DOI 10.1002/pmic.200600106 *Proteomics* 2006, *6,* 4716–4723

REVIEW

Metabolomics: Current technologies and future trends

Katherine Hollywood[1], *Daniel R. Brison*[2] *and Royston Goodacre*[1]

[1] School of Chemistry, The University of Manchester, Manchester, UK
[2] Department of Reproductive Medicine, St. Mary's Hospital, Manchester, UK

The ability to sequence whole genomes has taught us that our knowledge with respect to gene function is rather limited with typically 30–40% of open reading frames having no known function. Thus, within the life sciences there is a need for determination of the biological function of these so-called orphan genes, some of which may be molecular targets for therapeutic intervention. The search for specific mRNA, proteins, or metabolites that can serve as diagnostic markers has also increased, as has the fact that these biomarkers may be useful in following and predicting disease progression or response to therapy. Functional analyses have become increasingly popular. They include investigations at the level of gene expression (transcriptomics), protein translation (proteomics) and more recently the metabolite network (metabolomics). This article provides an overview of metabolomics and discusses its complementary role with transcriptomics and proteomics, and within system biology. It highlights how metabolome analyses are conducted and how the highly complex data that are generated are analysed. Non-invasive footprinting analysis is also discussed as this has many applications to *in vitro* cell systems. Finally, for studying biotic or abiotic stresses on animals, plants or microbes, we believe that metabolomics could very easily be applied to large populations, because this approach tends to be of higher throughput and generally lower cost than transcriptomics and proteomics, whilst also providing indications of which area of metabolism may be affected by external perturbation.

Received: February 10, 2006
Revised: April 21, 2006
Accepted: April 27, 2006

Keywords:
Chemometrics / Metabolomics

1 Introduction to metabolomics and systems biology

Since the 1950s, the central dogma of molecular biology was that there was a general unidirectional flow of information from gene to transcript to protein. Enzymes then affect metabolic pathways and hence lead to changes in the phenotype of the organism (Fig. 1A). However, this traditional 'linear' thinking is no longer true and the cellular processes are in reality intimately networked with many feedback-loops (Fig. 1B) and thus should be represented as dynamic protein complexes interacting with neighbourhoods of metabolites (Fig. 1C). The construction, visualisation and understanding of these networks [1] are certainly a big challenge for the life sciences, as is a full understanding of the fluxes through metabolic neighbourhoods and their control [2].

It is fair to comment that molecular biology has generally been bogged down by hypothetico-reductionist thinking where small parts of the jigsaw have been studied in isolation. For example, it was common to work on a 'favourite' gene in isolation using some pre-existing knowledge of its function. Whilst, this approach has yield useful information though in isolation, the whole picture, *i.e.* how that 'favourite' gene might interact with the whole cell, was often missed. The recent ability to analyse mRNA and proteins in a global way by using, for example, microarrays for transcriptomics and 2-DE with MS for proteomics, has allowed full functional analyses at these two levels and is becoming increasingly used for

Correspondence: Professor Royston Goodacre, School of Chemistry, The University of Manchester, PO Box 88, Sackville Street, Manchester, M60 1QD, UK
E-mail: Roy.Goodacre@manchester.ac.uk
Fax: +44-161-306-4519

Abbreviations: ANN, artificial neural networks; **DA**, discriminant analysis; **EC**, evolutionary computing; **GA**, genetic logarithms; **PLS**, partial least squares

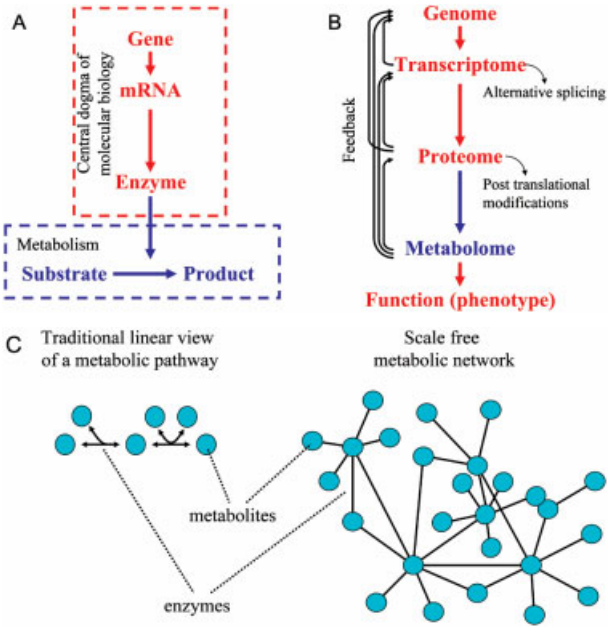

Figure 1. (A) Traditional central dogma of molecular biology where the flow of information goes from gene to transcript to protein, also shown is where enzymes act on metabolism. (B) General schematic of the 'omic organization where the flow of information is from genes to transcripts to proteins to metabolites to function (or phenotype). (C) Traditional linear view of a metabolic pathway and the now accepted view of scale-free connections in a metabolite neighbourhood; nodes are metabolites, whilst the connections represent enzymatic action.

knowledge discovery with functional genomics and biomarker generation [3–6]. The next functional level to analyse is that of the entire metabolite pool or metabolome, and this has more recently been generating much interest [7–10].

The metabolome is the quantitative complement of all the low-molecular weight molecules (typically <3000 m/z) present in cells in a particular physiological or developmental state. Whilst metabolomics is complementary to transcriptomics and proteomics, it may be seen to have special advantages. In particular, we know from the theory underlying metabolic control analysis (MCA) [2] as well as from experiment [11], that while changes in the levels of individual enzymes may be expected to have little effect on metabolic fluxes, they can and do have significant effects on the concentrations of a variety of individual metabolites. In addition, as the 'downstream' result of gene expression, changes in the metabolome are amplified relative to changes in the transcriptome and the proteome, which is likely to allow for increased sensitivity (Fig. 1B). Finally, it is known [12] that metabolic fluxes are not regulated only by gene expression but by post-transcriptional and post-translational events and as such, the metabolome can be considered closer to the phenotype.

In context with systems biology, even functional analyses have emphasized isolated investigations at the level of gene expression, protein translation and the metabolite network with little integration. The re-popularisation of systems biology constitutes a 'paradigm shift' for molecular biology and will initially be dominated by the integrative analyses of these 'omics to generate predictive and hypothesis generating mathematical models to better understand the cell at the systems level [13, 14].

2 Metabolome analyses

As the definition of the metabolome above suggests, in a metabolomics experiment one would like to quantify all the metabolites in a cellular system, which for present purposes can be defined as a cell or tissue in a given state, at a particular time point. For the analysis of mRNA and proteins one 'only' needs to know the genome sequence of the organism and exploit this information using nucleic acid hybridization or protein separation followed by MS (although PTM are problematic). However, the analysis of metabolites is not as straightforward. Whilst triple quadrupole MS instruments can be calibrated for accurate quantification of specific metabolites of known structure, in general, for unknown analytes there is a lack of simple automated analytical techniques that can measure many 100s to 1000s of metabolites quantitatively in a reproducible and robust way. In contrast with transcriptome analysis (but in common with protein analysis) methods are not available for amplification of metabolites and therefore sensitivity is a major issue. Metabolites are generally labile species, by their nature are chemically very diverse, and often present in a wide dynamic range. All of these challenges need to be adequately addressed by the analysis strategy employed. This is currently a very active area within metabolomics and in particular is presenting opportunities for novel analytical instrument manufacture. Finally, in contrast to transcripts or protein identification, metabolites are not organism specific (that is to say, sequence dependent), thus when one has learnt how to measure the metabolite once, the analytical protocol is equally applicable to prokaryotes, fungi, plants and animals.

Before any metabolome measurements are taken it is essential that metabolism is stopped as quickly as possible. This is because enzymes are still active; indeed, in yeast glycolysis metabolite turn over has been estimated to occur within seconds [15]. For unicellular organisms or biofluids this is usually achieved by spraying the biomass into very cold (<-40°C) 60% buffered methanol [16]. Whilst for animal and plant tissues, liquid N_2 is used to snap freeze the sample, this is followed by mechanical disruption, which is employed to release metabolites [17].

The next stage of the analysis is to extract the metabolites. There are many different methods [16, 18, 19] and the most common ones are: (i) acid extraction using perchloric acid ($HClO_4$), followed by freeze thawing, then neutralisation with potassium hydroxide (KOH); (ii) alkali extraction typically using sodium hydroxide (NaOH), followed by heating (80°C); and (iii) ethanolic extraction by boiling the sampling in ethanol (C_2H_5OH) at 80°C.

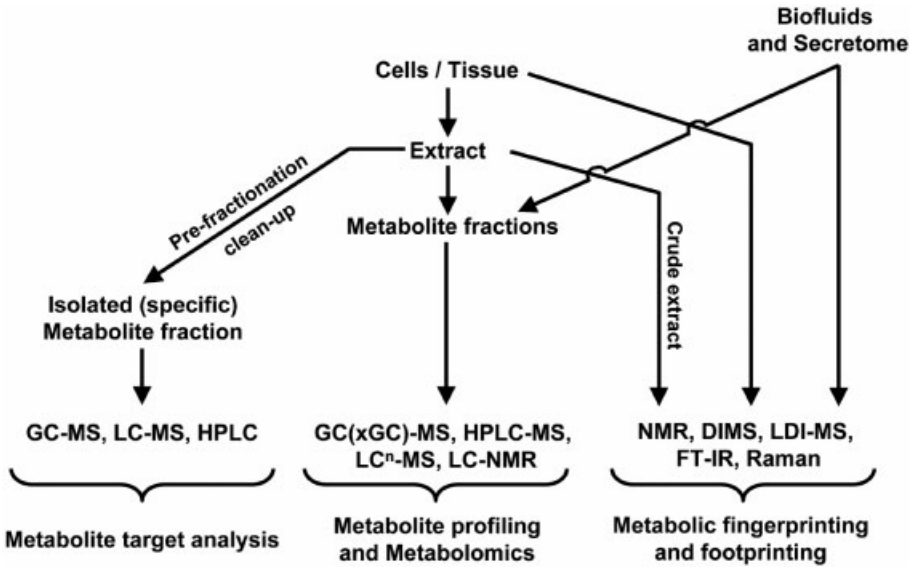

Figure 2. Technologies for metabolome analyses. For abbreviations and further explanation of the methods refer to Table 1.

The extract is now finally ready for analysis and there are many different methods and approaches that one could use. They are highlighted in Fig. 2 and listed in Table 1 (for excellent reviews of metabolomics technologies see [20, 21]). The choice of analytical tool is based on the level of chemical information required about the metabolites, remembering that there will be a chemical bias with respect to that method, and the speed of analysis is also another consideration. These issues are expanded in Table 2.

Armed with the appropriate metabolomics technology there are many different strategies that are being developed within the metabolomics field (Fig. 2). Whilst these strategies are not yet universally accepted and are evolving, the following four approaches are currently the most popular [8, 22, 23]: (i) metabolite target analysis, which is an approach that is restricted to metabolites of for example a particular enzyme system that would be directly affected by abiotic or biotic perturbation. (ii) Metabolite profiling, which is focused on a specific group of metabolites (*e.g.* lipids) or those associated with a specific pathway; within clinical and pharmaceutical analysis, this is often called metabolic profiling, which is used to trace the fate of a drug or metabolite. (iii) Metabolomics is the comprehensive analysis of the entire metabolome (all measurable metabolites), under a given set of conditions, this is often confused with metabonomics, which seeks to measure the fingerprint of biochemical perturbations caused by disease, drugs and toxins. (iv) Metabolic fingerprinting is used to classify samples based on provenance of either their biological relevance or origin by using a fingerprinting technology that is rapid but does not necessarily give specific metabolite information.

Table 1. Common analytical techniques applied to metabolomics

Abbreviation	Technique	Relevant reference(s)
GC-MS	Gas chromatography mass spectrometry	[43]
GCxGC-MS	2 dimensional GC coupled to MS	[79]
LC-EC	Liquid chromatography using an electrochemical array	[80]
HPLC-MS	High performance LC-MS	[18]
UPLC-MS	Ultra performance LC-MS	[81]
HILIC	Hydrophobic interaction chromatography	[82]
CE-MS	Capillary electrophoresis-MS	[83, 84]
NMR	nuclear magnetic resonance	[11, 85]
LC-NMR	LC coupled to NMR	[86]
FT-IR	Fourier transform infrared spectroscopy	[87, 88]
DIMS	Direct infusion ESI MS	[89, 90]
LDI-MS	Laser desorption ionisation MS	[91]
FT-ICR-MS	Fourier transform ion cyclotron resonance MS	[92]
SIDMAP	Stable isotope-based dynamic metabolic profiling, more commonly referred to as mass isotopomer analysis.	[76, 93]

Table 2. Considerations for metabolomics analyses

Consideration	Approach	Comments
Chemical information	MS	MS^n will provide some structural information. FT-ICR-MS can generate empirical formulae suggestions for $m/z < 500$.
	NMR	Gives detailed structural information, particularly using 2-D-NMR of isolated metabolites.
	Chromatography (GC, HPLC, CE)	On its own will not generally lead to metabolite identification. However, coupled with MS and NMR is very powerful for analyte identification.
	FT-IR, Raman	Provide limited structural information, but useful for identification of functional groups.
Chemical bias	GC-MS	Solvent extraction bias: non-polar versus polar analytes. Need for chemical derivatization
	LC-MS	Solvent bias means it is usually more applicable to polar compounds.
	NMR, FT-IR, Raman	These methods have little chemical bias and can be used directly on the sample.
Speed	Chromatography	Very useful for separation but typically take 10–30 min.
	NMR	Few minutes to hours. Depends on the strength of the magnet, sensitivity can be improved by magic angle spinning.
	ESI-MS	1–3 min in flow-injection (direct infusion) mode.
	FT-IR	10–60 s.

3 Metabolic footprinting: The exometabolome and secretome

The methods detailed above are generally aimed at measuring the cell or tissue samples directly where information on the intracellular metabolome is generated. By contrast, a completely non-invasive approach is to measure the extracellular metabolites, which is also referred to as the footprint or exometabolome [24–27]. The footprint of a cell in culture contains the medium components, less any substrate uptake, plus any secreted metabolites. Whilst the study of this secretome is only possible for cells in culture, it has the advantage that there is generally less variability in quenching and no metabolite extraction is needed. Information from the secretome can be valuable in understanding the behaviour and responses of cultured cells and has potential clinical applications. One example is the development in culture of preimplantation stage human embryos during *in vitro* fertilisation (IVF) treatment for infertility (Hollywood, in preparation). Non-invasive methods are urgently required to select the best IVF embryos for transfer back to the patient [28]. Since embryo metabolism is thought to be a critical determinant of viability [29], there is potential mileage in a metabolomics approach. Metabolic fingerprinting of the follicular fluid in which the oocyte develops is feasible [30] and may predict embryonic viability. Analysis of the amino acid composition of the culture medium in which the embryo develops has also provided predictive factors [31, 32]. Thus, it is reasonable to assume that a viable human embryo will possess a unique metabolic fingerprint, and this secretome will be expressed in culture medium as a metabolic footprint. Indeed, we have preliminary data (Hollywood, in preparation), demonstrating that metabolic footprinting can contain valuable information on the status of individual human embryos. Moreover, while the secretome contains delayed information and is normally an average of the biological system being studied, this is less of a problem with preimplantation embryos. Early stage embryos have very few (1–8) cells, which are synchronised in development, and individual embryos at various stages can be easily identified and analysed. Therefore, we regard the preimplantation embryo as a valuable model system in which to test the ability of metabolomics to describe a developing organism and its potential.

4 Understanding biological complexity and its environment

Whilst the human genome sequence has been completed, the vast array of commensal microorganisms (estimated to be >1000 different species) that live in symbiosis with man have not. It is estimated that there are ten times the number of prokaryotic cells as human cells in the body, and this microbiome (estimated to weigh 1 kg in an adult human) plays a vital role in our well-being. Nicholson and colleagues [33], who have reported that microbial metabolites from the gut microflora can be found in human serum and urine, have elegantly demonstrated this. Thus, understanding the interactions between different organisms in a single complex system is as important as understanding the impact that environmental effects may have on the hundreds of functionally specialized cell types found in man [34]. Furthermore, these authors advocate that one should also include temporal measurements to capture the full dynamics of the perturbation to a system.

One environmental factor that is very important is what happens to the health of an individual when it eats. The link between nutrition and health is obvious and the area of nutrigenomics [35] is aimed at defining tissue, cellular or biofluid-specific nutritional metabolomes. This would aid in the inter-

pretation of disease processes because a baseline healthy metabolome under different nutritional conditions can be defined leading to personalised metabolic assessment [36].

Finally, host-pathogen interactions are also amenable to metabolomics investigations [37]. This has been recently illustrated by Mur and colleagues [38] who used ESI-MS to identify discriminatory lipid metabolites, which are known to be important signals in plant defence. It was found that phosphatidic and phosphatidyl glycerol phospholipids were involved in rice blast disease when *Brachypodium distachyon* was infected by *Magnaporthe grisea*.

5 Informatics

As with all functional analyses, a typical metabolomics experiment can generate data torrents (samples *times* metabolites) and something has to be done to turn these data (information) into knowledge. In particular, we need well-curated databases, very good data to populate them, and even better algorithms to turn these metabolite data into knowledge.

5.1 Databases for metabolomics

Database curation is essential if metabolomics databases are to be useful for the wider community. As metabolomics data are multivariate (many measurements per sample) and hence complex, it is essential that the metabolite data are validated prior to being uploaded to a database. This starts by summarising the data about the metabolite data; these co-called metadata have to be captured and there have been two position papers detailing how this should be achieved. The first is Armet (architecture for metabolomics; www.armet.org), which contains details of the data schema for metabolomics and includes the basis for storage and transmission of data via UML [39], and the second is SMRS (standard metabolic reporting structure; www.smrsgroup.org), which contains details of what experimental data are needed [40]. There is obvious overlap between Armet and SMRS and the proponents of these approaches, plus the wider community, have been having active discussions to generate a unified approach to metabolomics databasing and data exchange.

Capturing the metadata is only the start of the database process and the question of precision and accuracy in multivariate metabolite data needs to be addressed. Clearly traceable standards are needed that will allow machine calibration, however, this is not so easy for 'omics measurements with hundreds or thousands of variables and where machine drift may be acute. It may be that standard cocktails made up freshly, together with the use of advanced transformations [41, 42] may help solve this problem. Again, the metabolomics community are addressing this and a working group headed by Fiehn will report on their findings soon. Workgroups' progress reports can be found at the Metabolomics Standards Iniative (http://msi-workgroups.sourceforge.net)

Finally, as many metabolites are yet to be described [43] there needs to be some agreed protocol of how one names unknown metabolites that have no detailed structure and thus how these can be recognised between different laboratories. As a step towards this goal, a naming protocol has been suggested by Bino and colleagues [44], which aims to capture the laboratory, analytical method used and any specific separation or *m/z* information. This will necessitate rigorous mass spectral libraries and deconvolution algorithms. An excellent overview of these resources has been provided by Weckwerth and Morgenthal [9].

5.2 Biomarker discovery

A very active research area with metabolomics is that of discovering which metabolites are indicative of disease. In this approach, metabolite profiling is used to generate quantitative lists of metabolites from control populations and test subjects that are diseased. Data analysis is then used to mine the metabolites and determine which are discriminatory for the disease and which of these could be used in predictive medicine. Whilst there are many different multivariate or chemometric analyses methods that could be used, the key objective is to make the analysis as valid as possible so stringent statistical validation is needed. It is essential that a good data generating experiment be designed to encompass the variation inherent in these experiments, which involve many measurements on complex biological systems. The first variation to be aware of is at the biological level. This is because even isogenic organisms show variation [43] and so many samples need to be analysed. The second source of variation is at the sample preparation level, both with respect to quenching metabolism and metabolite extractions. The last is of course any variation introduced by the analytical instrument itself. Thus, lots of data are needed and a good strategy to adopt is to collect three sets of data [45, 46]. The first set is called the 'training data' and is used to construct a mathematical model that will relate metabolite data with disease status; this can be categorical (diseased versus healthy) or quantitative (severity of disease or grade of cancer). The second set of data is usually termed the 'validation data' and is used to cross-validate the model generated by the training data. For example, some modelling processes (*e.g.* discriminant analysis [47] or linear regression methods [48]) involve the extraction of a certain number of latent variables: too few are inadequate whilst too many may cause the metabolite data to be over fitted (that is to say the model will include noise). Thus, the second set of validation data are used to tune this selection process. The final set of data is the 'test data' and is used to assess how well the modelling process described above has done. These data are "independent" and thus can be used to assess the predictive nature of the analysis.

For the detection of biomarkers, one can of course start with the easiest method of 'stare and compare!'. Whilst unlikely to give rise to any markers it is always wise to actually look at the data as this can also be used as a quality check. The next method is to use difference profiles where the average metabolite profile from the diseased subjects is subtracted from the average metabolite profile from healthy individuals. Finally, simple univariate analyses of ANOVA

(analysis of variance), Student's *t*-test or non-parametric equivalents can be used to ascertain if there is any statistically significant difference between individual metabolites for healthy versus diseased individuals.

In the unlikely event that these have worked, it probably means that the biomarker is obvious; however, single biomarkers are unlikely. The concept of multivariate biomarker profiles has become reality [49] and so more powerful supervised learning multivariate analysis methods are needed [50]. In supervised learning an algorithm (*vide infra*) is used to transform the multivariate data from metabolite profiles into something of biological interest, usually of much lower dimensionality, which as discussed above can be categorical (diseased *vs.* healthy) or quantitative (severity of disease). In supervised learning both metabolite data (inputs) and disease status (outputs or targets) are used, and these two types of data form pairs that are used in the calibration of the model. The goal of supervised learning is to find a "model" or "mapping" that will correctly associate the inputs with the targets. Below highlights the various algorithms that are currently employed to effect supervised learning:

Discriminant analysis (DA) is a particularly popular algorithm, which is a cluster analysis-based method and involves projection of test data into cluster space [47, 51]. This is a categorical method and loadings matrices can give an indication of important inputs (metabolites).

Partial least squares (PLS) is a very popular linear regression-based method [48]. The algorithm can be programmed in a quantitative way (PLS1) or categorical (PLS2 or PLS-DA), and as for DA, loadings matrices can give an indication of important metabolites.

Artificial neural networks (ANN) are very popular based machine learning methods, which in contrast to DA and PLS can learn non-linear as well as linear mappings [52]. The main algorithms used are multilayer perceptrons (MLP [53, 54]) and radial basis functions (RBF [55]). However, although popular, the mapping from input to output is largely opaque and whilst this can be improved by pruning or growing ANN, these methods are not very good for biomarker discovery.

Rule induction (RI)-based algorithms include CART (classification and regression trees [56]), FuRES (fuzzy rule-building expert system [57]) and C4.5/C5 [58]. These are categorical algorithms based on the growth of a decision tree, with predictive segregation of the data, which produces uni- or multi-variate decision boundaries, which can be used to discover which metabolites are important.

Evolutionary computation (EC) algorithms are based on concepts of Darwinian selection to generate and to optimize a desired mapping between input and output variables [59]. These include genetic algorithms (GA [60, 61]), genetic programming (GP [62–65]) and genomic computing [66]. EC algorithms are entirely data-driven and are thus especially appropriate for problems that are data-rich but hypothesis/information-poor. They generate predictive rules and hence hypotheses of which metabolites may serve as potential biomarkers.

We believe that a future trend where metabolomics will play a big role is for biomarker detection, as this approach will present the opportunity for objective automated diagnoses of disease. Currently, there are few examples of biomarker discovery but recent ones include the analysis of LC-electrochemical data with PLS-DA from patients with motor neuron disease [67], from GC-MS data from pregnant mothers with pre-eclampsia using genetic programming [68], and LC-MS data with PLS-DA, investigating dietary intake [69].

5.3 Metabolic networks

Recently, there has been a paradigm shift from classical metabolic pathways (Fig. 1C) to those based on metabolic networks and neighbourhoods [1, 13]. The elucidation and visualisation of metabolite neighbourhoods needs to be achieved to understand the structural properties of the network [70, 71]. This can only be done at the level of the metabolome since fluxes and hence relationships between metabolites through networks cannot be calculated accurately from transcripts or proteins. One approach that is being used is to collect 'omics data at the three different functional levels and then correlate each transcript, protein or metabolite with all the other 'omic data collected [72–75] to assess correlations within these data. In addition, one can use cluster analysis, including principal components analysis (PCA) and hierarchical cluster analysis (HCA) of the 'omic data to discover in a multivariate way groups of co-expressed mRNA, proteins, and metabolites. These data can then be combined and used to construct possible networks. As these are *in silico* predictions, they need to be validated and just as radiolabelling of substrates was used in the early part of the last century to discover novel pathways, a fluxomics-based approach has emerged based on mass isotopomer analysis. In this method, specific labelled (^{13}C or ^{15}N) metabolites are fed to tissues or bacterial or yeast cultures and the destination of these metabolites assessed using MS [76, 77] or NMR [78].

6 Outlook

Metabolomics is gaining increasing interest across a wide variety of disciplines, including functional genomics, integrative and systems biology, pharmacogenomics, and biomarker discovery for disease prognoses, diagnoses and therapy monitoring. There is great interest in biomarkers, and as many diseases are indeed a result of metabolic disorders, it makes a great deal of sense to measure metabolites directly. Once key metabolite markers are identified this will likely lead to a better understanding of the disease process as specific metabolic pathways (networks) will be highlighted as being important, and this will serve as a hypothesis starting point for therapeutic intervention and drug discovery. Finally, we believe non-invasive footprinting analysis of cells in culture has the potential to increase knowledge of the impact of *in vitro* systems and has many exciting clinical

applications, as has the use of metabolomics to help in the understanding of biological complexity as part of a systems biology programme.

K.H. thanks Stiefel Laboratories (UK) Ltd for her studentship and R.G. is indebted to the Engineering and Biological Systems Committee of the UK BBSRC for financial support.

7 References

[1] Barabasi, A.-L., Oltvai, Z. N., *Nat. Rev. Genet.* 2004, *5*, 101–113.

[2] Fell, D. A., *Understanding the Control of Metabolism*, Portland Press, London 1996.

[3] Hanash, S. M., Bobek, M. R., Rickman, D. S., Williams, T. *et al.*, *Proteomics* 2002, *2*, 69–75.

[4] Eymann, C., Homuth, G., Scharf, C., Hecker, M., *J. Bacteriol.* 2002, *184*, 2500–2520.

[5] Jones, M. B., Krutzsch, H., Shu, H. J., Zhao, Y. M. *et al.*, *Proteomics* 2002, *2*, 76–84.

[6] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[7] Fiehn, O., *Plant Mol. Biol.* 2002, *48*, 155–171.

[8] Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., Kell, D. B., *Trends Biotechnol.* 2004, *22*, 245–252.

[9] Weckwerth, W., Morgenthal, K., *Drug Discovery Today* 2005, *10*, 1551–1558.

[10] Hall, R. D., *New Phytologist* 2006, *169*, 453–468.

[11] Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N. S. *et al.*, *Nat. Biotechnol.* 2001, *19*, 45–50.

[12] ter Kuile, B. H., Westerhoff, H. V., *FEBS Lett.* 2001, *500*, 169–171.

[13] Kell, D. B., *Curr. Opin. Microbiol.* 2004, *7*, 296–307.

[14] Westerhoff, H. V., Palsson, B. O., *Nat. Biotechnol.* 2004, *22*, 1249–1252.

[15] de la Fuente, A., Snoep, J. L., Westerhoff, H. V., Mendes, P., *Eur. J. Biochem.* 2002, *269*, 4399–4408.

[16] Tweeddale, H., Notley-McRobb, L., Ferenci, T., *J. Bacteriol.* 1998, *180*, 5109–5116.

[17] Viant, M. R., Pincetich, C. A., de Ropp, J. S., Tjeerdema, R. S., *Metabolomics* 2005, *1*, 149–158.

[18] Buchholz, A., Takors, R., Wandrey, C., *Anal. Biochem.* 2001, *295*, 129–137.

[19] Villas-Boas, S. G., Højer-Pedersen, J., Akesson, M., Smedsgaard, J., Nielsen, J., *Yeast* 2005, *22*, 1155–1169.

[20] Dunn, W. B., Ellis, D. I., *Trends Anal. Chem.* 2005, *24*, 285–294.

[21] Dunn, W. B., Bailey, N. J. C., Johnson, H. E., *Analyst* 2005, *130*, 606–625.

[22] Fiehn, O., *Compar. Funct.l Genomics* 2001, *2*, 155–168.

[23] Harrigan, G. G. and Goodacre, R., *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Publishers, Boston 2003, pp. 335.

[24] Allen, J., Davey, H. M., Broadhurst, D., Heald, J. K. *et al.*, *Nat. Biotechnol.* 2003, *21*, 692–696.

Roy Goodacre is Professor of Biological Chemistry at The University of Manchester. The research group's (www.biospec.net) interests are broadly within bioanalytical chemistry, and in the combination of a variety of modern spectroscopy technologies (including MS, IR and Raman) and advanced chemometrics and machine learning to the explanatory analysis of complex biological systems within a metabolomics and proteomics context. He is Editor-in-chief of the journal *Metabolomics* and Founding Director of the Metabolomics Society.

[25] Kaderbhai, N. N., Broadhurst, D. I., Ellis, D. I., Goodacre, R., Kell, D. B., *Compar. Funct. Genomics* 2003, *4*, 376–391.

[26] Allen, J., Davey, H. M., Broadhurst, D., Rowland, J. *et al.*, *Appl. Environm. Microbiol.* 2004, *70*, 6157–6165.

[27] Kell, D. B., Brown, M., Davey, H. M., Dunn, W. B. *et al.*, *Nat. Rev. Microbiol.* 2005, *3*, 557–565.

[28] Leese, H. J., Conaghan, J., Martin, K. L., Hardy, K., *Bioessays* 1993, *15*, 259–264.

[29] Houghton, F. D., Leese, H. J., *Eur. J. Obstetrics Gynecol. Reprod. Biol.* 2004, *115 Suppl 1*, S92–S96.

[30] Thomas, N., Goodacre, R., Timmins, É. M., Gaudoin, M., Fleming, R., *Human Reprod.* 2000, *15*, 1667–1671.

[31] Houghton, F. D., Hawkhead, J. A., Humpherson, P. G., Hogg, J. E. *et al.*, *Human Reprod.* 2002, *17*, 999–1005.

[32] Brison, D. R., Houghton, F. D., Falconer, D., Roberts, S. A. *et al.*, *Human Reprod.* 2004, *19*, 2319–2324.

[33] Nicholson, J. K., Holmes, E., Lindon, J. C., Wilson, I. D., *Nat. Biotechnol.* 2004, *22*, 1268–1274.

[34] Nicholson, J. K., Wilson, I. D., *Nat. Rev. Drug Discov.* 2003, *2*, 668–676.

[35] van Ommen, B., Stierum, R., *Curr. Opin. Biotechnol.* 2002, *13*, 517–521.

[36] German, J. B., Hammock, B. D., Watkins, S. M., *Metabolomics* 2005, *1*, 3–9.

[37] Forst, C. V., *Drug Discov. Today* 2006, *11*, 220–227.

[38] Allwood, J. W., Ellis, D. I., Heald, J. K., Goodacre, R., Mur, L. A., *Plant J.* 2006, *46*, 351–368.

[39] Jenkins, H., Hardy, N., Beckmann, M., Draper, J. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1601–1606.

[40] Lindon, J. C., Nicholson, J. K., Holmes, E., Keun, H. C. *et al.*, *Nat. Biotechnol.* 2005, *23*, 833–838.

[41] Goodacre, R., Kell, D. B., *Anal. Chem.* 1996, *68*, 271–280.

[42] Goodacre, R., Timmins, E. M., Jones, A., Kell, D. B. *et al.*, *Anal. Chim. Acta* 1997, *348*, 511–532.

[43] Fiehn, O., Kopka, J., Dörmann, P., Altmann, T. *et al.*, *Nat. Biotechnol.* 2000, *18*, 1157–1161.

[44] Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J. *et al.*, *Trends Plant Sci.* 2004, *9*, 418–425.

[45] Kell, D. B., Sonnleitner, B., *Trends Biotechnol.* 1995, *13*, 481–492.

[46] Brown, M., Dunn, W. B., Ellis, D. I., Goodacre, R. *et al.*, *Metabolomics* 2005, *1*, 39–51.

[47] Manly, B. F. J., *Multivariate Statistical Methods: A Primer*, Chapman & Hall, London 1994, p. 215.

[48] Martens, H., Næs, T., *Multivariate Calibration*, John Wiley, Chichester 1989.

[49] van der Greef, J., *Curr. Opin. Chem. Biol.* 2004, *8*, 559–565.

[50] Hastie, T., Tibshirani, R., Friedman, J., *The elements of statistical learning: data mining, inference and prediction*, Springer-Verlag, Berlin 2001.

[51] Chatfield, C., Collins, A. J., *Introduction to Multivariate Analysis*, Chapman & Hall, London 1980.

[52] Bishop, C. M., *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford 1995.

[53] Rumelhart, D. E., Hinton, G. E., Williams, R. J., in: Rumelhart, D. E., McClelland, J., (Eds.), *Parallel Distributed Processing, Volume 1: Foundations*, MIT Press, Cambridge, MA 1986.

[54] Werbos, P. J., *The Roots of Back-Propagation: from Ordered Derivatives to Neural Networks and Political Forecasting*, John Wiley, Chichester 1994.

[55] Broomhead, D. S., Lowe, D., *Complex Syst.* 1988, *2*, 321–355.

[56] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., *Classification and Regression Trees*, Wadsworth, Inc., Pacific Grove, CA 1984.

[57] Harrington, P. B., *J. Chemometrics* 1991, *5*, 467–486.

[58] Quinlan, J. R., *C4.5: programs for machine learning*, Morgan Kaufmann, San Mateo, CA 1993.

[59] Bäck, T., Fogel, D. B., Michalewicz, Z., *Handbook of Evolutionary Computation*, IOPPublishing/Oxford University Press, Oxford 1997.

[60] Holland, J. H., *Adaption in natural and artifcial systems*, MIT Press, Cambridge, MA 1992.

[61] Reeves, C. R., *Genetic Algorithms – Principles and Perspectives: A Guide to GA Theory*, Kluwer Academic Publishers, Dordrecht 2002.

[62] Koza, J. R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA 1992, p. 819.

[63] Koza, J. R., *Genetic Programming II: Automatic Discovery of Reusable Programs*, MIT Press, Cambridge, MA 1994, p. 746.

[64] Koza, J. R., Bennett, F. H., Keane, M. A.. Andre, D., *Genetic Programming III: Darwinian Invention and Problem Solving*, Morgan Kaufmann,San Francisco 1999.

[65] Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W. *et al.*, *Genetic Programming: Routine Human-Competitive Machine Intelligence*, Kluwer, New York 2003.

[66] Kell, D. B., Darby, R. M., Draper, J., *Plant Physiol.* 2001, *126*, 943–951.

[67] Rozen, S., Cudkowicz, M. E., Bogdanov, M., Matson, W. R. *et al.*, *Metabolomics* 2005, *1*, 101–108.

[68] Kenny, L. C., Dunn, W. B., Ellis, D. I., Myers, J. *et al.*, *Metabolomics* 2005, *1*, 227–234.

[69] Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R. *et al.*, *Anal. Chem.* 2006, *78*, 567–574.

[70] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabási, A.-L., *Nature* 2000, *407*, 651–654.

[71] Wagner, A., Fell, D. A., *Proc. Royal Soc. London B* 2001, *268*, 1803–1810.

[72] Mendes, P., *Brief Bioinform.* 2002, *3*, 134–45.

[73] Weckwerth, W., Fiehn, O., *Curr. Opin. Biotechnol.* 2002, *13*, 156–160.

[74] Förster, J., Famili, I., Fu, P., Palsson, B. O., Nielsen, J., *Genome Res.* 2003, *13*, 244–253.

[75] Camacho, D., de la Fuente, A., Mendes, P., *Metabolomics* 2005, *1*, 53–63.

[76] Bederman, I. R., Reszko, A. E., Kasumov, T., David, F. *et al.*, *J. Biol. Chem.* 2004, *279*, 43207–43216.

[77] Villas-Bôas, S. G., Mas, S., Åkesson, M., Smedsgaard, J., Nielsen, J., *Mass Spectrom. Rev.* 2005, *24*, 616–646.

[78] Satake, M., Dmochowska, B., Nishikawa, Y., Madaj, J. *et al.*, *Invest. Ophthalmol. Vis. Sci.* 2003, *44*, 2047–2058.

[79] Welthagen, W., Shellie, R. A., Spranger, J. M., Ristow, M. *et al.*, *Metabolomics* 2005, *1*, 65–73.

[80] Gamache, P. H., *J. Am. Soc. Mass Spectrom.* 2004, *15*, 1717–1726.

[81] Wilson, I. D., Plumb, R., Granger, J., Major, H. *et al.*, *J. Chromatogr. B* 2005, *817*, 67–76.

[82] Tolstikov, V. V., Lommen, A., Nakanishi, K., Tanaka, N., Fiehn, O., *Anal. Chem.* 2003, *75*, 6737–6740.

[83] Soga, T., Ueno, Y., Naraoka, H., Ohashi, Y. *et al.*, *Anal. Chem.* 2002, *74*, 2233–2239.

[84] Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H. *et al.*, *J Proteome Res.* 2003, *2*, 488–494.

[85] Nicholson, J. K., Lindon, J. C., Holmes, E., *Xenobiotica* 1999, *29*, 1181–1189.

[86] Wolfender, J. L., Ndjoko, K., Hostettmann, K., *J. Chromatogr. A* 2003, *1000*, 437–455.

[87] Harrigan, G. G., LaPlante, R. H., Cosma, G. N., Cockerell, G. *et al.*, *Toxicol. Lett.* 2004, *146*, 197–205.

[88] Ellis, D. I., Goodacre, R., *Analyst* 2006, *131*, 875–885.

[89] Vaidyanathan, S., Kell, D. B., Goodacre, R., *J. Am. Soc. Mass Spectrom.* 2002, *13*, 118–128.

[90] Overy, S. A., Walker, H. J., Malone, S., Howard, T. P. *et al.*, *J. Exp. Botany* 2005, *56*, 287–296.

[91] Vaidyanathan, S., Jones, D., Broadhurst, D. I., Ellis, J. *et al.*, *Metabolomics* 2005, *1*, 243–250.

[92] Aharoni, A., Ric de Vos, C. H., Verhoeven, H. A., Maliepaard, C. A. *et al.*, *Omics* 2003, *6*, 217–234.

[93] Boros, L. G., *Metabolomics* 2005, *1*, 11–15.

