

viewpoint

In silico veritas

Data-mining and automated discovery: the truth is in there • by John F. Allen

Literacy underpins constitutions, civil rights and liberties. But E-mail is replacing the letter; the digital certificate, the written signature. Printing allows the recording and dissemination of observations, thoughts and ideas. But the Internet is competing with the printed page; the searchable database with the index of contents. If you are reading this article online, you may have retrieved the file because a search engine found a match to your query, indicating that there is something here you may wish to know. A contextual, semantic search will further confirm this and distil the essence of this article. Searching a genome database is exactly the same. Just as computers are transforming the way we communicate and store information, they are changing the way we discover things worth communicating. In the future, automated discovery will generate new knowledge, take over the process of doing science itself, and tell us what it is that we need to know and understand.

The search engines may, by now, be satisfied with this decoy. So, for those who read beyond titles and first paragraphs: do not believe a word of what you have read so far. The title of this article is irony and its introduction parody. But they describe

Computers dazzle and entertain us, but we should not give them credit for having the ideas in the first place

an attitude that seems to be gaining ground within the scientific community, namely that computers can do our thinking for us. I do not share this view. I rather think that excellent and exciting science is being done today by people who depend totally on computers, but that many are mistaken about the computers' role in doing science. My aim is neither to criticize this science, nor the reliance on computers in research, but I claim that knowledge does not arise *de novo* from computer-assisted analysis of biological data. Computers dazzle and entertain us, but we should not give them



Karl Popper 'The human mind is either a bucket....'. Courtesy of Sarah Allen. First published in BioEssays, **23**, 104–107 (2001), © John Wiley & Sons, Inc.

credit for having the ideas in the first place. The scientist is apt to produce his conclusion rather in the way that a conjurer produces a rabbit out of a hat. I assert that the conclusion, like the rabbit, was there all the time. Computerized data analysis just makes a particularly distracting hat.

Scientists have a record of being reluctant to describe where, and how, they get their ideas. The simplest reason may be that they do not know. Another reason may be that scientists are commendably careful to distinguish the results of an experiment from the preconceptions with which they designed it. Results are objective, public statements. Results are external and inclusive-anyone can inspect, interpret, repeat and confirm them. Preconceptions, unlike results, are subjective-just personal opinion, mere speculation. Preconceptions are internal and exclusive parts of our private thoughts, feelings and hopes. But, without preconceptions, there would be no way of deciding what to look for, no motive for doing the experiment and no basis for interpreting its results. There is a long history of pretence that there are no hypotheses in science, and it has some distinguished players. I suggest that the latest examples include *in silico* discovery, *ab initio* structure prediction and datamining. I shall choose two examples, DNA-microarrays and protein structural prediction, to illustrate this idea.

One of the central processes of biology is gene expression. Starting with the brilliant paradigm of the *lac* operon, many scientists have asked whether they can understand an organism's response to an environmental change by seeing whether and how such a change alters expression of one or more genes. The potential of global gene expression profiling—looking at the expression of many or even all genes in a cell or organism—is quite intoxicating. New possibilities are opening up for those who

Without preconceptions, there would be no way of deciding what to look for, no motive for doing the experiment and no basis for interpreting its results

hitherto laboured with northern blotting or transcriptional assays using individual, carefully chosen probes.

Pioneers of DNA-microarrays go further, and claim that their technique presents a different way of doing science. For instance, P.O. Brown and D. Botstein of Stanford University advocate collecting microarray data without preconception, and then exploring it. '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

viewpoint

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 something interesting would show up. If not, the chemical treatments might just as well have used reagents chosen at random. If you have no background

To be truly 'as model-independent as possible', why not assess the effect of Italian opera on yeast, too?

possible'. 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' (Brown and Botstein, 1999). We might add '...by computer' and convey the flavour of the more ambitious bioinformaticists or bioinformaticians. Their Latin tag might be *ex silica et data veritas*.

The philosophical approach implicit in the idea of converting 'data into information and [...] information into knowledge' is logical induction (Allen, 2001). Induction is the process of reasoning from the particular to the general. As a process comparable to deduction, it was discredited by David Hume in the 18th century and by Karl Popper in the 20th (Popper, 1972). Induction is not logical; it does not work. Here, I suggest that Brown and Botstein have not done what they think they have done. I intend no offence either to the investigators or to the potential of their marvellous technique. It is the philosophy that is wrong, not the science. But that can be the more damaging.

Consider the seminal paper of Eisen et al. (1998). For a large beautiful microarray the figure legend reads: 'Data from separate time courses of gene expression in the yeast S. cerevisiae were combined and clustered. Data were drawn from time courses during the following processes: the cell division cycle after synchronization by α -factor arrest (8 time points); centrifugal elutriation (14 time points), and with a temperature-sensitive cdc15 mutant (15 time points); sporulation (7 time points plus four additional samples); shock by high temperature (6 time points); reducing agents (4 time points) and low temperature (4 time points); and the diauxic shift (7 time points)'.

These experimental conditions were hardly selected at random. Even if the hypotheses being tested were only weak and ill-defined ones, they were there. There were reasons for expecting that knowledge at all, two profiles might just as well be selected from human tissue 'before and after listening to Verdi' (Brazma *et al.*, 2000). Or, to be truly 'as model-independent as possible', why not assess the effect of Italian opera on yeast, too? Actually, Ferea *et al.* (1999) chose instead to look in more detail at the longterm consequences of glucose-depletion far more sensible. Their results are fascinating, and summarised as: 'Alterations in



'...or a searchlight'. Courtesy of Sarah Allen. First published in BioEssays, **23**, 104–107 (2001), © John Wiley & Sons, Inc.

gene expression projected onto metabolic maps of central carbon metabolism' (Ferea *et al.*, 1999). The results show that gene expression responds, over many generations, to starvation, so that less glucose is fermented and more is respired. The conclusion is intuitively satisfying, and not at all something 'we neither knew nor expected' (Brown and Botstein, 1999). Furthermore, it is difficult to believe that the computer program that performed the 'statistical or visualization procedure' (Brown and Botstein, 1999) Protein structure prediction is another important and interesting area where it is sometimes claimed that computers have all the answers. In principle, there are two ways in which we might predict a protein's three-dimensional structure from its amino acid sequence. These are empirical association or correlation, and *ab initio* calculation.

The first method, empirical association, is to catalogue all known sequences and their corresponding structures, and then to compare any new sequence with the catalogue to see if it contains motifs associated with a particular fold or protein domain. There are many different approaches, and ways of weighting different motifs. Chemical properties of amino acids may be included as ab initio input. These approaches are mostly straightforward induction, and would have been recognized as such by Hume. He would also have seen at once that there can be no guarantee that the new sequence will conform to expectations based on previous experience. Some success has been achieved, however.

For example, programs such as TopPred and its successors predict membranespanning helices of intrinsic membrane proteins (Claros and von Heijne, 1994). However, there are underlying assumptions that are incorrect in many cases. For example, the hydrophobic, intrinsic-membrane protein, chloroplast light-harvesting complex II (LHC II), has three membranespanning α -helices (Kühlbrandt *et al.*, 1994). Yet TopPred predicts, with a high degree of confidence, that LHC II is not a membrane protein at all. The explanation is that two of the three helices each contain two polar, charged amino acid residues-glutamate and arginine-that cannot be accommodated into a hydrophobic helix, according to the assumptions of the program. In reality, the opposing charges on these residues compensate for each other, and provide two strong, ionic interactions that hold the two helices

It is difficult to believe that the computer program that performed the 'statistical or visualization procedure' rediscovered glycolysis and the tricarboxylic acid cycle for itself

rediscovered glycolysis and the tricarboxylic acid cycle for itself. The conclusion of Ferea *et al.* (1999) looks suspiciously like background biochemistry plus human intuition. None the worse for that. together in the membrane (Kühlbrandt *et al.*, 1994). Another interesting example is porin (Cowan *et al.*, 1992). TopPred predicts, correctly, that porin is a membrane protein, but for the wrong reason. What

viewpoint

appears to TopPred as a single membranespanning helix is, in fact, membrane extrinsic. What TopPred predicts as extrinsic to the membrane, because of low helix probability and hydrophobicity, is porin's large, anti-parallel β -barrel. The explanation is that TopPred is designed to look for hydrophobic α -helices. The β -barrel, however, forms a membrane-spanning channel that is hydrophobic on the outside, but has a hydrophilic inner surface. TopPred is blind to these novel and interesting features of porin.

The second method for prediction of protein structure, ab initio calculation, is an interesting case of in silico veritas. There is a prevalent view that we already know enough about amino acids and that it is insufficient raw computing power that now prevents us from predicting threedimensional protein structure. Consequently, solving the protein-folding problem is the next benchmark for supercomputers. It would be brave to pronounce on the outcome. From my own viewpoint I do not understand, in principle, how there can be a single, unique solution, arrived at completely ab initio, for the three-dimensional structure of a dipeptide. So much depends upon concentration, the solvent (if any), other solutes, physical parameters such as temperature and pressure and so on. In addition, there is surely no single structure for any quantity of dipeptide greater than a single molecule, at least in solution. And, apart from an astronomically increased number of possible interactions, a 'real' protein has a history, and may fail to adopt its functional tertiary and quaternary structure without the intervention of molecular chaperones, particularly in a living cell. I certainly do not object to the reductionist agenda of describing the structure and function of a protein in terms of the properties and interactions of its constituent amino acids-this is what we do. And, obviously, we will not get very far without computers to help us. Computers are necessary to analyse large data sets, but they are not sufficient-and their sufficiency is precisely what some influential voices now claim.

Creativity consists of a willingness to consider the relevance of observations that have no apparent connection with the problem as it is viewed conventionally. Look back at any great discovery in science, and you will see a leap of imagination. Darwin was willing to consider the breeding and domestication of pigeons as something to connect with biogeographical distribution of animals and plants, and with the fossil record. The orthodox assumption of the immutability of species did not fit, and therefore had to go. Mendel clearly began his experiments with a view that inheritance might be particulate, and that each parent contributed equally and independently to the particles of inheritance of the offspring. There was no prior evidence for that. Einstein was able to pursue the idea that there is no medium, or ether, through which light travels, and thus that the velocity of light is constant while all other motion is relative. Crick and Watson built on the implausible inference that genes were made of nucleic acid. They also had the tenacity to think that the chemical structure of DNA might in some way explain both Mendel's particles of inheritance and the X-ray diffraction patterns by Franklin and Wilkins.

Does creativity require something that computers do not possess? I think it does. We might mention vision, imagination, intellectual ambition even arrogance. It is impossible to understand the motive for creativity unless it includes dissatisfaction with conventional wisdom. Bringing together what orthodoxy regards as completely irrelevant factors also requires personal courage. After all, a scientist's career and reputation may depend upon a good or bad decision about what counts as relevant.

And computers? Even the best are dull, myopic and literal-minded devices. We say a computer 'has a mind of its own' when we have forgotten, or not understood, what we told it to do. At present, computer programs are iterative cycles of deduction, based on feedback from results and initiated at random or by what the user considers as relevant background knowledge. The program cannot decide this for itself. But there are ever more niches for what computers can do, and they will continue to surprise us. 'The Semantic Web' (Berners-Lee et al., 2001) is an intriguing vision that may produce a plausible impression of inanimate discovery of new knowledge. And we already have conventional computers that will learn from experience. The best of them can even beat the world's best chess playerand especially at chess.

We shall see. I think the title of this article and its first paragraph may be a

case in point. A computer precis of *in silico veritas* will be an interesting piece of evidence. But at least, and in contrast to the computer, a human will now be clear about where I stand. Useful as they may be at programmed exploration of patterns in data, I hope it is neither reactionary, nor incurably romantic, to suggest that computers do not yet have what it takes to make discoveries for us. And they certainly cannot tell us what it is that we need to understand.

References

- Allen, J.F. (2001) Bioinformatics and discovery: induction beckons again. *BioEssays*, 23, 104–107. Berners-Lee, T., Hendler, J. and Lassila, O. (2001)
- The semantic web. *Sci. Am.*, **284**(5), 29–37. Brazma, A., Robinson, A., Cameron, G. and Ashburner, M. (2000) One-stop shop for
- Ashburner, M. (2000) One-stop shop for microarray data. *Nature*, **403**, 699–700.Brown, P.O. and Botstein, D. (1999) Exploring the
- new world of the genome with DNA microarrays. *Nature Genet.*, **21**, 33–37.
- Claros, M.G. and von Heijne, G. (1994) TopPred-II— An improved software for membrane-protein structure predictions. *Comput. Appl. Biosci.*, 10, 685–686.
- Cowan, S.W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R.A., Jansonius, J.N. and Rosenbusch, J.P. (1992) Crystal-structures explain functional-properties of two *Escherichia coli* porins. *Nature*, **358**, 727–733.
- Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, 95, 14863–14868.
- Ferea, T.L., Botstein, D., Brown, P.O. and Rosenzweig, R.F. (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl Acad. Sci. USA*, 96, 9721–9726
- Kühlbrandt, W., Wang, D.N. and Fujiyoshi, Y. (1994) Atomic model of plant light-harvesting complex by electron crystallography. *Nature*, **367**, 614–621.
- Popper, K.R. (1972) Objective Knowledge. An Evolutionary Approach. Oxford University Press, Oxford, UK.



John F. Allen is Professor of Plant Cell Biology at Lund University. E-mail: john.allen@plantbio.lu.se

DOI: 10.1093/embo-reports/kve139