Metabolomics, machine learning and modelling: towards an understanding of the language of cells

D.B. Kell[1]
School of Chemistry, The University of Manchester, Faraday Building, Sackville Street, P.O. Box 88, Manchester M60 1QD, U.K.

Abstract

In answering the question 'Systems Biology – will it work?' (which it self-evidently has already), it is appropriate to highlight advances in philosophy, in new technique development and in novel findings. In terms of philosophy, we see that systems biology involves an iterative interplay between linked activities – for instance, between theory and experiment, between induction and deduction and between measurements of parameters and variables – with more emphasis than has perhaps been common now being focused on the first in each of these pairs. In technique development, we highlight closed loop machine learning and its use in the optimization of scientific instrumentation, and the ability to effect high-quality and quasi-continuous optical images of cells. This leads to many important and novel findings. In the first case, these may involve new biomarkers for disease, whereas in the second case, we have determined that many biological signals may be frequency rather than amplitude-encoded. This leads to a very different view of how signalling 'works' (equations such as that of Michaelis and Menten which use only amplitudes, i.e. concentrations, are inadequate descriptors), lays emphasis on the signal processing network elements that lie 'downstream' of what are traditionally considered the signals, and allows one simply to understand how cross-talk may be avoided between pathways which nevertheless use common signalling elements. The language of cells is much richer than we had supposed, and we are now well placed to decode it.

Introduction

> 'Progress in science depends on new techniques, new discoveries and new ideas, probably in that order'
> Sydney Brenner, Nature, June 5, 1980

Following Sydney Brenner's comment (above), though not in that order, and as part of a meeting entitled 'Systems Biology: will it work?', I have chosen to highlight three aspects of our current collaborative work. The first involves the philosophical underpinnings of our scientific strategy and of the systems biology agenda, which are each characterized by an iterative interplay [1,2] between a series of linked activities: these include data and ideas; theory, computation and experiment; and the iterative assessment of model parameters and variables. The second area relates to the development of analytical and computational technology, especially in metabolomics, to help provide both high-quality data and modelling. The third concentrates on conceptual developments following from our recent findings [3,4] to the effect that one way to look at biological signalling pathways is not so much in terms of changes in the concentrations of signalling intermediates, but in terms of the downstream 'signal processing elements' that respond to their dynamics. This gives a profoundly different view of the significance of networks in systems biology, and one that allows one a much better understanding of signalling as signal processing.

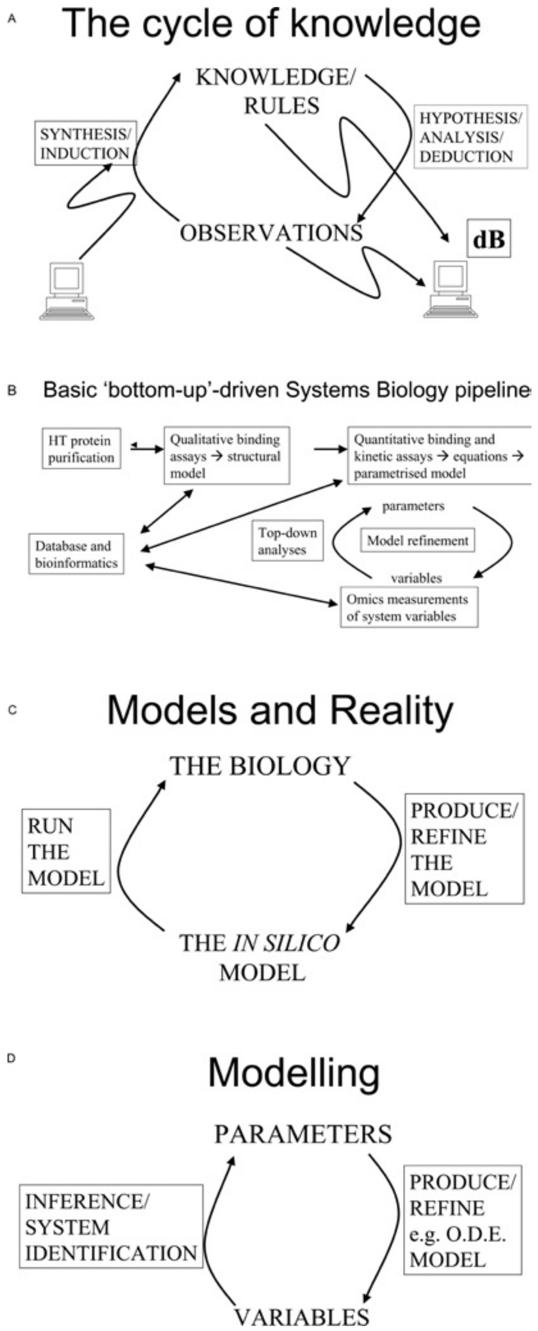
Philosophy of systems biology

Most commentators, including this author [5], take the systems biology agenda to include, in an iterative manner, both wet (experimental) and dry (computational and theoretical) work as part of an iterative cycle. (In this sense systems biology shares the same agenda as the long-established metabolic control analysis [6,7].) Figure 1 shows four views of this. Figure 1(A) stresses the importance of inductive methods of hypothesis generation; these have unaccountably had far less emphasis than they should have done because of the traditional obsession in 20th century biology with hypothesis testing. Principled hypothesis generation is clearly at least as important as hypothesis testing, and appropriate experimental designs ensure that the search for good candidate data is not an aimless fishing expedition but one which is likely to find novel answers in unexpected places [1,2,8,9]. Figure 1(B) stresses the importance of first getting the structural model (the fundamental building blocks of the 'language' of cells), then suitable equations that can represent then parametrize the kinetic data, as these can be used directly in forward models (e.g. [10,11]), whereas Figures 1(C) and 1(D) highlight the basic and iterative relations between computational models and reality on one hand and between changes in the model that are invoked and its subsequent dynamic behaviour. If the answer to 'Systems Biology – will it work?' is at least partly synonymous with the answer to 'can we make models that both simulate existing data and make exciting (and ultimately validated) predictions about the results of experiments not yet done?'. Then I think we can already answer that in the

Q1 **Key words:** computational technology, genetic programming, machine learning, metabolomics, signal processing elements, systems biology.
Abbreviations used: IκB, inhibitory κB; NF-κB, nuclear factor κB.

[1] **email dbk@manchester.ac.uk**

Figure 1 | Iterative elements of systems biology
(**A**) Science advances through an iterative interplay between ideas and experimental data. (**B**) A largely bottom-up view, as in the 'silicon cell' [67], of one approach to systems biology. First we seek a 'structural model' that defines the players and the qualitative nature of the interactions between them; then we seek equations that best describe the relationships, then finally we seek to parametrize those equations (recognizing that if errors occur in the earlier phases we may need to return and correct them in the light of further knowledge). (**C**) Modelling and comparison of the models with the reality as an iterative process. (**D**) Producing and refining a model: data on kinetic parameters allow one to run a forward model, whereas invoking such parameters from measured 'omics data (fluxes and concentrations) is an inverse or system identification problem.



affirmative [12,13] – but solving the inverse problem is by far the hardest of these challenges [14–16].

Metabolomics technology

If we consider metabolic systems, most analyses take discrete samples and provide 'metabolic snapshots' [6]. Typical model microbes such as baker's yeast [17] contain upwards of 1000 metabolites, most with $M_R < 1000$ [5]. An important area of metabolomics thus consists of maximizing the number of metabolites that may be measured reliably [18,19], as a prelude to exploiting such data via a chemometric and computational pipeline [20]. The problem here is that optimizing scientific instrumentation is a combinatorial problem that scales exponentially with the number of experimental parameters. Thus if there are 14 adjustable settings on an electrospray mass spectrometer, each of which can take ten values, the number of combinations to be tested by exhaustive search is 10^{14} [21]. Although heuristic methods that find good but not provably optimal solutions (such as methods based on genetic algorithms [21,22]) have proved successful, they are still slow because there is a human being in the loop, and the number of experiments that can be evaluated is correspondingly small.

More recently, in a manner related to the computationally driven supervised [23] and inductive [2] discovery of new biological knowledge [24], we have contributed to the Robot Scientist project [25]. Here a computational system is used (i) to hold background knowledge about a biological domain (amino acid biosynthesis, modelled as a logical graph), (ii) to use that knowledge to design the 'best' (most discriminatory) experiment to find the biochemical location in that graph of a specific genetic lesion, (iii) to perform that experiment using microbial growth tests, and to analyse the results, and (iv) on the basis of these to design, perform and evaluate the next experiment, the whole continuing in an iterative manner (i.e. in a closed loop, without human intervention) until only one possible hypothesis remains.

We have now combined these two set of ideas to use genetic search methods in an automated closed loop (the 'Robot Chromatographer') to maximize simultaneously the number of peaks observed and a signal/noise metric while also minimizing the run time [26]. Depending on the sample (serum [19] or yeast supernatant [27,28]), this has doubled or trebled the number of metabolite peaks that we can reliably observe using GC-MS [26], thereby allowing us to discover important new biomarkers for metabolic diseases.

Network motifs, sensitivity analysis, signal processing and credit assignment in systems biology models

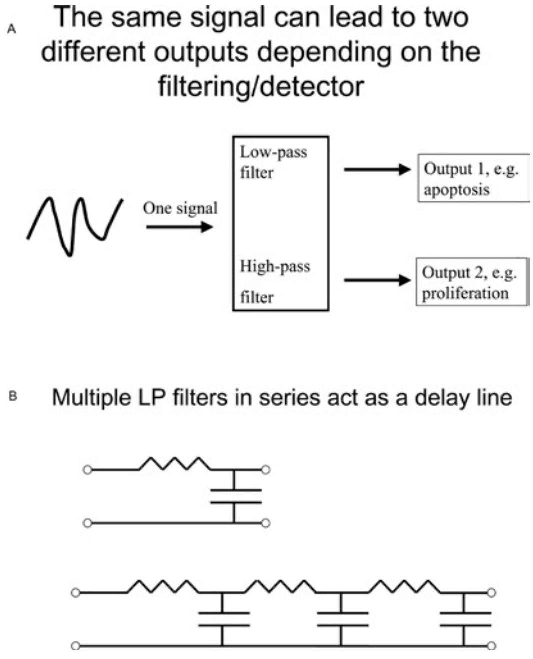
A hallmark of post-genomics is the development of high-throughput methods for the analysis of complex biological systems. In consequence, it is increasingly commonplace to have access to large datasets of variables ('omics data) against which to test a mathematical model of the system that might generate such data. In these cases, the model will usually be an ODE (ordinary differential equation) model, and finding a good model is a system identification problem [10].

Much less frequently, the kinetic and binding constants are available, and a reliable 'forward' model can be generated directly. One such case is the NF-κB (nuclear factor κB) signalling pathway [29]. NF-κB is a nuclear transcription factor that is normally held inactive in the cytoplasm by being bound to an IκB (inhibitory κB). When IκB is phosphorylated by a kinase (inhibition of κB kinase), it is degraded and free NF-κB can translocate to the nucleus, where it induces the expression of genes (including those such as IκB that are involved in its own dynamics). The NF-κB system is considered to be 'involved' in both cell proliferation and in apoptosis, although how a cell chooses which of these orthogonal processes to follow simply from the changes in the concentration of NF-κB in a particular location is neither obvious nor known. Earlier experimental measurements showed oscillations in NF-κB in single cells, though these were damped when assessed as an ensemble since individual cells were necessarily out of phase ([30], and see also [31,32] for a similar philosophy underpinning flow cytometry). More recently, with improved constructs and detector technology, the oscillations could clearly be measured accurately in individual cells alone [4].

Based on the model of Hoffmann et al. [29], and using Gepasi [10] we have modelled the 'downstream' parts of this pathway (there are 64 reactions and 23 variables), and performed sensitivity analysis (a generalized form of metabolic control analysis [33], useful in many other domains [34]). This showed [3] that only approx. 8 of the 64 reactions exerted any serious control over the timings and amplitudes of the oscillations in the NF-κB concentration, and most importantly that it was not so much the concentration of NF-κB but its dynamics that is responsible for controlling downstream activities [4]. This leads to a profound emphasis on the role of 'network motifs' [35–37] as downstream signal processing elements that can discriminate the dynamical properties of inputs that otherwise use the same components. Biological signalling is then best seen or understood as signal processing, a major field (mainly developed in areas such as data communications, image processing [38] and so on), in which we recognize that the structure, dynamics and performance of the receiver entirely determine which properties of the upstream signal are actually transduced into downstream (and here biological) events. The crucial point is that, in the signal processing world, these signals are separated by their dynamical, frequency-dependent properties. Normally, we model enzyme kinetics on the basis of a static concentration [e.g. the irreversible Michaelis–Menten reaction $v = V_m S/(S + K_m)$ includes only the 'instantaneous' concentration but not the dynamics of S]. However, if detectors have frequency-sensitive properties, this allows one in principle to solve the 'cross-talk problem'. (How do cells distinguish identical changes in the 'static' NF-κB concentration that might lead either to apoptosis or to proliferation, when these are in fact entirely orthogonal processes?) Although other factors can always contribute usefully (e.g. further transcription factors that act as a logical AND, OR or NOT [39]), encoding effective signals in the frequency domain allows one to separate signals independently of their amplitudes (i.e. concentrations).

Figure 2 | The importance of signal dynamics and of downstream signal processing in affecting biological responses

(**A**) A simple system illustrating how two different filters can transduce different features of the identical signal into two different events. (**B**) Simple RC filters (above) can act as a delay line when they are concatenated.


A
The same signal can lead to two different outputs depending on the filtering/detector
One signal
Low-pass filter
High-pass filter
Output 1, e.g. apoptosis
Output 2, e.g. proliferation
B
Multiple LP filters in series act as a delay line

In the most simplistic way, one could imagine a structure (Figure 2A) in which there was an input signal that could be filtered by a low-pass or high-pass filter before being passed downstream – a low-frequency signal would 'go one way' (i.e. be detected by only one 'detector' structure) and a high-frequency signal the other way. In this manner, the same components can change their concentrations, such that they may be at the same instantaneous levels while nevertheless having entirely different outcomes, solely because of the signal processing, frequency response characteristics of the detectors. Of course, the real system and its signal-processing elements will be much more complex than this. There is also precedent for the nonlinear frequency-selective (bandpass) responses of individual multistate enzymes to exciting alternating electrical fields [40–44].

Although the recognition that electrical circuit (signal processing) elements and biological networks are fundamentally similar representations is not especially new (e.g. [36,45–53]), Alon and co-workers [35,37,54] have developed these ideas particularly well. Thus, any element (Figure 2) in a metabolic or signal transduction pathway acts as a resistor-capacitor element [45] (as indeed do any 'relaxing' elements responding to an input, such as an alternating electrical signal

[55]), whereas a series of them acts as a delay line (see [56] or any other textbook of electrical filters, and in a biological context [57]). Similarly, a suitably configured ('coherent') feedforward network provides resistance to small input perturbations (noise – or at least an amount of signal not worth chasing) while transducing large ones (signals) into biological effects [58,59]. Other network structures – which are better seen as signal processing elements – exhibit robustness of their output to sometimes extreme variations in parameters [50,51,60–63].

Thus the recognition that we need to concentrate more on the dynamics of signalling pathways, rather than instantaneous concentrations of their components, means that we need to sample very frequently – preferably effectively in real time – and using single cell measurements to avoid oscillations and other more complex and functionally important dynamics being hidden through the combination signals from individual, out-of-phase cells. It also means that assays for signalling activity, for instance in drug development, should not concentrate just on the signalling molecules but on the structures that the cell uses to detect them.

Other areas and concluding remarks

In the short presentation at the meeting of which this represents the transaction, I covered only the three main aspects described above. However, there are many other pertinent areas that need to be stressed, including the need to integrate SBML models into post-genomic databases with schemas such as those for genomics (GIMS [64]), proteomics (PEDRo [65]) and metabolomics (ArMet [66]). Only then can we have a truly integrative Systems Biology.

I thank numerous colleagues, listed as coauthors in the references, for their participation in a variety of exciting collaborative projects, Q2 and Dr B. Todd for extremely useful discussions concerning electrical circuit theory. I also thank the RSC, EPSRC, BBSRC and NERC for financial support.

References

1 Kell, D.B. (2002) Trends Genet. **18**, 555–559
2 Kell, D.B. and Oliver, S.G. (2004) Bioessays **26**, 99–105
3 Ihekwaba, A., Broomhead, D.S., Grimley, R., Benson, N. and Kell, D.B. (2004) Systems Biol. **1**, 93–103
4 Nelson, D.E., Ihekwaba, A.E.C., Elliott, M., Gibney, C.A., Foreman, B.E., Nelson, G., See, V., Horton, C.A., Spiller, D.G., Edwards, S.W. et al. (2004) Science **306**, 704–708
5 Kell, D.B. (2004) Curr. Opin. Microbiol. **7**, 296–307
6 Kell, D.B. and Mendes, P. (2000) in Technological and Medical Implications of Metabolic Control Analysis (Cornish-Bowden, A, and Cárdenas, M.L., eds.), pp. 3–25 (and see http://dbk.ch.umist.ac.uk/WhitePapers/mcabio.htm), Kluwer Academic Publishers, Dordrecht
7 Westerhoff, H.V. and Palsson, B.O. (2004) Nat. Biotechnol. **22**, 1249–1252
8 Kell, D.B., Darby, R.M. and Draper, J. (2001) Plant Physiol. **126**, 943–951
9 Kell, D.B. (2002) Mol. Biol. Rep. **29**, 237–241
10 Mendes, P. and Kell, D.B. (1998) Bioinformatics **14**, 869–883
11 Mendes, P. and Kell, D.B. (2001) Bioinformatics **17**, 288–289
12 Teusink, B., Passarge, J., Reijenga, C.A., Esgalhado, E., van der Weijden, C.C., Schepper, M., Walsh, M.C., Bakker, B.M., van Dam, K., Westerhoff, H.V. et al. (2000) Eur. J. Biochem. **267**, 5313–5329
13 Pritchard, L. and Kell, D.B. (2002) Eur. J. Biochem. **269**, 3894–3904
14 Mendes, P. and Kell, D.B. (1996) Biosystems **38**, 15–28
15 Koza, J.R., Mydlowec, W., Lanza, G., Yu, J. and Keane, M.A. (2001) Pac. Symp. Biocomput. **6**, 434–445
16 Moles, C.G., Mendes, P. and Banga, J.R. (2003) Genome Res. **13**, 2467–2474
17 Förster, J., Famili, I., Fu, P., Palsson, B.Ø. and Nielsen, J. (2003) Genome Res. **13**, 244–253
18 Wilson, I.D. and Brinkman, U.A. (2003) J. Chromatogr. A **1000**, 325–356
19 Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004) Trends Biotechnol. **22**, 245–252
20 Brown, M., Dunn, W.B., Ellis, D.I., Goodacre, R., Handl, J., Knowles, J.D., O'Hagan, S., Spasic, I. and Kell, D.B. (2005) Metabolomics **1**, 35–46
21 Vaidyanathan, S., Broadhurst, D.I., Kell, D.B. and Goodacre, R. (2003) Anal. Chem. **75**, 6679–6686
22 Vaidyanathan, S., Kell, D.B. and Goodacre, R. (2004) Anal. Chem. **76**, 5024–5032
23 Kell, D.B. and King, R.D. (2000) Trends Biotechnol. **18**, 93–98
24 Langley, P., Simon, H.A., Bradshaw, G.L. and Zytkow, J.M. (1987) Scientific Discovery: Computational Exploration of the Creative Processes, MIT Press, Cambridge, MA
25 King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B. and Oliver, S.G. (2004) Nature (London) **427**, 247–252
26 O'Hagan, S., Dunn, W.B., Brown, M., Knowles, J.D. and Kell, D.B. (2005) Anal. Chem. **77**, 290–303
27 Allen, J.K., Davey, H.M., Broadhurst, D., Heald, J.K., Rowland, J.J., Oliver, S.G. and Kell, D.B. (2003) Nat. Biotechnol. **21**, 692–696
28 Allen, J., Davey, H.M., Broadhurst, D., Rowland, J.J., Oliver, S.G. and Kell, D.B. (2004) Appl. Environ. Microbiol. **70**, 6157–6165
29 Hoffmann, A., Levchenko, A., Scott, M.L. and Baltimore, D. (2002) Science **298**, 1241–1245
30 Nelson, G., Paraoan, L., Spiller, D.G., Wilde, G.J., Browne, M.A., Djali, P.K., Unitt, J.F., Sullivan, E., Floettmann, E. and White, M.R. (2002) J. Cell Sci. **115**, 1137–1148
31 Kell, D.B., Ryder, H.M., Kaprelyants, A.S. and Westerhoff, H.V. (1991) Antonie Van Leeuwenhoek **60**, 145–158
32 Davey, H.M. and Kell, D.B. (1996) Microbiol. Rev. **60**, 641–696
33 Kell, D.B. and Westerhoff, H.V. (1986) FEMS Microbiol. Rev. **39**, 305–320
34 White, T.A. and Kell, D.B. (2004) Comp. Funct. Genom. **5**, 304–327
35 Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Science **298**, 824–827
36 Wolf, D.M. and Arkin, A.P. (2003) Curr. Opin. Microbiol. **6**, 125–134
37 Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U. and Margalit, H. (2004) Proc. Natl. Acad. Sci. U.S.A. **101**, 5934–5939
38 Woodward, A.M., Rowland, J.J. and Kell, D.B. (2004) Analyst **129**, 542–552
39 Buchler, N.E., Gerland, U. and Hwa, T. (2003) Proc. Natl. Acad. Sci. U.S.A. **100**, 5136–5141
40 Westerhoff, H.V., Tsong, T.Y., Chock, P.B., Chen, Y. and Astumian, R.D. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 4734–4738
41 Westerhoff, H.V., Astumian, R.D. and Kell, D.B. (1988) Ferroelectrics **86**, 79–101
42 Woodward, A.M. and Kell, D.B. (1990) Bioelectrochem. Bioenerg. **24**, 83–100
43 Woodward, A.M., Jones, A., Zhang, X., Rowland, J. and Kell, D.B. (1996) Bioelectrochem. Bioenerg. **40**, 99–132
44 Kell, D.B., Woodward, A.M., Davies, E., Todd, R.W., Evans, M.F. and Rowland, J.J. (2004) in Nonlinear Dielectric Phenomena in Complex Liquids (Rzoska, S.J. and Zhelezny, V.P., eds.), pp. 335–344, Kluwer, Dordrecht
45 Mikulecky, D.C. (1983) Am. J. Physiol. **245**, R1–R9
46 Mikulecky, D.C. (2001) Comput. Chem. **25**, 369–391
47 Westerhoff, H.V. and van Dam, K. (1987) Thermodynamics and Control of Biological Free Energy Transduction, Elsevier, Amsterdam
48 Koza, J.R., Mydlowec, W., Lanza, G., Yu, J. and Keane, M.A. (2001) in Proc. GECCO-2001 (Spector, L., Goodman, E.D., Wu, A., Langdon, W.B., Gen, M., Sen, S., Dorigo, M., Pezeshk, S., Garzon, M.H. and Burke, E., eds.), pp. 57–65, Morgan Kaufmann, San Francisco, CA
49 Tyson, J.J., Chen, K. and Novak, B. (2001) Nat. Rev. Mol. Cell. Biol. **2**, 908–916

50 Csete, M.E. and Doyle, J.C. (2002) Science **295**, 1664–1669

51 Tyson, J.J., Chen, K.C. and Novak, B. (2003) Curr. Opin. Cell Biol. **15**, 221–231

52 Koza, J.R., Keane, M.A., Streeter, M.J., Mydlowec, W., Yu, J. and Lanza, G. (2003) Genetic Programming: Routine Human-Competitive Machine Intelligence, Kluwer, New York

53 Kramer, B.P., Fischer, C. and Fussenegger, M. (2004) Biotechnol. Bioeng. **87**, 478–484

54 Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U. (2004) Science **303**, 1538–1542

55 Pethig, R. and Kell, D.B. (1987) Phys. Med. Biol. **32**, 933–970

56 Chen, W.-K. (1986) Passive and Active Filters: Theory and Implementations, Wiley, New York

57 Rosenfeld, N. and Alon, U. (2003) J. Mol. Biol. **329**, 645–654

58 Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002) Nat. Genet. **31**, 64–68

59 Mangan, S. and Alon, U. (2003) Proc. Natl. Acad. Sci. U.S.A. **100**, 11980–11985

60 von Dassow, G., Meir, E., Munro, E.M. and Odell, G.M. (2000) Nature (London) **406**, 188–192

61 Aldana, M. and Cluzel, P. (2003) Proc. Natl. Acad. Sci. U.S.A. **100**, 8710–8714

62 Kitano, H. (2004) Nat. Rev. Genet. **5**, 826–837

63 Schmitt, B.M. (2004) ChemBioChem **5**, 1384–1392

64 Cornell, M., Paton, N.W., Hedeler, C., Kirby, P., Delneri, D., Hayes, A. and Oliver, S.G. (2003) Yeast **20**, 1291–1306

65 Garwood, K.L., McLaughlin, T., Garwood, C., Joens, S., Morrison, N., Taylor, C.F., Carroll, K., Evans, C., Whetton, A.D., Hart, S. et al. (2004) BMC Genomics **5**, 68

66 Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A.R., Taylor, J., Fiehn, O., Goodacre, R., Bino, R., Hall, R. et al. (2004) Nat. Biotechnol. **22**, 1601–1606

67 Westerhoff, H.V. (2001) Metab. Eng. **3**, 207–210

Received 14 January 2005