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.’

**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, the most significant being that it hinges upon the properties of networks, and is thus an issue of systems biology [8,19–23].

**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].
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].

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 achieved 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


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
dkb@manchester.ac.uk
www.dbkgroup.org