Defining the scientific method

The rise of ‘omics’ methods and data-driven research presents new possibilities for discovery but also stimulates disagreement over how science should be conducted and even how it should be defined.

> “Hypotheses aren’t simply useful tools in some potentially outmoded vision of science; they are the whole point.”
>
> Sean Carroll

> “Science, it turns out, is whatever scientists do.”
>
> David Goodstein

Modern biological research methods are powerful tools in biologists’ arsenal for investigating biology. But is the ability of these methods to amass extraordinary amounts of data altering the nature of scientific inquiry?

As schoolchildren we are taught that the scientific method involves a question and suggested explanation (hypothesis) based on observation, followed by the careful design and execution of controlled experiments, and finally validation, refinement or rejection of this hypothesis. Developed by thinkers including Bacon, Descartes and Pierce, this methodology has been credited with much of science’s success. Modern philosophers such as Feyerabend have argued that this is not how most science is conducted, but up to now most modern scientists have subscribed to the hypothesis-centric scientific method.

Scientists’ defense of this methodology has often been vigorous, likely owing to the historic success of predictive hypothesis-driven mechanistic theories in physics, the dangers inherent in ‘fishing expeditions’ and the likelihood of false correlations based on data from improperly designed experiments. For example, The Human Genome Project was considered by many at the time to be a serious break with the notion that proper biological research must be hypothesis-driven. But the project proceeded because others successfully argued that it would yield information vital for understanding human biology.

Methodological developments are now making it possible to obtain massive amounts of ‘omics’ data on a variety of biological constituents. These immense datasets allow biologists to generate useful predictions (for example, gene-finding and function or protein structure and function) using machine learning and statistics that do not take into account the underlying mechanisms that dictate design and function—considerations that would form the basis of a traditional hypothesis.

Now that the bias against data-driven investigation has weakened, the desire to simplify ‘omics’ data reuse has led to the establishment of minimal information requirements for different types of primary data. The hope is that this will allow new analyses and predictions using aggregated data from disparate experiments.

Last summer, the editor-in-chief of *Wired*, Chris Anderson, went so far as to argue that biology is too complex for hypotheses and models, and that the classical scientific method is dead. Instead, he called for these methods to be replaced by powerful correlative analyses of massive amounts of data gathered by new technologies similar to how Google Translate relies on only correlative analyses of documents on the internet.

This generated quite a response from the scientific community with California Institute of Technology physicist Sean Carroll arguing in *Edge* that “hypotheses aren’t simply useful tools in some potentially outmoded vision of science; they are the whole point. Theory is understanding, and understanding our world is what science is all about.”

Is the generation of parts lists and correlations in the absence of functional models science? Based on the often accepted definition of the scientific method, the answer would be a qualified no. But not everyone would agree. Carroll’s colleague, David Goodstein, previously stated in a Thesis article in *Nature Physics* that “science, it turns out, is whatever scientists do.” A philosopher would find this to be a circular and unfulfilling argument, but it is likely that many biologists who are more interested in the practical outcomes of their methods than their philosophical underpinnings would agree with this sentiment.

But the rise of methodologies that generate massive amounts of data does not dictate that biology should be data-driven. In a return to hypothesis-driven research, systems biologists are attempting to use the same ‘omics’ methods to generate data for use in quantitative biological models. Hypotheses are needed before data collection because model-driven quantitative analyses require rich dynamic data collected under defined conditions and stimuli.

So where does this leave us? It is likely that the high complexity of biology will actually make full biological understanding by purely correlative analysis impossible. This method works for Google because language has simple rules and low complexity. Biology has neither constraint. Correlations in large datasets may be able to provide some useful answers, but not all of them.

But ‘omics’ data can provide information on the size and composition of biological entities and thus determine the boundaries of the problem at hand. Biologists can then proceed to investigate function using classical hypothesis-driven experiments. It is still unclear whether even this marriage of the two methods will deliver a complete understanding of biology, but it arguably has a better chance than either method on its own.

Philosophers are free to argue whether one method is science and the other is not. Ultimately the public who funds the work and the biologists who conduct it want results that will materially impact the quality of life regardless of what the method is called.