# ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

---

## Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό

## Inference in Bayesian networks

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

---




## Άδειες Χρήσης

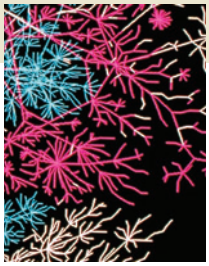
- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.


## Χρηματοδότηση


- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.




Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης





# Inference in Bayesian networks

Chris J Needham[1], James R Bradford[2], Andrew J Bulpitt[1] & David R Westhead[2]

**Bayesian networks are increasingly important for integrating biological data and for inferring cellular networks and pathways. What are Bayesian networks and how are they used for inference?**

How does one model a simple cell-signaling pathway? Consider a simple example consisting of a stimulant, an extracellular signal, an inhibitor of the signal, a G protein–coupled receptor, a G protein and the cellular response. The stimulant induces production of the extracellular signal in another cell or extracellularly. A Bayesian network can be constructed that expresses the relationships between these variables. For example:

- The stimulant may or may not generate a signal.
- The concentration of the signal may affect the level of the inhibitor.
- Whether the signal binds with the receptor depends on the concentrations of both the signal and the inhibitor.
- The G protein should become active if the receptor binds.
- An active G protein initiates a cascade of reactions that causes the cellular response.

## Conditional independence

Using this information one can identify which variables depend on which other variables and which variables are conditionally independent. If two variables are independent given the state of a third variable, then they are said to be conditionally independent. For example, consider two independent tests for a disease, $T_1$ and $T_2$. The tests are reasonably reliable, and a strong correlation is seen between $T_1$ and $T_2$. If the result of test $T_1$ is positive, it becomes more likely that $T_2$ will also be positive. However, if it is known that the person has or hasn't got the disease, then the result of $T_1$ has no effect on the expected value of $T_2$; the tests become conditionally independent.

The above relationships between the cell-signaling variables can be expressed by the graph structure shown in **Figure 1**; nodes represent variables, and the directed edges show the dependencies. Variables not connected by edges are conditionally independent. This forms a directed acyclic graph (DAG), that is, a graph with no loops. Feedback would result in a cyclic graph, for which the more general methods of probabilistic graphical models would need to be used. We will assume that all the variables are discrete, and take the following possible values:

- Stimulant (ST): present/not present
- Signal (SI): high/medium/low
- Inhibitor (IN): high/medium/low
- Receptor binds (RE): yes/no
- G protein (GP): active/not active
- Cellular response (CR): yes/no

## Conditional probability distributions

A model of the relationships between the variables can be built. In this discrete case, conditional probability tables can be formed to express the probability of the state of each variable given its parents (those it directly depends upon). For example, if the graph structure and conditional probability tables of the Bayesian network are taken to be as defined in **Figure 1**, then the conditional

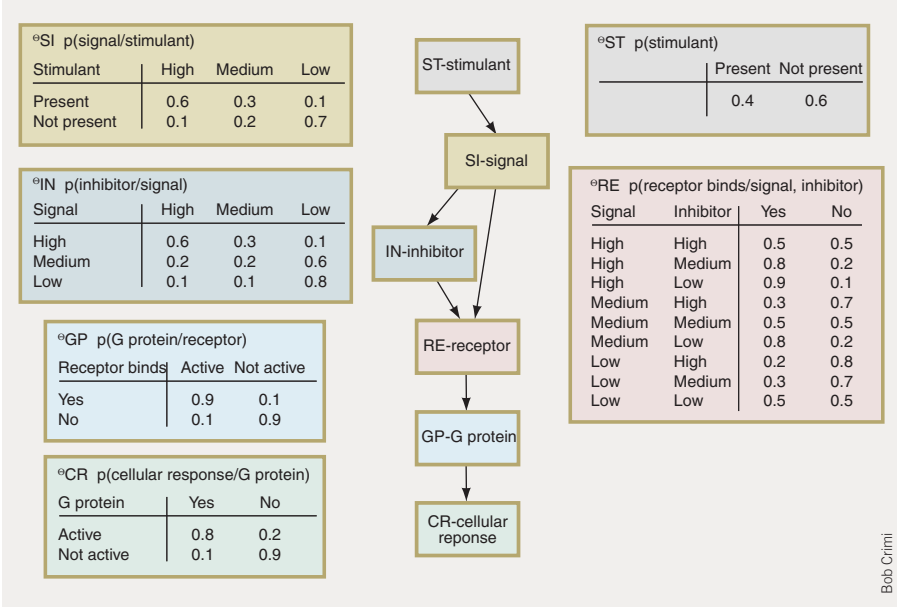



**Figure 1** Bayesian network of the cell-signaling pathway and example conditional probability tables

[1]*School of Computing, University of Leeds, Leeds, LS2 9JT, UK*
[2]*Institute of Molecular and Cellular Biology, Garstang Building, University of Leeds, Leeds, LS2 9JT, UK.*
*My email address is chrisn@comp.leeds.ac.uk*

probability that the signal is high given the stimulant is present is

$$p(SI = high \mid ST = present) = 0.6$$

(from **Fig. 1** $\theta_{SI}$) and the probability that the receptor binds given that the signal is high and the inhibitor is low (from **Fig. 1** $\theta_{RE}$) is

$$p(RE = yes \mid SI = high, IN = low) = 0.9$$

### Joint probability distributions

The joint probability distribution, p(ST,SI,IN,RE,GP,CR) can be expressed as a product of distributions over a smaller number of variables, through repeated application of the product rule of probability calculus

$$p(x, y) = p(x \mid y)p(y) \quad (1)$$

and by exploiting conditional independence relations described in the graph structure. Applying the product rule, and then using conditional independence gives equation 2 (**Box 1**). Continuing in this way for each variable, the joint probability over all the variables can be expressed as equation 3 (**Box 1**).

In the general case of Bayesian networks, consisting of a set of *n* nodes $x = \{x_1,\ldots,x_n\}$ organized in a directed acyclic graph, where each node $x_i$ has parents $\mathbf{pa}(x_i)$, the joint probability distribution is compactly expressed as

$$p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}(x_i))$$

The ability to express the joint probability in this way (exploiting conditional independencies) provides a concise representation in terms of simple component distributions (factors), thereby reducing the number of parameters to be estimated. In this example, to specify the full joint probability distribution as a joint probability table would require 143 parameters (one less than the product of the number of values each variable can take), whereas by exploiting conditional independence only 24 are required (the number of parameters in the conditional probability tables in **Figure 1** once the constraint that probabilities sum to one is taken into account). Although this may not seem that advantageous, consider a network with 100 nodes, each taking three possible values. If the graph was fully connected, the full probability distribution would require $>10^{47}$ parameters, compared with 1,800 parameters if each node had only two parents. This demonstrates just how powerful conditional independence can be. Not only is the parameter space smaller, but the parameters are easier for an expert to estimate as they involve fewer variables. So, we've seen how the joint probability distribution can be expressed. How do we use Bayesian networks for inference?

### What is the probability of the G protein being active, given that the stimulant is present?

Given evidence about the state of a variable, or set of variables, the state of other variables can be inferred. For example, to find the probability that the G protein is active given that it has been observed that a stimulant is present, that is, to find p(GP = active | ST = present), it is necessary to marginalize over the unknown parameters. This amounts to summing the probabilities of all routes through the graph, using the sum rule:

$$p(x) = \sum_y p(x,y)$$

where p(x, y) may be expanded using the product rule (equation 1). Thus the result is equation 4 (**Box 1**).

When evaluated with the conditional probabilities in **Figure 1,** p(GP=active|ST=not present)=0.592.
Note p(GP = active | ST = not present) = 0.5048.

### What is the probability the stimulant is present, given that the signal is high?

It is often of interest to calculate posterior probabilities such as the probability that the stimulant is present, given that the signal is high p(ST = present | SI = high) for which Bayes' rule may be applied:

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

Note also: $p(y) = \sum_x p(y \mid x)p(x)$

Thus the result is equation 5 (**Box 1**).

So within this neat representation of a Bayesian network, inference is easy. Inferences can be made about the value of any variable(s), given evidence about the state of other variable(s). Given a set of conditional probability distributions between a number of variables (as **Fig. 1**), the probability distributions between variables not explicitly specified may be inferred. For example, consider the prior probability that the stimulant is present p(ST = present) = 0.4. The inferred probability of the presence of a stimulant is dependent upon evidence about the other variables, for instance: p(ST = present | GP = active) = 0.44 and p(ST = present | GP = not active) = 0.35.

### Learning in Bayesian networks

The representation and use of probability theory make Bayesian networks suitable for learning from incomplete data sets, expressing causal relationships, combining domain knowledge and data and avoid overfitting a model to data. We can learn the parameters of the

## Box 1 Equations

$$p(ST, SI, IN, RE, GP, CR) = p(CR \mid ST, SI, IN, RE, GP)p(ST, SI, IN, RE, GP) = p(CR \mid GP)p(ST, SI, IN, RE, GP) \quad (2)$$

$$p(ST, SI, IN, RE, GP, CR) = p(CR \mid GP)p(GP \mid RE)p(RE \mid SI,IN)p(IN \mid SI)p(SI \mid ST)p(ST) \quad (3)$$

$$p(GP = active \mid ST = present) = \sum_x \sum_y \sum_z p(GP = active \mid RE = x)p(RE = x \mid IN = y, SI = z)p(IN = y \mid SI = z)p(SI = z \mid ST = present) \quad (4)$$

$$p(ST = present \mid SI = high) = \frac{p(SI = h \mid ST = p)p(ST = p)}{p(SI = h \mid ST = p)p(ST = p) + p(SI = h \mid ST = n)p(ST = n)} = \frac{0.6 \times 0.4}{0.6 \times 0.4 + 0.1 \times 0.6} = 0.8$$
(p, present; n, not present; h, high). (5)

Bayesian network from data. For example, the conditional probability tables could be constructed from empirical evidence. The parameters need not be discrete (as in the toy example), but may also be continuous and be modeled by a probability density function (commonly Gaussian distributions are used).

Structure learning, that is, learning the connections, is also possible. When the structure of the Bayesian network is unknown (that is, cannot be specified by prior knowledge), a heuristic search may be performed to look for 'good' structures. To learn the underlying causal model, one needs more than just structure learning, as many network structures are equivalent. To learn causal relationships between pairs of variables, patterns of dependency in the presence of a third variable must be observed in the context of interventions (fixing the values of particular variables).

Learning in Bayesian networks may use a point estimate of the parameters or Bayesian statistics[1] to average over possible model structures and parameters to provide an estimate of the posterior distribution of the variables, which avoids overfitting to the data, which may be noisy, limited, incomplete and uncertain. Heckerman[2] and Husmeier[3] have written in-depth tutorials on the subject.

## Applications in computational biology

Many applications in computational biology have taken advantage of Bayesian networks or, more generally, probabilistic graphical models. These include protein modeling, gene expression analysis, inferring cellular networks[4] and pathway modelling[5], biological data integration, protein-protein interaction and functional annotation, DNA sequence analysis, and genetics and phylogeny linkage analysis[6]. A recent application in computational biology is the inference of a Bayesian network to model a protein-signaling network from flow cytometry data[5]. This is an example of simultaneous observation of multiple signaling molecules in many thousands of cells in the presence of stimulatory cues and inhibitory interventions, which is necessary for identifying causal networks and potentially useful for understanding complex drug actions and dysfunctional signaling in diseased cells.


1. Eddy, S.R. What is Bayesian statistics? *Nat. Biotechnol.* **22**, 1177–1178, 2004.
2. Heckerman, D. A tutorial on learning with Bayesian networks. in *Learning in Graphical Models* (ed. Jordan, M.I.) 301–354 (Kluwer Academic, Boston 1998).
3. Husmeier, D., Dybowski, R. & Roberts, S (eds.). *Probabilistic Modeling in Bioinformatics and Medical Informatics* (Springer, London 2005).
4. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805, 2004.
5. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. & Nolan, G.P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529, 2005.
6. Beaumont, M.A. & Rannala, B. The Bayesian revolution in genetics. *Nat. Rev. Genet.* **5**, 251–261, 2004.