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 distill 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 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 credit for having the ideas in the first place.

Based on the 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 distill 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.

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

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.

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 membrane-spanning 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 membrane-spanning 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 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

© 2001 European Molecular Biology Organization
appears to TopPred as a single membrane-spanning 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 low helix probability and hydrophobicity, extrinsic to the membrane, because of spanning helix is, in fact, membrane protein structure, interesting features of porin. Side, 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, \textit{ab initio} calculation, is an interesting case of \textit{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 three-dimensional 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 \textit{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 particular, 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, intellec
tual 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 (Bermers-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 player—and 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 \textit{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**


