SGN-6156, Lecture 9
Modeling biological regulatory networks: Bayesian networks

Harri Lähdesmäki, harri.lahdesmaki@tut.fi

(Slides by Juha Kesseli 25.4.2007, updated by Harri Lähdesmäki)

Department of Signal Processing,

Tampere University of Technology

29.04.2008

Motivation

- Graphical models in general are a common way of representing qualitative biological information
  - E.g. regulatory interactions can be visualized by a graph in which the nodes represent genes and (directed) arcs the interactions: transcription factor A activates gene B
- Graphical models may be learned from limited data — a systematical approach of assessing the reliability is needed
- Bayesian networks provide a solution and can be used to model the interactions quantitatively as well
  - Including non-linearity and stochasticity

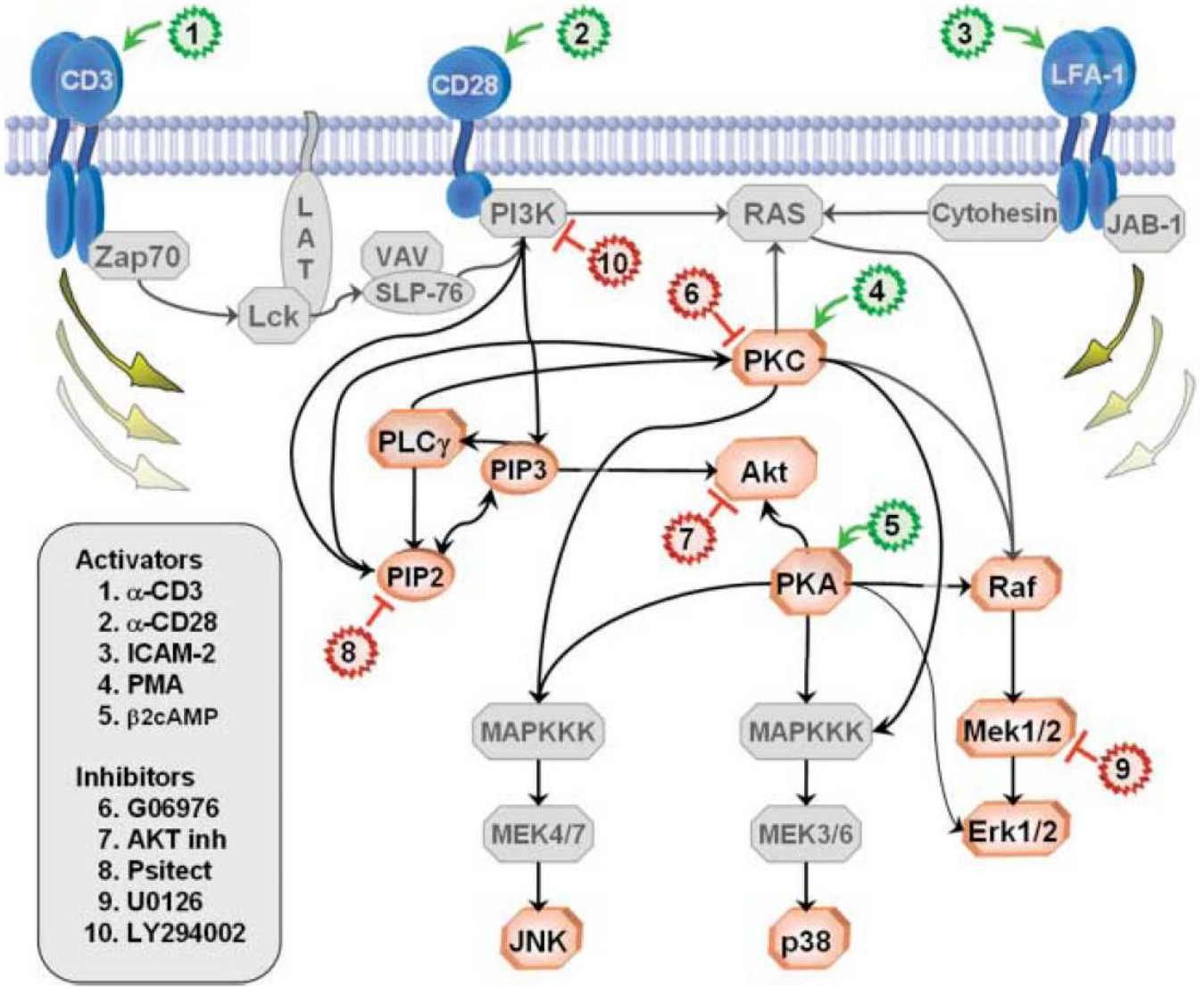


Figure from (Sachs et al., 2005)

Probability factorization

- Given a set of random variables $X = (X_1, \ldots, X_n)$, a Bayesian network is defined as a pair (S, θ), where
  - S is a directed acyclic graph (DAG), which is a graphical representation of the conditional independencies between variables in X
  - θ is the set of parameters for the conditional probability distributions of these variables
- In a Bayesian network, the probability of a state $x = (x_1, x_2, \ldots, x_n)^T$ is factored as

$$p(x) = p(x_1|\mathrm{pa}(x_1))p(x_2|\mathrm{pa}(x_2)) \cdot \ldots \cdot p(x_n|\mathrm{pa}(x_n)),$$

  where $\mathrm{pa}(x)$ denotes the parents of node x in the graph S
- This probability factorization represents the conditional (in)dependencies of the variables.

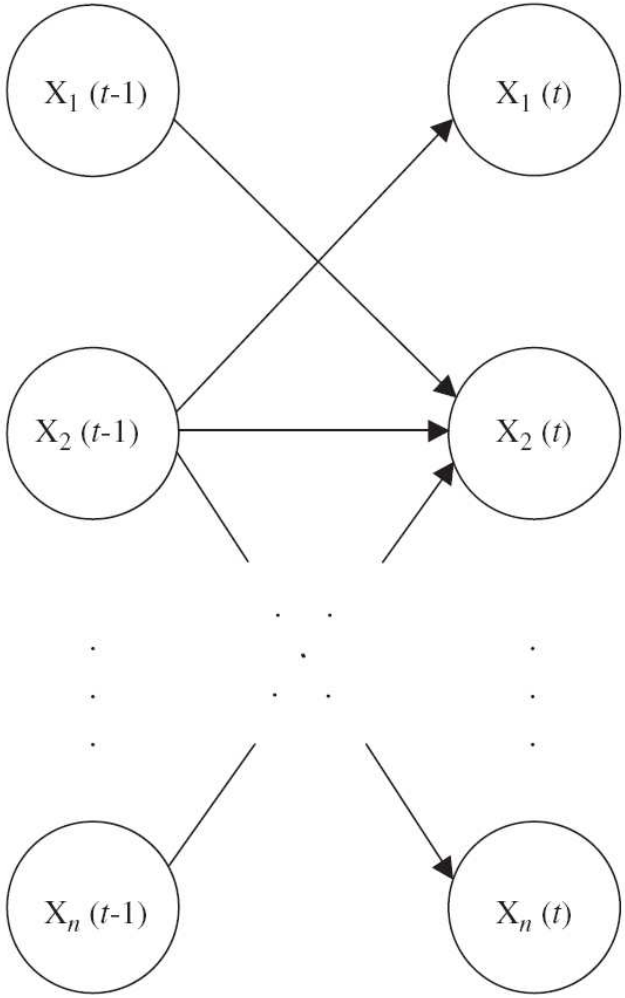
Graph modeling problems

- After observing a set of data, denoted by D, we may want to learn a graphical model
  - Estimate parameters θ for interactions of interest, given our a priori knowledge (knowledge before observing the data) about the structure (easier)
  - Estimate the structure of the network, S (more difficult)
  - Estimate both structure and parameters
- With a graphical models, we can also do inference, i.e. compute a posteriori probabilities for values of variables not seen in the data. In addition to the parameters, these could be future values in a dynamical model or variables simply not measured at all.
  - Note that in most other contexts, inference refers only to what is here called learning

Dynamic Bayesian networks

- Note that nowhere in the previous formulation was there any mention of time t
  - Bayesian networks, by default, are static — they do not consider time or causality but only conditional dependency of observations
  - Static networks, including Bayesian networks, are directed acyclic graphs (DAGs), which can be restricting
- Dynamic Bayesian Networks (DBNs) are temporal extensions of BNs, in which the probability factorization is performed for a discrete-time stochastic process $X(t) = (X_1(t), \ldots, X_n(t))^T$

- In the simplest case, we assume the process can be modeled as an unrolled version of a standard static Bayesian network
  - Parents of each node $X_i(t)$, $\mathrm{pa}(X_i(t))$, are among the nodes at the previous time slice $X(t-1)$
  - Process becomes a first order process
  - For discrete-valued networks, this corresponds to a discrete-state Markov chain
- Both static and dynamic networks can be considered for e.g. gene regulation



An illustration of the DBN model structure.

Dynamic Bayesian networks (cont.)

- In a first-order DBN, the probability factorization for a time series of length T can be written as

$$p(x(1),\ldots,x(T)) = p(x(1)) \prod_{t=2}^{T} p(x(t)|x(t-1))$$

$$= p(x(1)) \prod_{t=2}^{T} \prod_{i=1}^{n} p(x_i(t)|\mathrm{pa}(x_i(t-1))),$$

  where the parents of $x_i(t)$ show the conditional dependencies between the consecutive time steps

The Bayes formula

- Recall that the Bayes formula (Bayes' theorem) relates the conditional and marginal probabilities of events A and B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Alternatively, this can be viewed as updating prior probability $P(A)$ to posterior probability $P(A|B)$

- Similarly, in the case of two random variables x and y we have a connection between the conditional and marginal distributions (for continuous distributions as well as discrete):

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Bayesian framework

- In the Bayesian framework, both the data D and the parameters included in θ and structure S are modeled as random variables
  - Contrast with traditional estimation, where the parameters to be estimated are assumed to be unknown constants
  - The traditional approach can also be used to learn graphical models, resulting in Maximum Likelihood (ML) estimation
- We need to select probability distributions $p(S)$ and $p(\theta|S)$ to describe our a priori knowledge about the possible solutions

On learning the parameters

- The variables are independent conditioned on their parents
- In the simplest case, the conditional distributions (and their parameters) are assumed to be independent
  - The estimation problems for the parameters of each distribution are independent if we observe complete data
  - The posterior $p(\theta|D)$ of the parameters can be computed separately for each parameter
- For more complicated models, the computation of posteriors becomes more difficult

A discrete model

- Even though the amount of mRNA or protein levels, for example, can vary in a scale that is most conveniently modeled as continuous, we can still model the system by assuming that it operates with functionally discrete states
  - "activated"/"not activated" (2 states)
  - "under expressed"/"normal"/"over expressed" (3 states)
- Discretization of data values can be used to compromise between the
  - averaging out of noise
  - accuracy of the model
  - complexity/accuracy of the model/parameter learning
- Qualitative models can be learned even when the quality of the data is not sufficient for more accurate model classes

- As will be seen, with the discrete-valued observations the Bayesian network learning is relatively simple (in principle)
  - For now we assume here that the structure of the model is known

Summarizing the data

- Let N_{ijk} be the number of times we observe variable/node i in state k given parent node configuration j
- Summarize the number of total number of observations for variable i with parent node configuration j,

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$

- In frequentist setting, the well known ML estimate of multinomial probabilities is obtained by the normalized counts

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}}$$

- For the Bayesian estimation, we need a parameter prior

Dirichlet prior

- A convenient prior distribution to choose for the parameters in θ is given by the Dirichlet distribution,

$$(\theta_{ij1}, \ldots, \theta_{ijr_i}) \sim \text{Dirichlet}(\alpha_{ij1}, \ldots, \alpha_{ijr_i}).$$

- The Dirichlet distribution has PDF

$$f(\theta_{ij1}, ..., \theta_{ijr_i}; \alpha_{ij1}, ..., \alpha_{ijr_i}) = \frac{1}{B(\alpha_{ij})} \prod_{i=1}^{r_i} \theta_{ijr_i}^{\alpha_{ijr_i}-1},$$

  with $\theta_{ijr_i} \geq 0$, $\sum_i \theta_{ijr_i} = 1$, and hyperparameters $\alpha_{ijr_i} \geq 0$. α_{ij} summarizes the pseudocounts, $\alpha_{ij} = \sum_k \alpha_{ijk}$.

- The normalization constant, the Beta function, can be expressed using the gamma function,

$$B(\alpha_{ij}) = \frac{\prod_{k=1}^{r_i} \Gamma(\alpha_{ijr_i})}{\Gamma(\alpha_{ij})}$$

Conjugate prior

- The convenience arises from the fact that the distribution is conjugate to the multinomial distribution, i.e., if $p(\theta)$ is Dirichlet and $p(x|\theta)$ is multinomial, then $p(\theta|x)$ is Dirichlet as well
- The multinomial distribution is given (for $\sum_k N_{ijk} = N_{ij}$) by

$$f(N_{ij1}, \ldots, N_{ijr_i} | N_{ij}, \theta_{ij1}, \ldots, \theta_{ijr_i}) = \frac{N_{ij}!}{N_{ij1}! \ldots N_{ijr_i}!} \theta_{ij1}^{N_{ij1}} \ldots \theta_{ijr_i}^{N_{ijr_i}}$$

and is the distribution of observations in r_i classes if N_{ij} observations are selected as outcomes of independent selection from the classes with probabilities θ_{ijk}, $k = 1, \ldots, r_i$

Closed form solutions

- The a posteriori -distribution for the parameters θ_{ijk} is Dirichlet with updated hyperparameters $\alpha_{ijk} = \alpha_{ijk} + N_{ijk}$
- The maximum a posteriori and posterior mean parameter estimates are given as

$$\begin{aligned} \tilde{\theta}_{ijk} &= \frac{\alpha_{ijk} + N_{ijk} - 1}{\alpha_{ij} + N_{ij} - r_i} \\ \overline{\theta}_{ijk} &= \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}} \end{aligned}$$

- Using the Dirichlet prior we can obtain a Bayes score for the network structure analytically

Bayes scoring of networks

- In Bayesian context, the most natural score for a network structure S is the posterior probability given the observed data D:

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)},$$

  where we have made use of the Bayes formula

- Since probability $P(D)$ is not dependent on the structure, it is not needed to compare the scores of different networks

- What remains is thus

$$P(S|D) \propto P(D|S)P(S),$$

  containing a term describing our a priori knowledge of the structure and the marginal likelihood of the data which needs to be evaluated

Learning the network structure

- If we are only interested in the structures, we can obtain an analytically tractable form of the marginal likelihood (for the data given structure S):

$$\begin{aligned} P(D|S) &= \int_\theta p(D|\theta, S)p(\theta|S)\, d\theta \\ &= \dots \\ &= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \end{aligned}$$

- Efficient algorithms for finding optimal structures exist only for the simplest cases, e.g., a tree with at most one parent per node ($O(n^2 \log n)$)
- Finding the structure with maximal Bayes score is an NP hard problem even if we set a bound $k > 1$ for the maximum number of parents. Inference of variables given others is in general difficult as well

- For example, greedy optimization algorithms that change the structure towards a local optimum are often used as a heuristic solution
- Having an accurate structure makes a difference to the rest of the estimation
  - Missing edges in the model give a poor fit to data
  - Spurious edges lead to unnecessary parameters to estimate and lower estimation and predictive performance

Problems in practice

- As mentioned earlier, an exhaustive search and scoring approach for the different models will not work in practice (the number of networks increases super-exponentially, $2^{(n^2)}$ for dynamic Bayesian netwokrs
  - Heuristics are used to e.g. add parents to a node one at a time as long as the Bayesian score increases
- In addition, the case we have considered is simple in that all the variables are assumed to be observable
- Particularly in small sample settings the a posteriori -distribution may be rather flat
  - Looking for a single optimal model is not a good idea — we should consider the entire distribution, or in practice, several models with a good fit

Bayesian approach to structural properties

- In order to get more reliable results we can focus on features that can be inferred the most reliably
- for example, we can define a feature, an indicator variable f with value 1 if and only if the structure of the model contains a path between nodes A and B
- Looking at a set of models $\mathcal{S}$ with a good fit we can approximate the posterior probability of feature f by

$$P(f|D) = \sum_{S \in \mathcal{S}} f(S) P(S|D).$$

- With gene regulatory networks, one can look for only the most significant edges based on the scoring

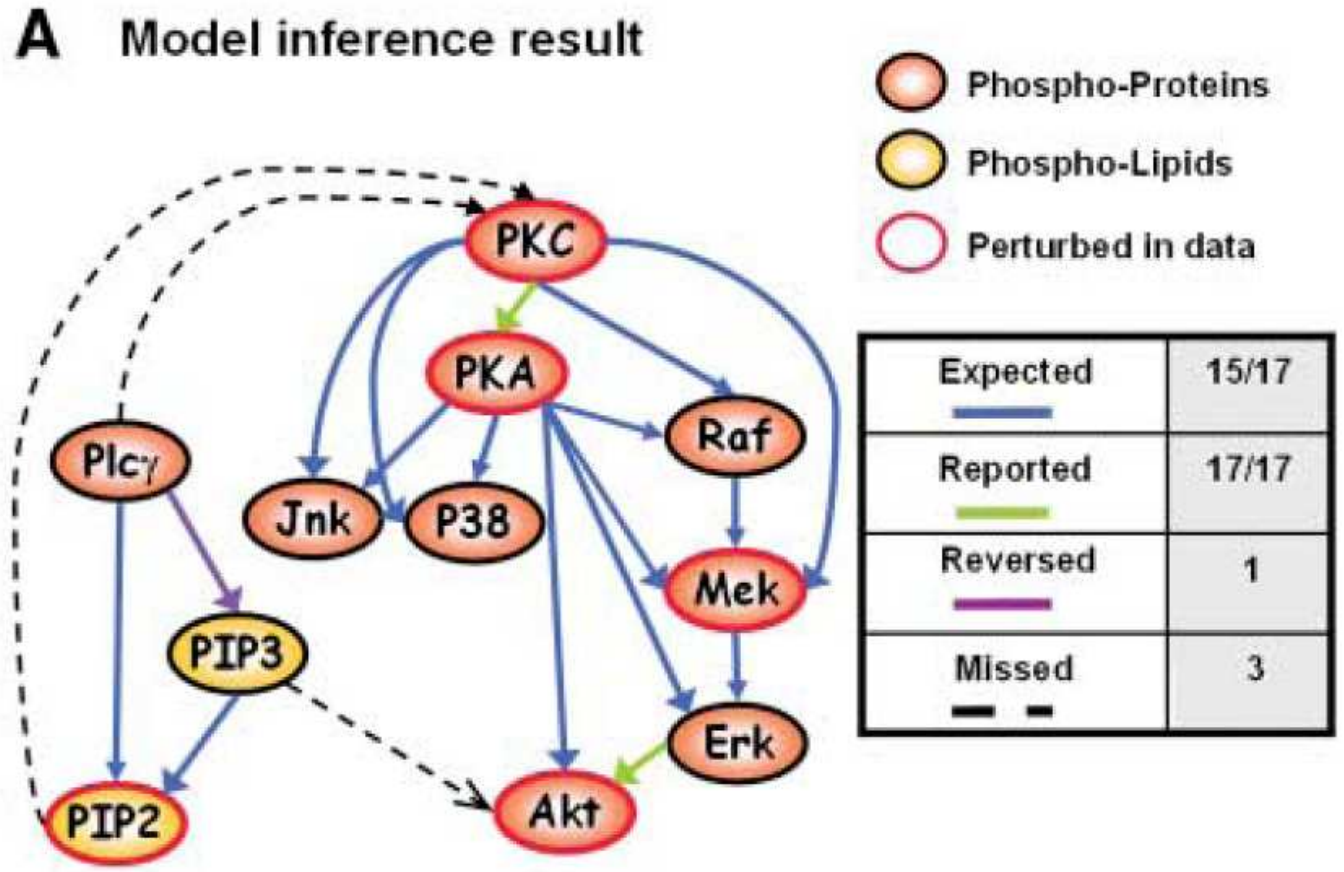


Figure from (Sachs et al., 2005)

Markov Chain Monte Carlo

- Since structures cannot be enumerated in general to compare their scores and posteriors can be difficult to compute, Markov Chain Monte Carlo (MCMC) sampling is often used
- A Markov chain is defined over Bayesian nets so that it approaches a steady-state distribution as it is being run, and the probabilities of the states (networks) correspond to their posterior probability
- Individual nets are created as states in the chain and after (assumed) convergence, samples S_i are taken
- Posterior probability of an edge can then be approximated with $P(f(S)|D) \approx \frac{1}{n} \sum_{i=1}^{n} f(S_i)$
- To get robust results (convergence of the chain), special methods need to be used. Real biological pathways have been reconstructed using Bayesian nets (with a subset of genes, hundreds of microarrays)

Hidden variables

- Hidden (non-observed) variables make the learning significantly more difficult
- Finding out hidden variables can significantly decrease the amount of parameters we need to estimate
- Incomplete data means that the marginal likelihood does not have an analytically tractable form and that the likelihood can have multiple maxima
- Expectation Maximization (EM) algorithm can be used to deal with incomplete data, iterating the following steps:
  - Generate expected data values for the hidden variables given observed data and current model parameters
  - Utilizing the complete data set thus obtained, learn parameters as with complete data

References

- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, Vol. 308, No. 5721, pp. 523-529.