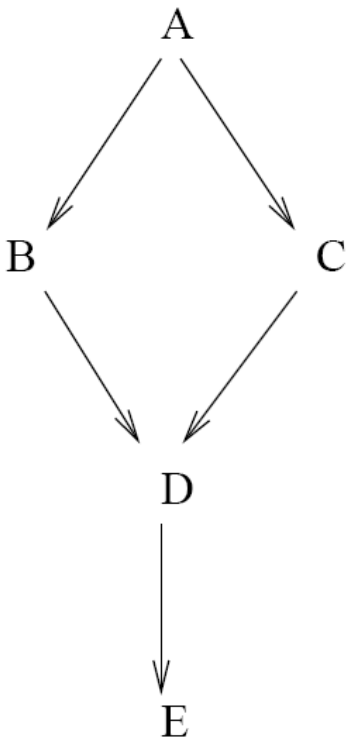


- A DBN represents random variables evolving over time.
- $X_{t+1} \perp X_{t-1} \mid X_t$ (Markov property)
- The random variables $\{X_t\}$ can be discrete or continuous.
- In general, X_t is multivariate and transitions are modeled by a Bayesian network. Thus, the DBN is an “unrolled BN”.
  - Sparse (factored) representation of state probabilities
  - Sparse transition matrices
- There can be hidden variables.

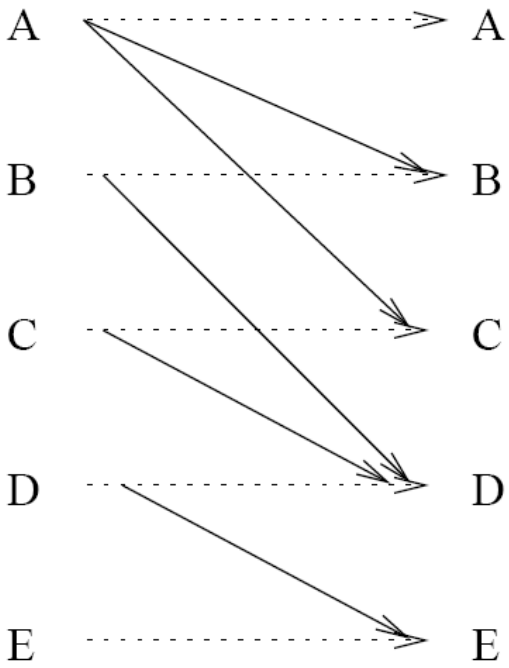
The HMM is a DBN



Unrolling Bayesian networks

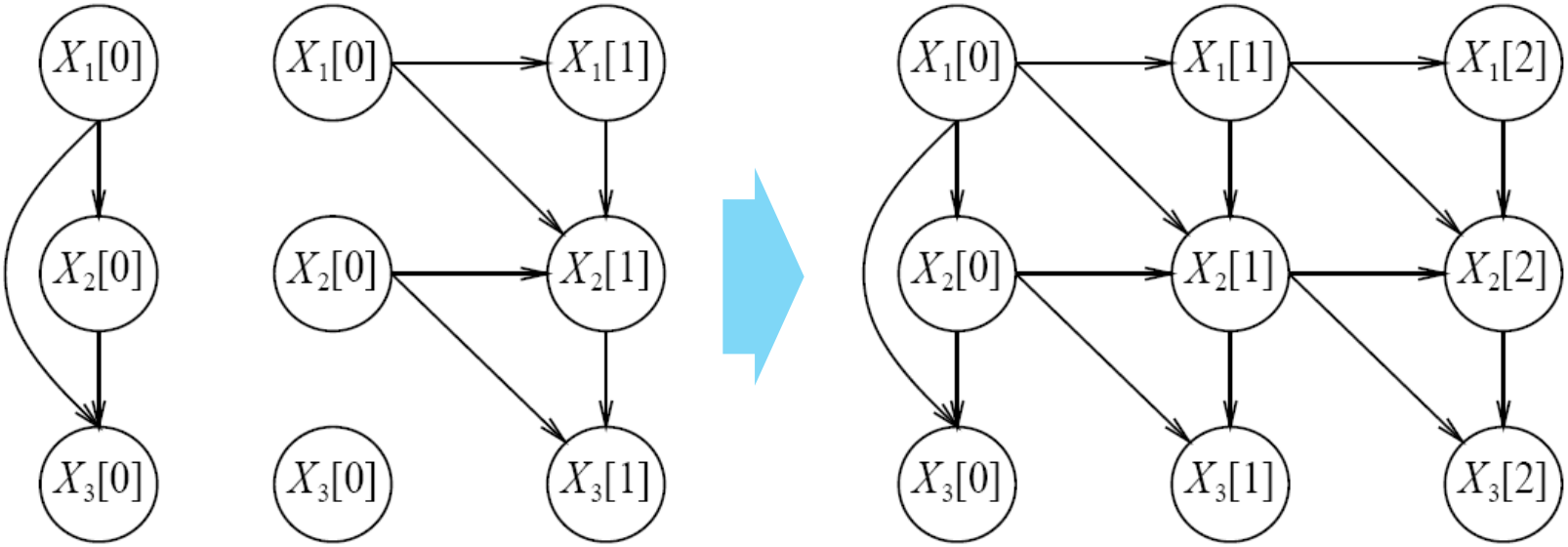


Bayesian network



Equivalent DBN

Definition of DBN



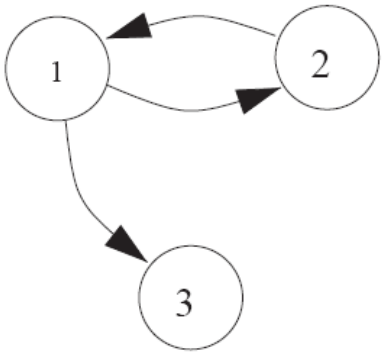
G_0, prior network

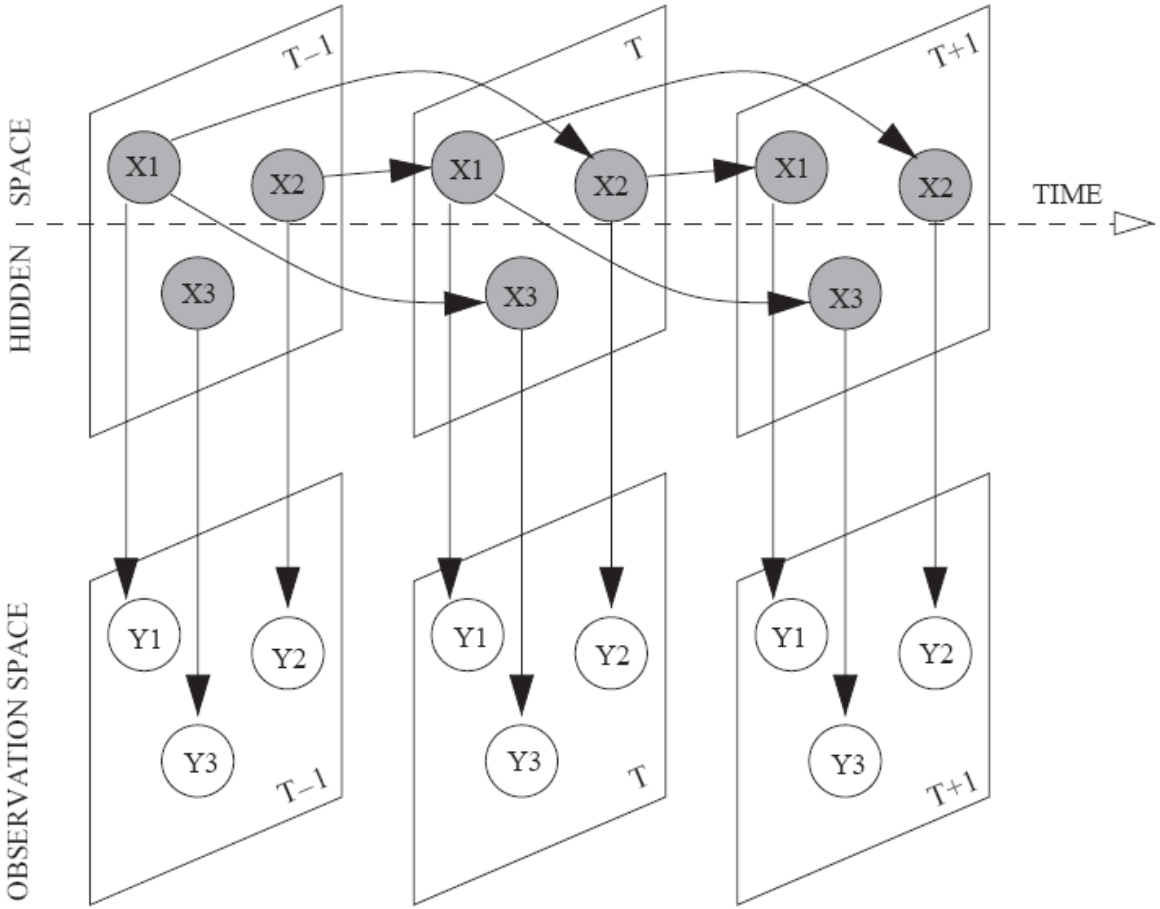
$G_\rightarrow$, transition network

G, DBN (unrolled network)

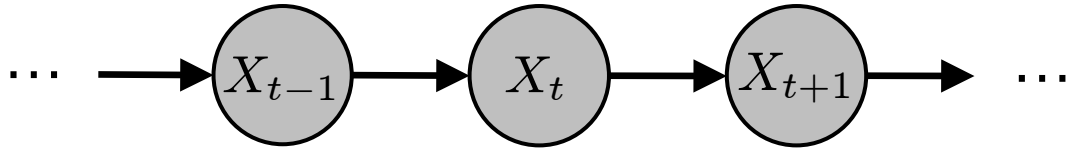
$$P(X[0],\ldots,X[T]) = P_0(X[0]) \prod_{t=0}^{T-1} P_\rightarrow(X[t+1] \mid X[t])$$

The DBN can resolve feedback loops





First-order auto-regressive time series model, AR(1)

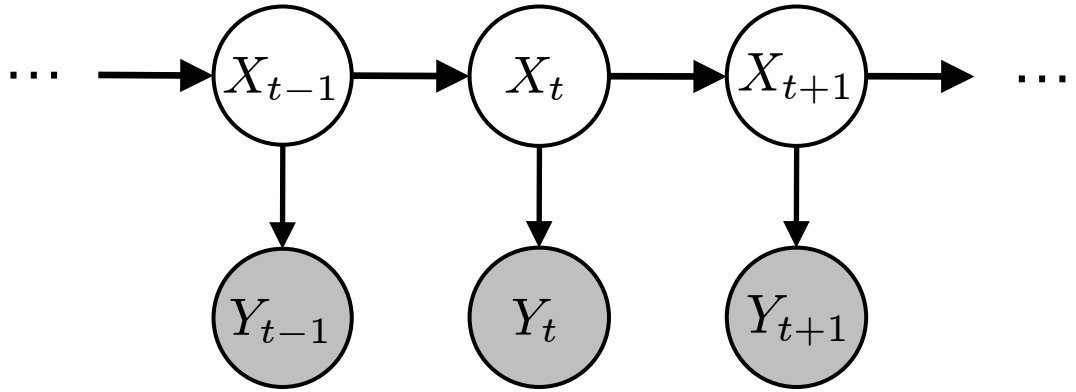


$$X_t = AX_{t-1} + \xi_t$$

$$\xi_t \sim \text{Norm}(0, \Gamma)$$

$$P(X_t = x_t \mid X_{t-1} = x_{t-1}) = \text{Norm}(x_t \mid Ax_{t-1}, \Gamma)$$

Linear dynamical system (LDS) (or Kalman filter, or state space model)



$$P(X_t = x_t \mid X_{t-1} = x_{t-1}) = \text{Norm}(x_t \mid Ax_{t-1}, \Gamma)$$

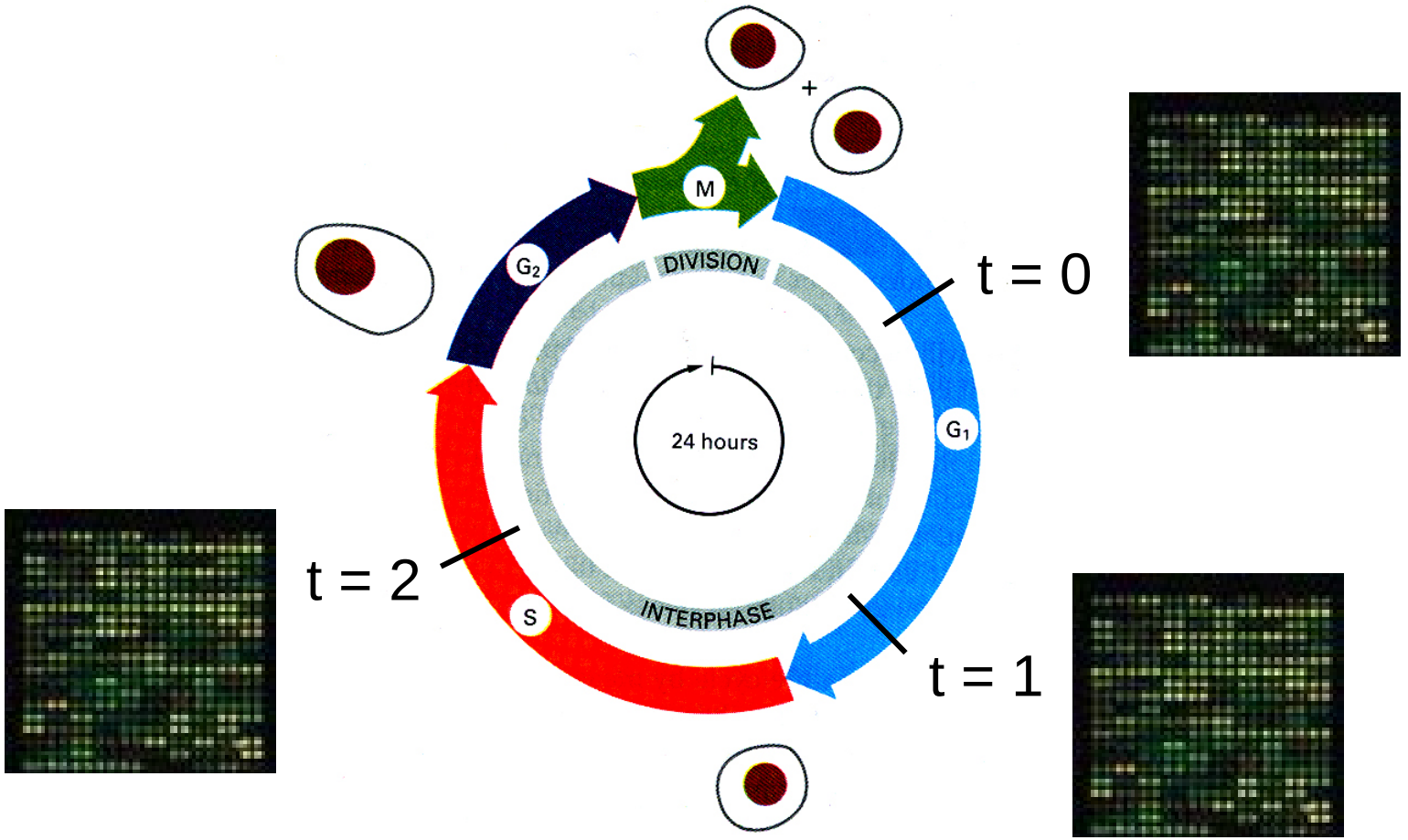
$$P(Y_t = y_t \mid X_t = x_t) = \text{Norm}(y_t \mid Cx_t, \Sigma)$$

Linear dynamical systems

- A LDS is a linear Gaussian model with HMM topology.
- As a linear Gaussian model, the joint, all marginals, and all conditionals are also Gaussian.
- Therefore, the MAP sequence $\mathbf{x}^*$ is equal to the sequence of MAP estimates x^*_t, unlike for HMMs. So we do not need a Viterbi algorithm for LDSs.
- Inference in LDS is efficient.
  - The LDS analogs of the forward and backward algorithms for HMMs are known as *Kalman filter* and *Kalman smoother*, respectively.

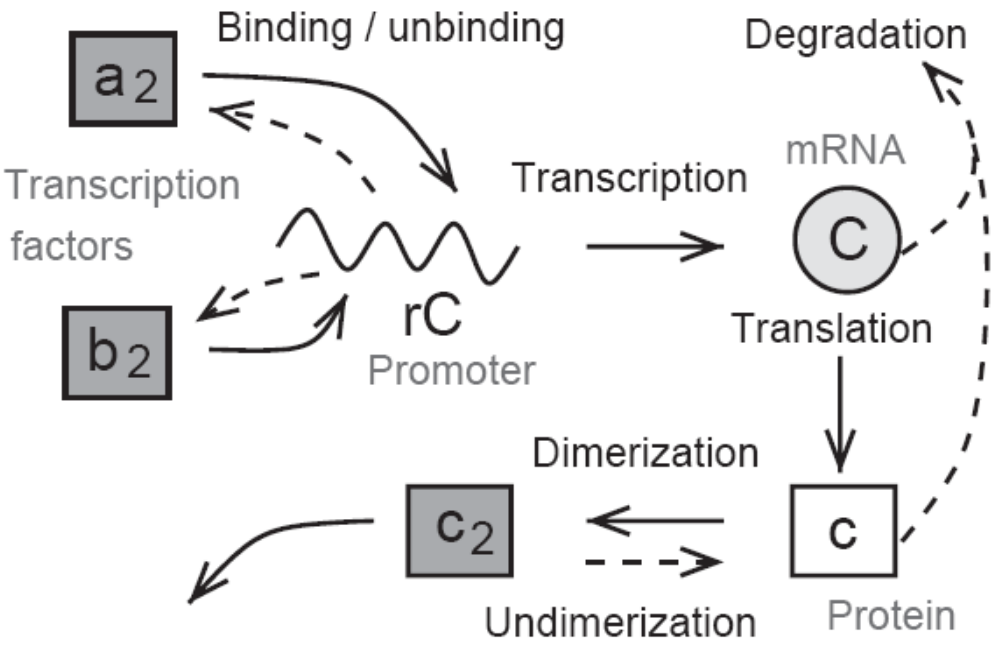
Rudolf E. Kálmán

Cell cycle: gene expression time series

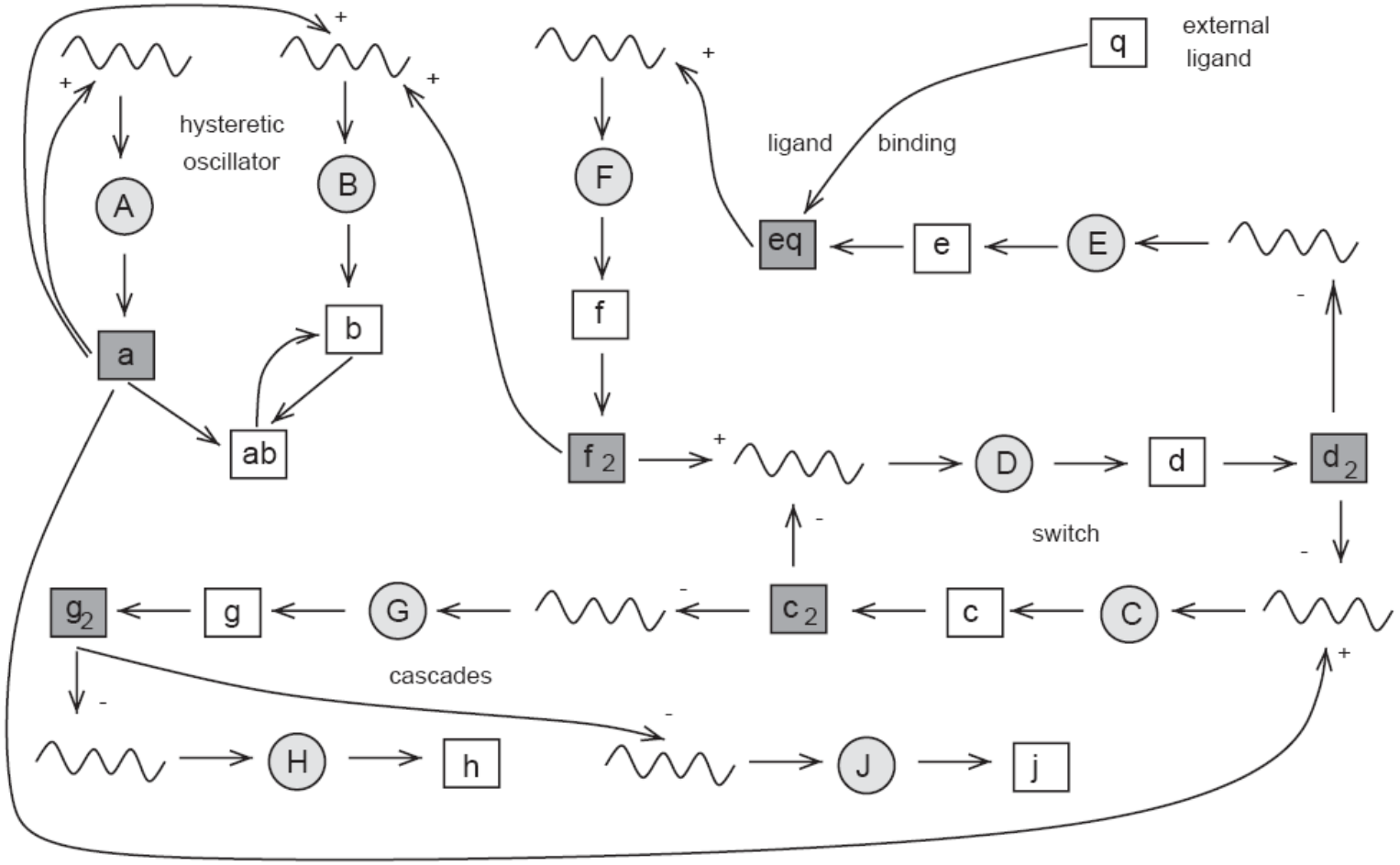


Simulation study

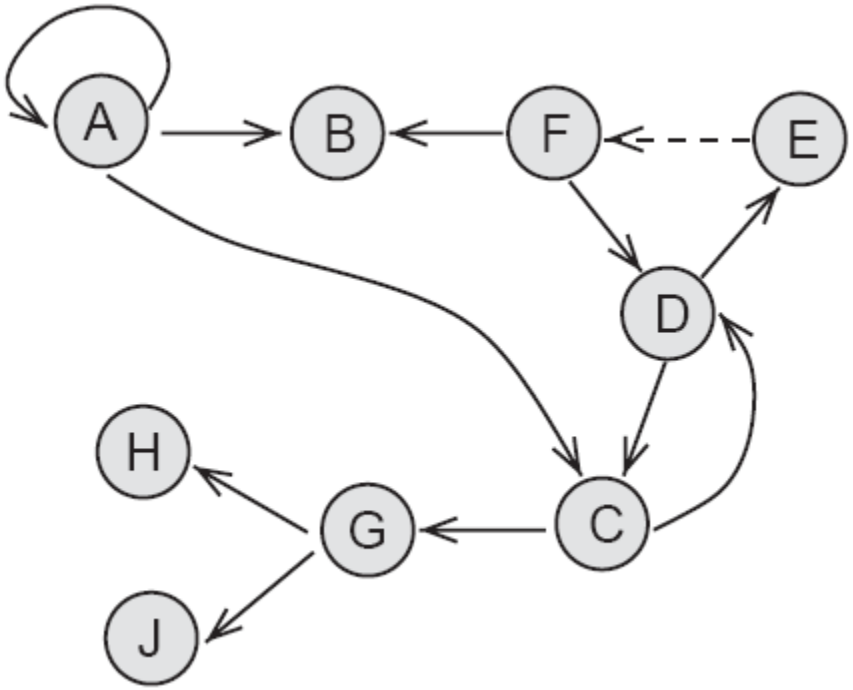
- Elementary processes
  - Transcription factor binding to promoter sequence
  - Transcription
  - Translation
  - Dimerization



Full deterministic model



Induced mRNA network



ODE model

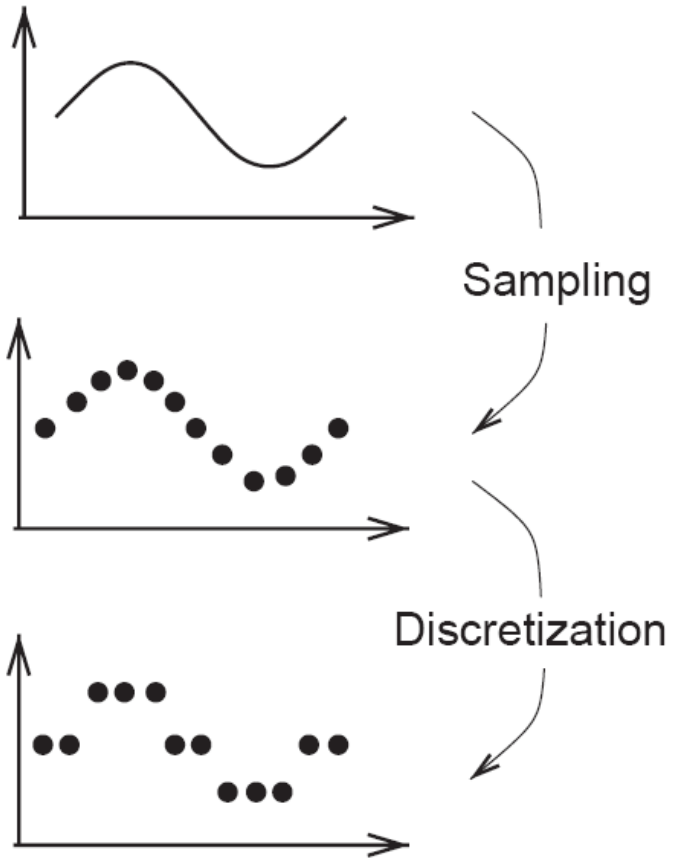
$$\frac{\mathrm{d}}{\mathrm{d}t}[a_2 \cdot rC] = \lambda^{+}_{a_2 \cdot rC}[a_2][rC] - \lambda^{-}_{a_2 \cdot rC}[a_2 \cdot rC],$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[C] = \lambda_{rC}[rC] + \lambda_{a_2 \cdot rC}[a_2 \cdot rC] + \lambda_{b_2 \cdot rC}[b_2 \cdot rC] - \lambda_C[C],$$

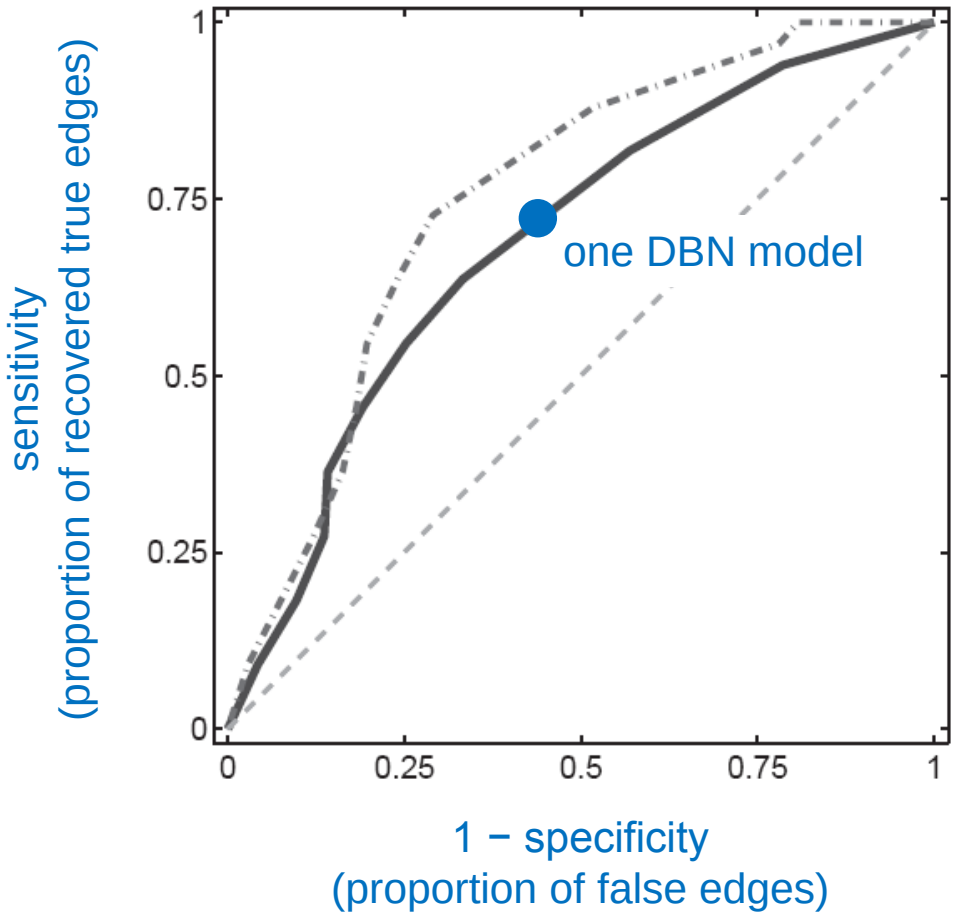
$$\frac{\mathrm{d}}{\mathrm{d}t}[c] = \lambda_{Cc}[C] - \lambda_c[c], \qquad \frac{d}{dt}[c_2] = \lambda^{+}_{cc}[c]^2 - \lambda^{-}_{cc}[c_2]$$

Sampling and discretization

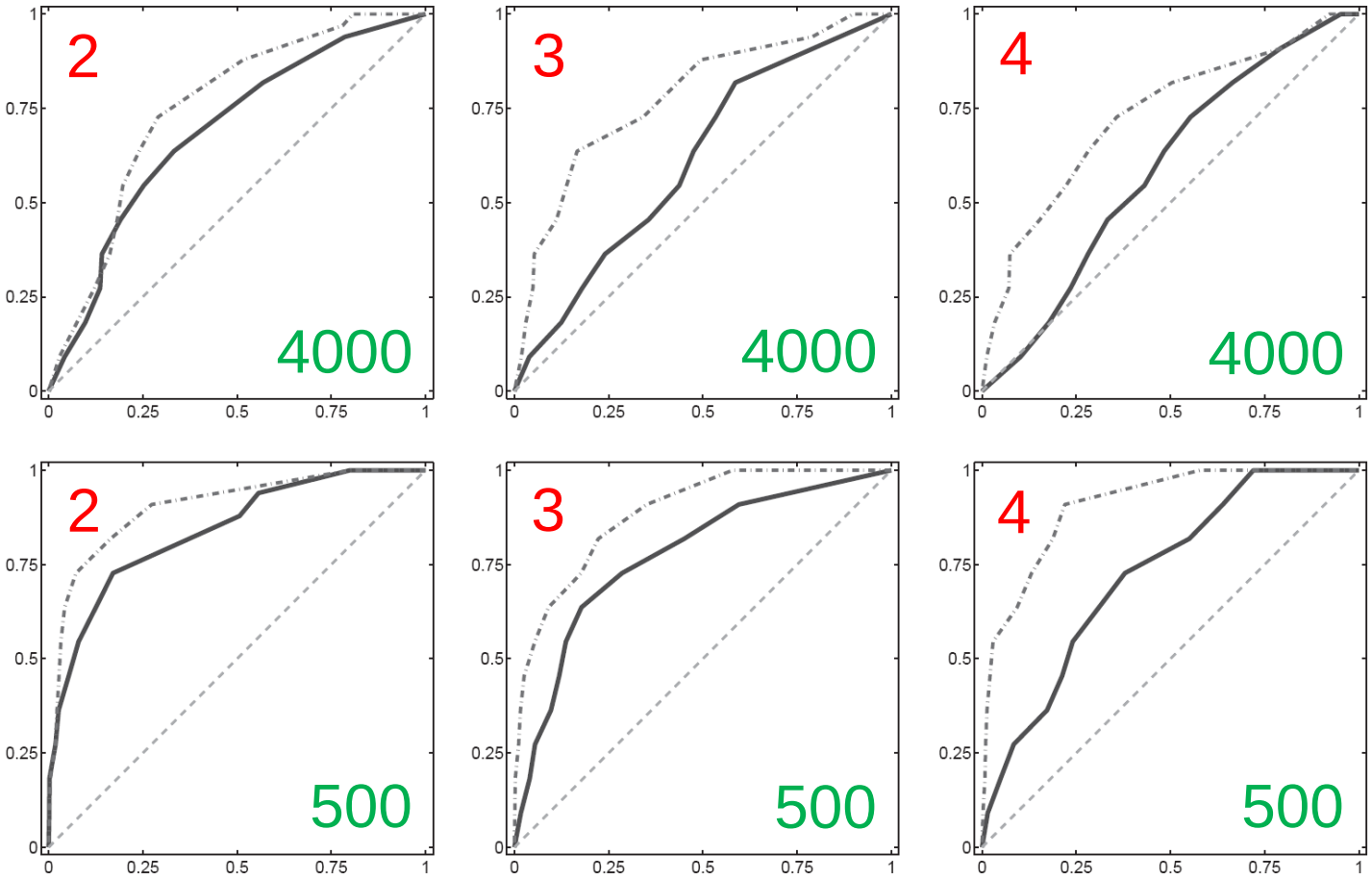
- First experiment:
  - Collect 12 data points over 4000 min after ligand injection
- Second experiment:
  - Collect 12 data points over 500 min after ligand injection
- Use MCMC (Metropolis-Hastings) to sample from P(G | $\mathcal{D}$).
- Different priors restricting the number of incoming edges (“fan-in”) are tested.



Performance measure: ROC curve



DBN Performance



Max. fan-in

Minutes after ligand infection

Additional sequence-based model component

Summary

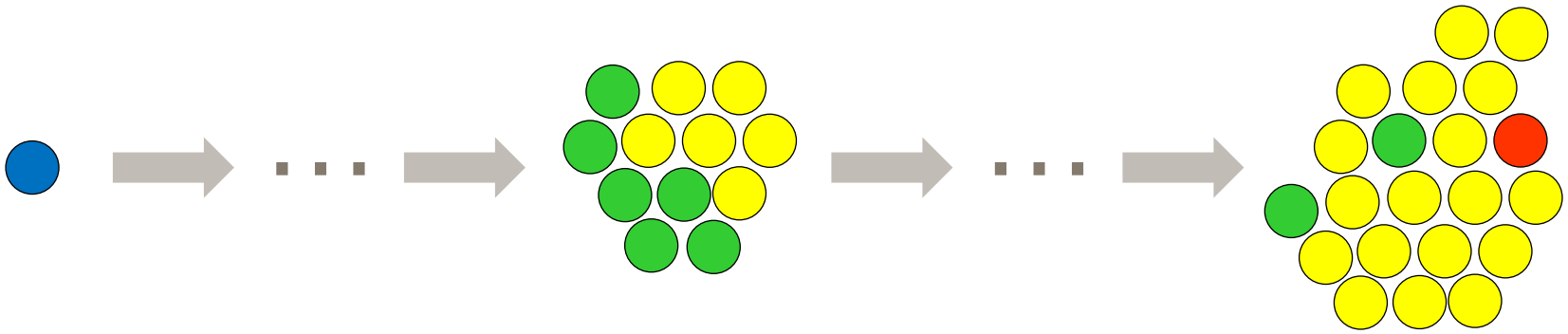
- Dynamic Bayesian networks can model multivariate (high-dimensional) longitudinal (time series) data.
- Inference in linear dynamical systems is particularly efficient (Kalman filter).

References

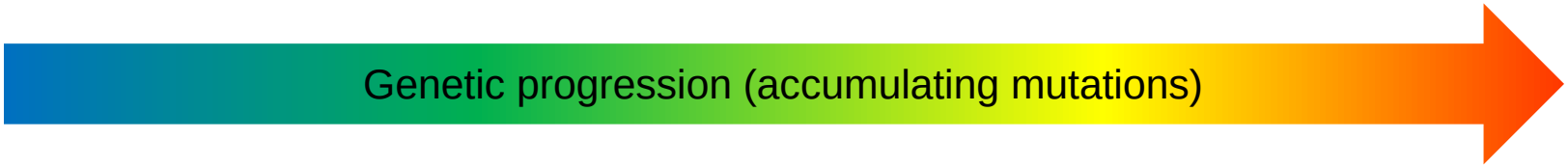
- Murphy K, Mian S. Modelling gene expression data using dynamic Bayesian networks. Technical Report, MIT Artificial Intelligence Laboratory, 1999.
- Bishop CM. Pattern Recognition and Machine Learning. Springer, 2007. Section 13.3.
- Husmeier D (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19(17):2271-2282.

Conjunctive Bayesian networks

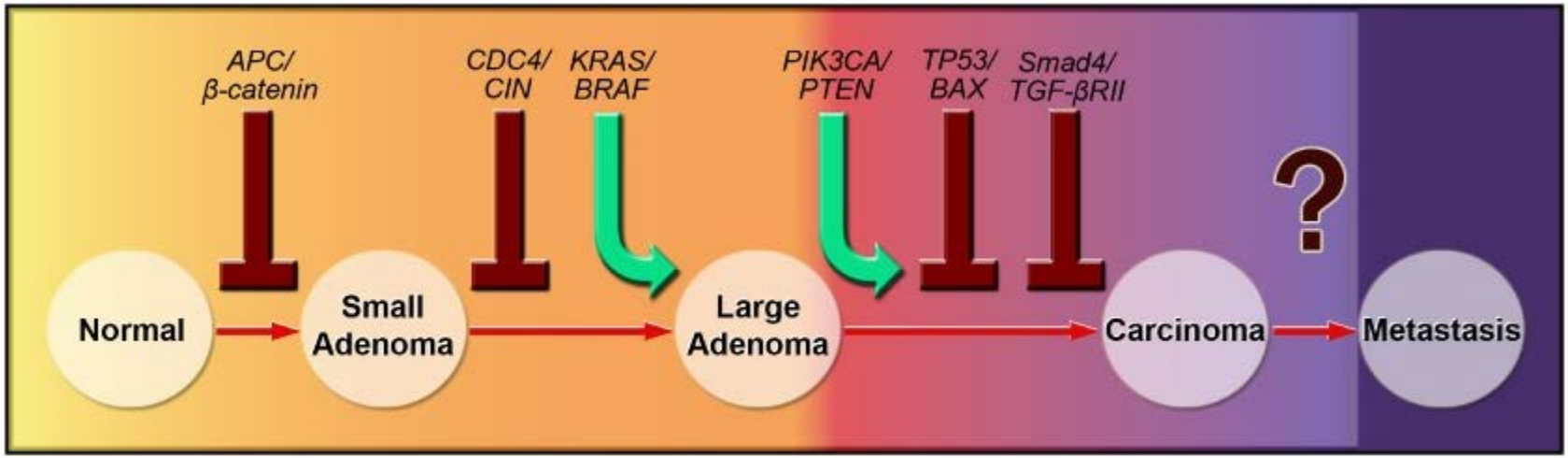
Cancer progression



origin adenoma carcinoma



Vogelgram: Linear genetic progression

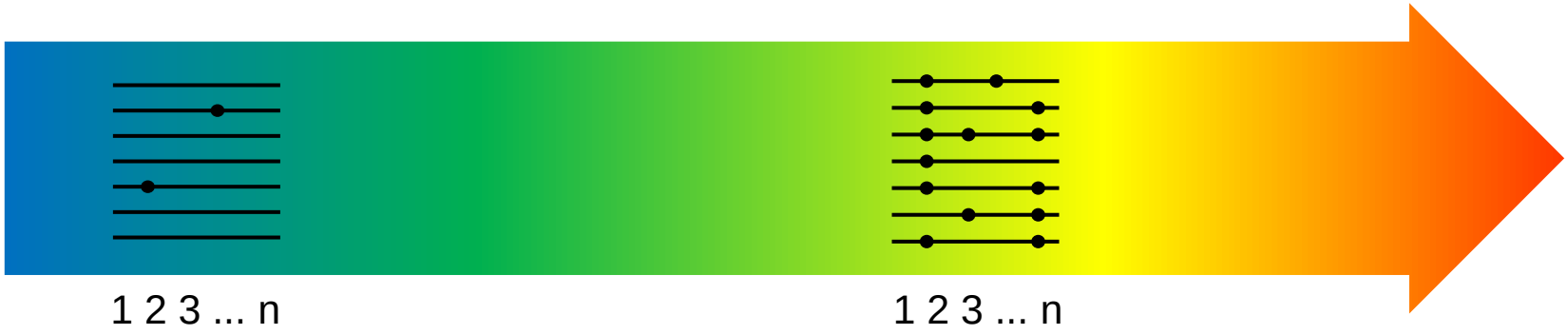


Vogelstein et al. 1988, Jones et al. 2008

- However, recent cancer genome sequencing projects indicate that mutational patterns are more complex and a linear model appears too simplified.

Modeling oncogenesis

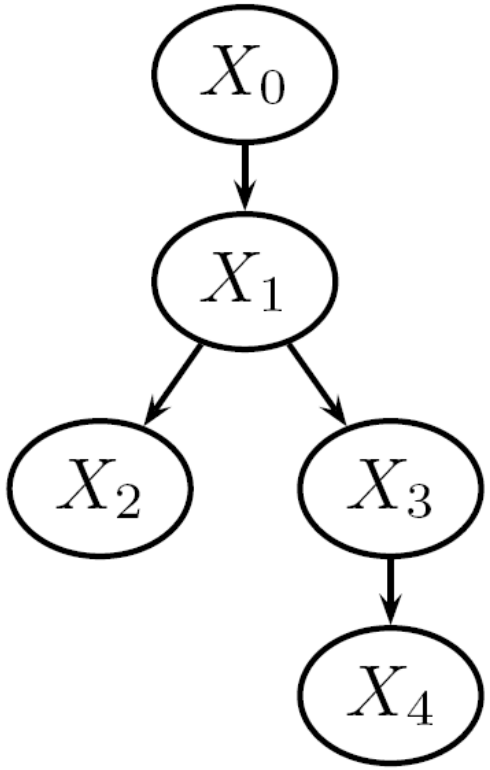
- Let $X = (X_1, ..., X_n)$ be binary random variables, each indicating one of n fixed genetic events.
- An observation x of X is called a genotype.
- We are interested in the dependencies among mutations.
- We assume non-reversibility of mutations
- We require that all predecessor events of a mutation have already occurred, before the mutation can occur.



Oncogenetic trees

- Let $\theta_{i,ab} = P(X_i = b \mid X_{pa(i)} = a)$, $i = 1, ..., n$.
- An oncogenetic tree is a Bayesian tree model for $X = (X_1, ..., X_n)$, such that

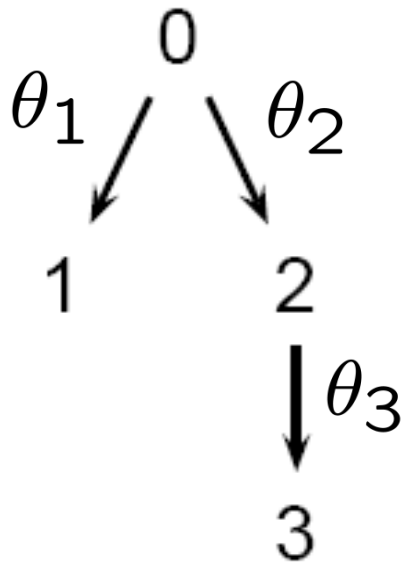
$$P(X = x) = \prod_{i=1}^{n} \theta_{i,x_{\mathrm{pa}(i)}x_i}$$



$$\theta_i = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1-\theta_{i,11} & \theta_{i,11} \end{pmatrix} \end{array}$$

and $X_0 = 1$.

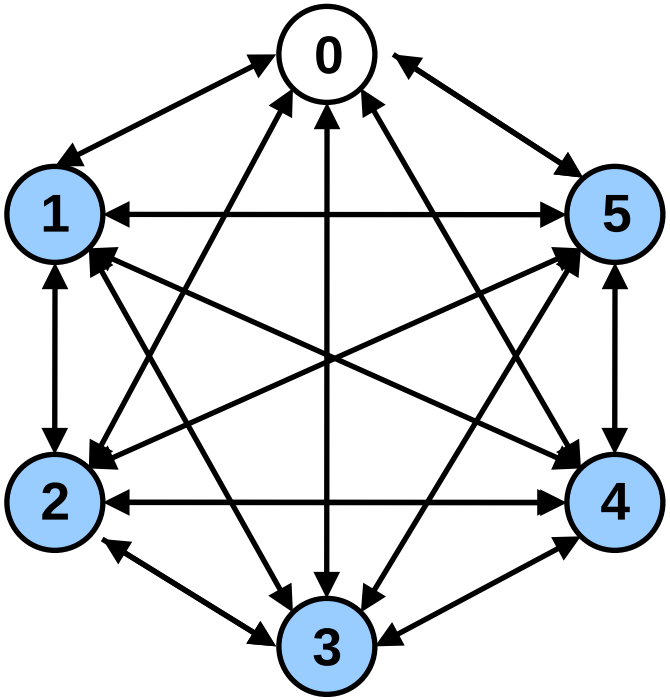
Example



$$
\begin{aligned}
P(000) &= (1-\theta_1)(1-\theta_2) \\
P(001) &= 0 \\
P(010) &= (1-\theta_1)\theta_2(1-\theta_3) \\
P(011) &= (1-\theta_1)\theta_2\theta_3 \\
P(100) &= \theta_1(1-\theta_2) \\
P(101) &= 0 \\
P(110) &= \theta_1\theta_2(1-\theta_3) \\
P(111) &= \theta_1\theta_2\theta_3
\end{aligned}
$$

Tree reconstruction

- Procedure (Desper et al, 1999):
  - Consider the complete weighted graph (G, w) on n+1 vertices
  - Find the maximum weight branching in G (Edmond's branching algorithm, O(|V||E|) time)
- **Theorem**: If P(X) is generated by an oncogenetic tree T, then the maximum weight branching recovers T.



$$w_{ij} = \log \left[\frac{P(X_i)}{P(X_i) + P(X_j)} \frac{P(X_i, X_j)}{P(X_i)P(X_j)} \right]$$

Tree reconstruction from observed data

- We need to compute the weights w_{ij} from observed data.
- The marginal probabilities involving single variables and pairs of variables can be estimated from (a moderate amount of) cross-sectional data.
- Nevertheless, the amount of data required to recover the true tree with high probability increases exponentially in the number of events, n.
- The maximum weight branching is a consistent estimator, but not the MLE.

Timed oncogenetic trees

- Oncogenetic trees can be interpreted as a process in time.
- For the waiting time, we assume independent Poisson processes

$$T_i \sim \text{Exp}(\lambda_i) \text{ for all } i, \text{ and}$$

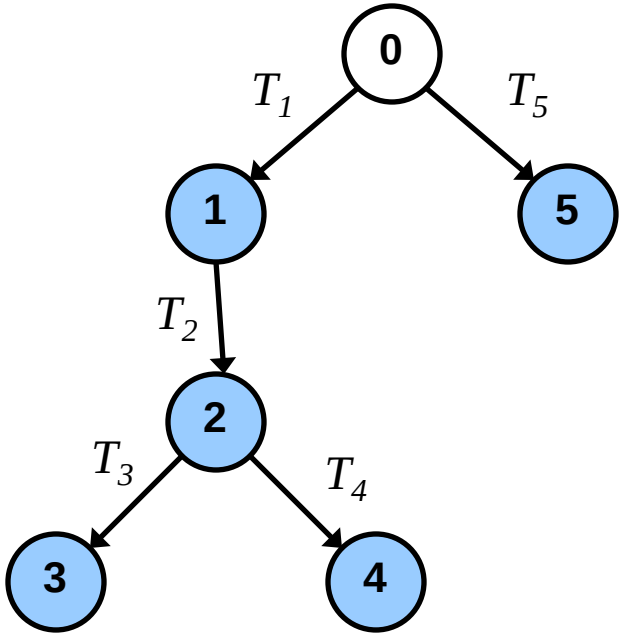
$$T_s \sim \text{Exp}(\lambda_s) \text{ for the sampling time}$$

and set

$$\theta_i = \frac{\lambda_i}{\lambda_i + \lambda_s}$$

(competing exponentials)

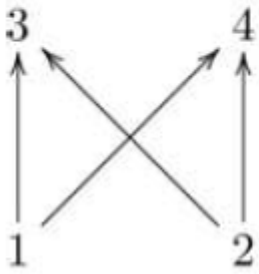
Example



Posets and order ideals

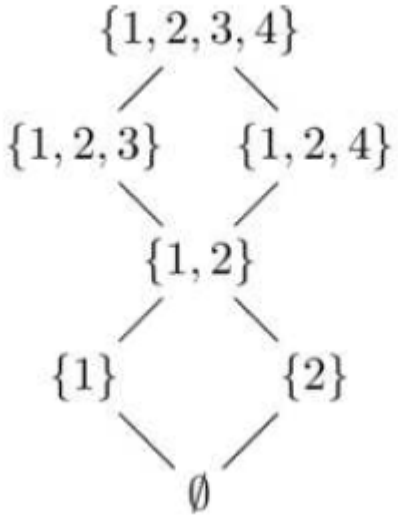
- We relax the tree assumption and consider partially ordered event sets. Let $\mathcal{E}$ be the event poset.
- An *order ideal* g is a subset of $\mathcal{E}$ that is closed downward w.r.t. the poset, i.e., $e_1 < e_2$ and $e_2 \in g$ implies $e_1 \in g$.
- The order ideals are exactly the genotypes that are compatible with the order contraints in $\mathcal{E}$.
- The set of order ideals $J(\mathcal{E})$ forms a distributive lattice. We call $\mathcal{G} = J(\mathcal{E})$ the genotype lattice.
- The *complement* of g is $g^c = \mathcal{E} \setminus g$.
- $e_1 < e_2$ is a *cover relation* if, for all $f \in \mathcal{E}$, $e_1 < f < e_2$ implies $e_1 = f$ or $e_2 = f$.

Example

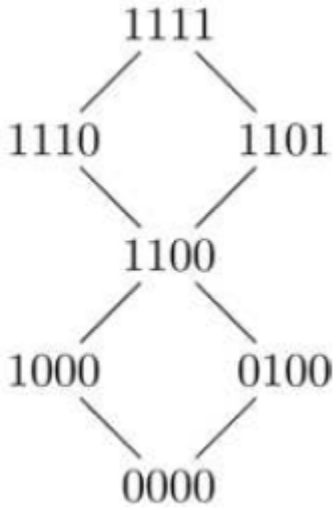


Poset, $\mathcal{E}$

$1 < 3$, $1 < 4$,
$2 < 3$, $2 < 4$



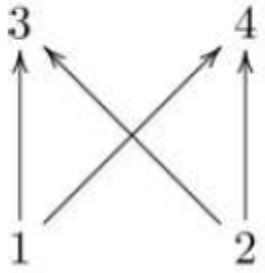
Lattice of order ideals, $J(\mathcal{E})$

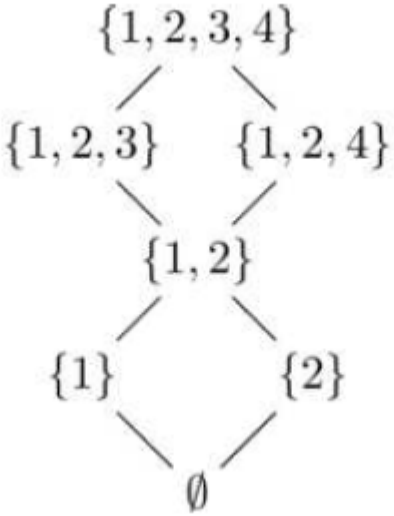


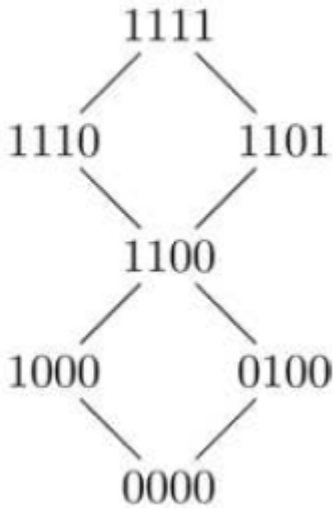
Genotype lattice, $\mathcal{G} = J(\mathcal{E})$

Exit sets

- For a genotype g, $\min(g^c)$ is the set of mutations that could happen next.
- We also call this set the *exit set* of g, $\mathrm{Exit}(g) = \min(g^c)$.
- Example: g = {1,2}, $\min(g^c)$ = {3,4}.







Conjunctive Bayesian networks (CBNs)

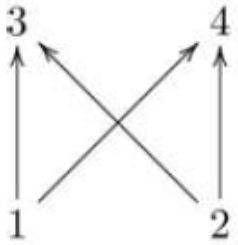
- For a poset $\mathcal{E}$ and parameters $\theta = (\theta_1, ..., \theta_n)$, the CBN is defined by

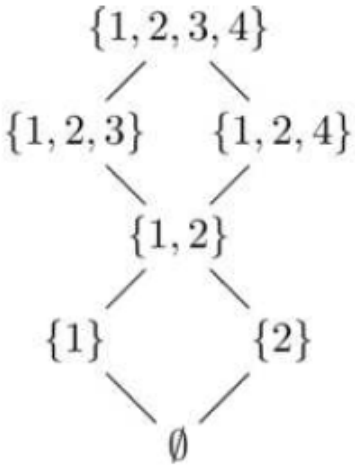
$$P(X = g) = \prod_{e \in g} \theta_e \cdot \prod_{e \in \mathrm{Exit}(g)} (1 - \theta_e)$$

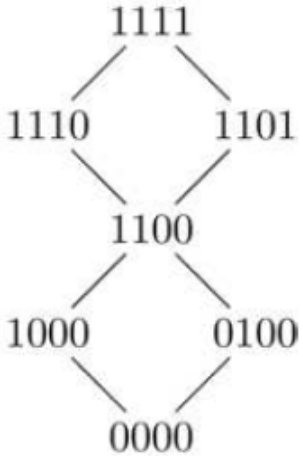
- Equivalently, the CBN is the Bayesian network model for the binary random variables $(X_e)_{e \in \mathcal{E}}$ whose graph has edges $e \to f$ for all cover relations $e < f$ in $\mathcal{E}$ and whose conditional probability tables are

$$[\Pr(X_e = b \mid X_{\mathrm{pa}(e)} = a)]_{a \in \{0,1\}^{\mathrm{pa}(e)}, b \in \{0,1\}} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 - \theta_e & \theta_e \end{bmatrix}$$

Example







$$P_{\varnothing}(\theta) = (1-\theta_1)(1-\theta_2), \qquad P_1(\theta) = \theta_1(1-\theta_2),$$

$$P_2(\theta) = \theta_2(1-\theta_1), \qquad P_{12}(\theta) = \theta_1\theta_2(1-\theta_3)(1-\theta_4),$$

$$P_{1234}(\theta) = \theta_1\theta_2\theta_3\theta_4, \qquad P_{123}(\theta) = \theta_1\theta_2\theta_3(1-\theta_4),$$

$$P_{124}(\theta) = \theta_1\theta_2\theta_4(1-\theta_3).$$

Parameter estimation

- Let u : $\mathcal{G} \to \mathbb{N}$ be the observed data, where u(g) = n_g denotes the frequency of genotype g in the data set.
- Let below(e) = {f $\in \mathcal{E}$ | f $\neq$ e and f < e}.

- **Theorem**: The maximum likelihood estimate (MLE) of θ is given by

$$\widehat{\theta}_e = \frac{\sum_{g:e\in g} u_g}{\sum_{g:\mathrm{below}(e)\subseteq g} u_g} \qquad for\ all\ e \in \mathcal{E}$$

- Proof is straightforward.

Model selection

- A genotype g is *compatible* with the poset $\mathcal{E}$, if g $\in$ J($\mathcal{E}$) = $\mathcal{G}$.
- The data u is compatible with the poset $\mathcal{E}$, if the support of u is a subset of the genotype lattice, supp(u) $\subset \mathcal{G}$.

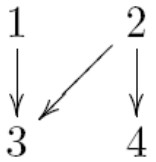
- **Theorem**: There is a unique largest poset $\mathcal{E}_u$ that is compatible with the data u. $\mathcal{E}_u$ is the ML poset.

- "largest poset" refers to poset refinements: $\mathcal{E}_1 \prec \mathcal{E}_2$ if all relations in $\mathcal{E}_1$ are also in $\mathcal{E}_2$.
- Note that $\mathcal{E}_1 \prec \mathcal{E}_2$ if and only if J($\mathcal{E}_1$) > J($\mathcal{E}_2$).
- Proof is straightforward, but more technical.

Continuous-time CBN (CT-CBN)

- Let T_i be the continuous random variable describing the waiting time to event *i*.
- The CT-CBN on T is defined on an event poset P as follows:

Partially ordered set, P

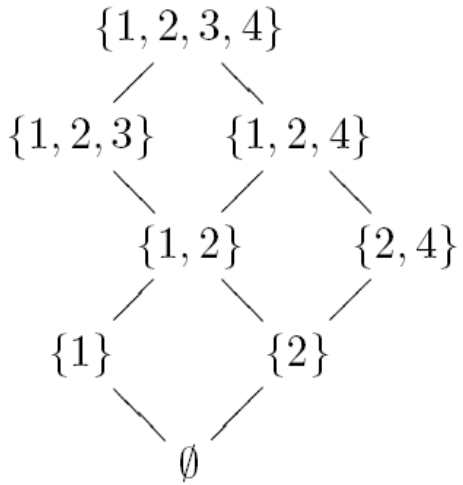


Waiting times, T

$$Z_i \sim \text{Exp}(\lambda_i), \quad i = 1, 2, 3, 4$$

$$T_1 = Z_1$$
$$T_2 = Z_2$$
$$T_3 = \max(T_1, T_2) + Z_3$$
$$T_4 = T_2 + Z_4$$

Lattice of order ideals, $J(P)$



Parameter estimation and model selection

Consider N observations $t_1, \ldots, t_N$, each consisting of n time points, $t_{k1}, \ldots, t_{kn}$.

Proposition. Let P be a partially ordered set. If all observations are compatible with P, then the maximum likelihood estimate of λ is given by

$$\widehat{\lambda}_i = \frac{N}{\sum_{k=1}^{N} (t_{ki} - \max_{j \in \mathrm{pa}(i)} t_{kj})}, \quad i \in [n].$$

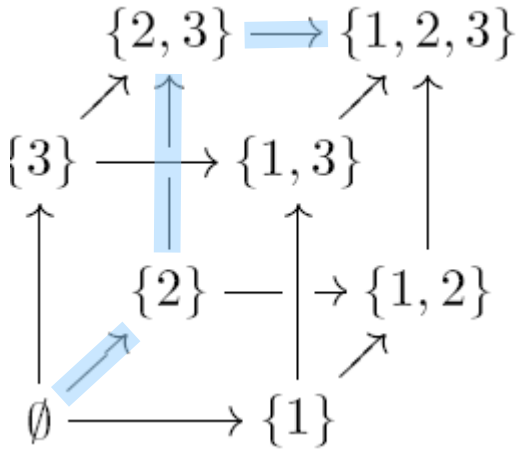
Otherwise the likelihood function is identically zero.

Theorem. The maximum likelihood poset is the largest poset that is compatible with the data.

Censoring

- Exponential sampling process, T_s ~ Exp(λ_s)
- EM algorithm
- In the E step, we have to compute the expectation of

$$\tau_k = \max_{j=1,\ldots,k} T_j$$

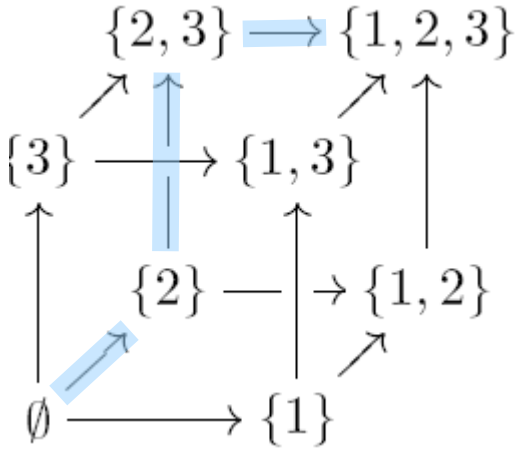


$$E[\tau_k] = \sum_{\pi:\, j_1 \to \cdots \to j_k} \mathrm{Prob}(\pi)\ \mathrm{ExpT}(\pi)$$

Probability and waiting time of a pathway

- Exit_i = possible mutations from genotype i
- The probability of a pathway $\pi = j_1 \to \ldots \to j_k$ is

$$\text{Prob}(\pi) = \prod_{i=1}^{k} \frac{\lambda_{j_i}}{\sum_{j \in \text{Exit}_i} \lambda_j}$$



- The expected waiting time of π is

$$\text{ExpT}(\pi) = \sum_{i=1}^{k} \frac{1}{\sum_{l \in \text{Exit}_{j_i}} \lambda_l}$$

- Can be computed by dynamic programming

Basic idea

$$\mathrm{E}[\max_{j \in P} T_j] =$$

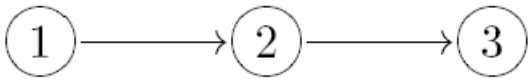
$$= \int_{\mathbb{R}^n_{\geq 0}} \max_{j \in P} t_j \cdot f(t)\, dt$$

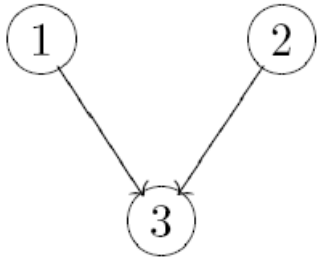
$$= \sum_{\sigma \in S_n} \int_{t_{\sigma_1}=0}^{\infty} \cdots \int_{t_{\sigma_n}=t_{\sigma_{n-1}}}^{\infty} t_{\sigma_n} f(t)\, dt$$

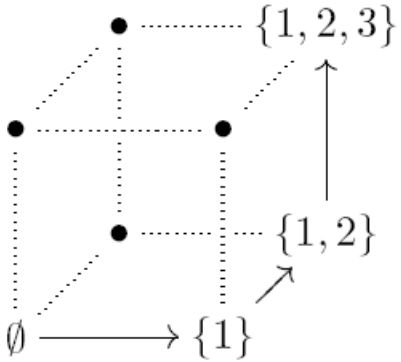
where $f(t) = 0$, unless $\sigma \in S_n$ is a linear extension of P

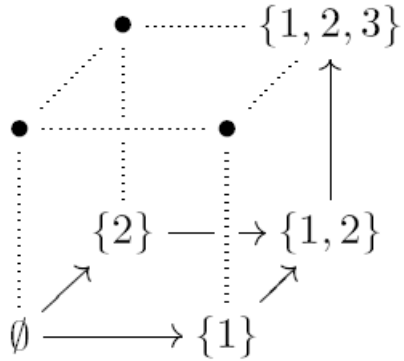
(S_n is the symmetric group on n letters)

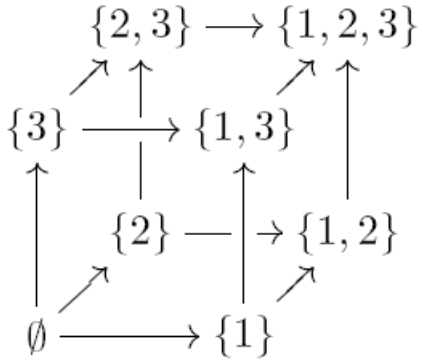
Posets define the geometry of genotype space





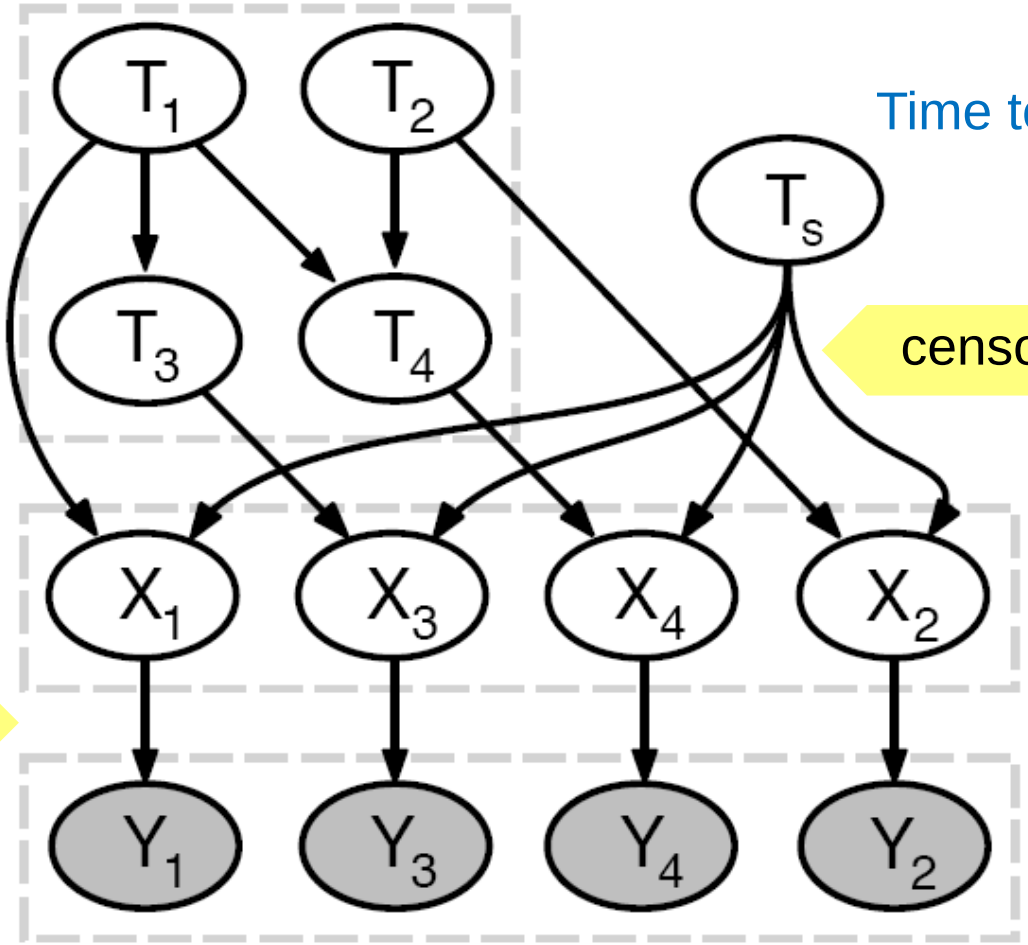






Hidden conjunctive Bayesian network (H-CBN)

Time to occurrence of mutations



Time to diagnosis

censoring

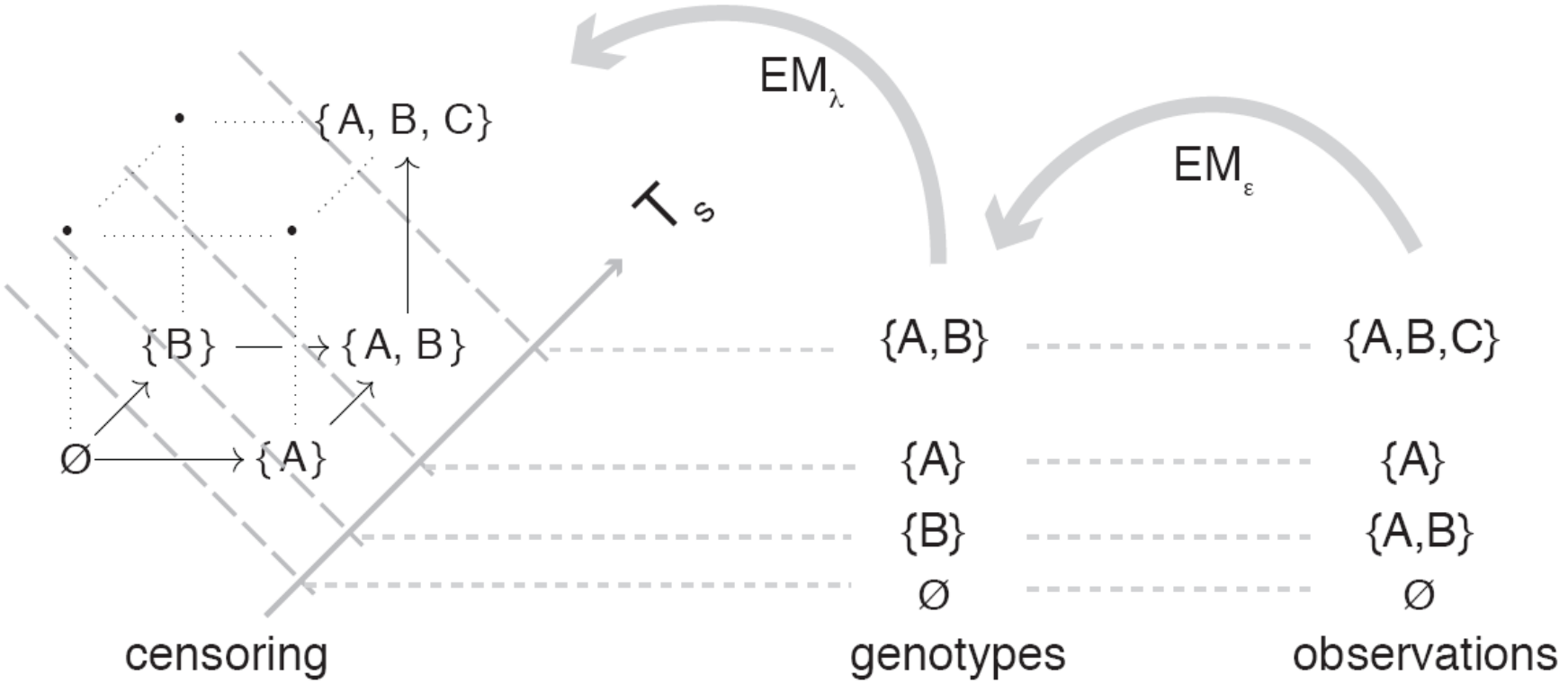
Genotype

noise

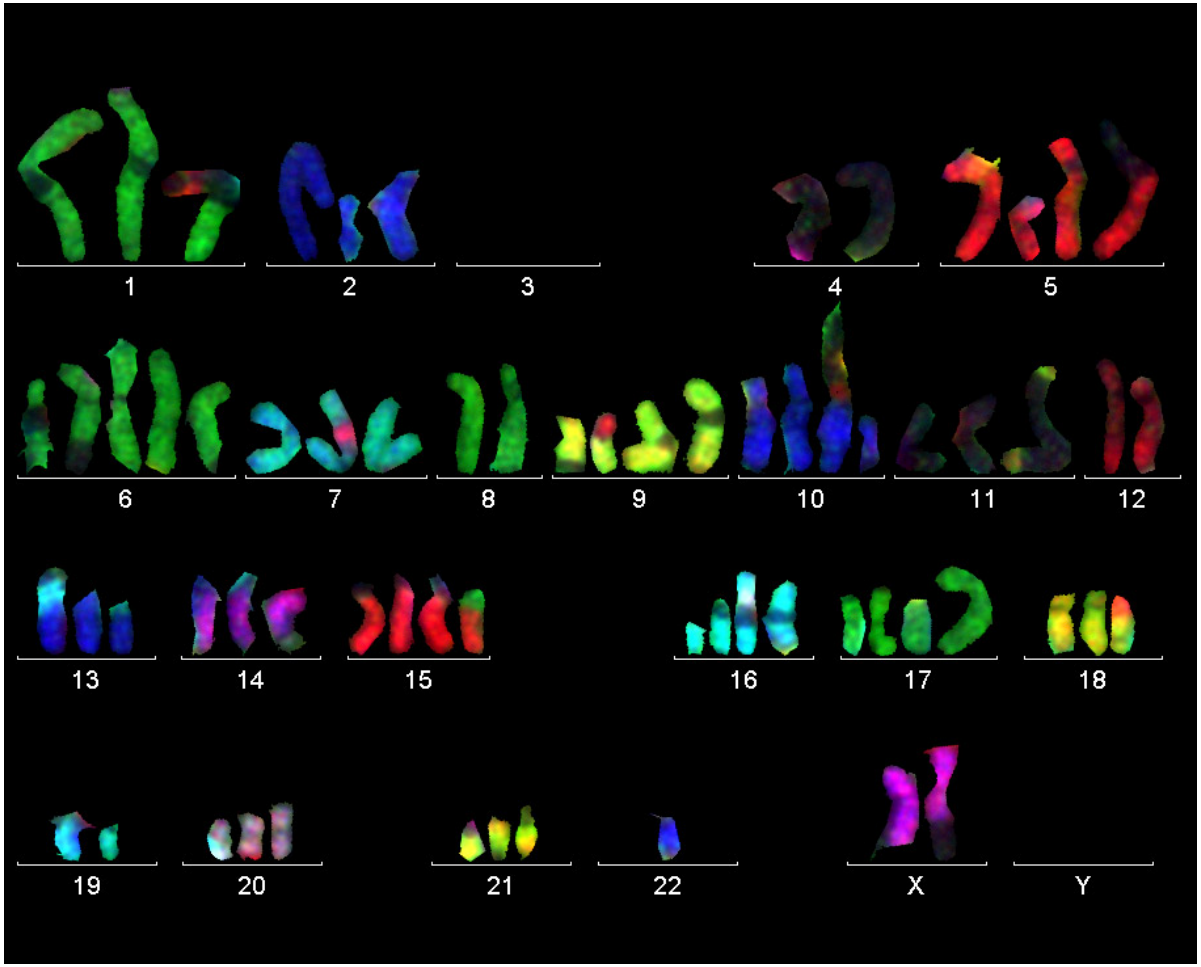
Observed genotype

Parameter estimation: Nested EM algorithm

- Hidden variables: X, T, T_s
- EM algorithm: Estimate genotypes X, then waiting times T, T_s

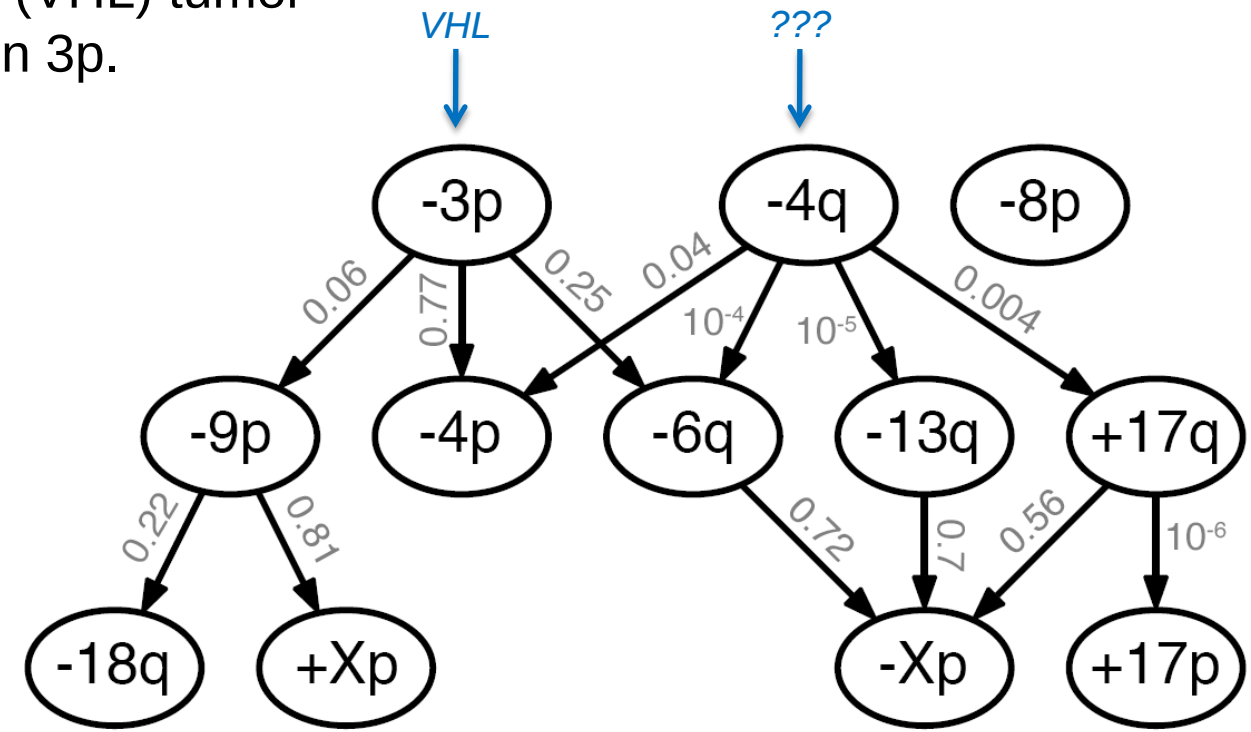


Comparative genome hybridization (CGH)



Example: Renal cell carcinoma

- N = 251 kidey cancer cases
- 12 most frequent mutations
- Von Hippel-Lindau (VHL) tumor suppressor gene on 3p.



Summary

- During tumorigenesis, mutations accumulate in the genomes of cancer cells. Oncogenetic tree models describe order constraints of this process. They can be estimated efficiently from cross-sectional data.
- Conjunctive Bayesian networks (CBNs) relax the tree assumption to a general DAG. Continuous-time CBNs explicitly model the waiting time for each mutation.

References

- Desper R et al (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol* 6:37–51.
- Beerenwinkel N et al (2005). Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol,* 12(6):584-598.
- Beerenwinkel N, Eriksson N, Sturmfels B (2007). Conjunctive Bayesian networks. *Bernoulli* 13(4):893–909.
- Beerenwinkel N, Sullivant S (2009). Markov models for accumulating mutations. *Biometrika* 96(3):645-861.