Network modelling in metabolic systems biology and biotechnology: Understanding the

Douglas B. Kell
(University of Manchester, UK)

Let me start by saying what I am not going to do, and that is to seek to define systems biology! However, it is widely recognized that the chief differences between Systems Biology and traditional molecular biology are (i) the concentration of systems biology (and systems biologists) not on the molecules involved, but on the dynamics of their interactions, and (ii) that systems biology should involve a judicious interplay between modelling, theory, experiment and technology development [1]. As the modelling element is really the key, it is this aspect that I stress here.

Although much of what I shall say also applies to signalling pathways, for reasons of focus I shall mainly concentrate on metabolic networks. These also have two especially useful properties over signalling networks, namely that they are subject to specific stoichiometric and thermodynamic constraints that offer considerable advantages in modelling them.

Parameters and variables

It is at once useful to distinguish the parameters and variables of a system (or model thereof)[2]. The parameters of a dynamical system are those properties of a system that are either inherent to the system of interest or whose values are controlled by an experimenter. In metabolic networks, these include the initial concentrations of enzymes and metabolites, and enzyme kinetic properties such as K_m, k_{cat} and K_i. The variables, by contrast, are those things that change during the time evolution of the system, typically concentrations of metabolites and metabolic fluxes. It is important to recognize that the parameters control the variables and not vice versa, although it is probably more common to measure the variables than the parameters. Especial virtue attaches to seeking to do both simultaneously (i.e. comparing modelled metabolic networks with their metabolomic properties[3,4]).

Metabolic network modelling

It is usual to recognize that the successful modelling of metabolic networks involves a four-stage process[1,5]. The first two stages are qualitative, with the first involving listing all the reactions that are known to occur in the system or organism of interest; nowadays these reaction lists are mainly derived from genomic annotations[6], with curation based on literature data[7], and an important feature being the use of principled descriptors for metabolites[5] and their disambiguation[8] from the many synonyms prevailing. A second qualitative stage adds known effectors, while the third and fourth stages add the known kinetic rate equations and the values of their parameters. Armed with such information, preferably encoded properly in a suitable manner, e.g. in the Systems Biology Markup Language (SBML; http://sbml.org/)[9], it is then possible to provide a stochastic or ordinary differential equation model of the entire metabolic network of interest. Running such a model (using software such as COPASI; www.copasi.org/[10,11]) provides the time evolution of the variables of the system that may be compared with experimental data on the variables. One can then seek to adjust the parameters of the network so that they more nearly reproduce the variables[12]. Methods in which one starts with the variables and seeks to infer the topology and other parameters of the system that generated them are known as *inverse methods* or *system identification* methods, and are considerably more demanding computationally (e.g. [13–15]). It is also usually the case that such systems are *underdetermined*, i.e. that many combinations of parameters can give rise to very similar values of the variables. This is in part due to the fact that natural evolution selected for robustness (especially in topology[3]), which has the advantage (from the experimenter's point of view) that one can then concentrate on those comparatively few (*combinations* of) parameters that have the greatest effect[16].

The purposes and benefits of modelling

I have set these out systematically elsewhere[17], and they include (i) testing whether the model can be made to reflect known experimental facts, (ii)

Key words: *human metabolic network model, metabolic network, modelling strategy, sensitivity analysis, Systems Biology Markup Language (SBML)*

why, how and whither

languages of cells

analysing the model to understand which parts of the system contribute most to some desired properties of interest (especially here the use of so-called *sensitivity analysis*), (iii) hypothesis generation and testing, allowing one rapidly to analyse the effects of manipulating experimental conditions in the model without having to perform complex and costly experiments (or to restrict the number that are performed) – so-called 'what if?' experiments, and (iv) testing what changes in the model would improve the consistency of its behaviour with experimental observations. Overall, given the ability to annotate models in a principled manner, including with the necessary literature references[5,18], metadata[19] and integrated links[20], it is at least arguable that it is the model itself that represents our knowledge of a biochemical system[21]. Certainly, the recognition that our knowledge is dispersed among multiple databases means that there is advantage to be had in joining them up in a loosely coupled manner[1], for which we have found the Taverna system (www.taverna.org.uk/) to be of considerable utility[22,23]. What might we then do with this knowledge

Exploiting our knowledge of biochemical systems properties in biotechnology and medicine

It has long been recognized that the optimization of biotechnological processes needs to be approached rationally[24] (such approaches contrasting with the very sluggish programmes of random mutation and selection that were traditional). The basic issue is that, in part because of the selection by evolution for robustness (something that contrasts with *human-made* networks such as transport networks, incidentally), it is normally necessary to modify the activities of *several* different enzymes in order to increase productivity significantly. This involves a purely (and fundamental) combinatorial problem that is much less easily attacked (initially) by experiment than by simulation. This follows because the number of combinations scales *exponentially* with the number of things one might wish to change, such that choosing combinations of one, two, three or four enzymes from a palette of 1000 involves 1000, 499 500, 1.66×10^8 and 4.14×10^{10} possibilities respectively. However, if one does need to change only (say)

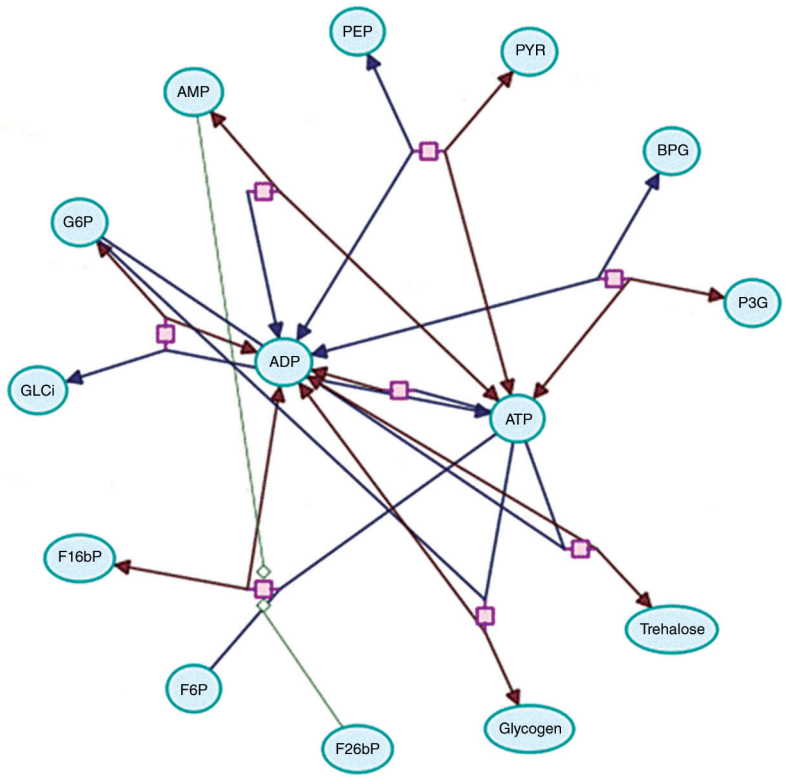

Figure 1. A metabolite-centric model of a metabolic network, here focusing on ADP as part of glycolysis. The yeast glycolytic network, encoded as SBML[53], was visualized using the Arcadia software[54]. PEP, phosphoenolpyruvate; PYR, pyruvate; BPG, bisphosphoglycerate; P3G, 3-phosphoglycerate; F26bP, fructose 2,6-bisphosphate; F6P, fructose 6-phosphate; F16bP, fructose 1,6-bisphosphate; GLCi, intracellular glucose; G5P, glucose 5-phosphate.

three enzymes, as may in fact quite commonly be the case (e.g. [25]), *in silico* analyses allow one to identify them fairly easily (i.e. the computational requirements are very modest, and, because the algorithms can be parallelized efficiently[12], actually scale close to linearly with the available processors). It is then a simple piece of molecular biology to make the necessary constructs. This fundamental relationship between a small number of important parameters and a very large number of *combinations* of those parameters means that the modelling strategy is necessarily highly efficient (and really the only sensible way to do industrial biotechnology in the modern era).

Having established which individual proteins

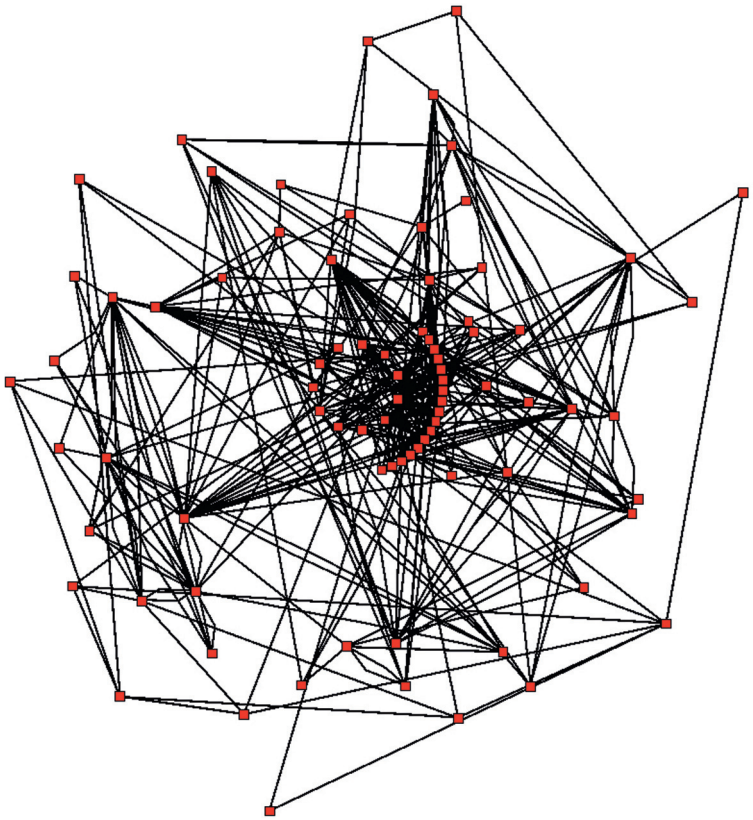

Figure 2. Metabolic network showing the links between enzymes and metabolites that interact with the *Arabidopsis* TCA cycle KEGG classification M00009. Enzymes and metabolites are the nodes (red), interactions are the lines. In total, 43 enzymes and 40 metabolites are shown. Created on Cytoscape using data from VirtualPlant 0.9. (Wiki)

might need improving, the same combinatorial issue pertains for their directed evolution. Thus the number of possible sequences of a protein of 300 amino acids is 20^{300} ($\sim 10^{390}$). The number of sequence variants for m substitutions in a given protein of n amino acids is $19^m \cdot n!/[(n-m)!m!]$. For a protein of 300 amino acids with changes in just one, two and three amino acids, this is 5700, ~16 million and ~30 billion respectively. However, evolutionary optimization methods[26] can speed up such searches considerably, and I might also point to a recent synthetic biology approach[27] in which we evolved efficient nucleic acid aptamers from a very small number (4×10^4) of those ($4^{30}\approx10^{18}$) possible with 30mers. This said, the advance of technology meant that in a related project we could screen all DNA 10mers to understand the nature of the protein sequence-activity landscape[28].

Biomedical applications remain an important focus of systems biology, and one of the goals of metabolic systems biology is the construction of a human metabolic network model[29], with encouraging progress already reported[30,31]. Note that some important aspects, such as iron metabolism[16,32,33], do not depend only on genetically encoded elements, whereas others, such as the transporter molecules important in the cellular uptake of pharmaceutical drugs, remain badly under-recognized[34–37]. A particular focus for purposes of drug discovery is the recognition that, as with the improvement of biotechnological processes, it is necessary for effective drugs to interact with multiple targets simultaneously[38] (whether with one polypharmacologically active drug[39] or with cocktails[40]). Abundant evidence suggests that successful drugs have been 'evolutionarily' selected accordingly[32], whether intentionally or otherwise.

Quo vadis for metabolic systems biology

The problem of biology is (and always has been) the problem of complexity. As we move to ultra-high-throughout measurements of genome sequences and of other '-omes', personalized medicine will soon be a reality. In agriculture, we can anticipate principled plant and animal breeding, as molecular markers for genotype–phenotype mapping[41] are then available for what amounts to every base. Of the many things one might wish to do with a metabolic network model[4], visualization remains a key element[42,43] (e.g. Figure 1), as well as the bringing together of our knowledge, for which automated text mining and related methods are going to be *de rigueur* (e.g. [20,44–46]). We still know much less than we would wish about the interactions between small molecules and proteins[47,48], and high-throughput mass spectrometric methods show promise here[49–51]. Automation is very important[52]. Comparative network analysis, tissue-dependent models (see, e.g., http://proteinatlas.org/), comparative metabolomics, genotype–phenotype mapping and inverse problem solving are likely to be among the chief areas of study, all of which contain substantial elements of modelling and computation. It is a truism that 3 months in the laboratory can save one a whole afternoon on the computer. ■

I thank many colleagues for useful discussions.

References

1 Kell, D.B. (2006) FEBS J. **273**, 873–894
2 Kell, D.B. and Oliver, S.G. (2004) BioEssays **26**, 99–105
3 Kell, D.B. (2004) Curr. Opin. Microbiol. **7**, 296–307
4 Kell, D.B. (2006) Drug Discovery Today **11**, 1085–1092
5 Herrgård, M.J., Swainston, N., Dobson, P. et al. (2008) Nat. Biotechnol. **26**, 1155–1160
6 Henry, C.S., Dejongh, M., Best, A.A., Frybarger, P.M.,

Linsay, B. and Stevens, R.L. (2010) Nat. Biotechnol. **28**, 977–982
7 Thiele, I. and Palsson, B.Ø. (2010) Nat. Protoc. **5**, 93–121
8 Ananiadou, S., Kell, D.B. and Tsujii, J.-i. (2006) Trends Biotechnol. **24**, 571–579
9 Hucka, M., Finney, A., Sauro, H.M. et al. (2003) Bioinformatics **19**, 524–531
10 Hoops, S., Sahle, S., Gauges, R. et al. (2006) Bioinformatics **22**, 3067–3074
11 Mendes, P., Hoops, S., Sahle, S., Gauges, R., Dada, J. and Kummer, U. (2009) Methods Mol. Biol. **500**, 17–59
12 Mendes, P. and Kell, D.B. (1998) Bioinformatics **14**, 869–883
13 Wilkinson, D.J. (2007) Brief. Bioinform. **8**, 109–116
14 Jayawardhana, B., Kell, D.B. and Rattray, M. (2008) Bioinformatics **24**, 1191–1197
15 Wilkinson, S.J., Benson, N. and Kell, D.B. (2008) Mol. Biosyst. **4**, 74–97
16 Kell, D.B. (2010) Arch. Toxicol. **577**, 825–889
17 Kell, D.B. and Knowles, J.D. (2006) in System Modeling in Cellular Biology: from Concepts to Nuts and Bolts (Szallasi, Z., Stelling, J. and Periwal, V., eds), pp. 3–18, MIT Press, Cambridge, MA, USA
18 Dobson, P.D., Smallbone, K. and Jameson, D. (2010) BMC Syst. Biol. **4**, 145
19 Hull, D., Pettifer, S.R. and Kell, D.B. (2008) PLoS Comput. Biol. **4**, e1000204
20 Attwood, T.K., Kell, D.B., McDermott, P., Marsh, J., Pettifer, S.R. and Thorne, D. (2009) Biochem. J. **424**, 317–333
21 Kell, D.B. and Mendes, P. (2008) J. Theor. Biol. **252**, 538–543
22 Li, P., Oinn, T., Soiland, S. and Kell, D.B. (2008) Bioinformatics **24**, 287–289
23 Li, P., Dada, J.O., Jameson, D. et al. (2010) Systematic integration of experimental data and models in systems biology. Bioinformatics **11**, 582
24 Kell, D.B. and Westerhoff, H.V. (1986) FEMS Microbiol. Rev. **39**, 305–320
25 Park, J.H., Lee, K.H., Kim, T.Y. and Lee, S.Y. (2007) Proc. Natl. Acad. Sci. U.S.A. **104**, 7797–7802
26 Handl, J., Kell, D.B. and Knowles, J. (2007) IEEE Trans. Comput. Biol. Bioinformatics **4**, 279–292
27 Knight, C.G., Platt, M., Rowe, W. et al. (2009) Nucleic Acids Res. **37**, e6
28 Rowe, W., Platt, M., Wedge, D., Day, P.J., Kell, D.B. and Knowles, J. (2010) J. R. Soc. Interface **7**, 397–408
29 Kell, D.B. (2007) IUBMB Life **59**, 689–695
30 Duarte, N.C., Becker, S.A., Jamshidi, N. et al. (2007) Proc. Natl. Acad. Sci. U.S.A. **104**, 1777–1782
31 Ma, H., Sorokin, A. and Mazein, A. (2007) Mol. Syst. Biol. **3**, 135
32 Kell, D.B. (2009) BMC Med. Genomics **2**, 2
33 Hower, V., Mendes, P., Torti, F.M. et al. (2009) Mol. Biosyst. **5**, 422–443
34 Dobson, P.D. and Kell, D.B. (2008) Nat. Rev. Drug Discovery **7**, 205–220
35 Dobson, P.D., Lanthaler, K., Oliver, S.G. and Kell, D.B. (2009) Curr. Top. Med. Chem. **9**, 163–184
36 Kell, D.B. and Dobson, P.D. (2009) In Proceedings of the International Beilstein Symposium on Systems Chemistry, pp. 149–168, Logos Verlag, Berlin
37 Giacomini, K.M., Huang, S.M., Tweedie, D.J. (2010) Nat. Rev. Drug. Discovery **9**, 215–236
38 Hopkins, A.L. (2008) Nat. Chem. Biol. **4**, 682–690
39 Hopkins, A.L., Mason, J.S. and Overington, J.P. (2006) Curr. Opin. Struct. Biol. **16**, 127–136
40 Zimmermann, G.R., Lehár, J. and Keith, C.T. (2007) Drug Discovery Today **12**, 34–42
41 Kell, D.B. (2002) Trends Genet. **18**, 555–559
42 Le Novère, N., Hucka, M., Mi, H. et al. (2009) Nat. Biotechnol. **27**, 735–741
43 Pettifer, S.R., Thorne, D., McDermott, P. et al. (2009) BMC Bioinformatics **10**, S19
44 Spasić, I., Simeonidis, E., Messiha, H.L., Paton, N.W. and Kell, D.B. (2009) Bioinformatics **25**, 1404–1411
45 Ananiadou, S., Pyysalo, S., Tsujii, J.i. and Kell, D.B. (2010) Trends Biotechnol. **28**, 381–390
46 Nobata, C., Dobson, P., Iqbal, S.A. et al. (2010) Metabolomics, doi:10.1007/s11306-010-0251-6
47 Leeson, P.D. and Springthorpe, B. (2007) Nat. Rev. Drug Discovery **6**, 881–890
48 Chang, R.L., Xie, L., Xie, L., Bourne, P.E. and Palsson, B.Ø. (2010) PLoS Comp. Biol. **6**, e1000938
49 Muckenschnabel, I., Falchetto, R., Mayr, L.M. and Filipuzzi, I. (2004) Anal. Biochem. **324**, 241–249
50 Li, X., Gianoulis, T.A., Yip, K.Y., Gerstein, M. and Snyder, M. (2010) Cell **143**, 639–650
51 Kell, D.B. (2011) Nat. Chem. Biol. **7**, in press
52 Swainston, N., Golebiewski, M., Messiha, H. (2010) FEBS J. **277**, 3769–3779
53 Pritchard, L. and Kell, D.B. (2002) Eur. J. Biochem. **269**, 3894–3904
54 Villéger, A.C., Pettifer, S.R. and Kell, D.B. (2010) Bioinformatics **20**, 1470–1471

Professor Douglas Bruce Kell graduated from St John's College, Oxford with a BA Hons in Biochemistry in 1975 and a Doctor of Philosophy (Oxon) in 1978 with a thesis entitled 'The Bioenergetics of Paracoccus denitrificans. *From 1978 to 2002, he worked at Aberystwyth University, moving to UMIST (now the University of Manchester) in 2002 as EPSRC/RSC Research Chair in Bioanalytical Sciences. He is currently Chief Executive of the Biotechnology and Biological Sciences Research Council (BBSRC). email: dbk@manchester.ac.uk*