BY KAREN SACHS

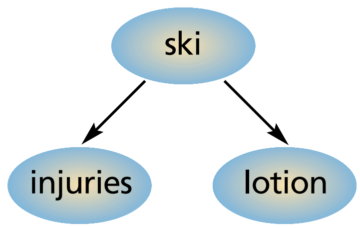
Bayesian Networks: A Quick Intro

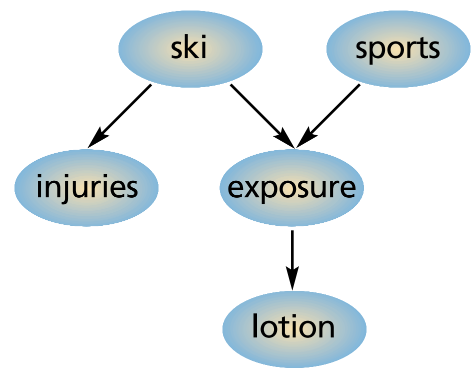
Advances in technology have brought to molecular biology datasets that are bigger, more sophisticated, and, unfortunately, more difficult to interpret than ever before. One computational analysis approach is called Bayesian networks, a machine learning tool that is able to automatically discover networks of dependencies and causal interactions among biomolecules of interest.

Bayesian networks are a form of graphical modeling, in which dependencies among variables are depicted in a graph, with the nodes representing variables (e.g., biomolecules such as proteins), and the edges (arrows) representing dependencies. Dependencies are *statistical* in nature, so an edge from **A** to **B** indicates that knowing **A** can help us predict **B**. This may or may not indicate a *causal* relationship, i.e. one in which **A** (directly or indirectly) affects **B**. Interventional data, in which biomolecules are specifically manipulated, can be used to discover causal connections.

The Bayesian network inference algorithm takes data in which biomolecules were quantified (and, ideally, also manipulated), and automatically reconstructs the underlying network of protein to protein influences that may have created the data.

How does this process work? Consider a ski resort, with skiers and non-skiers (hot-tub sitters). A study discovers a strong statistical correlation between sunscreen lotion use and skiing injuries. To further investigate this statistical dependency, a manipulation is performed on the **lotion** variable: all sunscreen lotions are secretly replaced with an ineffective placebo. This fails to affect the number of skiing **injuries**, and so it is determined that **lotion** use does not causally affect skiing **injuries**. The variable **ski** is also well-correlated. When the **ski** variable is manipulated (the ski slopes are closed for a day), both **lotion** use and **injuries** are greatly reduced or eliminated, thus implicating **ski** as the variable causally responsible for the other two. The study now expands to include a tropical island. The correlation between skiing and sunscreen use is weakened; however, when tropical **sports** are included in the study, the **ski** and **sports** variables together are able to predict sunscreen **lotion** use. If sun **exposure** is also included, it is found to be well predicted by **ski**ing and tropical **sports**, and is itself a good predictor of **lotion** use.





The Bayesian network works much like this example, examining correlations, determining which variables can be used to predict which other variables, and relying heavily on interventional data to determine causal connections and the directionality of node-to-node connections. It is able to find complex relationships beyond simple correlations; it can handle indirect relationships (e.g., **ski_lotion**, when **exposure** is not measured); and it can eliminate unnecessary edges (**ski_lotion** when **exposure** *is* measured). Therefore, it is potentially able to automatically construct a network much like the canonical pathways sketched out in biology text books. Our recent work (Sachs et al., 2005) shows an application of this approach to signaling proteins measured in single cells, demonstrating the ability of Bayes nets to find a first order map of a signaling pathway, and serve as an *in silico* generator of testable hypotheses. □

DETAILS

Karen Sachs is a graduate student in the Biological Engineering department at MIT, working on computational modeling of biological systems in Doug Lauffenburger's lab.