DATA MINING AND FISHING TRIPS IN THE EMPIRICAL SCIENCES

Nicholas H. Mann
Department of Biological Sciences
University of Warwick
COVENTRY CV4 7AL
e-mail: N.H.Mann@warwick.ac.uk

The Central Dogma of molecular biology proposes that the genetic information, constituting the genome of all organisms, resides in the sequence of bases within a nucleic acid molecule (usually DNA) and that this information is transcribed into RNA, which in turn determines the linear sequence of amino acids in proteins. Coupled with the assumed quasi-universality of the genetic code, the Central Dogma allows the molecular biologist to suppose that he or she can analyse the DNA from any organism and determine where the genes lie on that DNA molecule and also predict the amino acid sequence of the proteins they encode. Furthermore, by comparison with other conceptually translated proteins, inferences are made as to the biological properties of these proteins. Thus,"understanding nature's mute but elegant language of living cells" involves several, possibly inductive, steps. The large scale sequencing of DNA molecules is now commonplace and, in the sense that the information obtained is purely descriptive, might be regarded as the natural history of molecular biology. However, the explosion of sequence information that this has led to and, more importantly, its subsequent interpretation, has given rise to an absolute requirement for computerized approaches to the storage, organization, and indexing of sequence data and for specialized tools to view and analyze this information. These pressures have led to a novel field of science, Bioinformatics, in which biology, computer science, and information technology converge. A key feature of the new discipline is the development of novel algorithms to automatically annotate, interrogate and analyse large and complex data sets with the aims of permitting global comparisons between genomes and, some would say, permitting the automatic generation of testable hypotheses.

Some of these algorithms have undoubtedly been successful. The problem confronting us now is to critically examine these successes in the light of the well-known falsificationist objections to inductive reasoning.

INDUCTION AT THE COAL FACE

Douglas Kell
Department of Chemistry, Faraday Building, Sackville St
UMIST
MANCHESTER M60 1QD
e-mail: dbk@umist.ac.uk http://dbk.ch.umist.ac.uk

We have argued elsewhere [1] that the scientific process is best characterised by an iterative interplay between complementary hypothesis-generating and hypothesis-testing ('hypothesis-dependent') arms of a continuing cycle. In particular, especially in the "post-genomic" era of functional genomics, which tends to be data-rich but hypothesis-poor, arguably the more important arm presently in biology is that which is knowledge- or hypothesis-*generating*, and in which the direction of inference is from observations to ideas. Notwithstanding the well-known logical/philosophical insecurity of purely data-driven, inductive reasoning ('the sun rose yesterday and the day before, so I expect it to rise tomorrow') the hypothesis-generating arm, *as used by working scientists 'at the coalface'*, is essentially data-driven, and thus purely inductive (to rules) or abductive (to facts) in character. Many other activities of value to the scientific process, especially technology development, are free of specific hypotheses beyond that (view) which states that their outputs should at least be of value. Some sciences are especially data driven (epidemiology, 'whodunnit' forensic science).

A number of authors have described the evolution of ideas and knowledge, or of the optimal future behaviour that they might then govern, as a data-driven search on what amounts to a 'fitness' landscape (e.g. [2-5]). Similarly, from the scientific point of view, the "design", choice or evolution of the next experiment to do in a series is known as 'active learning' [6-8], and may again be purely data-driven. Machine learning methods, in which computer algorithms are used which improve their performance in the light of 'experience' [9], exemplify this. The 'Robot Scientist' [10] carries out an iterative cycle of active learning in an entirely closed loop manner (without human intellectual intervention) and is competitive with human reasoning in an important scientific domain.

[1] Kell, D. B. & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26**, 99-105.

[2] Kauffman, S., Lobo, J., & Macready, W. G. (2000). Optimal search on a technology landscape. *J. Econ. Behav. Organ.* **43**, 141-166.

[3] Goldberg, D. E. (2002). *The design of innovation: lessons from and for competent genetic algorithms*. Kluwer, Boston.

[4] Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., & Lanza, G. (2003). *Genetic programming: routine human-competitive machine intelligence*. Kluwer, New York.

[5] Vaidyanathan, S., Broadhurst, D. I., Kell, D. B., & Goodacre, R. (2003). Explanatory optimisation of protein mass spectrometry via genetic search. *Anal. Chem.* **75**, 6679-6686.

[6] Cohn, D. A., Atlas, L., & Ladner, R. (1994). Improving generalisation with active learning. *Machine Learning* **15**, 201-221.

[7] Raju, G. K. & Cooney, C. L. (1998). Active learning from process data. *AIChE Journal* **44**, 2199-2211.

[8] Bryant, C. H., Muggleton, S. H., Oliver, S. G., Kell, D. B., Reiser, P., & King, R. D. (2001). Combining inductive logic programming, active learning and robotics to discover the function of genes. *Electronic Transactions on Artificial Intelligence* **5**, 1-36 (www.ep.liu.se/ej/etai/2001/001/).

[9] Mjolsness, E. & DeCoste, D. (2001). Machine learning for science: state of the art and future prospects. *Science* **293**, 2051-5.

[10] King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247-252.

BIOINFORMATICS, DISCOVERY, AND DATA EXCAVATION: INDUCTION BECKONS AGAIN

John F. Allen
Center for Chemistry and Chemical Engineering
Lund University
Box 124
SE-221 00 LUND, Sweden
e-mail: john.allen@plantbio.lu.se http://plantcell.lu.se

In the life sciences there is 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 we can now expect computer programs to derive significance, relevance and meaning from chunks of information, be they nucleotide sequences or gene expression profiles. Increasingly, life scientists are advised to rely on computers, or on predetermined software algorithms, to do their thinking for them. A *Nature* editorial — "Can biological phenomena be understood by humans?" — provocatively implies that scientific discovery might well be carried out by machine. In contrast with this view, many are convinced that no purely logical process can turn observation into understanding. We owe this conviction to the work of Karl Popper. Here I argue that Popper was correct, and outline the way in which I think his philosophy applies to the newly data-rich areas of the Life Sciences, and to bioinformatics itself. I predict that even the formidable combination of computing power with ease of access to data does not amount to a qualitative shift in the way we do science: making hypotheses remains an indispensable component in the growth of knowledge.

Anon. (2000) Can biological phenomena be understood by humans? *Nature* **403**, 345

Allen, J. F. (2001) In silico veritas. Data-mining and automated discovery: the truth is in there. *EMBO Reports* **2**, 542-544

Allen, J. F. (2001) Bioinformatics and discovery: induction beckons again. *BioEssays* **23**, 104-107