Cause and effect

The experimental tractability of biological systems makes it possible to explore the idea that causal relationships can be estimated from observational data.

"Happy is he who is able to know the causes of things."

—Virgil

The idea that one needs to do an experiment—a controlled perturbation of a single variable—to assign cause and effect is deeply embedded in traditional thinking about the way scientific knowledge is obtained, but it is largely absent from everyday life. One knows, without doing an experiment, that the street is wet on a rainy day because the rain has fallen. To be sure, this form of causal reasoning requires prior knowledge. One has seen the co-occurrence of rain and the wet street many times and been taught that rain causes wetness. And although such relationships are, in the strict sense, merely very good correlations, human beings routinely, and necessarily, use them to assign cause and effect.

As discussed on this page a year ago, this form of thinking, at least as a starting point for hypothesis-making, is in practice not uncommon in scientific research as well. Even before our data-driven age, a testable idea often began with an observation. When the structure of a voltage-gated potassium channel was first solved, for instance, the physical basis for potassium selectivity was suggested from observing the disposition of the residues known to allow potassium, but not sodium, ions to pass. In another example a century or so earlier, Ramón y Cajal famously predicted many features of the operation of the nervous system, including the directionality of neuronal signaling, based on his observations of the organization of neurons in the brain. Experiments had to be designed to test these ideas, but the hypotheses about cause and effect were generated at least in part by observation.

Many areas of contemporary biology seek to learn causal relationships from biological data. In systems biology, for instance, researchers use measurements of gene expression, cellular protein amounts or metabolite levels, among other types of data, to assign causal or regulatory relationships in models describing the cell. In the context of large-scale systems data, it is usually not possible to assign such relationships just by looking at the data by eye. Statistical and visualization tools are needed, when, for example, one is looking at lists of expression data of thousands of genes and trying to determine which genes regulate what other genes. The methods used to assign causal arrows typically involve perturbation experiments. When unperturbed data are used, additional information such as change over time or prior biological knowledge has been used to order the data.

A Correspondence by Maathuis and colleagues published in this issue (p. 247), in contrast, explores the notion that it might be possible to estimate causal relationships simply by observing random variation in unperturbed data, with no other information added. Making use of gene expression data obtained either from single gene knockouts in yeast—a classical perturbation experiment—or from parallel control measurements on wild-type yeast, an unperturbed system in which there is presumably only random variation, the authors report that, under some assumptions, statistical analysis can be used to predict the strongest causal effects from the control data alone.

The idea that such prediction is theoretically possible is not in itself new and has received some interest in, among others, the social scientific, economic and medical spheres. But it is an idea that is not easy to test in a real-world setting. In a sense, then, the study in this issue exploits the unique properties of biological systems—their complexity, the availability of good tools for precise and ethical system manipulation, and the well-developed technology for acquiring large-scale unbiased data—to test an idea that could have interest and value outside the biological realm as well.

It is worth noting that the assumptions made—in its current iteration, the approach by Maathuis and colleagues provides no allowance for feedback and does not incorporate change over time—could pose serious obstacles for understanding biological as well as other systems. What is more, statistical inference will clearly not replace perturbation experiments in systems that are amenable to manipulation.

Nonetheless, causal inference from purely observed data could have practical value in the prioritization and design of perturbation experiments. Perturbations can be impossible, for instance, if the tools available are not specific enough, unethical, for example in human studies, or simply unfeasible owing to cost or impracticality. Observational data could be used to identify candidate causal relationships, which could then be the basis for the design of targeted perturbations or for further analysis.