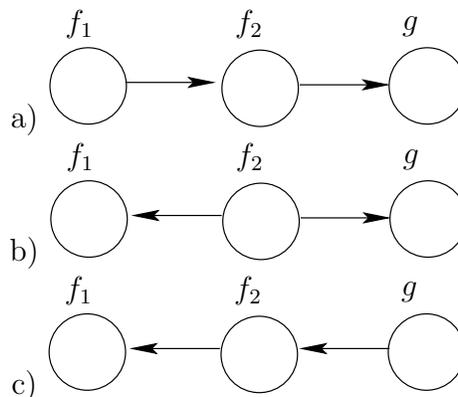


Acausal and causal models

Bayesian networks and causality

Probability models such as Bayesian networks do not inherently capture causal relations between variables. Bayesian networks, for example, attempt to model how variables depend on (are independent of) each other, not whether one causes the other. Saying that two expression profiles are correlated is obviously not a causal statement. Nevertheless, due to the fact that Bayesian networks are represented by acyclic directed graphs, there's often a temptation to interpret the graphs causally. We will begin here by discussing briefly why this is not appropriate in general.

Suppose our variables represent expression changes from one experiment to another and are discretized as $-1, 0, 1$ (down-, unchanged-, or up-regulated relative to the control, respectively). Suppose f_1 and f_2 are genes corresponding to known transcription factors and g is a particular gene (ORF). The graph that best models the data pertaining to these three proteins might look like figure a) below:



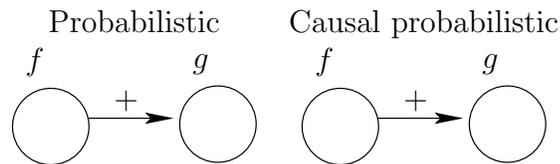
The Bayesian networks in figures b) and c) are, however, probabilistically equivalent – the graphs constrain the associated probability distribution over the three variables in exactly the same way (assuming we have imposed no prior constraints on how the variables can depend on each other). As a result, all three graphs would yield the same score based on the available data. While we cannot distinguish the three graphs as probability models, their (post-hoc) causal interpretations are radically different.

Note that this simple example holds for any type of data we might have available, including gene knock-outs that in principle could (should) be interpreted causally. We have to be

more careful in how Bayesian networks are estimated from data in order to ensure that the arrows can be interpreted causally.

Causal and non-causal probability models

Contrast: Let's start by contrasting predictions from two identically probability models that differ only in terms of whether we interpret the arrows causally:



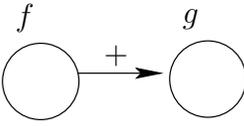
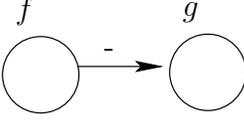
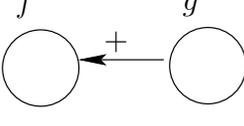
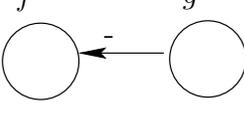
- | | | |
|---------------------------|------------------|---------------|
| 1) observational queries: | same predictions | |
| 2) knock-out of f : | g down | g down |
| 3) knock-out of g : | f down | f unchanged |

where by observational queries we mean questions like “If I see g down regulated in a particular experiment, what can I say about f ?”.

The models make different predictions only when we are interested in the parents (natural causes) in the graph following an intervention. By knocking out g we obviously bypass the natural mechanism where the expression of g is controlled by f . We shouldn't therefore try to “explain” the low expression of g as coming from changes in f . The causal probability model takes this into account appropriately. The acausal model, on the other hand, uses the estimated dependence between the two variables to answer a different (non-causal) question: “what can I say about f in cases where the expression of g is low?”.

Simple causal models

Consider the following set of simple causal models. We assume that the models are qualitative in the sense that there's little uncertainty in the predictions that they make in response to specific interventions such as knocking out g (setting $g = -1$).

model	prior probability	causal graph	prediction under $\text{set}(g = -1)$
M_0	$P(M_0) = 1/2$		f unchanged (0)
M_1	$P(M_1) = 1/8$		f unchanged (0)
M_2	$P(M_2) = 1/8$		f unchanged (0)
M_3	$P(M_3) = 1/8$		f down (-1)
M_4	$P(M_4) = 1/8$		f up (+1)

There are three problems with have to be able to solve with these models.

1. Represent prior uncertainty about the possible causal models. This is captured by the distribution $P(M_i)$ and is simple enough to define in the presence of only a few alternative models. The step is more challenging with large scale models.
2. Make predictions in light of the model uncertainty. In other words, given any specific experiment we should be able to provide a probability distribution over possible outcomes. This distribution is affected by both the model uncertainty as well as probabilistic predictions from individual models (our models are deterministic).
3. Revise the distribution over possible models based any specific experiment and observed outcome.

We will now solve the second and the third problem; a possible distribution $P(M_i)$ is already provided in the above table. Since the simple causal models make deterministic predictions

in response to knocking out either f or g , the only uncertainty in our predictions comes from not knowing which model is correct. So, for example, if we intervene and knock out g , formally $\text{set}(g = -1)$, we can calculate the predicted responses as follows:

$$P(f = 0|\text{set}(g = -1)) = P(M_0)P(f = 0|\text{set}(g = -1), M_0) + \dots + P(M_4)P(f = 0|\text{set}(g = -1), M_4)$$

where $P(f = 0|\text{set}(g = -1), M_i)$ is either zero or one depending on whether the deterministic prediction from model M_i agrees with $f = 0$. Quantitatively

$$P(f = 0|\text{set}(g = -1)) = \frac{1}{2} \cdot 1 + \dots + \frac{1}{8} \cdot 0 = \frac{3}{4}$$

We can similarly evaluate $P(f = -1|\text{set}(g = -1))$ and $P(f = 1|\text{set}(g = -1))$.

The third and the last problem can be solved simply by Bayes rule. Suppose the outcome of the experiment $\text{set}(g = -1)$ is indeed $f = 0$. Then

$$P(M_i|\text{set}(g = -1), f = 0) = \frac{P(M_i)P(f = 0|\text{set}(g = -1), M_i)}{P(f = 0|\text{set}(g = -1))}$$

where we have already evaluated the denominator. Only the first three models are consistent with $f = 0$ under experiment $\text{set}(g = -1)$ and thus

$$\begin{aligned} P(M_0|\text{set}(g = -1), f = 0) &= \frac{1/2 \cdot 1}{3/4} = \frac{2}{3} \\ P(M_1|\text{set}(g = -1), f = 0) &= \frac{1/8 \cdot 1}{3/4} = \frac{1}{6} \\ P(M_2|\text{set}(g = -1), f = 0) &= \frac{1/8 \cdot 1}{3/4} = \frac{1}{6} \end{aligned}$$

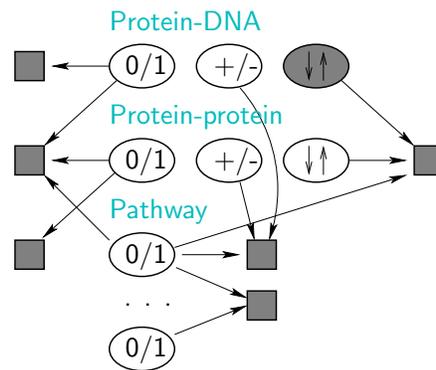
and the remaining posterior probabilities are zero. The number of viable models can therefore rapidly shrink in response to judiciously chosen experiments.

Large scale molecular interaction models

We can extend the reasoning presented in the simple case to models that attempt to capture annotated molecular interactions graphs (protein-protein or protein-dna and their properties) on a genomic scale. We will discuss here briefly how the three problems are solved in this case.

Problem 1. We can specify a probability distribution over possible annotated molecular interaction graphs by decomposing this distribution as a product of probabilities pertaining to individual features. So, for example, in the absence of any information about whether a factor is an activator or repressor of a gene, we would assign probability $1/2$ to both $+$ and $-$ annotation of the corresponding protein-dna interaction. The interaction itself has a certain probability of being present and so on.

In addition to the annotations we have to also provide ways of linking such causal models to the data. As before, the effect of a single knock-out is assumed to propagate downstream of (oriented) interactions but the path(s) that are active in terms of carrying the effect need to be chosen. Thus the model involves also variables pertaining to (a priori) uncertain selection of active pathways. Only pathways related to actual experiments need to be considered. In the figure below, the ovals represent different types of variables needed to specify this type of annotated interaction graphs (and how they relate to data); the square nodes indicate available evidence or constraints about the values of the variables (e.g., interactions along a pathway have to exist and need to be oriented appropriately).



Problem 2. The solution is formally the same as before but cannot be in general solved exactly due to the large number (and the type) of possible interaction graphs. There are however a number of approximate algorithms (e.g., belief propagation) that can be used to generate predictions from this class of models, i.e., effectively carrying out the summation over the large number of possible models.

Problem 3. This step is again formally identical but carried out approximately in a distributed manner. In other words, we don't necessarily need to know the posterior probability of each possible graph but rather the probabilities that different features are present (a specific protein-protein interaction, a factor functions as an activator, etc). Such prob-

abilities are averages relative to the full posterior. For example:

$$P(\text{Kss1} \rightarrow \text{Ste12}|\text{data}) = \sum_{\text{models } M_i \text{ for which } \text{Kss1} \rightarrow \text{Ste12}} P(M_i|\text{data})$$

These marginal probabilities are often easy to evaluate approximately.