Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era

Douglas B. Kell[1]* and Stephen G. Oliver[2]

Summary

It is considered in some quarters that hypothesis-driven methods are the only valuable, reliable or significant means of scientific advance. Data-driven or 'inductive' advances in scientific knowledge are then seen as marginal, irrelevant, insecure or wrong-headed, while the development of technology—which is not of itself 'hypothesis-led' (beyond the recognition that such tools might be of value)—must be seen as equally irrelevant to the hypothetico-deductive scientific agenda. We argue here that data- and technology-driven programmes are not alternatives to hypothesis-led studies in scientific knowledge discovery but are complementary and iterative partners with them. Many fields are data-rich but hypothesis-poor. Here, computational methods of data analysis, which may be automated, provide the means of generating novel hypotheses, especially in the post-genomic era. *BioEssays* 26:99–105, 2004.

''Simply gathering data without having any specific question in mind is an approach to science that many people are doubtful about. Modern science is supposed to be mostly hypothesis-driven'....My first studies of the worm lineage didn't require me to ask a question (other than 'What happens next?'). They were pure observation, gathering data for the sake of seeing the whole picture.... This kind of project suits me—it's never bothered me that it doesn't involve bold theories or sudden leaps of understanding, or indeed that it doesn't usually attract the same level of recognition as they do.'' John Sulston.[1]

Introduction

The generation and testing of hypotheses is widely considered to be the primary method by which Science progresses. So much so that it is still common, in some circles, to find a scientific proposal or an intellectual argument damned on the grounds that ''it has no hypothesis being tested'', ''it is merely a fishing expedition'', and so on. Extreme versions run ''if there is no hypothesis, it is not Science'', the clear implication being that hypothesis-driven programmes (as opposed to data-driven studies or technology development) are the only contributor to the scientific endeavour. In our view, such divisive or exclusive views—possibly based on a misreading of Popper[2] and/or more readable commentators such as Medawar[3]—misrepresent the complex intellectual and social intricacies that more correctly characterise the generation of knowledge and understanding from the study of natural phenomena and laboratory experiments.

A discussion of some of these important issues[4] was initiated in this journal by John Allen,[5] and elicited some further debate.[6–10] However, the somewhat polemical starting position[5] inevitably organised the combatants into an either-or view that is too simplistic. The purpose of this essay is to promote the view that the hypothesis-driven and inductive modes of reasoning are not competitive but complementary (see also Ref. 11). Our motivation, in part, is to understand the failure of the prevailing scientific practices to have predicted the existence of so many genes (many of them essential) that were uncovered by the systematic genome sequencing programs,[12] and to rehearse the relative roles of inductive expression profiling methods, technology development and scientific hypothesis testing in post-genomic systems biology.

Abstractions and data

It is commonplace in philosophy to distinguish the world of the mind, knowledge, ideas, thoughts, hypotheses, rules and other mental constructs from physical and material reality as perceived by our senses or measured by our instruments (data or observations). To make things simple, we refer to these two elements as Ideas and Data, respectively. This is the first important distinction to make (Fig. 1), and recasts our questions in terms of the nature of the form of the relationship between Ideas and Data. It is (we hope) obvious that (i) the logical means of going from one to the other depend on the

[1]Department of Chemistry, UMIST, Manchester, UK.
[2]School of Biological Sciences, University of Manchester, UK
*Correspondence to: Douglas B. Kell, Faraday Building, Sackville Street, PO Box 88, Manchester M60 1QD, UK.
E-mail: dbk@umist.ac.uk
DOI 10.1002/bies.10385
Published online in Wiley InterScience (www.interscience.wiley.com).

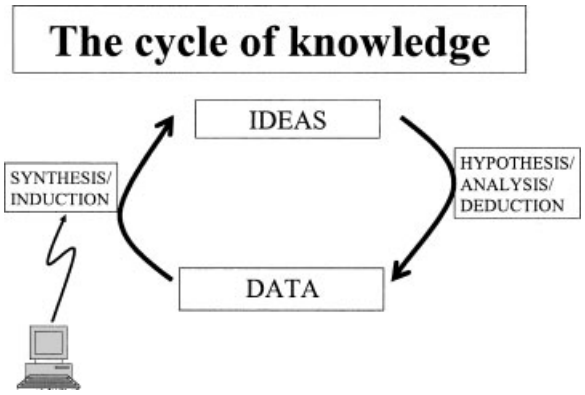

Figure 1. Scientific advance may be seen as an iterative cycle linking knowledge and observations. The hypothetico-deductive mode of reasoning uses background knowledge to construct a hypothesis that is tested experimentally to produce observations. This is only half the story, however, as the inductive mode of reasoning is based purely on generalising rules (or hypotheses) from examples, i.e. it is purely data-driven (and the hypothesis is the end, not the beginning). Because of the high dimensionality of typical data, computer-intensive methods are required to turn the data into knowledge. We argue here that scientific advance should exploit both deductive and inductive modes of reasoning in an iterative cycle.

direction involved (i.e. one is not simply the reverse of the other,[13] and (ii) the process is to be seen as an iterative cycle.

Logical inference: deduction, induction and abduction

Inference is the derivation of new facts from existing facts or premises by any acceptable form of reasoning. Three main types of logical inference are deduction, induction and abduction. The direction of (hypothetico-)deductive reasoning is from Ideas to Data;[14,15] an experimenter has an idea, designs and performs a controlled experiment with a predicted outcome that leads (for a well-designed experiment) to data that are either consistent or inconsistent with the hypothesis.[2] The distinction between abduction and induction is not settled[16] and, for our purposes, we combine the two under one heading—induction, and note the key point that they go from Data to Ideas. This is generalisation from cases, and can also be seen as going from effect to cause, a process referred to as 'inverse entailment'. Thus, by deduction, we can say IF it rained (cause), THEN the grass will be wet (effect). However, we cannot with certainty invert the argument to read IF the grass is wet (effect) THEN it has rained (cause) as the wetting might have been done with a garden hose.

The reason that Deduction appears to enjoy preferred philosophical status then seems to be that if the axiom and the observation are correct the logical inference must be correct (all whales are blue; George is a whale; therefore George is blue). By contrast, induction is seen as being insecure philosophically as it falls to counter-examples. If George is a whale and is blue, Anne is a whale and is blue, Percy is a whale and is blue, and so on, we can induce the idea (hypothesis) that all whales are blue. When Moby Dick comes along and is a whale but white, the inductively generated hypothesis is found to be false. The problem with this cosy distinction is that the appearance of Moby Dick also falsifies the deductive version ('the great tragedy of Science: the slaying of a beautiful hypothesis by an ugly fact'—T.H. Huxley). Of course, in the real world, we know that preferred hypotheses survive any number of inconvenient facts.[17] We thus see that the great philosophical preference for deduction has no genuinely secure basis, but seems to be rooted in a qualitative logical system that is based on a search for certainty or inevitability. Neither of these is noticeably a property of the world of complex, non-linear systems such as those that are the hallmark of modern biology.

Cause and effect in post-genomic science and systems biology

Parameters and variables

In a dynamical system, the parameters are the parts of the system, which have values that are either controlled by the experimenter or are invariant during the experiment. In metabolic biochemistry, these might be parameters such as the pH, or the k_{cat} of an enzyme. Variables are those things that change during an experiment as a result of a change in the parameters. In metabolic biochemistry, the variables include metabolic fluxes and concentrations. Although it is often very desirable to be able to estimate the parameters from the variables (see, for example, Refs. 18,19 and see later), the parameters are the causes and the variables the effects. (Note however that the time elapsed during an experiment is often regarded as an 'honorary' variable.)

Pre- and post-genomics

The cause–effect relationship for genetics/genomics and observable phenotypes is, of course, that the phenotype is caused by the genotype, not vice versa, although it is possible to infer the genotype from the phenotype. Similarly, our perception of the relationship between gene and function (however defined[20]) depends, as in Fig. 1, on the direction involved. Pre-genomic molecular biology tended to be 'function first' and sought genes that were involved in providing that function. Post-genomics starts with, nominally, all the genes, for many of which there is no corresponding biochemical activity or function known, and is thus 'gene first'.[21] Why, then, did the hypothesis-driven mode of reasoning fail to find the approximately 40% of the genes that were uncovered, even in well-worked model organisms, after whole-genome sequencing methods were applied? We think that the main

reason is that classical molecular genetics was both reductionist and qualitative.

Understanding complex systems—holistic and reductionist strategies

At least two strategies for understanding complex systems can be envisaged. The reductionist view would have it that if we can break the system into its component parts and understand them and their interactions in vitro, then we can reconstruct the system physically or intellectually. This might be seen as a 'bottom-up' approach. The holistic approach takes the opposite view, that the complexities and interactions in the intact system mean that we must study the system as a whole. Although these ideas are far from new,[22–26] such strategies are nowadays often referred to as 'systems biology'.[27–32] The molecular biology agenda was explicitly reductionist. The other chief attribute of the molecular biology of the last 50 years is that it was largely qualitative.[33,34] The aim was to make statements that were either true or false (strains X and Y do or do not carry mutations in the same gene, 'UUU does or does not code for phenylalanine'). Experiments were designed to have qualitative readouts (e.g. growth or no growth, red or white colonies etc.). Plausibly, many of the genes 'missed' before systematic genome sequencing were missed because their mutation had only quantitative effects, and thus could not be detected via the classically qualitative readouts of molecular biology.

Rather than being a Johnny-come-lately, the systems approach is the normal starting point in engineering. An engineer would be very surprised to be told that to understand a complex system like a car s/he should fragment it in a blender, centrifuge the bits to separate items of different relative density, and then separate them further on the basis of size and charge in a 2D gel. Having done this, s/he is informed that the study of each of the bits would then allow understanding of how the car really worked. We find it useful to deploy this style of *reductio ad absurdam* when conveying to biologists how unnatural the bottom-up reductionist approach is to an engineer. Of course, both approaches are of value and can be seen as complementary (Fig. 2), but the specific point in the present context is that, by and large, engineering strategies and (by extension) Systems Biology do not represent hypothesis-driven science.

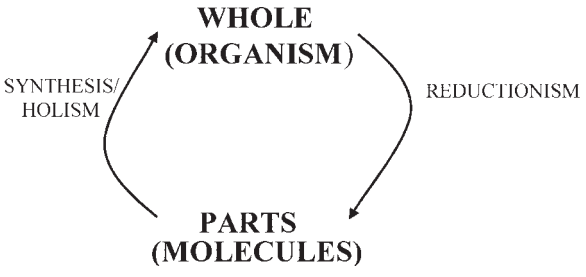

Figure 2. Holism and reductionism as complementary and iterative strategies for understanding complex systems.

Quantitative expression profiling methods and scientific hypothesis testing

A now-common strategy in post-genomic biology is to measure, quantitatively, the action of all (or as many as possible) of the genes at the level of the transcriptome, proteome, metabolome and phenotype (see, for example Ref. 35), and to use computerised methods to infer gene function via various kinds of pattern recognition techniques.[36,37] Such activities are seen as lacking in hypotheses, and are an explicit target of Professor Allen.[5] Actually nothing is completely value-free, and a linkage back to the world of ideas can always be traced; what is meant by Professor Allen is that there is no specific hypothesis, as clearly one can always cast the hypothesis in terms of a view ('hypothesis') that generating such data from a specific set of samples will at least be of value. Thus, throughout, we use 'hypothesis' to mean a specific proposition about the behaviour of a (biological or other) system, based on a logical reasoning that leads to an experimentally verifiable prediction that is either confirmed to be consistent with it or otherwise.

Hypothesis-free science

Epidemiology

The science of epidemiology (see, for example Refs. 38,39) seeks to find the genetic, environmental or other characteristics that are differentially prevalent in those with diseases relative to those who are classed as being well. It holds a special place as a well-established science that is essentially data-driven, and in which hypotheses are the result of the epidemiological study of interest, not its starting point. Nonetheless, it is possible to cast epidemiology into a similarly hypothesis-driven mode as that given above (e.g. ''the hypothesis is that there will be associations between an environmental or genetic precondition, as in the old 'biochemical individuality'[40,41] or the new 'pharmacogenomics',[42] and the prevalence of a disease or response in that subset of the population exhibiting the 'behaviour' or the genotype''). However, by all common sense, the same criticism can be levelled at epidemiology as at expression profiling by those who complain that 'there is no hypothesis'. In a similar vein, we comment that almost all kinds of data mining (see, for example, Refs. 43–45) equivalently search for patterns, and generalise rules as inductive inferences from associations or patterns that occur regularly. Indeed, data mining is practically synonymous with 'knowledge discovery' in databases.[46,47] To this extent, a significant part of the scientific discovery process involves establishing regularities of this type.

Are there many other important biological advances that have been hypothesis-free?

The classical inductive learning example comes from astronomy and is represented by Kepler's Laws. Based on a series of paired data listing the orbital periods, P, and the semi-major axis of their elliptical orbits round the Sun, a, of the planets then known, Kepler induced a mathematical relationship between them, which became known as Kepler's (Third) Law of planetary motion: $P^2 \propto a^3$ or $P^2 = \text{constant} \times a^3$. No hypothesis was being tested, although subsequently (as in Fig. 1) Newton showed (hypothesised?) that an inverse square law of gravitational attraction could largely account for the equations of planetary motions proposed by Kepler. Are there examples of important hypothesis-free advances in biology? Although it is not necessarily easy to prove an 'absence' (of a hypothesis), we certainly think so, and some examples follow.

First, we mention technological developments. These are certainly important in science and that they are science is attested by the fact that many have won Nobel science prizes for their creators. In biological chemistry, the development of methods for sequencing proteins and nucleic acids by Sanger (see for example Ref. 48) or of the polymerase chain reaction by Mullis[49] and of soft-ionisation mass spectrometric methods (see, for example Refs. 50,51) are three obvious examples. Beyond a recognition (hypothesis?) that the ability to perform such analyses would be scientifically valuable, it is not reasonable to claim that these advances were 'hypothesis-driven', and certainly not to suggest that these authors might, at the time of their invention, have imagined the importance of these developments to the human genome project nor the significance of mass spectrometry in proteomics. A recent UK initiative in 'Basic Technology' (see http://www.rcuk.ac.uk/basictech/) explicitly recognises that the results of the technology development that it is promoting are not hypothesis-driven, but that excellent hypothesis-driven science could result from it (again, much as in Fig. 1).

Perutz[52] comments "I have also found that scientific advances are not made by any one single method. Some arise following Popper's hypothetico-deductive one; others are the result of induction from observation that Newton prescribed. In practice, scientific advances often originate from observation, made either by accident or design, without any hypothesis or paradigm in mind. The discovery of pulsars by Tony Hewish and his colleagues was accidental and came as a surprise. The idea that radio pulses might be emitted by rotating neutron stars arose afterwards." The discovery of the cosmic microwave background by Penzias and Wilson was equally serendipitous, and clearly—by definition—serendipitous scientific discoveries[53] arise from observations made with no hypothesis (or at least no hypothesis directed to that specific end) being tested.

Modern biology rests on three major pillars—the Theory of Evolution by Natural Selection, Mendel's Laws of Inheritance, and the double helical structure of DNA. We will now examine how these pillars were built and whether hypothetico-deductive or inductive reasoning was involved.

Neither Darwin nor Wallace, at the time they started to collect specimens and make observations of the living world in far-flung parts of the globe, sought to test any specific hypothesis. Their aim, and the principal item on the agenda of Victorian biology, was to catalogue all the living organisms on the planet (a goal that has yet to be achieved). It was only when they started to organise their specimens, and make sense of their observations, that they entered upon the grand *synthesis* that is the Theory of Evolution by Natural Selection; Mayr discusses this extensively.[54,55] Mendel's Laws of Inheritance provided the mechanism that was missing from Darwinian theory, and certainly appear to have been the product of hypothetico-deductive reasoning. Thus the initial theoretical basis of biology is a telling example of the complementarity of the inductive and deductive approaches. However, what of the structure of DNA itself, the very foundation of molecular biology?

Erwin Chargaff discovered (see, for example, Ref. 56) using thin-layer chromatographic methods, that—within the precision available—the ratio G:C and A:T was approximately unity in all organisms' DNA, whatever the total G + C content the DNA of any specific organism. He did not set much store by this, stating[56] that "A comparison of the molar proportions reveals certain striking, but perhaps meaningless, regularities". Chargaff made his measurements, not to test any specific hypothesis, but "to gain an insight into the differences in composition, and therefore, presumably, in nucleotide sequence, distinguishing . . . DNAs . . . derived from different species."[57] The 1952 paper[57] also did not attach any particular significance to the G:C and A:T ratios, as opposed to the A:G and T:C ones; nevertheless the data provided an important clue that enabled Watson and Crick[58] to solve the structure of DNA. Indeed, Watson and Crick themselves, had no specific hypothesis—other than the conviction that the molecular structure of the genetic material should provide some clue as to gene function. Rosalind Franklin, whose X-ray diffraction data they used, certainly had no hypothesis about the structure of DNA—believing firmly that it would fall out of the Patterson calculations.[59]

When one of us (SGO) embarked upon the sequencing of the DNA molecule of *Saccharomyces cerevisiae* chromosome III (the first chromosome to be sequenced from any organism), he had no specific hypothesis in mind. Yeast chromosomes were of a size that meant that their sequencing was achievable using the current (then manual) technology. Most importantly, Carol Newlon's isolation and cloning of a ring derivative of chromosome III[60] meant that a chromosome-specific library of clones was available. Before the sequencing of chromosome III became a European endeavour, SGO would give seminars in which he stated that he 'expected to find the

unexpected'. He did not predict that there would be approx. five times as many genes on the chromosome as had been found by classical genetics, nor that the relationship between physical and genetic distances would vary so much across the chromosome, nor the existence of transposition 'hot-spots', nor a new class of transposons,[61] nor a number of other phenomena that have proved fertile areas for subsequent enquiries using the hypothetico-deductive approach.

Observational biology

Before the advent of reductionist molecular biology, biology was largely an observational science. It is not obvious that much of observational biology was hypothesis-driven: finding an organelle and calling it a mitochondrion is hypothesis-free (although seeking the function of mitochondria by inferring the consequences of inhibiting their function can well be). Brent points out that much of post-genomic biology is, in this sense, observational in character.[62,63]

The value of data

Intellectual activity, including that which produces patentable inventions and other outcomes commonly recognised as 'intellectual property', can be seen as the navigation of a complex search space or 'landscape' in search of ideas or material inventions that are, in a quasi-evolutionary sense, 'better' or 'fitter' than those pre-existing.[64–68] The only hypotheses here, then, are that a knowledge of the landscape will help in guiding the search,[69] and that there are tools which can improve the chances of getting to the top of Everest rather than being stuck on Snowdon. The mere generation and dissemination of data, the latter now of course to be done electronically and via the Web, is then seen—when viewed in the correct context—as a highly valuable component of the scientific process, even when no hypothesis was involved in the generation of those data.

The role of computers

Can computational activity generate new knowledge?

Going back to the early days of 'artificial intelligence', projects such as DENDRAL and METADENDRAL[70–73] (and see http://smi-web.stanford.edu/projects/history.html#DENDRAL and http://smi-web.stanford.edu/projects/history.html#META-DENDRAL) sought explicitly to enquire as to whether scientific reasoning could be mechanised. Specifically, this involved the computational analyses of paired datasets—in this case mass spectra and the structure of the chemicals from which they came—with a view to determining (i) whether one could infer one (the structure) from the other (the spectrum), and (ii) what rules underlying any such successful inferences had been discovered by (in) the computer. Although these particular projects are largely of historical interest, their positive outcomes made it clear, for instance, that induction could be automated as heuristic search. Other summaries of the use of computational intelligence in relation to creativity or to scientific discovery appear in Refs. 15,67,68,74–81.

The future: intelligent search

Active Learning[82–86] "studies the closed-loop phenomenon of a learner selecting actions or making queries that influence what data are added to its training set".[82] Most modern strategies for navigating complex combinatorial optimisation landscapes involve some kind of active learning as so defined. Studies on active learning in functional genomics have been initiated using logic programming.[87,88] Other strategies are often based on evolutionary algorithms,[89–94] and truly 'closed loop' strategies—in which the next experiment is iteratively selected, performed and analysed entirely by computational means without human intervention—have been in existence or mooted for some time.[95–99] It seems obvious that automating the processes of Fig. 1 in a closed-loop manner will form at least part of the scientific landscape of the future.

In conclusion, we would like to stress that hypothesis-driven and data-driven science are not in competition with each other but are complementary and best carried out iteratively. The development of suitable technologies and the generation of relevant data sets are activities of great scientific value in their own right, and the computer-assisted generation of new knowledge should not be seen as inimical to the creative brilliance of scientists but as a tool that can greatly assist it.

Acknowledgments

We thank Ross King for useful discussions and the BBSRC, EPSRC, NERC and the Wellcome Trust for funding our efforts in both hypothesis-driven and inductive biology.

References

1. Sulston J, Ferry G. The common thread: a story of science, politics, ethics and the human genome. London: Bantam Press; 2002, p. 44.
2. Popper KR. Conjectures and refutations: the growth of scientific knowledge, 5th ed. London: Routledge & Kegan Paul; 1992.
3. Medawar P. Pluto's republic. Oxford: Oxford University Press; 1982.
4. Wilkins AS. Editorial: why the philosophy of science actually does matter. Bioessays 2001;23(1):1–2.
5. Allen JF. Bioinformatics and discovery: induction beckons again. Bioessays 2001;23(1):104–107.
6. Kelley LA, Scott M. On Allen's critique of induction. Bioessays 2001; 23(9):860–861.
7. Gillies DA. Popper and computer induction. Bioessays 2001;23(9):859–860.
8. Allen JF. Hypothesis, induction and background knowledge. Data do not speak for themselves. Replies to Donald A Gillies, Lawrence A Kelley and Michael Scott. Bioessays 2001;23(9):861–862.
9. Allen JF. In silico veritas—Data-mining and automated discovery: the truth is in there. EMBO Rep 2001;2(7):542–544.
10. Smalheiser NR. Informatics and hypothesis-driven research. EMBO Rep 2002;3:702.
11. Sternberg MJE, King RD, Lewis RA, Muggleton S. Application of machine learning to structural molecular biology. Phil Trans R Soc London B 1994;344(1310):365–371.

12. Goffeau A, et al. Life With 6000 Genes. Science 1996;274(5287):546–567.
13. Kell DB, Welch GR. No turning back, Reductonism and Biological Complexity. Times Higher Educational Supplement 1991; 9th August:15.
14. Oldroyd D. The arch of knowledge: an introduction to the history of the philosophy and methodology of science. New York: Methuen; 1986.
15. Langley P, Simon HA, Bradshaw GL, Zytkow JM. Scientific Discovery: computational exploration of the creative processes. Cambridge, MA: MIT Press; 1987.
16. Flach PA, Kakas AC, editors. Induction and abduction: essays on their relation and integration. Dordrecht: Kluwer Academic Publishers; 2000.
17. Gilbert GN, Mulkay M. Opening Pandora's box: a sociological analysis of scientists' discourse. Cambridge: Cambridge University Press; 1984.
18. Mendes P, Kell DB. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 1998;14:869–883.
19. Pearl J. Causality: models, reasoning and inference. Cambridge: Cambridge University Press; 2000.
20. Kell DB, King RD. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. Trends Biotechnol 2000;18(3):93–98.
21. Oliver SG. From DNA sequence to biological function. Nature 1996;379:597–600.
22. von Bertalanffy L. General System Theory. New York: George Braziller; 1969.
23. Iberall AS. Toward a general science of viable systems: McGraw-Hill; 1972.
24. Kacser H, Burns JA. The control of flux. In: Davies DD, editor. Rate Control of Biological Processes. Symposium of the Society for Experimental Biology Vol 27. Cambridge: Cambridge University Press; 1973. p 65–104.
25. Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. Eur J Biochem 1974;42:89–95.
26. Lazebnik Y. Can a biologist fix a radio?—or, what I learned while studying apoptosis. Cancer Cell 2002;2:179–182.
27. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2001;2:343–372.
28. Csete ME, Doyle JC. Reverse engineering of biological complexity. Science 2002;295(5560):1664–1669.
29. Forst CV. Modeling systems biology for research and target prioritization. Pharmacogenomics 2002;3(6):739–743.
30. Kitano H. Systems biology: a brief overview. Science 2002;295(5560):1662–1664.
31. Steven Wiley H, Shvartsman SY, Lauffenburger DA. Computational modeling of the EGF-receptor system: a paradigm for systems biology. Trends Cell Biol 2003;13(1):43–50.
32. Kitano H. Computational systems biology. Nature 2002;420(6912):206–210.
33. Maddox J. Is Molecular Biology yet a Science? Nature 1992;355(6357):201–201.
34. Maddox JS. Towards more measurement in biology. Nature 1994;368:95.
35. Hughes TR, et al. Functional discovery via a compendium of expression profiles. Cell 2000;102(1):109–126.
36. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Berlin: Springer-Verlag; 2001.
37. Duda RO, Hart PE, Stork DE. Pattern classification, 2nd ed. London: John Wiley; 2001.
38. Rothman KJ, Greenland S. Modern epidemiology, 2nd ed. Philadelphia: Lippincott, Williams & Wilkins; 1998.
39. Rothman KJ. Epidemiology: an introduction. Oxford: Oxford University Press; 2002.
40. Williams RJ. Biochemical Individuality. New York: John Wiley; 1956.
41. Davey HM, Kell DB. Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analysis. Microbiol Rev 1996;60:641–696.
42. Evans WE, Johnson JA. Pharmacogenomics: the inherited basis for interindividual differences in drug response. Annu Rev Genomics Hum Genet 2001;2:9–39.
43. Adamo J-M. Data mining for association rules and sequential patterns. New York: Springer-Verlag; 2001.
44. Hand D, Mannila H, Smyth P. Principles of data mining. Cambridge, MA: MIT Press; 2001.
45. Freitas AA. Data mining and knowledge discovery with evolutionary algorithms. Berlin: Springer-Verlag; 2002.
46. Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in Knowledge Discovery and Data Mining, Boston: AAAI/MIT Press; 1996.
47. Piatetsky-Shapiro G, Frawley WJ, editors. Knowledge Discovery in Databases. Boston: MIT Press; 1991.
48. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 1977;74:5463–5467.
49. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 1987;155:335–350.
50. Karas M, Bahr U, Hillenkamp F. UV laser matrix desorption ionization mass spectrometry of proteins in the 100 000 Dalton range. Int J Mass Spectrom Ion Process 1989;92:231–242.
51. Fenn JB. Nobel lecture: electrospray wings for molecular elephants. http://www.nobel.se/chemistry/laureates/2002/fenn-lecture.html 2002.
52. Perutz M. I wish I'd made you angry earlier. Oxford: Oxford University Press; 1998.
53. Roberts RM. Serendipity: accidental discoveries in science. New York: Wiley; 1989.
54. Mayr E. The growth of biological thought: diversity, evolution and inheritance. Harvard: Bellknap; 1982.
55. Mayr E. What evolution is. New York: Basic Books; 2001.
56. Vischer E, Zamenhof S, Chargaff E. Microbial nucleic acids: the desoxypentose nucleic acids of avian tubercle bacilli and yeast. J Biol Chem 1949;177:429–438.
57. Zamenhof S, Brawerman G, Chargaff E. On the desoxypentose nucleic acid from several microorganisms. Biochim Biophys Acta 1952;9:402–405.
58. Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 1953;171:737–738.
59. Maddox B. Rosalind Franklin: dark lady of DNA. London: Harper Collins; 2002.
60. Newlon CS, et al. Analysis of a circular derivative of Saccharomyces cerevisiae chromosome III: a physical map and identification and location of ARS elements. Genetics 1991;129(2):343–357.
61. Oliver SG, et al. The complete DNA sequence of yeast chromosome III. Nature 1992;357(6373):38–46.
62. Brent R. Genomic biology. Cell 2000;100(1):169–183.
63. Brent R. Functional genomics: Learning to think about gene expression data. Current Biology 1999;9(9):R338–R341.
64. Goldberg DE. The design of innovation: lessons from and for competent genetic algorithms. Boston: Kluwer; 2002.
65. Kauffman S, Lobo J, Macready WG. Optimal search on a technology landscape. Journal of Economic Behavior & Organization 2000;43(2):141–166.
66. Kauffman SA. Investigations. Oxford: Oxford University Press; 2000.
67. Koza JR, Keane MA, Streeter MJ. Evolving inventions. Sci Am 2003;288(2):52–59.
68. Koza JR, Keane MA, Yu J, Bennett FH III, Mydlowec W. Automatic creation of human-competitive programs and controllers by means of genetic programming. Genetic Progr Evolvable Machines 2000;1:121–164.
69. Corne DW, Oates MJ, Kell DB. Landscape State Machines: tools for evolutionary algorithm performance analyses and landscape/algorithm mapping. In: Cagnoni S, et al. editor. Evoworkshops 2003. Volume LNCS 2611. Berlin: Springer; 2003. p 187–198.
70. Feigenbaum EA, Buchanan BG. DENDRAL and META-DENDRAL: Roots of knowledge systems and expert system applications. Artificial Intelligence 1993;59:223–240.
71. Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J. DENDRAL—a Case study of the first expert system for scientific hypothesis formation. Artificial Intelligence 1993;61(2):209–261.
72. Buchanan BG, Feigenbaum EA, Lederberg J. On Gray interpretation of the DENDRAL project and programs—myth or mythunderstanding. Chemometr Intell Lab Syst 1988;5(1):33–35.

73. Buchanan BG, Feigenbaum EA. DENDRAL and META-DENDRAL: their application dimensions. Artif Intell 1978;11:5–24.
74. Thompson A. Hardware evolution: automatic design of electronic circuits in reconfigurable hardware by artificial evolution. Berlin: Springer; 1998.
75. Bentley PJ, editor. Evolutionary design by computers. San Francisco: Morgan Kaufmann; 1999.
76. Valdés-Pérez RE. Discovery tools for science apps. Commun ACM 1999; 42(11):37–41.
77. Langley P. The computational support of scientific discovery. Int J Human-Comput Studies 2000;53:393–410.
78. Koza JR, Bennett FH, Keane MA, Andre D. Genetic Programming III: Darwinian Invention and Problem Solving. San Francisco: Morgan Kaufmann; 1999.
79. Munakata T. Knowledge discovery. Commun ACM 1999;42(11):26–29.
80. Koza JR, Keane MA, Streeter MJ, Mydlowec W, Yu J, Lanza G. Genetic programming: routine human-competitive machine intelligence. New York: Kluwer; 2003.
81. Dreyfus HL. What computers still can't do: a critique of artificial reason. Boston: MIT Press; 1992.
82. Cohn DA, Ghabhramani Z, Jordan MI. Active learning with statistical models. J Artif Intell Res 1996;4:129–145.
83. Lin F-R, Shaw MJ. Active training of backpropagation neural networks using the learning by experimentation methodology. Ann Oper Res 1997; 75:105–122.
84. Gillies D. Artificial intelligence and scientific method. Oxford: Oxford University Press; 1996.
85. Hasenjäger M, Ritter H. Active learning with local models. Neural Proc Lett 1998;7:110–117.
86. Kell DB, Mendes P. Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era. In: Cornish-Bowden A, Cárdenas ML, editors. Technological and Medical Implications of Metabolic Control Analysis. Dordrecht: Kluwer Academic Publishers; 2000. p 3–25 (and see http://qbab.aber.ac.uk/dbk/mca99.htm).
87. Bryant CH, Muggleton SH, Oliver SG, Kell DB, Reiser P, King RD. Combining inductive logic programming, active learning and robotics to discover the function of genes. Electronic Transactions on Artificial Intelligence 2001;5(B):1–36 (http://www.ep.liu.se/ej/etai/2001/001/).
88. Reiser P, King RD, Kell DB, Muggleton SH, Bryant CH, Oliver SG. Developing a logical model for yeast metabolism. Electronic Trans on Artif Intell 2001;5:223–244 http://www.ep.liu.se/ej/etai/2001/013/.
89. Reeves CR, editor. Modern heuristic techniques for combinatorial problems. London: McGraw Hill; 1995.
90. Bäck T, Fogel DB, Michalewicz Z, editors. Handbook of evolutionary computation. Oxford: IOPPublishing/Oxford University Press; 1997.
91. Corne D, Dorigo M, Glover F, editors. New ideas in optimization. London: McGraw Hill; 1999.
92. Michalewicz Z. Genetic algorithms + data structures = evolution programs. Berlin: Springer-Verlag; 1994.
93. Michalski RS, Bratko I, Kubat M, editors. Machine learning and data mining. Methods and applications. Chichester: Wiley; 1998.
94. Michalewicz Z, Fogel DB. How to solve it: modern heuristics. Heidelberg: Springer-Verlag; 2000.
95. Zytkow JM, Zhu J, Hussam A. Automated discovery in a chemistry laboratory. 1990. Boston: AAAI Press; pp. 889–894.
96. Judson RS, Rabitz H. Teaching lasers to control molecules. Phys Rev Lett 1992;68(10):1500–1503.
97. Daniel C, Full J, Gonzalez L, Lupulescu C, Manz J, Merli A, Vajda S, Woste L. Deciphering the reaction dynamics underlying optimal control laser fields. Science 2003;299(5606):536–539.
98. Vaidyanathan S, Broadhurst DI, Kell DB, Goodacre R, Explanatory optimisation of protein mass spectrometry via genetic search. Anal Chem 2003; (in press).
99. King RD, Whelan KE, Jones FM, Reiser PGK, Bryant CH, Muggleton SH, Kell DB, Oliver SG. A robot scientist: automated hypothesis generation and experimentation for functional genomics. Nature 2003; (in press).