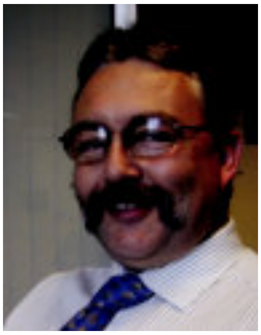
Expert Review of Molecular Diagnostics

ISSN: 1473-7159 (Print) 1744-8352 (Online) Journal homepage: informahealthcare.com/journals/iero20

Metabolomic biomarkers: search, discovery and validation

Douglas B Kell



Douglas B Kell
University of Manchester,
School of Chemistry & Manchester
Interdisciplinary Biocentre,
131 Princess Street, Manchester
M1 7DN, UK
dbk@manchester.ac.uk
www.dbkgroup.org

Metabolomic biomarkers: search, discovery and validation

'… we will have good systems biology models of metabolism and metabolomics long before the same can be said of gene or protein networks.'

Expert Rev. Mol. Diagn. 7(4), 329–333 (2007)

Why metabolomics?

With the mainstream concentration during the reductionist molecular biology era being on qualitative studies of macromolecules, metabolism has become the Cinderella subject of this period [1]. However, this strategy was latterly seen as a partial failure (as evidenced by genomics), since it failed to uncover the existence (let alone the function) of approximately half the genes in even well-worked organisms, such as *Escherichia coli* and baker's yeast. Subsequently, this led to a data-driven rather than hypothesis-dependent strategy [2,3], accompanied by a much greater emphasis on the phenotype at a global omics level, which has now passed from a focus on the transcriptome via the proteome to the metabolome [4–18]. There are many reasons why it is appropriate to concentrate on the metabolome (BOX 1), the most significant being that it hinges upon the properties of networks, and is thus an issue of systems biology [8,19–23].

'There are many reasons why it is appropriate to concentrate on the metabolome, the most significant being that it hinges upon the properties of networks, and is thus an issue of systems biology.'

The metabolome is amplified relative to the transcriptome or the proteome

Although some of its roots can be found earlier, it was the genius of Kacser and Burns [24] and of Heinrich and Rapoport [25] to recognize that metabolic networks were – and needed to be treated as – systems of interacting components that could not be understood solely in isolation, and that various important and mathematically provable theorems followed from the formalism that they developed, termed metabolic control analysis (MCA). These theorems are known as the flux-control and concentration-control summation theorems [26,27]. These theorems effectively demonstrate that, while small changes in the activities of individual enzymes (hence in their expression as the proteome and transcriptome) have little effect on metabolic fluxes, they can and do have substantial changes on metabolite concentrations. This is why the metabolome is normally amplified relative to the transcriptome and the proteome. In extreme cases, concentrations of metabolites can change without any change in flux at all [28]. A tutorial on MCA (largely written by Pedro Mendes) is available on my website [101], while other reviews of MCA include [29–31]. The converse of these analyses is that if one wishes to increase fluxes while minimizing changes in metabolite concentrations, it is necessary to manipulate the activities of many pathway enzymes simultaneously [32]. Finally, it should be noted that MCA is really a version of a local sensitivity analysis for small changes in parameters, and although this can be of substantial value [33–35], there are many other strategies that may be more global and more powerful (albeit while sacrificing the summation theorems) [36–39].

  ISSN 1473-7159

A pipeline for metabolomic biomarkers

Central to modern experimental design and bioinformatics is the concept of a pipeline or workflow of individually linked steps that must be performed correctly to achieve the desired results [14,15,40–42]. In metabolomics, such a pipeline (largely illustrated rather shamelessly here with our own work) includes [43]:

- Design of the experiment (to include adequate sample sizes without confounding variables [44])
- Optimization of the instruments that perform the measurements [45,46]
- Various kinds of data preprocessing, such as deconvolution, normalization and outlier removal [47,48]
- Data storage in well-architected databases obeying international standards [49,50]
- A variety of supervised and unsupervised schemes for classifying the samples into different groups [7,47,51–56]

Finally, it is vital to note that the methods of multivariate statistics and machine learning that are employed for this are at once both very powerful and very dangerous [57,58], and it is all too easy to produce clusters or models that are simply statistical artifacts [59–63]. Only the methods of external validation can overcome this [44,64], and frankly, the literature is absolutely full of complete rubbish resulting from a combination of over-optimism in the face of ostensibly positive findings, statistical ignorance and the fear of journals to scrutinize data too carefully lest they find something unpleasant. Our view is that we can only hope to see a serious improvement in the situation when all the data and meta-data from which conclusions are drawn are made publicly available in electronic form [44]. Marking data properly with suitable ontologies or other semantic markups [65,66] is also vital to allow enhanced reasoning over the internet [67–71]. With apologies to Marshall McLuhan, we consider that 'the Markup is the Model' [13,72].

'…we can only hope to see a serious improvement in the situation when all the data and meta-data from which conclusions are drawn are made publicly available in electronic form.'

Metabolomic biomarkers are increasingly becoming available

In a sense, metabolomics is only chemical pathology writ large, since metabolites are, of course, widely used in disease diagnosis today; however, the number of such metabolites presently used is pathetically small (e.g., glucose, cholesterol, creatinine, urea, uric acid and triglycerides). By contrast, the number of metabolites we know about in humans is continually climbing [15,73,74], albeit that there are many molecules considered or known to be produced by humans that are not yet in these databases (for one unexpected example, see [75], and for another recent one, see [76]). Some areas of metabolism, such as transmembrane transport and metabolic transactions involving metals, are especially poorly represented. Our own experience is that, in many cases, considerable numbers of metabolites that are not previously recognized or used in disease diagnoses will be found when modern methods of metabolomics are applied [77–79].

Ushering in the future: metabolomic biomarkers meet systems biology

One property of biological systems is that they are controlled by their parameters. In the case of metabolic networks, these are the concentrations (activities) of enzymes, the concentrations of fixed flux-generating metabolites, and the kinetic and binding constants (e.g., K_m and k_{cat}) of the enzymes and their effectors. The variables of the system, which can be modeled in any number of modeling packages (e.g., Gepasi [80–82] and Copasi [83]), are then the metabolite concentrations and fluxes over time and in the steady state (should such exist), and variables are the effects and not the causes of a system's behavior. It is curious then that we are concentrating on measuring variables rather than parameters [20], since this then leads to an 'inverse problem' [84] or 'system identification' problem [85] in which we seek to infer the parameters from the variables [86,87]. We must attack these problems from both sides, simultaneously creating the parameterized metabolic network models while constraining their possible forms and

Box 1. Why metabolomics?

- It is downstream: changes in the metabolome (metabolite concentrations, not fluxes) are amplified relative to changes in the transcriptome and the proteome, and are numerically more tractable.
- There is no need for whole genome sequences or large expressed sequence tag databases for each species.
- Metabolic profiling is much cheaper with very much higher throughput compared with proteomics and transcriptomics, making it feasible to examine large numbers of samples from organisms that have been grown under or exposed to a wide range of conditions.
- The technology is generic as a given metabolite (unlike a transcript or protein) is the same in every organism that contains it.
- Metabolic networks have thermodynamic and stoichiometric constraints that can make them easier to understand than, for example, signalling networks.
- Metabolomic methods have already been shown to be highly effective.
- Compendia of genome-wide metabolomes and metabolic networks are available.

values using the measured metabolomes as constraints. Bringing these disparate data together will best be acheived by adopting workflow strategies in an environment such as Taverna [14,15,40,41,88], because:

- Metabolic networks have major thermodynamic and stoichiometric constraints [23,89]
- Amplification is inherent in metabolomics [5,90]
- Metabolomics experiments are cheap and can thus be performed on many samples with many replicates [91]

I am confident that we will have good systems biology models of metabolism and metabolomics long before the same can be said of gene or protein networks. Provided that the search and validation steps are performed correctly, the prospects for finding metabolomic biomarkers are excellent.

Acknowledgements

I thank the Biotechnology and Biological Sciences Research Council, Engineering and Physical Sciences Research Council and British Heart Foundation for financial support, the Royal Society/Wolfson Foundation for a Research Merit Award, and many colleagues for very enjoyable and stimulating discussions. This is a contribution from the Manchester Centre for Integrative Systems Biology (www.mcisb.org).

References

1 Brenner S. *Loose ends.* Current Biology, London, UK (1997).

2 Kell DB, Oliver SG. 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 (2004).

3 Brent R, Lok L. A fishing buddy for hypothesis generators. *Science* 308(5721), 504–506 (2005).

4 Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 16(9), 373–378 (1998).

5 Raamsdonk LM, Teusink B, Broadhurst D *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 19, 45–50 (2001).

6 Harrigan GG, Goodacre R (Eds). *Metabolic Profiling: its Role in Biomarker Discovery and Gene Function Analysis*. Kluwer Academic Publishers, MA, USA (2003).

7 Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22, 245–252 (2004).

8 Kell DB. Metabolomics and systems biology: making sense of the soup. *Curr. Op. Microbiol.* 7, 296–307 (2004).

9 Vaidyanathan S, Harrigan GG, Goodacre R (Eds). *Metabolome Analyses: Strategies for Systems Biology*. Springer, NY, USA (2005).

10 Tomita M, Nishioka T (Eds). *Metabolomics: the Frontier of Systems Biology*. Springer, Tokyo, Japan (2005).

11 Dunn WB, Ellis DI. Metabolomics: current analytical platforms and methodologies. *Trends. Anal. Chem.* 24(4), 285–294 (2005).

12 Dunn WB, Bailey NJC, Johnson HE. Measuring the metabolome: current analytical technologies. *Analyst* 130(5), 606–625 (2005).

13 Kell DB, Brown M, Davey HM *et al.* Metabolic footprinting and systems biology: the medium is the message. *Nat. Rev. Microbiol* 3(7), 557–565 (2005).

14 Kell DB. 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 (2006).

15 Kell DB. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Disc. Today* 11(23–24), 1085–1092 (2006).

16 van der Greef J, Hankemeier T, McBurney RN. Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics* 7(7), 1087–1094 (2006).

17 Lenz EM, Wilson ID. Analytical strategies in metabonomics. *J. Proteome Res.* 6(2), 443–458 (2007).

18 Lindon JC, Nicholson JK, Holmes E (Eds). *The Handbook of Metabonomics and Metabolomics*. Elsevier, Amsterdam, The Netherlands (2007).

19 Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. *Systems Biology in Practice: Concepts, Implementation and Clinical Application.* Wiley/VCH, Berlin, Germany (2005).

20 Kell DB, Knowles JD. The role of modeling in systems biology. In: *System Modeling in Cellular Biology: from Concepts to Nuts and Bolts.* Szallasi Z, Stelling J, Periwal V (Eds). MIT Press, Cambridge, UK 3–18 (2006).

21 Alon U. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, London, UK (2006).

22 Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. *Brief Bioinform.* 7(2), 140–150 (2006).

23 Palsson BØ. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, Cambridge, UK (2006).

24 Kacser H, Burns JA. The control of flux. In: *Rate Control of Biological Processes. Symposium of the Society for Experimental Biology Vol 27.* Davies DD (Ed.). Cambridge University Press, Cambridge, UK 65–104 (1973).

25 Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.* 42, 89–95 (1974).

26 Fell DA. *Understanding the Control of Metabolism*. Portland Press, London, UK (1996).

27 Heinrich R, Schuster S. *The Regulation of Cellular Systems.* Chapman & Hall, NY, USA (1996).

28 Mendes P, Kell DB, Westerhoff HV. Why and when channeling can decrease pool size at constant net flux in a simple dynamic channel. *Biochim. Biophys. Acta* 1289(2), 175–186 (1996).

29 Fell DA. Increasing the flux in metabolic pathways: a metabolic control analysis perspective. *Biotechnol. Bioeng.* 58(2–3), 121–124 (1998).

30 Cornish-Bowden A, Cárdenas ML (Eds). *Technological and Medical Implications of Metabolic Control Analysis*. Kluwer Academic Publishers, Dordrecht, The Netherland (2000).

31 Cascante M, Boros LG, Comin-Anduix B *et al.* Metabolic control analysis in drug discovery and disease. *Nat. Biotechnol.* 20(3), 243–249 (2002).

32 Fell DA, Thomas S. Physiological control of metabolic flux: the requirement for multisite modulation. *Biochem. J.* 311(Pt 1), 35–39 (1995).

33 Nelson DE, Ihekwaba AEC, Elliott M *et al.* Oscillations in NF-κB signalling control the dynamics of target gene expression. *Science* 306, 704–708 (2004).

34 Ihekwaba AEC, Broomhead DS, Grimley R, Benson N, Kell DB. Sensitivity analysis of parameters controlling oscillatory signalling in the NF-κB pathway: the roles of IκK and IκBα. *Systems Biol.* 1(1), 93–103 (2004).

35 Ihekwaba AEC, Broomhead DS, Grimley R *et al.* Synergistic control of oscillations in the NF-κB signalling pathway. *IEE Systems Biol.* 152(3), 153–160 (2005).

36 Alis OF, Rabitz H. Efficient implementation of high dimensional model representations. *J. Math. Chem.* 29(2), 127–142 (2001).

37 Saltelli A, Tarantola S, Campolongo F, Ratto M. *Sensitivity Analysis in Practice: a Guide to Assessing Scientific Models*. Wiley, NY, USA (2004).

38 Saltelli A, Ratto M, Tarantola S, Campolongo F. Sensitivity analysis for chemical models. *Chem. Rev.* 105(7), 2811–2827 (2005).

39 Yue H, Brown M, Knowles J *et al.* Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-κB signalling pathway *Mol. Biosyst.* 2, 640–649 (2006).

40 Oinn T, Greenwood M, Addis M *et al.* Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice & Experience* 18(10), 1067–1100 (2006).

41 Oinn T, Li P, Kell DB *et al.* Taverna/myGrid: aligning a workflow system with the life sciences community. In: *Workflows for E-science: Scientific Workflows for Grids.* Taylor IJ, Deelman E, Gannon DB, Shields M (Eds.). Springer, Guildford, UK, 300–319 (2007).

42 Taylor IJ, Deelman E, Gannon DB, Shields M. *Workflows for E-science: Scientific Workflows for Grids*. Taylor IJ, Deelman E, Gannon DB, Shields M (Eds.). Springer, Guildford, UK (2007).

43 Brown M, Dunn WB, Ellis DI *et al.* A metabolome pipeline: from concept to data to knowledge. *Metabolomics* 1, 35–46 (2005).

44 Broadhurst D, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2(4), 171–196 (2006).

45 O'Hagan S, Dunn WB, Brown M, Knowles JD, Kell DB. 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 (2005).

46 O'Hagan S, Dunn WB, Broadhurst D *et al.* Closed-loop, multi-objective optimisation of two-dimensional gas chromatography (GCxGC-tof-MS) for serum metabolomics. *Anal. Chem.* 79(2), 464–476 (2007).

47 Allen JK, Davey HM, Broadhurst D *et al.* High-throughput characterisation of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* 21(6), 692–696 (2003).

48 van der Greef J, Smilde AK. Symbiosis of chemometrics and metabolomics: past, present, and future. *J. Chemometrics* 19(5–7), 376–386 (2005).

49 Jenkins H, Hardy N, Beckmann M *et al.* A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* 22(12), 1601–1606 (2004).

50 Spasic I, Dunn WB, Velarde G *et al.* MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinformatics* 7, 281 (2006).

51 Goodacre R, Kell DB, Bianchi G. Neural networks and olive oil. *Nature* 359, 594 (1992).

52 Goodacre R, Neal MJ, Kell DB. Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Z. Bakteriol.* 284, 516–539 (1996).

53 Kell DB, Darby RM, Draper J. Genomic computing: explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.* 126, 943–951 (2001).

54 Kell DB. Genotype:phenotype mapping: genes as computer programs. *Trends Genet.* 18(11), 555–559 (2002).

55 Allen J, Davey HM, Broadhurst D *et al.* Discrimination of the modes of action of antifungal substances by use of metabolic footprinting. *Appl. Env. Micr.* 70, 6157–6165 (2004).

56 Goodacre R. Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *J. Exp. Bot.* 56(410), 245–254 (2005).

57 Duda RO, Hart PE, Stork DE. *Pattern Classification, 2nd Ed.* John Wiley, London, UK (2001).

58 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, Berlin, Germany (2001).

59 Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat. Rev. Cancer* 4(4), 309–314 (2004).

60 Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat. Rev. Cancer* 5(2), 142–149 (2005).

61 Ioannidis JP. Why most published research findings are false. *PLoS Med* 2(8), e124 (2005).

62 Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294(2), 218–228 (2005).

63 Kirschenlohr HL, Griffin JL, Clarke SC *et al.* Proton NMR analysis of plasma is a weak predictor of coronary artery disease. *Nat. Med.* 12, 705–710 (2006).

64 Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 3201–3212 (2005).

65 Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. *Trends Biotechnol.* 24, 571–579 (2006).

66 Ananiadou S, McNaught J (Eds). *Text Mining in Biology and Biomedicine*. Artech House, London, UK (2006).

67 Fensel D, Hendler J, Lieberman H, Wahlster W (Eds). *Spinning the Semantic Web*. MIT Press, Cambridge, MA, USA (2003).

68 Stevens R, Bodenreider O, Lussier YA. Semantic webs for life sciences. *Pac. Symp. Biocomput.* 112–115 (2006).

69 Berners-Lee T, Hendler J. Publishing on the semantic web. *Nature* 410(6832), 1023–1024 (2001).

70 Hendler J. Science and the semantic web. *Science* 299(5606), 520–521 (2003).

71 Baker CJO, Cheung K-H (Eds). *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. Springer, NY, USA (2007).

72 McLuhan M, Fiore Q. *The Medium is the Massage*. Penguin Books, London, USA (1971).

73 Duarte NC, Becker SA, Jamshidi N *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA* 104(6), 1777–1782 (2007).

74 Wishart DS, Tzur D, Knox C *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35(Database issue), D521–D526 (2007).

75 Boettcher C, Fellermeier M, Boettcher C, Drager B, Zenk MH. How human neuroblastoma cells make morphine. *Proc. Natl Acad. Sci. USA* 102(24), 8495–8500 (2005).

76 Williams PJ, Gumaa K, Scioscia M, Redman CW, Rademacher TW. Inositol phosphoglycan P-type in preeclampsia: a novel marker? *Hypertension* 49(1), 84–89 (2007).

77 Kenny LC, Dunn WB, Ellis DI *et al.* Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics* 1(3), 227–234 (2005).

78 Underwood BR, Broadhurst D, Dunn WB *et al.* Huntington's disease patients and transgenic mice have similar pro-catabolic serum metabolite profiles. *Brain* 129(4), 877–886 (2006).

79 Dunn WB, Broadhurst DI, Sasalu D *et al.* Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate. *Metabolomics* (2007) (In Press).

80 Mendes P. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22, 361–363 (1997).

81 Mendes P, Kell DB. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14, 869–883 (1998).

82 Mendes P, Kell DB. MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous cellular systems. *Bioinformatics* 17, 288–289 (2001).

83 Hoops S, Sahle S, Gauges R *et al.* COPASI: a COmplex PAthway SImulator. *Bioinformatics* 22(24), 3067–3074 (2006).

84 Mendes P, Kell DB. On the analysis of the inverse problem of metabolic pathways using artificial neural networks. *Biosystems* 38, 15–28 (1996).

85 Ljung L. *System Identification: Theory for the User*. Prentice Hall, Englewood Cliffs, NJ, USA (1987).

86 Moles CG, Mendes P, Banga JR. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* 13(11), 2467–2474 (2003).

87 Rodriguez-Fernandez M, Mendes P, Banga JR. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* 83(2–3), 248–265 (2006).

88 Oinn T, Addis M, Ferris J *et al.* Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20(17), 3045–3054 (2004).

89 Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol* 2(11), 886–897 (2004).

90 Castrillo JI, Zeef LA, Hoyle DC *et al.* Growth control of the eukaryote cell: a systems biology study in yeast. *J. Biol* 6(1), 4 (2007).

91 Catchpole GS, Beckmann M, Enot DP *et al.* Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl Acad. Sci. USA* 102(40), 14458–14462 (2005).

Website

101 The Metabolic Control Analysis Web http://dbkgroup.org/mca_home.htm

Affiliation

- *Douglas B Kell*
  University of Manchester, School of Chemistry & Manchester Interdisciplinary Biocentre,
  131 Princess Street, Manchester M1 7DN, UK
  dbk@manchester.ac.uk
  www.dbkgroup.org