Bioinformatics and discovery: induction beckons again

John F. Allen

With the flood of information from genomics, proteomics, and microarrays, what we really need now is the computer software to tell us what it all means. Or do we?

Introduction

In the life sciences, there has recently been a strong resurgence of the view that there is a direct route from observation to understanding. By this route, knowledge can flow securely from data without the human and fallible intervention of guesswork, imagination or hypothesis. Information technology now puts oceans of data at our immediate disposal, and even the ubiquitous personal computer can process and analyse these data at huge speed. Surely, the thinking goes, we can now expect computer programs to derive significance, relevance and meaning from chunks of information, be they nucleotide sequences or gene expression profiles. A *Nature* editorial,[1] for instance, discusses biologists' increasing reliance on computers to do their thinking for them. The editorial is rather kind to the biologists. Its title—"Can biological phenomena be understood by humans?"—provocatively implies that scientific discovery might well be carried out, instead, by machine. In contrast with this view, many are convinced that no purely logical process can turn observation into understanding. We owe this conviction, first and foremost, to the work of Karl Popper. [2–4] Here I argue that Popper was correct, and outline the way in which I think his philosophy applies to bioinformatics. I predict that even the formidable combination of computing power with ease of access to data cannot a produce a qualitative shift in the way that we do science: the making of hypotheses remains an indispensable component in the growth of knowledge.

The problem of induction

"Logical deduction" is a process by which the truth of a general statement entails the truth of a particular statement. For example, if it is true that "all men are mortal", then we can deduce from the statement "Socrates is a man" that "Socrates is mortal". The reverse process, a logical route from the particular to the general, has been called "logical induction", but it has never been clear how this might work. The possibility of logical induction was dismissed by the Scottish philosopher David Hume, in the eighteenth century.[5] One of Hume's concerns was the idea of causality — how can we know that "*a*" causes "*b*", when all we can say with certainty is that we have observed that "*b*" follows "*a*" on a number of occasions? How many times do we have to observe that "*b*" follows "*a*" in order for us to be sure that "*a*" causes "*b*"? Hume's answer is that we never can be sure. And what are we doing when we make predictions about future events? For example, why do we believe that the sun will rise tomorrow? Admittedly, we have seen it rise many times before, but extrapolation is always uncertain, and we feel that "knowledge" must be more secure than this. Hume believed that we can never really know that the sun will rise tomorrow. Our expectation that it will, like our idea of causality, has, according to Hume, no rational foundation.

Bertrand Russell put the consequences thus: "It is important to discover whether there is any answer to Hume. If not, it follows that there is no intellectual distinction between sanity and insanity. The lunatic who believes that he is a poached egg is to be condemned solely on the ground that he is in a minority, ... or on the ground that the government does not agree with him".[6] Russell also pointed to the stark consequences of having no rational basis for the resolution of conflicting theories. Writing in 1944, Russell put it thus: "The growth of unreason throughout the nineteenth century and what has passed of the twentieth is a natural sequel to Hume's destruction of empiricism".[6]

Induction and verifiability

In the early twentieth century, logical positivists proposed that there was an answer to Hume, and that there was indeed a logical route to certain knowledge. This route was "scientific method". Science, and science alone, could tell us whether "a" causes "b", and allow us to predict when the sun will rise. According to the philosophy of logical positivism, a general statement or theory can be arrived at by inductive reasoning. Positivists also thought that such a theory, if it is verified by observation or experiment, can be promoted to a "law". Indeed, positivists required that a theory must be verifiable in order to count as "scientific". *Verifiability* was the criterion of what is, and is not, science. Thus, in the view of positivists,

Plant Biochemistry, Lund University, Box 117, SE-221 00 Lund, Sweden. E-mail: john.allen@plantbio.lu.se

Funding agencies: Crafoord Foundation; Swedish Natural Sciences Research Foundation.

BioEssays 23:104–107, © 2001 John Wiley & Sons, Inc.

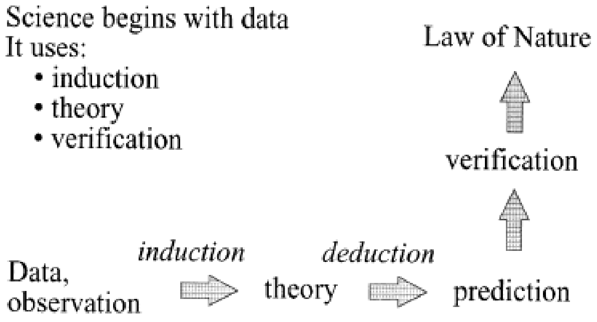

Figure 1. Science starts with the collection of data. Induction turns data into a theory, which can be verified by observation or experiment. Theories which are verified become Laws of Nature. Science produces certain truth.

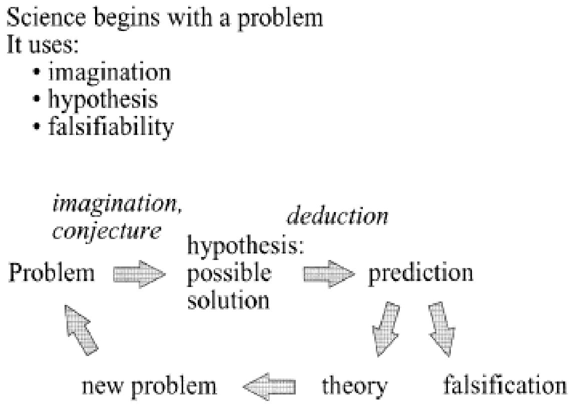

Figure 2. Science starts with a problem: with an inconsistency within existing knowledge or an inconsistency between theory and observation. Imagination is required to envisage a possible solution to the problem, that is, to make an hypothesis. The hypothesis is testable, that is, falsifiable: it makes predictions about observations that must have the capacity to prove that it is false. Science produces theories of increasing explanatory power, but not certain truth. Even if any theory were true, we could never know it.

science produces certified laws of nature by the twin processes of induction and experimental verification (Fig. 1). Physicists, in particular, were once prominent in endorsing logical positivism. I remember that my "A" level physics textbook was very explicit in endorsing this analysis, with a diagram of a sort of tree of truth which arose, by inexorable logic, from a messy mire of observation and experiment. Karl Popper, in contrast, provided a radically different view of the source of scientific knowledge. Popper's ideas allow increased understanding to be a rational and progressive process, yet one that does not require induction. We can then have advancement of knowledge without a requirement that statements about specific observations may entail the truth of any general conclusion. If Popper is correct, induction doesn't work, verification is impossible, and positivism destroys the validity of the science it seeks to defend.

Falsifiability and imagination

Popper had a different answer to Hume. Scientific progress occurs, according to Popper, because hypotheses or theories make predictions by means of which they may be proved false (Fig. 2). A scientific hypothesis is not one that can be verified, but one that is capable of being disproved.(2) The real criterion is not *verifiability*, but *falsifiability*.

The fame of a simple example indicates how widely Popper's philosophy has reached. Consider the statement "All swans are white". Popper agreed with Hume: this statement cannot be proved to be true, no matter how many white swans you observe. However, just one black swan proves it to be false. More generally, truth can be transmitted only in one direction, from a general statement to a particular statement — from an hypothesis to its specific predictions, by logical deduction. However, falsehood can be transmitted in the reverse direction. If the truth of "*p*" entails the truth of "*q*", then, where "*q*" is false, "*p*" must also be false. No matter how many times you observe "*q*" to be true, however, you cannot draw the conclusion that "*p*" is also true. You can make the assumption that "*p*" is true if you wish, but you have no rational basis for doing so. Hume knew this, but had nothing to offer except that our tendency to draw general conclusions from particular observations is habit and laziness. From Russell's standpoint, if Popper has solved the problem of induction, then it would not be an exaggeration to say that Popper saved rationality. Would rationality be destroyed again by the sort of induction machine sought by certain post-genomic biologists? I do not prohibit the search on moral grounds. I merely argue here that it is a waste of time, for there is no such machine.

Popper's theory has had enormously wide implications.(7,8) The art historian Ernst Gombrich(9) and the financier and philanthropist George Soros,(10) for example, both count Popper's influence as decisive in their respective spheres. In contrast to the positivist inclinations of older physicists, Popper's early and successful scientific champions included many distinguished and influential biologists, including Peter Medawar,(11) Jacques Monod,(12) and Peter Mitchell.(13) Enumerating supporters, no matter how illustrious, proves nothing, of course, but it points again to the wide influence of Popper's ideas. I must declare my position, in case it is not already clear. As a life scientist, I, too, am an enthusiastic and partisan spectator of this debate. I was introduced to Popper's books—more blown away, actually—in the early 1970s, by the passionate advocacy of a university biophysics tutor, Colin McClare, and this has been a continuing inspiration in experimental work, research and teaching. I believe that any significant work that I have done conforms precisely to Popper's model.

I had thus thought the matter was essentially settled. Physics was once held to be the exemplary science by logical positivists, but many physicists are now Popperians, and have reservations about genomicists' claims for hypothesis-free, computer-generated understanding.[1] Physicists now value imagination, hypothesis-testing, and Popper's criterion of falsifiability.[14] As I understand it, the Large Hadron Collider is justified in precisely such terms: if the standard theory of matter is true, there must be a Higgs boson, and we need the LHC to see if it is there.[15] If it is not, our theory is wrong, and we do not currently understand the nature of matter. In contrast to physics, molecular biology, once the domain of Popperians, has spawned a sort of genomic hubris: there is now so much data out there that it must surely contain deep understanding and explanatory models, if only we could devise an algorithm or computer program to tell us where these lie. It is as if we are too busy with the all-important task of generating more data, and have come to view thinking as a distracting waste of time. It seems that physics and biology have changed places.

Data mining and discovery in silico

My assertion is that biology is now threatened with a new dark age of positivism. Consider "data mining". There are numerous jobs available for those who claim they can do it and even a journal of *Data Mining and Knowledge Discovery*.[16] A definition of data mining is quoted with approval by leaders in gene expression profiling.[17] "Data mining has been described as the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules".[18] This seems to me to be asking for the epistemological equivalent of a perpetual motion machine. I wonder also whether data mining is a special case of "in silico discovery", an objective of the bioinformatics department of a major drug company's research base. The authors of a recent e-mail flyer take the bull by the horns, and use the I-word explicitly in a description of what they seek from a postdoc: "The object is to use inductive logic programming and other data mining algorithms to find rules to assist in the prediction of protein structure from sequence".[19]

Apart from the mountains of sequence data waiting for an inductivist key to unlock their secrets, gene expression now has hard-line neo-positivists speaking up for wholesale abandonment of hypothesis testing. "Exploration means looking around, observing, describing and mapping undiscovered territory, not testing theories or models. The goal is to discover things we neither knew nor expected, and to see relationships and connections among the elements, whether previously suspected or not. It follows that this process is not driven by hypothesis and should be as model-independent as possible".[20] Furthermore, "The ultimate goal is to convert data into information and then information into knowledge. Knowledge discovery by exploratory data analysis is an approach in which the data 'speak for themselves' after a statistical or visualization procedure is performed".[20]

There is no question that DNA-array data, like sequence data, are potentially of enormous value, and there is every reason to support proposals to make them as available and accessible as possible.[21] When you search a database, however, you must have something to search with, and a reason for wanting to know whether what you're looking for—or something like it—is actually there. It is difficult to understand how the data themselves might tell you which search string to use, why you are using it, or what, for you, is an acceptable degree of similarity between the string and what is retrieved. *Every search for a sequence feature embodies an hypothesis.* A search that contained only wild cards could be regarded as hypothesis-free, but even then you would have to have a reason for choosing a particular database. And what could such a search possibly return?

The prediction

I predict that induction and data-mining, uninformed by ideas, can themselves produce neither knowledge nor understanding. This is itself a Popperian prediction, and I now outline a potentially falsifying result—all good predictions are prohibitions of something. Everything I have said above would be disproved by a single demonstration that a purely logical (i.e. hypothesis-free) process, when applied to data alone, is sufficient to produce a gain in understanding. Anyone who claims to have obtained this result should publish it, without delay, and in the most conspicuous place. If the result is repeatable, I will retract my hypothesis that such an event is impossible, and conclude that Popper was wrong.

I must concede that I can foresee a practical problem with repeatability, however, since repeatability requires independent testing. Any gain in understanding that can be demonstrated to arise de novo will carry implications that run far, far beyond the experimental system from which the data is obtained, and there may thus be an understandable conflict of interest in free publication of the techniques used to produce the result. Consider the intellectual property implications of the breakthrough that shows how to generate understanding of the natural world by an automated process. Perhaps the understanding so produced would itself be patentable, along with the process itself. One could then corner the market in knowledge, and be paid royalties every time anyone understood something. By comparison, the mediaeval alchemists' hopes were modest. It is no wonder that we fail to hear of positive results from initial experiments in "in silico discovery"—what valuable intellectual property it must be. Developments in this exciting field must obviously be shrouded in the greatest secrecy, for sound commercial reasons.

Lose no sleep, however. Consider a simpler explanation: discovery cannot occur in silico. There is no induction machine. The only harm will be waste of resources spent on the search itself. The search should nevertheless be encouraged. Companies and research councils with plenty of money to spare will provide a valuable service by investing in teraflop devices and the development of induction algorithms, since there can be no other way to obtain evidence that induction can't work. Even if data mining is doomed from the start, this conclusion, like everything else, will never be proved beyond doubt. The question is ultimately one of where we wish to put our time and money. Personally, I'd favour another strategy, namely to invest in people. Although they are occasionally irrational, humans have a good track record in generating knowledge. They also tend to want to share what they have found.

Acknowledgments

I thank J. Gallant, C. G. Kurland, W. Martin, and A. S. Wilkins for discussion and suggestions, and G. Ratcliffe for comments on the manuscript.

References

1. Can biological phenomena be understood by humans? Nature 2000; 403:345.
2. Popper KR. The Logic of Scientific Discovery. London: Hutchinson. 1968.
3. Popper KR. Conjectures and refutations. The growth of scientific knowledge. 4th Edition. London: Routledge & Kegan Paul. 1972.
4. Popper KR. Objective knowledge. An evolutionary approach. Oxford: Oxford University Press. 1972.
5. Hume D. An enquiry concerning human understanding. 1748. Oxford: Oxford University Press. 1999.
6. Russell B. History of Western Philosophy. London: Allen & Unwin. 1961.
7. Magee B. Popper. London: Fontana/Collins. 1973.
8. The Karl Popper Web. http://www.eeng.dcu.ie/~tkpw/
9. Gombrich EH. Art and Illusion. Oxford: Phaidon. 1980.
10. Soros G. The crisis of global capitalism: The Open Society Endangered. London: Little, Brown & Co. 1998.
11. Medawar P. Pluto's Republic. Oxford: Oxford University Press. 1982.
12. Monod J. Chance and Necessity. London: Collins. 1972.
13. Mitchell P. The culture of imagination. J Roy Inst Cornwall New Series 1980;8:173–191.
14. Bondi H. The philosopher for science. Nature 1992;358:363.
15. Maddox J. The case for the Higgs boson. Nature 1993;362:785.
16. Data Mining and Knowledge Discovery. http://www.digimine.com/usama/datamine
17. Bassett DE, Eisen MB, Boguski MS. Gene expression informatics—it's all in your mine. Nature Genetics Supplement 1999;21:51–55.
18. Berry MJA, Linoff G. Data Mining Techniques for Marketing, Sales, and Customer Support. New York: John Wiley & Sons. 1997.
19. University of York Machine Learning Group. http://www.cs.york.ac.uk/mlg/
20. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nature Genetics Supplement 1999;21:33–37.
21. Brazma A, Robinson A, Cameron G, Ashburner M. One-stop shop for microarray data. Nature 2000;403:699–700.