ELSEVIER

Metabolomics and systems biology: making sense of the soup

Douglas B Kell

Novel techniques for acquiring metabolomics data continue to emerge. Such data require proper storage in suitably configured databases, which then permit one to establish the size of microbial metabolomes (hundreds of major metabolites) and allow the nature, organisation and control of metabolic networks to be investigated. A variety of algorithms for metabolic network reconstruction coupled to suitable modelling algorithms are the ground substances for the development of metabolic network and systems biology. Even qualitative models of metabolic networks, when subject to stoichiometric constraints, can prove highly informative, and are the first step to the quantitative models, which alone can allow the true representation of complex biochemical systems.

Addresses
Department of Chemistry, UMIST, Faraday Building, Sackville St, PO Box 88, Manchester M60 1QD, UK
e-mail: dbk@umist.ac.uk

Current Opinion in Microbiology 2004, **7**:296–307

This review comes from a themed issue on
Techniques
Edited by Simon Foster and Andrew Camilli

Available online 10th May 2004

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2004.04.012

Abbreviations

FTIR	Fourier transform infrared (spectrometry)
GC-MS	gas chromatography-mass spectrometry
GC-TOF	gas chromatography time-of-flight (mass spectrometry)
LC-MS	liquid chromatography-mass spectrometry
MCA	metabolic control analysis

Introduction

'There is an epoch in the growth of a science during which facts accumulate faster than theories can accommodate them.'

Medawar, P. (1982) in *Pluto's Republic*. Oxford University Press, Oxford, p. 29.

'But one thing is clear: to understand the whole one must study the whole.'

Kacser, H. (1986) in The Organization of cell metabolism, ed., G.R. Welch and J.S. Clegg, Plenum Press, New York, p. 327.

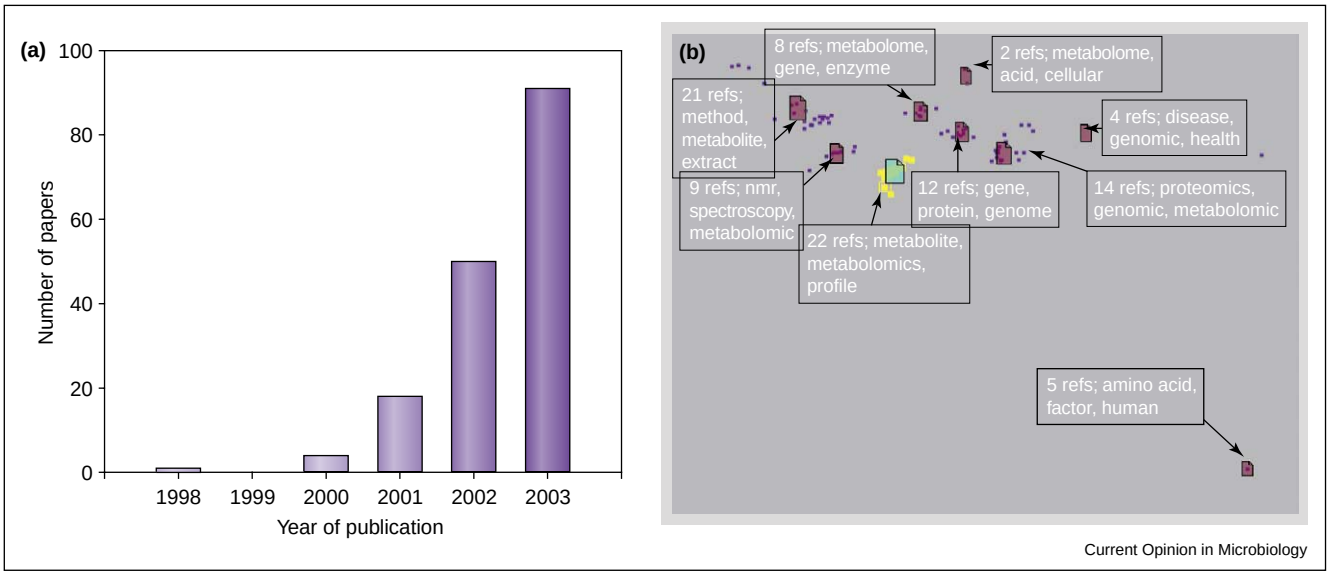
It is becoming increasingly apparent that our ability to generate large quantities of metabolomic or metabolic profiling data will help to open up many previously inaccessible areas of biology. However, such data are merely the inputs or ground substance to systems designed to provide understanding or knowledge, and affecting this may require substantial changes in the conventional and purely hypothesis-dependent, reductionist thinking that has heretofore been common [1••,2]. Metabolomics is a burgeoning field (Figure 1), which produces voluminous data that, like other 'omics' data, should be seen as a resource that contributes specifically to the former half of an iterative cycle of hypothesis-generating and hypothesis-testing phases [2,3•,4,5••] (Figure 2).

In this review, I highlight advances in the way we both gather and use metabolomic data for the large-scale reconstruction of biological systems and for the generation of both testable hypotheses and the predictive models that lie at the heart of systems biology. To 'make sense of the soup', we should concentrate on the questions 'who is there', 'who are they talking to', 'how are metabolic networks organized' and 'what does it mean for our understanding of the cell or organism?' For readers seeking recent reviews of the general field of metabolomics, the following useful surveys have appeared during the review period [6•–8•,9••,10•,11,12•–14•,15••,16•].

Getting the data

The first requirement is to have available techniques that are as comprehensive as possible for metabolic analyses. As the chemistry of different metabolites is very heterogeneous, isolating and measuring them all together ('true metabolomics') is very hard, and most metabolic studies are really 'metabolic profiling' of subsets of chemical classes [8•]. Favoured instrumental methods in different fields (especially plants and microbes versus animals) have differed, largely for historical reasons, but there is increasing convergence to use as many as possible for all samples as their complementarity is appreciated. As well as increasingly refined gas chromatography-mass spectrometry (GC-MS) methods, especially those using gas chromatography time-of-flight mass spectrometry (GC-TOF) instruments that allow much better deconvolution than do most GC-MS instruments because they can record spectra, and thus sample, very quickly, several recent methods appear to show promise. Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometry [17,18•] is a very high-resolution mass spectral method (10^5 – 10^6, with mass accuracy better than 1 ppm), which allows separation and empirical formula measurement of potentially thousands of metabolites

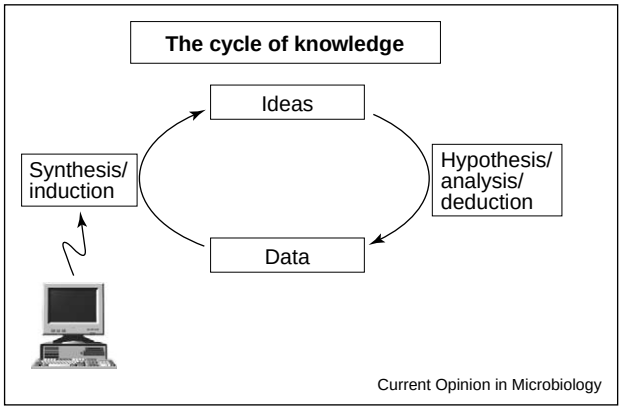
Figure 1



Bibliometric and text mining analysis of the recent metabolomics literature (to the end of 2003). **(a)** The growth in metabolomics papers judged by searching titles and abstracts of Web of Knowledge using 'metabolom'* as the search term. **(b)** Text mining analysis of the metabolomics literature. The text mining tool Ref Viz (Adept Scientific) was used to cluster, according to their keywords, papers that had 'metabolome' or 'metabolomics' in their title. The main groupings seem to be based on whether the emphasis is on technologies, on integration with other omics, or in predicting higher order properties such as disease.

without chromatographic separation. Although comparatively expensive, these sophisticated instruments will have a significant role in the development of metabolomics, especially when problems with ion–ion interferences at high concentrations have been overcome. While their sensitivity and resolution mean that they can be prone to artefacts (gratuitous interferents), there is a clear implication from a series of metabolomic studies that we have only just begun to recognise how many metabolites a typical cell can contain or produce [9••]. For instance, the pioneering studies by Fiehn and colleagues of *Arabidopsis thaliana* [19] indicated the presence of some 326 metabolites; better deconvolution of the data has raised this to over 1000 (Fiehn, personal communication). Thus multidimensional separation methods are coming to the fore [20•] as they can routinely separate more than 1000 compounds [21,22].

Figure 2



Scientific advance may be seen as an iterative cycle linking knowledge/ideas/mental constructs ('ideas') and observations/data ('data'). The hypothetico-deductive mode of reasoning uses background knowledge to construct a hypothesis that is tested experimentally to produce observations. This is only a part of the story, however, as the inductive and abductive modes of reasoning are purely data-driven, i.e. are based purely on generalising rules (or ideas or hypotheses) from examples (and the hypothesis is the end, not the beginning). Because of the high dimensionality of typical data, computer-intensive methods are required to turn the data into knowledge. Scientific advances should (and often do) exploit both deductive and inductive modes of reasoning in an iterative cycle [2,5••].

Liquid-chromatography-mass spectrometry (LC-MS) methods typically have somewhat lower chromatographic resolution than do GC-MS methods, but can access much greater mass ranges because volatilisation (and hence derivatisation) for the chromatographic step is unnecessary. Liquid-phase methods such as LC-MS can be subject to matrix effects (numerous causes often being bundled together under the somewhat unhelpful and inaccurate catch-all term 'ion suppression'), a major one being the presence of non-volatiles, which may reduce the evaporation of volatile ions during the electrospray process [23]. Although reverse phase methods are widely used, normal phase methods can be highly useful in the separation of more polar compounds. However, ion-exchange methods require the use of salts,

which can interfere with mass spectrometry methods, and Tolstikov, Fiehn and colleagues [24,25•] have developed an excellent new hydrophilic interaction liquid chromatography (HILIC) method for this, coupled to electrospray mass spectrometry. In the liquid phase, capillary electrophoresis coupled to mass spectrometry has been used to separate a variety of anionic metabolites from *Bacillus subtilis* [26•], many of which could be quantified accurately by mass spectrometry. Coulometry is another sensitive and somewhat underutilised detection method; Matson and colleagues give an example in which some 600 metabolites could be discriminated [27].

A new development involves the direct injection of the sample into a low-resolution electrospray mass spectrometer [28,29,30••], where quality data are obtainable in less than 1 min, opening up the use of such methods in high-information, high-throughput screening. Matrix-assisted laser desorption ionisation (MALDI) mass spectrometry — the other soft-ionisation mass spectrometric method of choice in proteomics — uses a matrix, typically of aromatic acids, whose molecular weight is similar to those of typical metabolites and thus disallows the mass spectrometric measurement of the latter. Siuzdak and colleagues have neatly avoided this problem via the development of the matrix-free 'desorption ionisation on silicon' (DIOS) method [31]. They have recently shown that when interfaced with a tandem time-of-flight mass spectrometer it provides an extremely convenient tool for metabolite detection [32•], and that the quantitative reproducibility of the method is significantly enhanced by using electrospray deposition [33]. Mass spectrometer conditions are normally adjusted rather arbitrarily, but the search space of possible conditions is huge. The use of a genetic search algorithm to improve substantially the quality of mass spectra from complex mixtures showed both that excellent improvements could be obtained in a comparatively short time and that matrix effects could be reduced or eliminated by varying the mass spectral conditions systematically [34•].

Although it will probably never get close to matching mass spectrometry for sensitivity, NMR continues to improve in resolution and sensitivity [14•,35–37], and has uncovered novel inborn errors of metabolism [38]. Its chief virtue is arguably its non-invasive nature, which can allow one to obtain spatially resolved metabolic profiles and to investigate metabolomics *in vivo* [39], while the nominal independence of the magnitude of the response from different non-exchangeable protons is helpful. For historical reasons it is more commonly applied than mass spectrometry to mammalian samples, and some large-scale studies show reasonably good inter-laboratory reproducibility [40]. Additional specificity can be obtained by using fluorinated metabolites (e.g. [41]). LC-NMR and LC-NMR-MS are newly (re)emerging techniques, reviewed briefly in [42,43] and extensively in [44].

Like NMR, vibrational spectroscopies such as Raman and Fourier transform infrared (FTIR) spectrometry are comparatively insensitive, but the latter especially allows high-throughput screening and classification of biological samples [45], and equally fits the 'omics philosophy' of providing unbiased, whole-system measurements. Among recent examples is its use in detecting the microbial spoilage of meat [46•], where the onset of proteolysis could be clearly identified. FTIR is also a valuable method for the high-throughput screening of mutant strains for different levels of target metabolites [47], a typical recent example being a study of plant cell wall mutants [48].

For fundamental reasons outlined in the theory of metabolic control analysis (MCA), changes in individual enzyme (or presumably transcript) levels have little effect on fluxes but major effects on metabolite concentrations (e.g. [8•,49,50]). Following on from this, where the transcriptome and metabolome have been compared on the same samples, the metabolome indeed seems to be more discriminatory [51].

Finally, we note that most metabolome measurements are 'metabolic snaphots' [49] (and see http://dbk.ch.umist.ac.uk/WhitePapers/mcabio.htm), and what is really desired are methods for reporting, non-invasively and without modifying them, changes in metabolite concentrations in living cells in real time, for which optical strategies are likely to be required. As well as purely spectroscopic methods (UV/Vis, Raman, infrared), molecular biological methods can provide *in situ* sensors, an excellent example being that of the work of Fehr *et al.* [52•].

Making sense of raw metabolomic data

Deconvoluting raw metabolomic data can mean at least three things: (i) working out the signal from metabolites that are imperfectly separated using a 'hyphenated' chromatographic method such as GC-MS, and hence their concentration, (ii) providing a chemical identity for metabolites reproducibly recognised as being present as judged by for example their retention index and mass spectrum, and (iii) using the metabolomic data to reconstruct the metabolic networks in which they participate. (For reasons of space, and although it is an important area, the use of chemometric methods for manipulating the very high-dimensional metabolomic data so as to classify samples according to some scheme of interest (e.g. [53]), is not considered in any detail here.) The first is very important, and improved (and preferably public, non-hidden, non-proprietary) algorithms are required here, with easy-to-use interfaces and high but controllable degrees of automation that process the raw 'hyphenated' data as inputs and produce lists of metabolites as the output. Sumner and colleagues [54••] have produced MSFACTS (metabolomics

spectral formatting, alignment and conversion tools) for this purpose. A high-speed peak alignment algorithm was devised by Synovec's group [55], while an interesting paper by Brereton and colleagues compared a series of methods commonly used for identifying the number of substances in a complex peak [56•], an important prerequisite to optimal deconvolution.

The second of these ('spectrum-to-structure') has not been adequately attacked, and needs automating, since in plants some 80% of metabolites recognised by mass spectrometry have mass spectra that do not appear in the standard libraries, which have concentrated more on organic chemical than on natural biological metabolites. Curiously, one of the earliest forays into computational machine learning, the DENDRAL and METADENDRAL projects [57–59] sought explicitly to enquire as to whether this problem could be automated, but comparatively little in the mass spectral world has happened since. While the picture for NMR data is somewhat rosier [60,61,62•,63,64], this 'spectrum-to-structure' problem remains an important component of 'making sense of the soup', as it is hard to argue that we understand a metabolic system when we do not even know what most of its metabolites are.

The third of these 'deconvolution' issues, metabolic network reconstruction, is dealt with below.

However, another important feature of post-genomics data is represented by their storage, along with the relevant metadata, in suitable databases constructed according to a public and standardised data model [6•,7•,65•], as has already happened for transcriptomic [66,67] and proteomic data [68•,69].

Developments in metabolic modelling

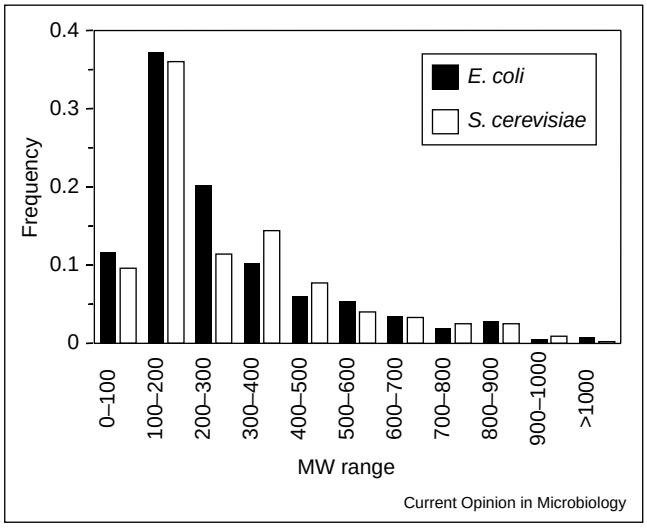
Although we ultimately need to solve the inverse problem [70], metabolic modelling in the 'forward' direction, typically using linked ODE solvers such as *Gepasi* [71] (and see links to others at http://dbk.ch.umist.ac.uk/sysbio.htm#links), continues to provide important knowledge. An especially useful development is the production of data standards, especially the Systems Biology Markup Language (SBML http://www.sbml.org) [72••], which will allow interoperability between different models, including their metadata, and modelling packages. Regarding the latter, the well-known E-cell system [73] has now been revamped and extended to work not just under Linux but also on several operating systems including MS-Windows [74•] (and see http://www.e-cell.org/), which will undoubtedly extend its user base substantially. The FluxAnalyzer [75•] is a MATLAB-based system for the modelling of flux distributions in metabolic pathways, while advances in the important area of metabolome visualization include GenMAPP [76].

Metabolic network organisation and reconstruction

Recognising that metabolites and metabolomes are system variables, the problem of reconstructing metabolic networks, which themselves constitute part of the parameters of the system, becomes an 'inverse problem' or one of system identification [12•,77,78,79••,80•]. Several general questions immediately arise about the size, nature, structure and organisation of metabolic networks [81••]. As to the ostensible size of the 'natural' metabolome, we should comment that many organisms can and will metabolise xenobiotics for non-nutritional purposes, making the potential size of the metabolome practically infinite, and it is possible to read values such as 200 000 for the claimed number of (mainly secondary) metabolites in the plant kingdom [13•]. However, annotated genomic data alone can provide the baseline of reactions, which are more or less well known or may be presumed to occur under at least some conditions in the organism of interest. Although these are available in the general metabolic databases such as KEGG (Kyoto encyclopaedia of genes and genomes; http://www.genome.ad.jp/kegg/) [82], it is necessary to produce organism-specific ones. Important steps in these directions include the Ecocyc project [83,84] (note license conditions) (see http://biocyc.org:1555/ECOLI/class-subs-instances?object=Pathways) and the related MetaCyc sites (http://metacyc.org/), while the latest *E. coli* model has 931 unique biochemical reactions [85•,86•]. The yeast metabolic reaction scheme [87••] is especially useful, as it also exploits biochemical and physiological knowledge (and not merely reconstructions based solely on genomic data), sequence matching and reaction guessing. These models give numbers of metabolites in the hundreds, which are comparatively easy to handle, and while these will be underestimates due to (i) imperfect knowledge, (ii) the lack of specificity among enzymes, and (iii) the production of substances at very low concentrations that we do not routinely detect, they form a very good starting point. An analysis (Figure 3) of the distribution of molecular weights (MWs) of the metabolomes of *E. coli* [88••] and *S. cerevisiae* [87••] suggests that most of the major metabolites are under MW 300. A candidate mouse metabolome is already available [89], and one for humans must be imminent. A very nice example of the use of post-genomic metabolic reconstruction concerns the bacterium *Tropheryma whipplei*, the causative agent of Whipple's disease, whose presence was originally detectable only by molecular methods of nucleic acid hybridisation. Until recently the organism could be grown only in fibroblasts. However, an analysis of its genome showed that it lacked the ability to produce several amino acids, which allowed Renesto *et al.* [90] to design a medium that would support its growth axenically.

Solving inverse problems of metabolic network reconstruction from metabolome data (even with a well-observed time series) in general is hard, but there are

Figure 3



Histogram of molecular weights of typical microbial metabolites. *E. coli* data are from [88••] and were kindly supplied by Dr Irilenia Nobeli (who points out that some of the larger molecular weight metabolites such as lipids are not included as they were not then in Ecocyc). *S. cerevisiae* data were extracted from [87••]. Although the assumptions are rather different it is clear that most of the common microbial metabolites are of relatively low molecular weight.

several constraints that can help to make it easier, and here real progress is being made. First, several important studies suggest that metabolic networks often show a scale-free [81••,91•,92•] or 'small world' [93] kind of organisation. Second, there are important stoichiometric constraints and 'elementary modes' [94,95,96•,97••,98••] that restrict both the networks and their regulations which are possible, although further restrictions are necessary to make even these analyses manageable [99•]. As with blood vessels in higher organisms, the metabolic fluxes in *E. coli* are arranged into a major 'backbone' that dominates the main fluxes [100••]. However, we note that many of these representations use graph theory, which is somewhat 'static', and may fail to capture the richness of the dynamics and control of such networks [101•].

This said, a particularly elegant strategy combines the networks that are reconstructed qualitatively from the genomic data with the constraints imposed by (quantitative) mass balancing [102,103••]. This allows one, subject to additional constraints about how cells 'choose' or have evolved to regulate their fluxes in terms of growth rate or growth yield (most likely the former [104]), to make some extremely powerful and successful predictions of whole-cell behaviour at the metabolic and physiological levels from such *in silico* analysis alone [105••,106,107•,108••, 109,110••]. The success of this endeavour highlights the very great importance of the topological structure of metabolic networks (independently of their kinetic properties), to their effective functioning.

An especially helpful analysis of metabolic networks in terms of 'network motifs' was introduced by Alon and colleagues [111••,112••] (and see also much work over the years more specifically on the design of metabolic and, especially, genetic networks by Savageau and colleagues, for example [113,114•]). Network motifs, by loose analogy with protein structural motifs, are arrangements of reactions, including feedback structures, which regularly occur in biology and are therefore assumed to have functional use [115•]. Codifying them is thus a very important activity. Of all the possible feedback arrangements between n separate elements, a very restricted subset is found to occur regularly in nature (i.e. to have been selected by evolution). One example is the series of feed-forward loop network motifs [116•]. As well as those bestowed by nature, it is possible to produce artificial regulatory networks. To this end, Guet *et al.* [117] used combinatorial methods to vary the 'logical phenotypes' exhibited by various strains of *E. coli*, while Pilpel and colleagues did the same for regulatory networks [118].

A development of MCA called co-response analysis [119] has proved of value both in pattern recognition analysis of the metabolome [120] and in recognising that the co-variation of pairs of metabolites from the same organism under different conditions can provide very useful information of their 'connectedness'. Thus Kose *et al.* [121], drawing on elements of graph theory, developed clique correlation analysis, and especially the use of maximal cliques, for the visualisation of metabolomic data. More recently, Steuer *et al.* [122•] have sought to relate metabolite covariance matrices, while statistical and machine learning methods have proved useful — as with transcriptomics and proteomics data — in deconstructing the highly multivariate data that metabolomics provides [123•]. In an analysis of *E. coli* mutants, a genetic algorithm analysis of direct injection mass spectra identified just two or three peaks that served to pinpoint the nature of the mutation involved [124]. Indeed in our hands genetic algorithms and genetic programming have proved extremely successful strategies for discovering which metabolites are most important for 'explaining' some biological process of interest (e.g. [3•,30••,46•125–128,129•]). In many cases we simply do not know the kinetic properties of the system's components, and genetic programming has also proved useful in the analysis of the inverse problem of metabolism referred to above [130••]. Another very useful strategy is to approximate the kinetic properties with generic rate equations, for example by power laws [113], while Heijnen and colleagues (e.g. [131•] have more recently exploited lin-log kinetic equations for this purpose.

Perturbing, rather than merely observing, metabolic networks is much closer to the (not so) new 'systems' biology that is largely aimed at solving the inverse problem of genetic and metabolic networks. However, deciding

which perturbation to make is a very interesting area (known as 'active learning' [132,133]), as this can greatly improve the quantity and quality of information available from a particular experiment [5••]. Vance *et al.* [134•] presented a computational strategy for working out the causal connectivities of metabolic networks from such perturbations, and a somewhat related strategy was used by Kholodenko and colleagues for genetic and signalling networks [135•]. This, and a slightly different strategy by de la Fuente, Mendes and colleagues [136•] (which also used linear perturbation theories as in MCA) are summarised in a readable way by Stark, Callard and Hubank [101•], while Ross gives a differently focused overview of recent advances in deconstructing chemical networks from time series data [137•]. In a similar vein, the site of interaction of mitomycin C in a small subnetwork of the *E. coli* SOS system was identified [138•].

Integrating metabolomics into systems biology

Although such ideas are very far from being new [139,140], there is a huge interest in understanding complex biological systems from the more holistic 'systems' point of view. Such strategies — contrasted with those of the qualitative and more reductionsist molecular biology era [1••,141] — are usually considered to involve both quantitative experimentation and mathematical simulation/modelling (see above) in an iterative fashion [13•,78,142,143]. Although metabolomics measurements have a major role to play in metabolic network reconstruction, true systems biology will require the integration of metabolomic measurements with measurements of the time-dependent concentrations of other types of components. The availability of protein microarrays consisting of all the cloned proteins of an organism will allow them to be screened systematically for drug or metabolite binding [144,145], and this will be a major contribution to integrative biology. A particularly exciting possibility is that of purifying proteins directly from target organisms in a mass spectrometer without cloning them, and thereby producing microarrays directly [146].

Some applications

While the metabolomic data constitute the 'ground substance' for inferring knowledge, and are the focus of this microbiologically oriented review, several very significant findings of basic or applied biomedical interest have emerged from recent metabolomic studies. A major driver is discovering biomarkers (for definition see [147]) or disease status [9••,148,149•]. Thus Nicholson, Grainger and colleagues [150•] used ^{1}H-NMR measurements together with discriminant partial least squares analysis to distinguish various forms of coronary heart disease via blood samples, although only a small number of samples was used. As well as assessing biomarkers for clinical diseases, it is also vital to gain an understanding of the normal human serum metabolome in health [151], where interesting diet-dependent changes could be observed [152].

There is much interest in determining the mode or site of action of compounds, in functional genomics, in target discovery and in toxicity assessment, which are thereby interrelated. Ott and colleagues [153,154•] combined NMR profiles and neural networks with great success in detecting the mode of action of crop protection substances in aqueous extracts of plants, while Nicholson and colleagues have been particularly active in developing the use of metabolite measurements for assessing drug toxicity [10•]. It is obvious that these methods are generic, and that metabolomics will increase in importance in toxicology, mode of action analysis and functional genomics.

Many new signals are waiting to be discovered via the metabolomics approach. Shi *et al.* [155•] used a generalised purification and metabolomics approach to detect a novel signalling metabolite, uric acid, released from dying mammalian cells.

On the biotechnological front, Askenazi and colleagues [156] integrated transcriptome and (limited) metabolome profiles, the latter measured using LC-MS, to improve the yields of lovastatin and (+)-geodin from *Aspergillus terreus*, and it is clear that improved understanding of the metabolic pathways and fluxes [157] to products of biotechnological interest should be revolutionised by the exploitation of the types of metabolomics and network biology methods reviewed here.

Future perspective and concluding remarks

Metabolomics is a burgeoning science (Figure 1) that brings together analytical technology, genomics and computation, and lies at the core of the systems biology agenda. Major areas for development will involve improving the sensitivity, universality and discrimination of our instruments, and this will involve new approaches and better deconvolution. Structural (i.e. chemical) identification of the many uncharacterised metabolites is still a very important and routinely unsolved problem, while integrating metabolic models in genomically characterised organisms with their experimentally determined metabolomes will allow an iterative improvement of our understanding of the latter. This can be seen as the hallmark and purpose of the systems biology agenda.

Acknowledgements

I thank Katya Tarasova and Warwick Dunn for assistance with literature gathering and yeast metabolite histograms, respectively; Warwick Dunn and Irena Spasic for commenting on the manuscript; many colleagues for very useful and stimulating discussions; and the BBSRC, EPSRC and NERC for financial support.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Lazebnik Y: **Can a biologist fix a radio? - or, what I learned while** •• **studying apoptosis**. *Cancer Cell* 2002, **2**:179-182.
A wonderful and amusing piece extolling the virtues of a top-down and quantitative approach to understanding complex biological networks. The radio engineer succeeds because s/he has a wiring diagram, not merely a list of isolated components that may be present but without any knowledge of their values and what they interact with.

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* 2004, **26**:99-105.

3. Kell DB: **Genotype:phenotype mapping: genes as computer** • **programs**. *Trends Genet* 2002, **18**:555-559.
Genetic programming methods can provide a straightforward link between the genotype and for example metabolomic phenotype of an organism, which maybe useful in both genetic and other nonlinear mappings. The strategy could clearly be used to determine important connectivities in nonlinear metabolic networks.

4. Sweetlove LJ, Last RL, Fernie AR: **Predictive metabolic engineering: a goal for systems biology**. *Plant Physiol* 2003, **132**:420-425.

5. King RD, Whelan KE, Jones FM, Reiser PGK, Bryant CH, •• Muggleton SH, Kell DB, Oliver SG: **Functional genomic hypothesis generation and experimentation by a robot scientist**. *Nature* 2004, **427**:247-252.
The authors implemented an entirely closed loop system of automated experimentation in which a computer containing background knowledge proposed the 'best' (most discriminating) experiment, performed it and evaluated it *seriatim* before performing the next cycle, without human intervention. The method was applied to the identification of mutants in aromatic amino acid biosynthesis in baker's yeast, using growth and nutrient feeding experiments. The strategy could be applied to many areas of post-genomic biology and elsewhere.

6. Mendes P: **Emerging bioinformatics for the metabolome**. • *Brief Bioinform* 2002, **3**:134-145.
Metabolome data differ from transcriptome and proteome data and have special bioinformatics needs, which are reviewed here. Special emphasis is placed on the important issue of data visualisation.

7. Sumner LW, Mendes P, Dixon RA: **Plant metabolomics:** • **large-scale phytochemistry in the functional genomics era**. *Phytochemistry* 2003, **62**:817-836.
A first-class and easy-to-read review of the technology and science of metabolomics, concentrating on plants, and including a helpful discussion of database issues.

8. Fiehn O: **Metabolomics: the link between genotypes and** • **phenotypes**. *Plant Mol Biol* 2002, **48**:155-171.
A very nice review, including a detailed discussion of the differences between metabolomics, metabolic profiling and metabolic fingerprinting and methods for performing them. GC-MS remains the most common and 'gold standard' method, but other methods are being introduced.

9. Harrigan GG, Goodacre R (Ed): *Metabolic profiling: its role in* •• *biomarker discovery and gene function analysis* Boston: Kluwer Academic Publishers; 2003.
A useful compendium of articles from many of the main players in metabolomics, covering many issues in more detail than is possible (or likely) in a single review. The cover, which is based on an iceberg, illustrates the inevitable tendency to believe that we have measured everything when we have merely measured all that we, or our chemometrics, can see.

10. Nicholson JK, Connelly J, Lindon JC, Holmes E: **Metabonomics:** • **a platform for studying drug toxicity and gene function**. *Nat Rev Drug Discov* 2002, **1**:153-161.
A review of metabonomics (metabolomics) with specific reference to its use in assessing mechanisms of toxicity from the metabolic profiles in urine. A particularly useful feature of this review is the helpful figure relating specific markers in the profiles to the major target sites.

11. Watkins SM, German JB: **Metabolomics and biochemical profiling in drug discovery and development**. *Curr Opin Mol Ther* 2002, **4**:224-228.

12. Fiehn O, Weckwerth W: **Deciphering metabolic networks**. • *Eur J Biochem* 2003, **270**:579-588.
A review that aims to concentrate on methods aimed at exploiting metabolomic measurements for deciphering metabolic networks, but also covering many of the existing or emerging methods for measuring the metabolome itself.

13. Weckwerth W: **Metabolomics in systems biology**. *Annu Rev* • *Plant Biol* 2003, **54**:669-689.
A review of metabolomics concentrating on its contribution to the emerging field of systems biology.

14. Lindon JC, Holmes E, Nicholson JK: **So what's the deal with** • **metabonomics? Metabonomics measures the fingerprint of biochemical perturbations caused by disease, drugs, and toxins**. *Anal Chem* 2003, **75**:384A-391A.
A review of the 'metabonomics' agenda, concentrating on high-resolution NMR and pattern recognition, although somewhat misrepresenting the origins and usage of the term 'metabolomics'. Stress is laid on the use of automation in making such methods rapid and robust, which means the cost-per-sample can be low even when using rather expensive equipment.

15. Nicholson JK, Wilson ID: **Understanding 'global' systems** •• **biology: Metabonomics and the continuum of metabolism**. *Nat Rev Drug Discov* 2003, **2**:668-676.
A review that stresses the relationship between diet, the gut microflora, and the host's genome, metabolome and disease. It includes a figure illustrating 'metabolic space' with over 5000 samples and provides a clear statement of the authors' distinction between the metabolome and the 'metabonome', the latter of which is said to relate to the metabolic responses to pathophysiological stimuli or genetic modification.

16. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB: • **Metabolomics by numbers: acquiring and understanding global metabolite data**. *Trends Biotechnol*. 2004: in press.
An introductory overview to the field of metabolomics that lays particular stress on handling the data generated.

17. Marshall AG, Hendrickson CL: **Fourier transform ion cyclotron resonance detection: principles and experimental configurations**. *Int J Mass Spectrom* 2002, **215**:59-75.

18. Aharoni A, Ric de Vos CH, Verhoeven HA, Maliepaard CA, • Kruppa G, Bino R, Goodenowe DB: **Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry**. *OMICS* 2002, **6**:217-234.
An FTICR instrument was used to separate metabolites following electrospray or chemical ionisation solely on the basis of their accurate masses, and to identify nine specific metabolites that changed following transfection with a specific MYB transcription factor.

19. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L: **Metabolite profiling for plant functional genomics**. *Nat Biotechnol* 2000, **18**:1157-1161.

20. Guttman A, Varoglu M, Khandurina J: **Multidimensional** • **separations in the pharmaceutical arena**. *Drug Discov Today* 2004, **9**:136-144.
One way to improve separations, and hence improve the effective dynamic range, is to use multidimensional methods in which a first dimension effects a partial separation and imperfectly separated peaks are then resolved in a second dimension. This review gives a very useful overview of existing methods and trends.

21. Lu X, Cai JL, Kong HW, Wu M, Hua RX, Zhao MY, Liu JF, Xu GW: **Analysis of cigarette smoke condensates by comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. 1 acidic fraction**. *Anal Chem* 2003, **75**:4441-4451.

22. Blumberg LM: **Comprehensive two-dimensional gas chromatography: metrics, potentials, limits**. *J Chromatogr A* 2003, **985**:29-38.

23. Niessen WM: **Progress in liquid chromatography-mass spectrometry instrumentation and its impact on high-throughput screening**. *J Chromatogr A* 2003, **1000**:413-436.

24. Tolstikov VV, Fiehn O: **Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry**. *Anal Biochem* 2002, **301**:298-307.

25. Tolstikov VV, Lommen A, Nakanishi K, Tanaka N, Fiehn O: • **Monolithic silica-based capillary reversed-phase liquid**

chromatography/electrospray mass spectrometry for plant metabolomics. *Anal Chem* 2003, **75**:6737-6740.
Monolithic C18 silica allows the use of long columns that give much better separation and, together with peak alignment and deconvolution software, reveal qualitatively many more peaks in the polar metabolome than had previously been construed.

26. Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M, Nishioka T:
• **Quantitative metabolome analysis using capillary electrophoresis mass spectrometry**. *J Proteome Res* 2003, **2**:488-494.
A substantial improvement on this group's previous work combining capillary electrophoresis and low-resolution mass spectrometry. The method (which uses three columns for cationic, anionic and nucleotide/CoA metabolites) was validated with 352 metabolite standards. It was suggested that one can discriminate 1692 metabolites from *Bacillus subtilis* extracts (of which 150 were identified and 83 assigned), a number rather greater than that assessed from genomics. Attractions of the method include its quantitative nature and speed (20 min).

27. Kristal BS, Vigneau-Callahan KE, Matson WR: **Simultaneous analysis of the majority of low-molecular-weight, redox-active compounds from mitochondria**. *Anal Biochem* 1998, **263**:18-25.

28. Goodacre R, Vaidyanathan S, Bianchi G, Kell DB: **Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils**. *Analyst* 2002, **127**:1457-1462.

29. Castrillo JI, Hayes A, Mohammed S, Gaskell SJ, Oliver SG: **An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry**. *Phytochemistry* 2003, **62**:929-937.

30. Allen JK, Davey HM, Broadhurst D, Heald JK, Rowland JJ,
•• Oliver SG, Kell DB: **High-throughput characterisation of yeast mutants for functional genomics using metabolic footprinting**. *Nat Biotechnol* 2003, **21**:692-696.
Ignoring the widely proclaimed potential 'ion suppression effects', the authors showed that rapid and reproducible low-resolution (i.e. unit) mass spectra could be obtained from yeast supernatants grown in a suitable medium. Numerical processing of the data, including the use of genetic programming, allowed them to use such 'metabolic footprinting' data to discriminate *S. cerevisiae* strains lacking just a single gene.

31. Wei J, Burlak JM, Siuzdak G: **Desorption-ionization mass spectrometry on porous silicon**. *Nature* 1999, **399**:243-246.

32. Go EP, Prenni JE, Wei J, Jones A, Hall SC, Witkowska HE,
• Shen Z, Siuzdak G: **Desorption/ionization on silicon time-of-flight/time-of-flight mass spectrometry**. *Anal Chem* 2003, **75**:2504-2506.
The matrix-less desorption ionisation on silicon method can be applied with advantage to metabolomics, and a tandem (dual time-of-flight) mass spectrometer allows rapid identification of the metabolites in question.

33. Go EP, Shen Z, Harris K, Siuzdak G: **Quantitative analysis with desorption/ionization on silicon mass spectrometry using electrospray deposition**. *Anal Chem* 2003, **75**:5475-5479.

34. Vaidyanathan S, Broadhurst DI, Kell DB, Goodacre R: **Explanatory**
• **optimisation of protein mass spectrometry via genetic search**. *Anal Chem* 2003, **75**:6679-6686.
The authors identified 14 settings of an electrospray mass spectrometer that could be modified to improve mass spectral quality. If each could take ten values the search space would be 10^{14}. However, a genetic algorithm allowed novel and high quality areas of the (highly nonlinear and epistatic) search space to be identified using $\sim 10^{12}$ fewer experiments. A genetic programming analysis of the data was used to identify a major and novel reason for the improvement.

35. Viant M: **Improved methods for the acquisition and interpretation of NMR metabolomic data**. *Biochem Biophys Res Commun* 2003, **310**:943-948.

36. Griffin JL: **Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis**. *Curr Opin Chem Biol* 2003, **7**:648-654.

37. Grivet J-P, Delort A-M, Portais J-C: **NMR and microbiology: from physiology to metabolomics**. *Biochimie* 2003, **85**:823-840.

38. Moolenaar SH, Engelke UF, Wevers RA: **Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism**. *Ann Clin Biochem* 2003, **40**:16-24.

39. Wang Y, Bollard ME, Keun H, Antti H, Beckonert O, Ebbels TM, Lindon JC, Holmes E, Tang H, Nicholson JK: **Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning 1H nuclear magnetic resonance spectroscopy of liver tissues**. *Anal Biochem* 2003, **323**:26-32.

40. Lindon JC, Nicholson JK, Holmes E, Antti H, Bollard ME, Keun H, Beckonert O, Ebbels TM, Reily MD, Robertson D *et al.*: **Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project**. *Toxicol Appl Pharmacol* 2003, **187**:137-146.

41. Nishikawa Y, Dmochowska B, Madaj J, Xue J, Guo Z, Satake M, Reddy DV, Rinaldi PL, Monnier VM: **Vitamin C metabolomic mapping in experimental diabetes with 6-deoxy-6-fluoro-ascorbic acid and high resolution 19F-nuclear magnetic resonance spectroscopy**. *Metabolism* 2003, **52**:760-770.

42. Lindon JC, Nicholson JK, Wilson ID: **Directly coupled HPLC-NMR and HPLC-NMR-MS in pharmaceutical research and development**. *Journal of Chromatogr* 2000, **748**:233-258.

43. Lindon JC: **HPLC–NMR–MS: past, present and future**. *Drug Discov Today* 2003, **8**:1021-1022.

44. Albert K: *On-line LC-NMR and related techniques*. New York: Wiley; 2002.

45. Oliver SG, Winson MK, Kell DB, Baganz F: **Systematic functional analysis of the yeast genome**. *Trends Biotechnol* 1998, **16**:373-378.

46. Ellis DI, Broadhurst D, Kell DB, Rowland JJ, Goodacre R: **Rapid**
• **and quantitative detection of the microbial spoilage of meat using Fourier-Transform infrared spectroscopy and machine learning**. *Appl Environ Microbiol* 2002, **68**:2822-2888.
In the spirit of making sense of the (chicken) soup, the authors acquired FTIR spectra from contaminated meat, then used genetic programming to find rules that correlated with a bacterial load of 10^7 g^{-1}. The 'top' rule obtained made clear that the main biochemical indicator of spoilage was proteolysis.

47. Winson MK, Goodacre R, Timmins ÉM, Jones A, Alsberg BK, Woodward AM, Rowland JJ, Kell DB: **Diffuse reflectance absorbance spectroscopy taking in chemometrics (DRASTIC). A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction**. *Anal Chim Acta* 1997, **348**:273-282.

48. Mouille G, Robin S, Lecomte M, Pagant S, Hofte H: **Classification and identification of *Arabidopsis* cell wall mutants using Fourier-Transform InfraRed (FT-IR) microspectroscopy**. *Plant J* 2003, **35**:393-404.

49. Kell DB, Mendes P: **Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era**. In *Technological and Medical Implications of Metabolic Control Analysis*. Edited by Cornish-Bowden A., Cárdenas ML: Kluwer Academic Publishers; 2000:3-25 (and see http://dbk.ch.umist.ac.uk/WhitePapers/mcabio.htm)

50. Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ, Lee PW: **Metabolic control analysis in drug discovery and disease**. *Nat Biotechnol* 2002, **20**:243-249.

51. Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR: **Parallel analysis of transcript and metabolic profiles: a new approach in systems biology**. *EMBO Rep* 2003.

52. Fehr M, Lalonde S, Lager I, Wolff MW, Frommer WB: ***In vivo***
• **imaging of the dynamics of glucose uptake in the cytosol of COS-7 cells by fluorescent nanosensors**. *J Biol Chem* 2003, **278**:19127-19133.
The authors used a glucose-binding system engineered with two GFP variants, whose ability to perform fluorescence resonance energy transfer (FRET) was a function of the glucose concentration. This allowed optical imaging of the glucose concentration. A striking finding for metabolomics was that in COS-7 cells, cytoplasmic glucose concentrations can vary over at least two orders of magnitude.

53. Jonsson P, Gullberg J, Nordström A, Kusano M, Kowalczyk M, Sjöström M, Moritz T: **A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS**. *Anal Chem* 2004:in press.

54. Duran AL, Yang J, Wang L, Sumner LW: **Metabolomics spectral formatting, alignment and conversion tools (MSFACTs)**. *Bioinformatics* 2003, **19**:2283-2293.
•• MSFACTs contains two tools. RTAlign takes peak lists and aligns them, somewhat like image registration does in 2D gel analysis, and RICExtract takes raw 'hyphenated' (e.g. GC-MS) data in ASCII form and reconstructs an aligned series of chromatograms based on retention times with values for the 'peaks' that can subsequently be identified. It is available via http://www.noble.org/PlantBio/MS/MSFACTs/MSFACTs.html.

55. Johnson KJ, Wright BW, Jarman KH, Synovec RE: **High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis**. *J Chromatogr A* 2003, **996**:141-155.

56. Wasim M, Hassan MS, Brereton RG: **Evaluation of chemometric methods for determining the number and position of components in high-performance liquid chromatography detected by diode array detector and on-flow 1H nuclear magnetic resonance spectroscopy**. *Analyst* 2003, **128**:1082-1090.
• Comparison of a series of chemometric methods for deconvoluting peaks in methods such as LC-NMR and LC-DAD (diode array detection).

57. Feigenbaum EA, Buchanan BG: **DENDRAL and META-DENDRAL: Roots of knowledge systems and expert system applications**. *Artif Intell* 1993, **59**:223-240.

58. Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J: **DENDRAL — a Case study of the first expert system for scientific hypothesis formation**. *Artif Intell* 1993, **61**:209-261.

59. Buchanan BG, Feigenbaum EA: **DENDRAL and META-DENDRAL: their application dimensions**. *Artif Intell* 1978, **11**:5-24.

60. Steinbeck C: **SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry**. *J Chem Inf Comput Sci* 2001, **41**:1500-1507.

61. Meiler J, Sanli E, Junker J, Meusinger R, Lindel T, Will M, Maier W, Kock M: **Validation of structural proposals by substructure analysis and 13C NMR chemical shift prediction**. *J Chem Inf Comput Sci* 2002, **42**:241-248.

62. Meiler J, Will M: **Genius: a genetic algorithm for automated structure elucidation from 13C NMR spectra**. *J Am Chem Soc* 2002, **124**:1868-1870.
• A continuation of earlier work from the same group [61] in which a genetic algorithm generates candidate chemical structures (the empirical formula is assumed known from high-resolution mass spectrometry), a neural network is trained to predict chemical shifts, and further candidate structures are evolved predominantly from those whose calculated spectra best match those observed.

63. Fontana P, Pretsch E: **Automatic spectra interpretation, structure generation, and ranking**. *J Chem Inf Comput Sci* 2002, **42**:614-619.

64. Elyashberg ME, Blinov KA, Williams AJ, Martirosian ER, Molodtsov SG: **Application of a new expert system for the structure elucidation of natural products from their 1D and 2D NMR data**. *J Nat Prod* 2002, **65**:693-703.

65. Hardy N, Fuell H: **Databases, data modeling and schemas: database development in metabolomics**. In *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Edited by Harrigan GG, Goodacre R: Kluwer Academic Publishers; 2003:277-291.
• Very useful discussion of schemas, data standards and data models as applied to metabolomics, concentrating on plant metabolomics.

66. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al.*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data**. *Nat Genet* 2001, **29**:365-371.

67. Brazma A, Sarkans U, Robinson A, Vilo J, Vingron M, Hoheisel J, Fellenberg K: **Microarray data representation, annotation and storage**. *Adv Biochem Eng Biotechnol* 2002, **77**:113-139.

68. Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba–Garcia I *et al.*: **A systematic approach to modelling capturing and disseminating proteomics experimental data**. *Nat Biotechnol* 2003, **21**:247-254.
• The first proposal for a data model for proteomics data, described as a UML (unified modelling language) model, together with XML (extensible markup language) and SQL (structured query language) implementations, as well as a description of a suitable proteome repository (PEDRo).

69. Orchard S, Hermjakob H, Apweiler R: **The proteomics standards initiative**. *Proteomics* 2003, **3**:1374-1376.

70. Mendes P, Kell DB: **On the analysis of the inverse problem of metabolic pathways using artificial neural networks**. *Biosystems* 1996, **38**:15-28.

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

72. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A *et al.*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics* 2003, **19**:524-531.
•• Systems biology markup language (SBML) provides a medium for exchanging metabolic models and their metadata regardless of the (SBML-compliant) simulator that was used to produce and run the model.

73. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC *et al.*: **E-CELL: software environment for whole-cell simulation**. *Bioinformatics* 1999, **15**:72-84.

74. Takahashi K, Ishikawa N, Sadamoto Y, Sasamoto H, Ohta S, Shiozawa A, Miyoshi F, Naito Y, Nakayama Y, Tomita M: **E-cell 2: Multi-platform E-Cell simulation system**. *Bioinformatics* 2003, **19**:1727-1729.
• A multi-platform (Windows 98/NT/2000/XP/ Linux) update of the e-cell simulation system for modelling metabolic networks.

75. Klamt S, Stelling J, Ginkel M, Gilles ED: **FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps**. *Bioinformatics* 2003, **19**:261-269.
• The FluxAnalyzer is a software tool in MATLAB, which *inter alia* allows users to perform metabolic flux analysis, flux optimisation, detection of topological features and pathway analysis by elementary flux modes or extreme pathways. Licenses are free to academic users.

76. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data**. *Genome Biol* 2003, http://genomebiology.com/2003/4/1/R7.

77. Weckwerth W, Fiehn O: **Can we discover novel pathways using metabolomic analysis?** *Curr Opin Biotechnol* 2002, **13**:156-160.

78. Noble D: **The future: putting Humpty-Dumpty together again**. *Biochem Soc Trans* 2003, **31**:156-158.

79. Moles CG, Mendes P, Banga JR: **Parameter estimation in biochemical pathways: a comparison of global optimization methods**. *Genome Res* 2003, **13**:2467-2474.
•• Parameter estimation of metabolic networks is recognised explicitly as an ill-conditioned inverse problem. A case study concerning the estimation of 36 parameters of a three-step, nonlinear biochemical dynamic model is used as a benchmark model. Of a series of algorithms, one based on an evolutionary computing method was the most robust and successful for this dataset.

80. Mendes P, Sha W, Ye K: **Artificial gene networks for objective comparison of analysis algorithms**. *Bioinformatics* 2003, **19(Suppl 2)**:II122-II129.
• Arguably the biggest problem with developing algorithms for solving the inverse problem of metabolic and genetic networks is that we do not know what the correct answer is! The obvious way round this is to simulate the time-dependent behaviour of networks of known properties (with added noise) so as to determine which algorithms do best at reconstructing the generating networks successfully. This is what is done here using an artificial gene network generator. Various models are available at http://www.vbi.vt.edu/~mendes/AGN.

81. Barabási A-L, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**:101-113.
•• A superb review of current knowledge about the general properties of networks, including a substantial focus on metabolic networks. The scale-free nature of metabolic networks probably has an evolutionary basis, in that the growth of networks, and especially hubs, occurs by accretion, with more 'linked' nodes being more likely to acquire new

partners. The network motifs seen in metabolic networks, especially triangles, differ greatly from those seen in electrical circuits for example, where squares are common. A must-read paper for network biologists.

82. Wixon J, Kell DB: **The Kyoto Encyclopedia of Genes and Genomes - KEGG - http://www.genome.ad.jp/kegg**. *Yeast* 2000, **17**:48-55.

83. Karp PD, Kummenacker M, Paley S, Wagg J: **Integrated pathway-genome databases and their role in drug discovery**. *Trends Biotechnol* 1999, **17**:275-281.

84. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S: **The EcoCyc Database**. *Nucleic Acids Res* 2002, **30**:56-58.

85. • Reed JL, Vo TD, Schilling CH, Palsson BØ: **An expanded genome-scale model of *Escherichia coli* K12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4**:R54 (open access at http://genomebiology.com/2003/2004/2009/R2054).
This describes an up-to-date metabolic model of *E. coli* (available at http://systemsbiology.ucsd.edu/organisms/ecoli.html). It is both elementally and charge-balanced, and contains 931 reactions and 625 metabolites. A nomenclature for *in silico* models based on conventions for naming plasmids, and reflected in the paper's title is also introduced.

86. • Reed JL, Palsson BØ: **Thirteen years of building constraint-based *in silico* models of *Escherichia coli***. *J Bacteriol* 2003, **185**:2692-2699.
Another useful review of the present state of constraint-based modelling in *E. coli*, complementary to [85•].

87. •• Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: **Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network**. *Genome Res* 2003, **13**:244-253.
The author's combine genomic, biochemical and physiological knowledge to reconstruct a metabolic network for the yeast *Saccharomyces cerevisiae*, leading to a network containing 1155 biochemical reactions and 584 metabolites. Very useful supplementary information is available at http://www.cpb.dtu.dk/models/yeastmodel.html.

88. •• Nobeli I, Ponstingl H, Krissinel EB, Thornton JM: **A structure-based anatomy of the *E. coli* metabolome**. *J Mol Biol* 2003, **334**:697-719.
A very interesting paper in which the authors seek to catalogue the metabolome of *E. coli* from a physico-chemical point of view. The numbers of known 'basal' metabolites are in the hundreds, with a large and effectively continuous spread of physical properties (knowledge of which is useful for designing metabolomics analysis strategies), and most are of fairly low molecular weight (Figure 3). A small family of substructures can be used to build up this metabolome, and a certain level of clustering of metabolites according to physico-chemical properties is then possible.

89. Bono H, Nikaido I, Kasukawa T: **RIKEN GER Group, GSLMembers, Hayashizaki Y, Okazaki Y: Comprehensive analysis of the mouse metabolome based on the transcriptome**. *Genome Res* 2003, **13**:1345-1349.

90. Renesto P, Crapoulet N, Ogata H, La Scola B, Vestris G, Claverie JM, Raoult D: **Genome-based design of a cell-free culture medium for *Tropheryma whipplei***. *Lancet* 2003, **362**:447-449.

91. • Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**:1551-1555.
The metabolic network of 43 organsisms is analysed, with a focus on *E. coli*. Metabolic networks are scale-free networks, and consist of modular building blocks organised hierarchically, with the degree of clustering following a power law. In *E. coli*, networks can also be clustered on the basis of the biochemical type or function of the molecules involved.

92. • Alm E, Arkin AP: **Biological networks**. *Curr Opin Struct Biol* 2003, **13**:193-202.
A useful review of the emerging recognition that biological systems are best viewed as networks and not lists of parts, concentrating on techniques, structures and modules/motifs.

93. Wagner A, Fell DA: **The small world inside large metabolic networks**. *Proc R Soc Lond B Biol Sci* 2001, **268**:1803-1810.

94. Schuster S, Dandekar T, Fell DA: **Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering**. *Trends Biotechnol* 1999, **17**:53-60.

95. Klamt S, Stelling J: **Two approaches for metabolic pathway analysis?** *Trends Biotechnol* 2003, **21**:64-69.

96. • Schuster S, Hilgetag C, Woods JH, Fell DA: **Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism**. *J Math Biol* 2002, **45**:153-181.
Provides a formal definition of elementary modes in a metabolic network, and provides and proves an algorithm that can generate all the elementary flux modes of an arbitrary network containing reversible or irreversible reactions or both.

97. •• Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED: **Metabolic network structure determines key aspects of functionality and regulation**. *Nature* 2002, **420**:190-193.
As well as their formal use, elementary modes show that even in the absence of quantitative parameters the essential structure of a metabolic network alone contains otherwise hidden constraints that can provide very useful predictions on matters such as flexibility (robustness), functionality, regulation and even cell viability.

98. •• Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BØ: **Metabolic pathways in the post-genome era**. *Trends Biochem Sci* 2003, **28**:250-258.
Metabolic pathways are really networks, which can be analysed formally via a variety of mathematically exact methods. This paper provides a useful, systematic and readable account of them, including elementary modes and extreme pathways.

99. • Burgard AP, Nikolaev EV, Schilling CH, Maranas CD: **Flux coupling analysis of genome-scale metabolic network reconstructions**. *Genome Res*. 2004:1-12.
Elementary mode and extreme pathway analysis can lead to many ways of deconvolving real pathways. This paper studies the extent to which fluxes are qualitatively coupled, and shows *inter alia* that there is a reaction 'core' that is necessary for biomass formation (amounting to 28% and 14% of the total reactions for *E. coli* and *S. cerevisiae*, respectively).

100. •• Almaas E, Kovács B, Vicsek T, Oltvai ZN, Barabási A-L: **Global organization of metabolic fluxes in the bacterium *Escherichia coli***. *Nature* 2004, **427**:839-843.
A global analysis of metabolic fluxes in *E. coli* shows that while most metabolic reactions have small fluxes, the overall metabolic activity is dominated by a comparatively small subset of reactions with very high fluxes. The organism responds to changes in growth conditions largely by reorganising the rates of selected fluxes within this high-flux backbone. This may have significant implications for metabolic engineering and the 'design' of high-yielding strains.

101. • Stark J, Callard R, Hubank M: **From the top down: towards a predictive biology of signalling networks**. *Trends Biotechnol* 2003, **21**:290-293.
A readable review concerning reconstructing genetic networks from transcriptomic data, concentrating on [135•] and [136•].

102. Schilling CH, Edwards JS, Palsson BØ: **Toward metabolic phenomics: Analysis of genomic data using flux balances**. *Biotechnol Prog* 1999, **15**:288-295.

103. •• Kauffman KJ, Prakash P, Edwards JS: **Advances in flux balance analysis**. *Curr Opin Biotechnol* 2003, **14**:491-496.
A very helpful review summarising the historical and present development of flux-balancing methods for analysing complex metabolic systems.

104. Westerhoff HV, Hellingwerf KJ, van Dam K: **Thermodynamic efficiency of microbial growth is low but optimal for maximal growth rate**. *Proc Natl Acad Sci USA* 1983, **80**:305-309.

105. •• Famili I, Forster J, Nielson J, Palsson BØ: ***Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network**. *Proc Natl Acad Sci USA* 2003, **100**:13134-13139.
By combining the genomic and metabolic network data in their earlier model [87••], these authors were able to predict the growth yield, P/2e-ratio, maintenance energy yield and metabolite secretion in *S. cerevisiae*, which also allowed them to propose transcriptome levels. The *in silico* results were consistent with the phenotypes observed ~75% of the time.

106. Förster J, Famili I, Palsson BØ, Nielsen J: **Large-scale evaluation of *in silico* deletions in *Saccharomyces cerevisiae***. *OMICS* 2003, **7**:193-202.

107. • Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metab Eng* 2003, **5**:264-276.

Sometimes the constraint-based modelling strategy can lead to quite different flux distributions with equally fast growth rates. This paper shows that these depend substantially on the network characteristics. Clearly an iterative interplay between calculations and experiment is necessary here.

108. Ibarra RU, Edwards JS, Palsson BØ: ***Escherichia coli* K12**
•• **undergoes adaptive evolution to achieve *in silico* predicted optimal growth**. *Nature* 2003, **420**:186-189.
In continuous cultures, there is a selective pressure to evolve to ever-greater growth rates. The question then arises as to how organisms might manage this from a metabolic network point of view. This excellent article shows that *E. coli* evolved experimentally in 700 generations (40 days) from using a sub-optimal network to one consistent with that predicted by the *in silico* constraint-based prediction.

109. Segrè D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proc Natl Acad Sci USA* 2002, **99**:15112-15117.

110. Burgard AP, Pharkya P, Maranas CD: **OptKnock: A bilevel**
•• **programming framework for identifying gene knockout strategies for microbial strain optimization**. *Biotechnol Bioeng* 2003, **84**:647-657.
A variant of flux balance analysis (MOMA – method of minimisation of metabolic adjustment), which recognises that knockout or other strains that are not growing at their maximum growth rate may not have evolved optimal fluxes but such strains may have metabolic fluxes which have minimal redistribution with respect to the flux configuration of the wild type, a suggestion borne out by data.

111. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U:
•• **Network motifs: simple building blocks of complex networks**. *Science* 2002, **298**:824-827.
An analysis of different types of networks shows that metabolic networks and food web networks share related but distinct motifs, and that those observed regularly in nature occur far more often than would be expected for random networks. A 'feed-forward loop' has the useful property of being activated only if the input is persistent.

112. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in**
•• **the transcriptional regulation network of *Escherichia coli***. *Nat Genet* 2002, **31**:64-68.
In *E. coli*, three motifs are particularly prevalent: the feed-forward loop, the 'single-input module', and the 'densely overlapping regulon'. Each of these motifs has a specific type of behaviour and can be argued to have a particular function in gene expression networks.

113. Savageau M: *Biochemical systems analysis: a study of function and design in molecular biology*. Reading, MA: Addison-Wesley; 1976.

114. Wall ME, Hlavacek WS, Savageau MA: **Design of gene circuits:**
• **Lessons from bacteria**. *Nat Rev Genet* 2004, **5**:34-42.
A convenient and clearly expressed review of the behaviour of different genetic networks, concentrating on their quantitative stability, robustness and responsiveness, with emphasis on the role of different transcription factors. The 'optimal' designs — usually but not exclusively those found or exploited by evolution – differ for different kinds of gain (amplification) and whether the system is based on induction or repression.

115. Wolf DM, Arkin AP: **Motifs, modules and games in bacteria**.
• *Curr Opin Microbiol* 2003, **6**:125-134.
A very useful survey of various network motifs and modules and their 'functions' (i.e. behaviour) as switches, oscillators, filters, amplifiers and other activities typical of electrical circuits.

116. Mangan S, Alon U: **Structure and function of the feed-forward**
• **loop network motif**. *Proc Natl Acad Sci USA* 2003, **100**:11980-11985.
Feed-forward network loop motifs come in eight types; four are 'incoherent' ones that act as 'sign-sensitive accelerators', i.e. they speed up the response time of their target genes to a stimulus when it is changed in one direction but not the other. The other four are sign-sensitive delays.

117. Guet CC, Elowitz MB, Hsing W, Leibler S: **Combinatorial synthesis of genetic networks**. *Science* 2002, **296**:1466-1470.

118. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements**. *Nat Genet* 2001, **29**:153-159.

119. Hofmeyr JH, Cornish-Bowden A: **Co-response analysis: a new experimental strategy for metabolic control analysis**. *J Theor Biol* 1996, **182**:371-380.

120. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh M, Berden JA, Brindle KM, Kell DB, Rowland JJ *et al.*: **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations**. *Nat Biotechnol* 2001, **19**:45-50.

121. Kose F, Weckwerth W, Linke T, Fiehn O: **Visualizing plant metabolomic correlation networks using clique-metabolite matrices**. *Bioinformatics* 2001, **17**:1198-1208.

122. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and**
• **interpreting correlations in metabolomic networks**. *Bioinformatics* 2003, **19**:1019-1026.
The authors used a simple Pearson correlation between pairs of metabolites and proposed that these reflected the enzymatic arrangement of the system via the Jacobian of the system. However, although some correlations make sense, an example from a simulation of glycolysis is used to show that these correlations may reflect only poorly the actual pathway(s) that generated them.

123. Taylor J, King RD, Altmann T, Fiehn O: **Application of**
• **metabolomics to plant genotype discrimination using statistics and machine learning**. *Bioinformatics* 2002, **18(Suppl 2)**:S241-S248.
Machine learning methods using metabolomics data as the inputs can be used to discriminate plant ecotypes. Deconvolution allows the major metabolites involved to be pinpointed.

124. Kaderbhai NN, Broadhurst DI, Ellis DI, Goodacre R, Kell DB: **Functional genomics via metabolic footprinting: Monitoring metabolite secretion by *Escherichia coli* tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry**. *Comp Funct Genomics* 2003, **4**:376-391.

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

126. Kell DB: **Defence against the flood: a solution to the data mining and predictive modelling challenges of today**. *Bioinformatics World (part of Scientific Computing News)* 2002, **Issue 1**:16-18 http://www.abergc.com/biwpp16-18_as_publ.pdf.

127. Kell DB: **Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules**. *Mol Biol Rep* 2002, **29**:237-241.

128. Goodacre R, Kell DB: **Evolutionary computation for the interpretation of metabolome data**. In *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Edited by Harrigan GG, Goodacre R: Kluwer Academic Publishers; 2003:239-256.

129. Goodacre R: **Explanatory analysis of spectroscopic data using**
• **machine learning of simple, interpretable rules**. *Vib Spectrosc* 2003, **32**:33-45.
A useful review (albeit in a journal unlikely to be seen routinely by microbiologists) of the use of infrared and Raman spectroscopies in metabolomics and of evolutionary computing methods for analysing the data obtained.

130. Koza JR, Keane MA, Streeter MJ, Mydlowec W, Yu J, Lanza G:
•• *Genetic programming: routine human-competitive machine intelligence*. New York: Kluwer; 2003.
A tour de force in Koza's series on genetic programming (GP), showing how GP can be used to solve many problems in which one needs to synthesise a network which has some desired properties. One of the examples comes from metabolism.

131. Visser D, Heijnen JJ: **Dynamic simulation and metabolic**
• **re-design of a branched pathway using lin-log kinetics**. *Metab Eng* 2003, **5**:164-176.
Describes the use of lin-log kinetics for approximating metabolic networks over a much wider range of validity than usually available in metabolic control analysis (of which this strategy is an extension). Given the present difficulty of measuring kinetic rate equations on a genomic scale, this strategy should be valuable for modellers generally.

132. Raju GK, Cooney CL: **Active learning from process data**. *AIChE J* 1998, **44**:2199-2211.

133. Bryant CH, Muggleton SH, Oliver SG, Kell DB, Reiser P, King RD: **Combining inductive logic programming, active learning and robotics to discover the function of genes**.

Electronic Transactions on Artificial Intelligence 2001, **5**:1-36 (http://www.ep.liu.se/ej/etai/2001/2001/).

134. Vance W, Arkin A, Ross J: **Determination of causal connectivities**
• **of species in reaction networks**. *Proc Natl Acad Sci USA* 2002, **99**:5816-5821.
The propagation of changes through a network is largely determined by the connectivities and their kinetic constants. The former can be used to predict the latter, and the result may be used to propose strategies for the perturbations that are likely to be most informative.

135. Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag E,
• Westerhoff HV, Hoek JB: **Untangling the wires: a strategy to trace functional interactions in signaling and gene networks**. *Proc Natl Acad Sci USA* 2002, **99**:12841-12846.
This paper describes a method (with roots in metabolic control analysis) based on a mathematical derivation that demonstrates how the topology and strength of network connections can be retrieved from experimentally measured network responses to successive perturbations of all modules.

136. de la Fuente A, Brazhnik P, Mendes P: **Linking the genes: inferring**
• **quantitative gene networks from microarray data**. *Trends Genet* 2002, **18**:395-398.
A comparatively straightforward method for identifying not only the genes involved in regulatory networks but the strength of these relationships.

137. Ross J: **New approaches to the deduction of complex reaction**
• **mechanisms**. *Acc Chem Res* 2003, **36**:839-847.
A useful review, highlighting the canonical metric construction method, and others based on genetic algorithms, for constraining the possible chemical reaction (or metabolic) networks consistent with a particular time series of metabolome data. Glycolysis is used as a specific example.

138. Gardner TS, di Bernardo D, Lorenz D, Collins JJ: **Inferring genetic**
• **networks and identifying compound mode of action via expression profiling**. *Science* 2003, **301**:102-105.
Systematic transcriptional perturbations were used to identify regulatory interactions in a nine-gene sub-network of the *E. coli* SOS system. The model permitted the major transcriptional targets of mitomycin C to be identified.

139. von Bertalanffy L: *General System Theory*. New York: George Braziller; 1969.

140. Iberall AS: *Toward a general science of viable systems*: McGraw-Hill; 1972.

141. Maddox JS: **Towards more measurement in biology**. *Nature* 1994, **368**:95.

142. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology**. *Annu Rev Genomics Hum Genet* 2001, **2**:343-372.

143. Kitano H: **Computational systems biology**. *Nature* 2002, **420**:206-210.

144. Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA, Snyder M: **Analysis of yeast protein kinases using protein chips**. *Nat Genet* 2000, **26**:283-289.

145. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al.*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**:2101-2105.

146. Ouyang Z, Takáts Z, Blake TA, Gologan B, Guymon AJ, Wiseman JM, Oliver JC, Davisson VJ, Cooks RG: **Preparing protein microarrays by soft-landing of mass-selected ions**. *Science* 2003, **301**:1351-1354.

147. Biomarkers Definitions Working Group: **Biomarkers and surrogate endpoints: preferred definitions and conceptual framework**. *Clin Pharmacol Ther* 2001, **69**:89-95.

148. Rashed MS: **Clinical applications of tandem mass spectrometry: ten years of diagnosis and screening for inherited metabolic diseases**. *J Chromatogr B Biomed Sci Appl* 2001, **758**:27-48.

149. van der Greef J, Davidov EJ, Verheij E, Vogels J, van der Heijden R,
• Adourian AS, Marple EW, Naylor S: **The role of metabolomics in systems biology: a new vision for drug discovery and development**. In *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Edited by Harrigan GG, Goodacre R: Kluwer; 2003:171-198.
A very clear exposition of the metabolomics strategy for determining possible biomarkers that might be useful in the early stages of disease progression, toxicity evaluation, drug efficacy, and other purposes.

150. Brindle JT, Antti H, Holmes E, Tranter G, Nicholson JK, Bethell HW,
• Clarke S, Schofield PM, McKilligin E, Mosedale DE *et al.*: **Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics**. *Nat Med* 2002, **8**:1439-1444.
The authors used ^{1}H-NMR of blood samples and discriminant partial least squares analysis to discriminate patients with various forms of coronary heart disease. The method consequently shows great promise for providing a replacement for angiography. However, the numbers of samples were rather low, especially when prevalence is taken into account, and the ongoing larger study will be of great interest.

151. Lenz EM, Bright J, Wilson ID, Morgan SR, Nash AFP: **A ^{1}H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects**. *J Pharm Biomed Anal* 2003, **33**:1103-1115.

152. Solanky KS, Bailey NJC, Beckwith-Hall BM, Davis A, Bingham S, Holmes E, Nicholson JK, Cassidy A: **Application of biofluid 1H nuclear magnetic resonance-based metabonomic techniques for the analysis of the biochemical effects of dietary isflavones on human plasma profile**. *Anal. Biochem* 2003, 323.

153. Aranibar N, Singh BK, Stockton GW, Ott K-H: **Automated mode-of-action detection by metabolic profiling**. *Biochem Biophys Res Commun* 2001, **286**:150-155.

154. Ott K-H, Aranibar N, Singh B, Stockton GW: **Metabonomics**
• **classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts**. *Phytochemistry* 2003, **62**:971-985.
NMR spectra of extracts of plants treated with compounds with different modes of action were used as the inputs to trained artificial neural networks. As possibly expected from the discrimination of single-gene knockouts, other substances could then be classified according to their mode of action with significant success.

155. Shi Y, Evans JE, Rock KL: **Molecular identification of a danger**
• **signal that alerts the immune system to dying cells**. *Nature* 2003, **425**:516-521.
The authors purified a molecule from dying cells that had strong adjuvant activity in antigenic stimulation. The molecule turned out to be uric acid. Eliminating uric acid *in vivo* inhibited the immune response to antigens associated with injured cells, but not to antigens presented by activated dendritic cells.

156. Askenazi M, Driggers EM, Holtzman DA, Norman TC, Iverson S, Zimmer DP, Boers ME, Blomquist PR, Martinez EJ, Monreal AW *et al.*: **Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains**. *Nat Biotechnol* 2003, **21**:150-156.

157. Holms H: **Flux analysis and control of the central metabolic pathways in *Escherichia coli***. *FEMS Microbiol Rev* 1996, **19**:85-116.