The Importance of Mathematics in Systems Biology

Douglas B. Kell, School of Chemistry and The Manchester Interdisciplinary Biocentre, University of Manchester and Chief Executive of the BBSRC

Modern biology is predictive and quantitative, and thereby necessarily mathematical [1]. Even the largely qualitative era of molecular biology and genetics could not have happened without mathematics and computation – witness the importance of the Fourier transform in X-ray crystallography [2] and indeed in modern mass spectrometries (see e.g.[3]). Modern genetics is highly quantitative (e.g. [4; 5] – and needs to be [6] http://blogs.bbsrc.ac.uk/index.php/2008/12/when-genetics-meets-the-environment/). As part of this special issue on 'Whither Mathematics in the UK?', I have chosen to pick Systems Biology as an exemplar with which to illustrate some of the challenges we face in biology, and how mathematics (which I take to include related numerical disciplines such as statistics, probability, computation, algorithmics and so on) can assist us in meeting them and thereby helping us to understand biological systems in a quantitative and predictive manner. Note the existence of chemometrics (e.g. [7; 8]) as an **entire field** that covers the application of mathematical methods to problems of (bio)chemistry. The balkanisation of the science and its literature [9-13] means that chem(o)informatics [14; 15] is seen (and has developed) as a discipline largely separate from chemometrics!

Graph theory

Most biochemical networks can be seen as graphs (in the mathematical sense) [16-18], since they consist of actors (nodes) that interact via other players (edges) that serve to connect them. Even molecules can be seen as graphs in that they consist of atoms (nodes) connected by bonds (e.g. [19-26].) Graph theory is therefore an area of great potential significance to biology (both at the cellular and higher levels, including ecology). In one sense systems biology **is** network biology. Note however, that some of the interactions can be indirect.

Biochemical networks

Because 'to understand the whole one must study the whole' [27], modern methods of analysing metabolic networks are applied at the genomic scale [28-31]. However, let us look at a typical metabolic pathway, illustrated by yeast glycolysis (see figure 1), 'starting' with glucose and 'finishing' with ethanol and other products (see e.g. [32; 33]). This is a graph that connects molecules (nodes) that are transformed into other molecules using enzymes that catalyse the relevant reactions (the edges). Sometimes nodes (molecules) that are not directly connected to edges can affect the rates of the reactions catalysed by those edges [34]. As with any simple system, it consists of parameters and variables. In this case the parameters are the 'fixed' concentrations of 'external' metabolites, the ('fixed') concentrations of enzymes, the kinetic rate equations that describe their behaviour as a function of the concentrations of the molecules with which they interact, and the parameters of those equations. The variables are represented by the time-dependent changes in the concentrations of metabolite molecules and the rate of transformation of glucose into ethanol (the 'flux').

While these systems are almost always highly nonlinear (see [35]), and therefore lack analytical solutions (cf. [36]), it is straightforwardly possible [37] to model this kind of system using a numerical ODE integrator or stochastic simulator (a number have been written that have specific domain knowledge, e.g. [38-42]) and thereby obtain the time series of variables that result, as well as any steady state that might be obtained. This is true of both metabolic systems and of signalling systems such as that involving NF-κB with which we have been concerned (e.g. [43-48]). Recognising biochemical systems as nonlinear dynamical systems allows us to bring to bear the concepts and mathematical technologies developed for understanding such nonlinear dynamic systems, albeit that we normally lack **analytical** solutions. One example was the discovery of deterministic chaos in microbiology [49]; another involves the analysis of sensitivities [50-53]. I have little doubt that there is scope for improved mathematical representations of these kinds of system (e.g. [54]).

As well as modelling the system in the 'forward direction' (given the parameters, predict the variables), we may also be interested in solving the 'inverse problem' ('system identification' [55]) by which we seek to estimate the parameters that best account for the behaviour of the variables given the equations (e.g. [56-61]). This is typically an underdetermined and NP-hard [62] optimisation problem (and sometimes we do not even have the equations [63]). Although present algorithms scale poorly (see e.g. the results of the DREAM competitions [64; 65]), Bayesian methods (e.g. [66]) are becoming popular (e.g. [67-71]), and – given the baseline – we may anticipate considerable progress. Note too that biochemical models have unusual 'structures' for their uncertainties [72], and even changing the objective function of what it is that one is trying to fit ('recasting the problem') can bring considerable benefits (e.g.[63]). Other more general mathematical questions that might be asked of a model [73] include whether a **given** structure and set of equations with **any** parameters **could** produce a particular behaviour. These kinds of question also pertain in the emerging field of synthetic biology (e.g. [74; 75]) – a modernised form of metabolic engineering – where one is looking to **design** a system that performs a specific function i.e. exhibits a particular output in response to a given input. There is also considerable interest in optimising models with regard to the fluxes that they generate [76; 77], or (equivalently) finding combinations of sites or drugs that have the maximal effect for the minimum off-target effects [78-80]; these are classical combinatorial optimisation problems. It is also worth mentioning the extensive synergy between the methods of mathematics and those of modern computer science and bioinformatics here; regarding the latter, it is of great importance that we have XML-based standards such as SBML for representing biochemical networks [81] (http://www.sbml.org/), since when combined

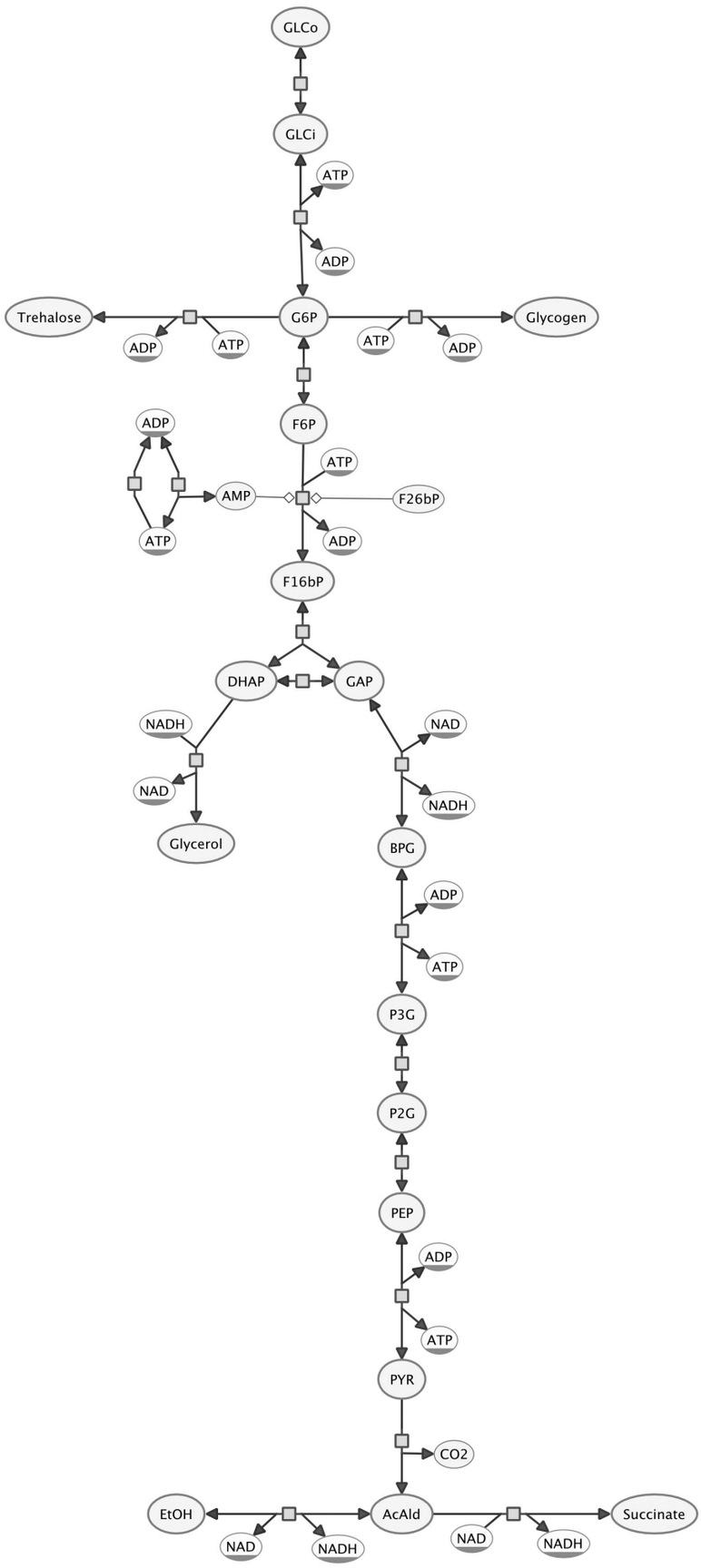

Figure 1: A biochemical network: glycolysis as an example
Acknowledgment: I thank Dr Alice Villéger for providing this figure

with semantic annotation [33; 82; 83] these allow both unambiguous representation and convenient and interoperable manipulation of these kinds of model. In a similar vein, the ability to represent chemical entities in digital format (e.g. [84; 85]) makes it considerably easier to analyse biochemical networks [10; 12] and the molecules they contain [26; 86], and to make use of the distributed knowledge available over the interweb [87-92]. This ability to bring together data, mathematical processing tools and knowledge from disparate sources (e.g. [93-98]) will have considerable impact [99] in reducing the awful rate of attrition [100; 101] in the pharmaceutical industry.

'Omics'

The suffix 'omics' refers to measurements of 'everything' at a particular level of biological organisation e.g. all transcripts (transcriptomics), proteins (proteomics) or metabolites (metabolomics), and comes form the concept of a genomic or genome-wide scale. These concentrations represent the parameters and variables to which we need to fit our models. DNA sequencing itself has increased in speed and power at an astonishing rate since Sanger's paper [102] setting down one of the more popular methods. Modification, parallelisation and automation of the Sanger method was used to generate run sizes ('read lengths' or 'reads') of up to about 500 base pairs at the time of the draft human genome sequencing effort (2000), that then had to be assembled from random fragments ('shotgun sequencing'). Unless there are repetitive sequences this is not too difficult (albeit computationally demanding for large genomes). However, many of the so-called next generation methods (see e.g. [103-106]) generate only short reads, and thus many more fragments per genome. There is a considerable need for suitable assembly algorithms here, a recent example being 'velvet' [107] which – astonishingly – appears to scale as $O(n)$. Other recent examples (see also a summary [108]) include [109-111]. The storage, retrieval and analysis of these kinds of dataset also provide major issues, since as one goes from Gbytes to Tbytes the amount of data involved are not best transferred using conventional Internet protocols; we are entering the era of data-intensive science [112; 113]

Omics methods are not like most other analytical methods, as their high dimensionality (hundreds or even thousands of variables [114]) can be subjected to unusual systematic errors and drift, often spatial, that are hard to model analytically and thereby to correct. Artificial neural networks can be used for this purpose [115; 116] but they are highly empirical. Wavelets have been used to advantage in chemistry [117], and in the cases of registering (aligning) images of protein gels we have exploited [118] the innovative **complex** wavelet transforms developed by Kingsbury (e.g.[119; 120]). A similar problem exists in the alignment of 2-dimensional (chromatography-spectrometry) traces, and despite some very useful progress (e.g. [121-124]), this is still an unsolved area that would benefit considerably from further mathematical input.

Data mining

The high dimensionality [125], the somewhat 'fuzzy'/inaccurate nature[126], and the dynamic range [127] of omics data bring many challenges, both in exploratory data analysis [128; 129] and more directed ('supervised') analyses[130; 131]. A typical scenario involves a series of expression profiles of 1000s of genes performed over potentially hundreds of conditions. Heat maps [132] are the commonest way of displaying these biclusters, but finding the optimal arrangement, even as to the optimal **number** of clusters (e.g. [133]) according to some 'closeness' criterion/a is again NP hard. Any clustering algorithm will **produce** clusters, and so cluster validation is another major issue [134]. There is a trend towards **multiobjective** clustering [135] and optimisation [136]. However, overfitting biological data mining for marker discovery is rampant [137-139]. Nevertheless, data mining (e.g. [131; 140; 141]) and indeed text mining [142-152]) in general are and will continue to be highly important to modern biology.

Optimisation

Almost everything we might do in science (including biology) involves, explicitly or otherwise, some kind of optimisation. Even the choice of the next experiment to perform (and remember that much of experimental design was driven by agriculture [153]) is an optimisation over the space of **possible** experiments. There is an exciting trend towards 'active learning' here, where a suitable algorithm analyses what has gone before (the 'landscape' that has been sampled) and on this basis designs the next experiment in a series in a principled manner (see e.g. [114; 154-161]). This is also being done, for instance, in chemical optimisation for drug discovery via 'fragment-based' methods (e.g. [162] [163-165]), and is starting to be applied in network pharmacology [166] http://blogs.bbsrc.ac.uk/index.php/2009/01/network-pharmacology-meets-systems-biology/. Note of course that analysis of any NP-hard problems – as these tend to be – necessarily requires some kind of heuristic strategy (e.g. [167-169]). This, in my view, is a major area for improving the speed, efficiency or other outputs of any chosen style of algorithm. Recognising the fact that the best algorithm is domain-specific (there is 'no free lunch' [170; 171]) means that one approach is to seek and learn regularities in the particular data structures to hand, and thereby choose the optimal algorithm 'on the fly' [172].

How best to bring together biology and mathematics?

It is obvious from the above that there is barely any part of modern biology that has not benefited, and would not benefit more, from the application of mathematics. Because the problems of biologists are both highly complex and nonlinear, often involving stochastic elements such as mutations, they also offer the potential to mathematicians to develop new mathematics. Most approaches to date have been numerical rather than analytical, but when we can find analytical solutions (as in [36]) this is of course very satisfying. I recognise that many of these problems will be seen as 'applied maths', especially focussing on nonlinear dynamics, optimisation, algorithmics and data processing *sensu lato*, but until the 'pure maths' community engages more it is hard to know what benefits will accrue to either party. The answer to the question that provides this section's title is almost certainly by bringing together biologists and mathematicians, preferably by colocalisation. Short or long term sabbatical visits are also highly beneficial. Much has been written about the balkanisation of scientific disciplines (and the literature [9; 11; 13]), the unwisdom of organising research into silos on the basis of academic units designed for undergraduate admissions, and of the baleful influence of the RAE in perpetuating these inward-looking cultures. Nonetheless, wishing something to change does not necessarily deliver that change without considerable insight [173]. However, the recognition that most world-class research is multi- and inter-disciplinary is causing funders to find mechanisms with

which to try to break down these needless intellectual barriers. I trust that this brief survey might strongly assist that process.❑

Acknowledgments and disclaimer.
Douglas Kell is Chair of Bioanalytical Science at the University of Manchester and Chief Executive of the Biotechnology and Biological Sciences Research Council (BBSRC). The above does not **necessarily** represent the views of the BBSRC, though the sentiments expressed are entirely in accord with its 10-year vision:
http://www.bbsrc.ac.uk/publications/policy/bbsrc_vision.html.
dbk@manchester.ac.uk; http://dbkgroup.org/; http://blogs.bbsrc.ac.uk/

REFERENCES

1 Doyle, F. J., 3rd & Stelling, J. (2006). Systems interface biology. *J R Soc Interface* **3,** 603-16.
2 Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* **185,** 422-7.
3 Marshall, A. G. & Hendrickson, C. L. (2002). Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *Int. J. Mass Spectrom.* **215,** 59-75.
4 Montana, G. (2006). Statistical methods in genetics. *Brief Bioinform* **7,** 297-308.
5 Gunter, C. (2008). Quantitative genetics. *Nature* **456,** 719.
6 Maher, B. (2008). The case of the missing heritability. *Nature* **456,** 18-21.
7 Brereton, R. G. (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Wiley, New York.
8 Lavine, B. & Workman, J. (2006). Chemometrics. *Anal Chem* **78,** 4137-45.
9 Kostoff, R. N. (2002). Overcoming specialization. *Bioscience* **52,** 937-941.
10 Dobson, P. D. & Kell, D. B. (2008). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat Rev Drug Discov* **7,** 205-220.
11 Hull, D., Pettifer, S. R. & Kell, D. B. (2008). Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biol* **4,** e1000204. doi:10.1371/journal.pcbi.1000204.
12 Dobson, P., Lanthaler, K., Oliver, S. G. & Kell, D. B. (2009). Implications of the dominant role of cellular transporters in drug uptake. *Curr Top Med Chem* **9,** 163-184.
13 Kell, D. B. (2009). Iron behaving badly: inappropriate iron chelation as a major contributor to the aetiology of vascular and other progressive inflammatory and degenerative diseases. *BMC Medical Genomics* **2,** 2 - doi:10.1186/1755-8794-2-2.
14 Gasteiger, J. (2003). Handbook of Chemoinformatics: From Data to Knowledge. Wiley/VCH, Weinheim.
15 Oprea, T. I. (2004). Chemoinformatics in drug discovery. Wiley/VCH, Weinheim.
16 Haggarty, S. J., Clemons, P. A. & Schreiber, S. L. (2003). Chemical genomic profiling of biological networks using graph theory and combinations of small molecule perturbations. *J Am Chem Soc* **125,** 10543-5.
17 Barabási, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5,** 101-113.
18 Fischer, E. & Sauer, U. (2005). Large-scale *in vivo* flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat Genet* **37,** 636-40.
19 Bader, R. W. F. (1990). *Atoms in molecules: a quantum theory*: Oxford University Press, Oxford.
20 Willett, P., Barnard, J. M. & Downs, G. M. (1998). Chemical similarity searching. *J Chem Inf Comp Sci* **38,** 983-996.
21 Popelier, P. L. A. (2000). *Atoms in molecules. an introduction*. Pearson Education, London.
22 Rücker, G. & Rücker, C. (2001). On finding nonisomorphic connected subgraphs and distinct molecular substructures. *J Chem Inf Comput Sci* **41,** 314-20.
23 Raymond, J. W. & Willett, P. (2002). Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J Comput Aided Mol Des* **16,** 59-71.
24 Raymond, J. W. & Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Computer-Aided Mol Design* **16,** 521-533.
25 Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. (2004). Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org Biomol Chem* **2,** 3256-66.
26 Dobson, P. D., Patel, Y. & Kell, D. B. (2009). "Metabolite-likeness" as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Disc Today* **14,** 31-40.
27 Kacser, H. (1986). On parts and wholes in metabolism. In *The organization of cell metabolism* (ed. G. R. Welch and J. S. Clegg), pp. 327-337. Plenum Press, New York.
28 Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srvivas, R. & Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci* **104,** 1777-1782.
29 Kell, D. B. (2007). The virtual human: towards a global systems biology of multiscale, distributed biochemical network models. *IUBMB Life* **59,** 689-95.
30 Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O. & Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3,** 135.
31 Feist, A. M. & Palsson, B. Ø. (2008). The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* **26,** 659-667.
32 Pritchard, L. & Kell, D. B. (2002). Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis. *Eur. J. Biochem.* **269,** 3894-3904.
33 Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Le Novère, N., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasić, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kirdar, B., Penttilä, M., Klipp, E., Palsson, B. Ø., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J. & Kell, D. B. (2008). A consensus yeast metabolic network obtained from a community approach to systems biology. *Nature Biotechnol.* **26,** 1155-1160.
34 Cornish-Bowden, A. (1995). Fundamentals of enzyme kinetics, 2nd ed. Portland Press, London.
35 Westerhoff, H. V. & Kell, D. B. (1987). Matrix method for determining the steps most rate-limiting to metabolic fluxes in biotechnological processes. *Biotechnol. Bioeng.* **30,** 101-107.
36 Smallbone, K., Simeonidis, E., Broomhead, D. S. & Kell, D. B. (2007). Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J* **274,** 5576-5585.
37 Kell, D. B. & Knowles, J. D. (2006). The role of modeling in systems biology. In *System modeling in cellular biology: from concepts to nuts and bolts* (ed. Z. Szallasi, J. Stelling and V. Periwal), pp. 3-18. MIT Press, Cambridge.
38 Mendes, P. (1997). Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* **22,** 361-363.
39 Mendes, P. & Kell, D. B. (2001). MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous cellular systems. *Bioinformatics* **17,** 288-289.
40 Klipp, E., Herwig, R., Kowald, A., Wierling, C. & Lehrach, H. (2005). *Systems biology in practice: concepts, implementation and clinical application*. Wiley/VCH, Berlin.
41 Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. & Kummer, U. (2006). COPASI: a COmplex PAthway SImulator. *Bioinformatics* **22,** 3067-74.
42 Wilkinson, D. J. (2006). *Stochastic modelling for systems biology*. Chapman and Hall/CRC, London.
43 Nelson, D. E., Ihekwaba, A. E. C., Elliott, M., Gibney, C. A., Foreman, B. E., Nelson, G., Sée, V., Horton, C. A., Spiller, D. G., Edwards, S. W., McDowell, H. P., Unitt, J. F., Sullivan, E., Grimley, R., Benson, N., Broomhead, D. S., Kell, D. B. & White, M. R. H. (2004). Oscillations in NF-κB signalling control the dynamics of target gene expression. *Science* **306,** 704-708.
44 Ihekwaba, A. E. C., Broomhead, D. S., Grimley, R., Benson, N. & Kell, D. B. (2004). Sensitivity analysis of parameters controlling oscillatory signalling in the NF-κB pathway: the roles of IKK and IκBα. *Systems Biology* **1,** 93-103.
45 Ihekwaba, A. E. C., Broomhead, D. S., Grimley, R., Benson, N., White, M. R. H. & Kell, D. B. (2005). Synergistic control of oscillations in the NF-κB signalling pathway. *IEE Systems Biology* **152,** 153-160.
46 Yue, H., Brown, M., Knowles, J., Wang, H., Broomhead, D. S. & Kell, D. B. (2006). Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-kappaB signalling pathway *Mol Biosyst* **2,** 640-649.
47 Yue, H., Brown, M., He, F., Jia, J. & Kell, D. B. (2008). Sensitivity analysis and robust experimental design of a signal transduction pathway system. *Int J Chem Kinet* **40,** 730-741.
48 Ashall, L., Horton, C. A., Nelson, D. E., Paszek, P., Ryan, S., Sillitoe, K., Harper, C. V., Spiller, D. G., Unitt, J. F., Broomhead, D. S., Kell, D. B., Rand, D., Sée, V. & White, M. R. H. (2009). Pulsatile stimulation determines timing and specificity of NFkappa-B-dependent transcription. *Science* **324,** 242-246.
49 Davey, H. M., Davey, C. L., Woodward, A. M., Edmonds, A. N., Lee, A. W. & Kell, D. B. (1996). Oscillatory, stochastic and chaotic growth rate fluctuations in permittistatically-controlled yeast cultures. *Biosystems* **39,** 43-61.
50 Kell, D. B. & Westerhoff, H. V. (1986). Metabolic control theory: its role in microbiology and biotechnology. *FEMS Microbiol. Rev.* **39,** 305-320.
51 Lüdtke, N., Panzeri, S., Brown, M., Broomhead, D. S., Knowles, J., Montemurro, M. A. & Kell, D. B. (2008). Information-theoretic Sensitivity Analysis: a general method for credit assignment in complex networks *J Roy Soc Interface* **5,** 223-235.
52 Rand, D. A. (2008). Mapping global sensitivity of cellular network dynamics: sensitivity heat maps and a global summation law. *J R Soc Interface* **5 Suppl 1,** S59-69.
53 Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. (2008). *Global sensitivity analysis: the primer*. WileyBlackwell, New York.
54 Boguñá, M., Krioukov, D. & Claffy, K. C. (2009). Navigability of complex networks. *Nature Physics* **5,** 74-80.
55 Ljung, L. (1987). *System identification: theory for the user*. Prentice Hall, Englewood Cliffs, NJ.

56 Mendes, P. & Kell, D. B. (1998). Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14,** 869-883.
57 Moles, C. G., Mendes, P. & Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res* **13,** 2467-74.
58 Bongard, J. & Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *Proc Natl Acad Sci U S A* **104,** 9943-8.
59 Banga, J. R. & Balsa-Canto, E. (2008). Parameter estimation and optimal experimental design. *Essays Biochem* **45,** 195-209.
60 Wilkinson, S. J., Benson, N. & Kell, D. B. (2008). Proximate parameter tuning for biochemical networks with uncertain kinetic parameters. *Mol Biosyst* **4,** 74-97.
61 Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet* **10,** 122-33.
62 Garey, M. & Johnson, D. (1979). *Computers and intractability: a guide to the theory of NP-completeness.* Freeman, San Francisco.
63 Schmidt, M. & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science* **324,** 81-5.
64 Stolovitzky, G. & Califano, A. (2006). *Reverse engineering biological networks: opportunities and challenges in computational methods for pathway inference*. New York Academy of Sciences, New York.
65 Stolovitzky, G., Kahlem, P. & Califano, A. (2009). *The challenges of systems biology: community efforts to harness biological complexity*. New York Academy of Sciences, New York.
66 Kennedy, M. C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *J Royal Stat Soc B* **63,** 425-450.
67 Needham, C. J., Bradford, J. R., Bulpitt, A. J. & Westhead, D. R. (2006). Inference in Bayesian networks. *Nat Biotechnol.* **24,** 51-53.
68 Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform* **8,** 109-16.
69 Vyshemirsky, V. & Girolami, M. A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics* **24,** 833-9.
70 Jayawardhana, B., Kell, D. B. & Rattray, M. (2008). Bayesian inference of the sites of perturbations in metabolic pathways via Markov Chain Monte Carlo. *Bioinformatics* **24,** 1191-1197.
71 Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* **6,** 187-202.
72 Schaber, J., Liebermeister, W. & Klipp, E. (2009). Nested uncertainties in biochemical models. *IET Syst Biol* **3,** 1.
73 Kell, D. B. (2006). Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Disc Today* **11,** 1085-1092.
74 Andrianantoandro, E., Basu, S., Karig, D. K. & Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. *Mol Syst Biol* **2,** 2006 0028.
75 Leonard, E., Nielsen, D., Solomon, K. & Prather, K. J. (2008). Engineering microbes with synthetic biology frameworks. *Trends Biotechnol* **26,** 674-81.
76 Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S. & Palsson, B. Ø. (2004). Comparison of network-based pathway analysis methods. *Trends Biotechnol* **22,** 400-5.
77 Palsson, B. Ø. (2006). *Systems biology: properties of reconstructed networks*. Cambridge University Press, Cambridge.
78 Lehár, J., Zimmermann, G. R., Krueger, A. S., Molnar, R. A., Ledell, J. T., Heilbut, A. M., Short, G. F., 3rd, Giusti, L. C., Nolan, G. P., Magid, O. A., Lee, M. S., Borisy, A. A., Stockwell, B. R. & Keith, C. T. (2007). Chemical combination effects predict connectivity in biological systems. *Mol Syst Biol* **3,** 80.
79 Zimmermann, G. R., Lehár, J. & Keith, C. T. (2007). Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov Today* **12,** 34-42.
80 Lehár, J., Stockwell, B. R., Giaever, G. & Nislow, C. (2008). Combination chemical genetics. *Nat Chem Biol* **4,** 674-81.
81 Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, II, Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novere, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. & Wang, J. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19,** 524-31.
82 Kell, D. B. & Mendes, P. (2008). The markup is the model: reasoning about systems biology models in the Semantic Web era. *J Theoret Biol* **252,** 538-543. .
83 Pettifer, S. R., Thorne, D., McDermott, P., Marsh, J., Villéger, A., Kell, D. B. & Attwood, T. K. (2009). Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics* **10**, S19.
84 Weininger, D. (1988). SMILES, a chemical language and information system .1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28,** 31-36.
85 Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S. & Zhang, Y. (2005). Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic & Biomolecular Chemistry* **3,** 1832-1834.
86 Kell, D. B. & Dobson, P. D. (2009). The cellular uptake of pharmaceutical drugs is mainly carrier-mediated and is thus an issue not so much of biophysics but of systems biology. In *Proceedings of the International Beilstein Symposium on Systems Chemistry*, pp. 149-168. Logos Verlag, Berlin.
87 Murray-Rust, P., Rzepa, H. S., Stewart, J. J. & Zhang, Y. (2005). A global resource for computational chemistry. *J Mol Model* **11,** 532-41.
88 Casher, O. & Rzepa, H. S. (2006). SemanticEye: a semantic web application to rationalize and enhance chemical electronic publishing. *J Chem Inf Model* **46,** 2396-411.
89 Knox, C., Shrivastava, S., Sothard, P., Eisner, R. & Wishart, D. S. (2007). Biospider: a web server for automating metabolome annotations. *Pac Symp Biocomputing* **12,** 145-156.
90 Klinger, R., Kolářik, C., Fluck, J., Hofmann-Apitius, M. & Friedrich, C. M. (2008). Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics* **24,** i268-76.
91 Williams, A. J. (2008). Internet-based tools for communication and collaboration in chemistry. *Drug Discov Today* **13,** 502-6.
92 Williams, A. J. (2008). A perspective of publicly accessible/open-access chemistry databases. *Drug Discov Today* **13,** 495-501.
93 Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. & Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20,** 3045-54.
94 Kell, D. B. (2006). Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher lecture. *FEBS J* **273,** 873-894.
95 Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A. & Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice & Experience* **18,** 1067-1100.
96 Oinn, T., Li, P., Kell, D. B., Goble, C., Goderis, A., Greenwood, M., Hull, D., Stevens, R., Turi, D. & Zhao, J. (2007). Taverna / myGrid: aligning a workflow system with the life sciences community. In *Workflows for e-Science: scientific workflows for Grids* (ed. I. J. Taylor, E. Deelman, D. B. Gannon and M. Shields), pp. 300-319. Springer, Guildford.
97 Li, P., Oinn, T., Soiland, S. & Kell, D. B. (2008). Automated manipulation of systems biology models using libSBML within Taverna workflows. *Bioinformatics* **24,** 287-289.
98 Li, P., Castrillo, J. I., Velarde, G., Wassink, I., Soiland-Reyes, S., Owen, S., Withers, D., Oinn, T., Pocock, M. R., Goble, C. A., Oliver, S. G. & Kell, D. B. (2008). Performing statistical analyses on quantitative data in Taverna workflows: an example using R and maxdBrowse to identify differentially expressed genes from microarray data. *BMC Bioinformatics* **9,** 334.
99 Waller, C. L., Shah, A. & Nolte, M. (2007). Strategies to support drug discovery through integration of systems and data. *Drug Discov Today* **12,** 634-9.
100 Kola, I. & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* **3,** 711-5.
101 Leeson, P. D. & Springthorpe, B. (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov* **6,** 881-90.
102 Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74,** 5463-5467.
103 Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol* **26,** 1135-45.
104 Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53-9.
105 Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9,** 387-402.
106 Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clin Chem* **55,** 641-58.
107 Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18,** 821-9.
108 Bateman, A. & Quackenbush, J. (2009). Bioinformatics for next generation sequencing. *Bioinformatics* **25,** 429.

109 Chaisson, M. J., Brinza, D. & Pevzner, P. A. (2008). *De novo* fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res* **19,** 336-346.
110 Malhis, N., Butterfield, Y. S., Ester, M. & Jones, S. J. (2009). Slider—maximum use of probability information for alignment of short sequence reads and SNP detection. *Bioinformatics* **25,** 6-13.
111 Rausch, T., Koren, S., Denisov, G., Weese, D., Emde, A. K., Doring, A. & Reinert, K. (2009). A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads. *Bioinformatics* **25,** 1118-24.
112 Szalay, A. & Gray, J. (2006). 2020 computing: science in an exponential world. *Nature* **440,** 413-4.
113 Bell, G., Hey, T. & Szalay, A. (2009). Beyond the data deluge. *Science* **323,** 1297-8.
114 Zelena, E., Dunn, W., Broadhurst, D., Francis-McIntyre, S., Carroll, K., Begley, P., O'Hagan, S., Knowles, J. D., Halsall, A., HUSERMET Consortium, Wilson, I. & Kell, D. (2009). Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Anal Chem* **81,** 1357-1364.
115 Goodacre, R. & Kell, D. B. (1996). Correction of mass spectral drift using artificial neural networks. *Anal. Chem.* **68,** 271-280.
116 Goodacre, R., Timmins, E. M., Jones, A., Kell, D. B., Maddock, J., Heginbothom, M. L. & Magee, J. T. (1997). On mass spectrometer instrument standardization and interlaboratory calibration transfer using neural networks. *Anal. Chim. Acta* **348,** 511-532.
117 Alsberg, B. K., Woodward, A. M. & Kell, D. B. (1997). An introduction to wavelet transforms for chemometricians: a time-frequency approach. *Chemometrics and Intelligent Laboratory Systems* **37,** 215-239.
118 Woodward, A. M., Rowland, J. J. & Kell, D. B. (2004). Fast automatic registration of images using the phase of a complex wavelet transform: application to proteome gels. *Analyst* **129,** 542-552.
119 Kingsbury, N. (2001). Complex wavelets for shift invariant analysis and filtering of signals. *Appl. Comp. Harmonic Anal.* **10,** 234-253.
120 Selesnick, I. W., Baraniuk, R. G. & Kingsbury, N. G. (2005). The dual-tree complex wavelet transform. *IEEE Sig Proc Mag* **22,** 123-151.
121 Katajamaa, M., Miettinen, J. & Orešič, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **22,** 634-6.
122 Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* **78,** 779-87.
123 Dunn, W. B., Broadhurst, D., Brown, M., Baker, P. N., Redman, C. W. G., Kenny, L. C. & Kell, D. B. (2008). Metabolic profiling of mammalian serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system. *J Chromatogr B* **871,** 288-298.
124 Benton, H. P., Wong, D. M., Trauger, S. A. & Siuzdak, G. (2008). XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. *Anal Chem* **80,** 6382-9.
125 Schena, M. (2000). Microarray biochip technology. Eaton Publishing, Natick, MA.
126 Piper, M. D., Daran-Lapujade, P., Bro, C., Regenberg, B., Knudsen, S., Nielsen, J. & Pronk, J. T. (2002). Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. *J Biol Chem* **277,** 37001-8.
127 Hoyle, D. C., Rattray, M., Jupp, R. & Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics* **18,** 576-584.
128 Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley, Reading, MA.
129 Duda, R. O., Hart, P. E. & Stork, D. E. (2001). *Pattern classification, 2nd ed.* John Wiley, London.
130 Kell, D. B. & King, R. D. (2000). On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends Biotechnol.* **18,** 93-98.
131 Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning: data mining, inference and prediction*. Springer-Verlag, Berlin.
132 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95,** 14863-14868.
133 De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B. & Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* **18,** 735-746.
134 Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21,** 3201-3212.
135 Handl, J. & Knowles, J. (2007). An evolutionary approach to multiobjective clustering. *IEEE Trans Evol Comput* **11,** 56-76.
136 Handl, J., Kell, D. B. & Knowles, J. (2007). Multiobjective optimization in bioinformatics and computational biology. *IEEE Trans Comput Biol Bioinformatics* **4,** 279-292.
137 Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med* **2,** e124.
138 Broadhurst, D. & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2,** 171-196; DOI: 10.1007/s11306-006-0037-z.
139 Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology* **19,** 640-8.
140 Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of data mining*. MIT Press, Cambridge, MA.
141 Wu, X. D., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems* **14,** 1-37.
142 Nenadic, G., Mima, H., Spasic, I., Ananiadou, S. & Tsujii, J. (2002). Terminology-driven literature mining and knowledge acquisition in biomedicine. *Int J Med Inform* **67,** 33-48.
143 Ananiadou, S. & McNaught, J. (2006). Text mining in biology and biomedicine. Artech House, London.
144 Ananiadou, S., Kell, D. B. & Tsujii, J.-i. (2006). Text Mining and its potential applications in Systems Biology. *Trends Biotechnol* **24,** 571-579.
145 Jensen, L. J., Saric, J. & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* **7,** 119-29.
146 Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics* **8,** 358-375.
147 Agarwal, P. & Searls, D. B. (2008). Literature mining in support of drug discovery. *Brief Bioinform* **9,** 479-92.
148 Cohen, K. B. & Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol* **4,** e20.
149 Krallinger, M., Valencia, A. & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* **9 Suppl 2,** S8.
150 Rzhetsky, A., Seringhaus, M. & Gerstein, M. (2008). Seeking a new biology through text mining. *Cell* **134,** 9-13.
151 Spasić, I., Schober, D., Sansone, S.-A., Rebholz-Schuhmann, D., Kell, D. B. & Paton, N. (2008). Facilitating the development of controlled vocabularies for metabolomics with text mining. *BMC Bioinformatics* **9,** S5.
152 Spasić, I., Simeonidis, E., Messiha, H. L., Paton, N. W. & Kell, D. B. (2009). KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways. *Bioinformatics* **25,** 1404-1411.
153 Fisher, R. A. (1926). *The design of experiments*. Oliver & Boyd, Edinburgh.
154 Sacks, J., Welch, W., Mitchell, T. & Wynn, H. (1989). Design and analysis of computer experiments (with discussion). *Statist Sci* **4,** 409-435.
155 Jones, D. R., Schonlau, M. & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *J. Global. Opt.* **13,** 455-492.
156 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.
157 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.
158 O'Hagan, S., Dunn, W. B., Brown, M., Knowles, J. D. & Kell, D. B. (2005). Closed-loop, multiobjective optimisation of analytical instrumentation: gas-chromatography-time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Anal Chem* **77,** 290-303.
159 Knowles, J. (2006). ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Trans Evol Comput* **10,** 50-66.
160 O'Hagan, S., Dunn, W. B., Broadhurst, D., Williams, R., Ashworth, J. A., Cameron, M., Knowles, J. & Kell, D. B. (2007). Closed-loop, multi-objective optimisation of two-dimensional gas chromatography (GCxGC-tof-MS) for serum metabolomics. *Anal Chem* **79,** 464-476.
161 King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E. & Clare, A. (2009). The automation of science. *Science* **324,** 85-9.
162 Rees, D. C., Congreve, M., Murray, C. W. & Carr, R. (2004). Fragment-based lead discovery. *Nat Rev Drug Discov* **3,** 660-72.
163 Hajduk, P. J. & Greer, J. (2007). A decade of fragment-based drug design: strategic advances and lessons learned. *Nat Rev Drug Discov* **6,** 211-9.
164 Jhoti, H., Cleasby, A., Verdonk, M. & Williams, G. (2007). Fragment-based screening using X-ray crystallography and NMR spectroscopy. *Curr Opin Chem Biol* **11,** 485-93.
165 Hubbard, R. E. (2008). Fragment approaches in structure-based drug discovery. *J Synchrotron Res* **15,** 227-230.
166 Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* **4,** 682-690.
167 Corne, D., Dorigo, M. & Glover, F. (1999). New ideas in optimization. McGraw Hill, London.
168 Michalewicz, Z. & Fogel, D. B. (2000). *How to solve it: modern heuristics*. Springer-Verlag, Heidelberg.
169 Kind, T. & Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8,** 105.
170 Radcliffe, N. J. & Surry, P. D. (1995). Fundamental limitations on search algorithms: evolutionary computing in perspective. *Computer Science Today* **1995,** 275-291.
171 Wolpert, D. H. & Macready, W. G. (1997). No Free Lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1,** 67-82.
172 Wedge, D. & Kell, D. B. (2008). Rapid prediction of optimum population size in genetic programming using a novel genotype - fitness correlation. *GECCO 2008 (M. Keizer et al., eds)*, 1315-1322.
173 Barber, M. (2007). *Instruction to deliver: fighting to transform Britain's public services.* Methuen, London.