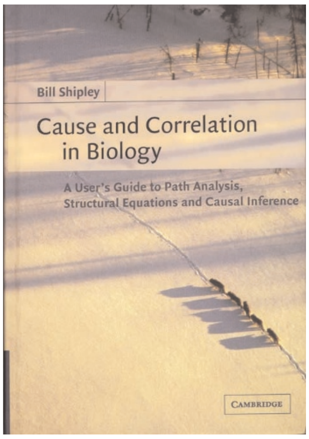
Conditional truths



Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference
by Bill Shipley

Cambridge University Press • 2001 (2nd edition)
Hardback, £45/$70

Johan Paulsson

Between 1920 and 1970, the Dow Jones Industrial Average rose and fell with the lengths of skirts in fashion magazines. As skirts became shorter, stock prices shot higher. Glamorous as this may sound, it does not imply that designers and investors influenced one another — they could also have been subject to the same external cues. In statistical terms, correlation does not imply causation.

Most of us accept this claim as logically indisputable, but we might have to reconsider its practical value. Statistical patterns can allow us to make educated guesses about causal mechanisms that otherwise would remain obscure. The pragmatic question is then not if correlation implies causation, but how. How should experiments be designed and models be tested when we only see the statistical shadows cast by the underlying processes? These questions are sharpened and to some extent answered in Bill Shipley's *Cause and Correlation in Biology*.

Most experiments in molecular biology are set up so that sophisticated theory is superfluous. By varying one factor and keeping the others as constant as possible, one hopes for straightforward interpretations. Randomization, as when assigning drugs or placebos to subjects of a clinical study, solves the same problem: it can filter out known and unknown factors in one sweep and with a statistical confidence that can be calculated in advance. The real problem arises when your experiments can be neither controlled nor randomized. In some cases the solution is to exercise statistical control by organizing the data set: conditioning on a variable can mimic the effect of physically keeping it constant. For instance, if A and B are independent apart from a shared dependence on C, conditioning on C reveals their independence. By contrast, if A and B both affect C, conditioning on C will make them observationally dependent even if they are causally independent.

The examples above illustrate the idea behind Shipley's D-separation test. Starting with a box-and-arrow causal model, the test derives an exhaustive list of implied conditional independencies. Because independent variables should be uncorrelated, the list can be checked with standard statistical methods. The great appeal of the approach is that virtually all details can be left unspecified. The shapes of distributions are irrelevant and only the existence or absence of causal effects matters. As Shipley proceeds to more sophisticated methods, the scope narrows accordingly. Structural equations modelling (SEM), the main topic of the book, typically assumes linear associations between normally distributed variables. The basic idea is to translate a causal graph into a full probabilistic model, derive the covariances and statistically compare model predictions with observed data. So how is this unusual? Entire disciplines are already occupied with explaining statistical fluctuations in terms of probabilistic mechanisms. Even if few of these models are drawn as graphs, they also reflect causality relations and can be tested statistically. The justification for SEM is that its idealizations give you something in return. Most non-systematic models are developed without descriptions for how they can be rigorously tested, and data is typically handled with little concern for the physical phenomena. SEM takes the unusual and commendable stance of actively merging probabilistic modelling with statistical data analysis. By keeping within modelling limitations, the model is automatically suitable for rigorous statistical tests, including systematic treatments of unmeasured variables, measurement errors and hierarchically structured data. The book concludes with algorithms for discovering models that fit experiments. The strategy is now the inverse. Known statistical correlations are used to generate a list of conditional independencies that, using the notion of d-separation, in turn can be used to infer causal graphs. An inconclusive outcome of the algorithm indicates that an unmeasured variable has a significant influence on a process, something that can also be tested statistically. The methods can, therefore, not only track down models in a given set of variables, but also detect hidden variables.

The methods presented are relevant in all fields of biology, but perhaps not as urgently as in ecology and biometrics where Shipley picks his examples. Most cell- or molecular-level experiments are designed to look at averages or qualitative features, whereas statistics is used mainly to estimate measurement errors. This could be about to change. Probabilistic modelling is up-and-coming and high-throughput methods are generating a wealth of statistical data. There will then be a greater demand for evaluation tools, but the future use of SEM is still not obvious. Much of theoretical biology focuses on dynamics, and probabilistic models are often formulated using birth-and-death processes or stochastic differential equations. The book barely mentions time and instead presents a static picture where variables are expressed as sums of other variables. Shipley points out that static models could describe equilibrium states, but even these are often more attainable from dynamic models where the change is set to zero.

Addressing students and practising biologists, Shipley does a terrific job of making mathematical ideas accessible. He introduces all methods from scratch, walks through realistic examples and generously supplies historical and philosophical notes. The writing style is overall enthusiastic and playful, but some parts feel half-baked with overly detailed anecdotes. What the book lacks in austerity, it makes up for by an unusual intellectual curiosity and diversity. *Cause and Correlation in Biology* is a nontechnical and honest introduction to statistical methods for testing causal hypotheses. □

Johan Paulsson is in the Department of Molecular Biology, Princeton University, Princeton, New Jersey, 08544, USA
email: paulsson@princeton.edu