Review TRENDS in Biotechnology Vol.22 No.5 May 2004

Full text provided by www.sciencedirect.com

Metabolomics by numbers: acquiring and understanding global metabolite data

Royston Goodacre[1], Seetharaman Vaidyanathan[1], Warwick B. Dunn[1], George G. Harrigan[2] and Douglas B. Kell[1]

[1]Department of Chemistry, UMIST, P.O. Box 88, Sackville Street, Manchester, M60 1QD, UK
[2]Global HTS, Pfizer, 700 Chesterfield Parkway, Chesterfield, MO 63198, USA

In this postgenomic era, there is a specific need to assign function to orphan genes in order to validate potential targets for drug therapy and to discover new biomarkers of disease. Metabolomics is an emerging field that is complementary to the other 'omics and proving to have unique advantages. As in transcriptomics or proteomics, a typical metabolic fingerprint or metabolomic experiment is likely to generate thousands of data points, of which only a handful might be needed to describe the problem adequately. Extracting the most meaningful elements of these data is thus key to generating useful new knowledge with mechanistic or explanatory power.

Since the completion of the first whole-genome sequence of a free-living organism (that of the bacterium *Haemophilus influenzae* [1], although the sequencing of a human mitochondrion long predates it [2]), we began to realize the paucity of our knowledge with respect to the existence, let alone the function, of the novel genes thereby uncovered. Sequencing of the microbiologist's pet organism, *Escherichia coli*, revealed that a staggering 38% of the total 4288 open reading frames had not been observed or studied before [3]. More recently, completion of the human genome sequence [4,5] has accelerated further the demand for determining the biochemical function of orphan genes and for validating them as molecular targets for therapeutic intervention.

The search for biomarkers that can serve as indicators of disease progression or response to therapeutic intervention has also increased. Functional studies have thus emphasized analyses at the level of gene expression (transcriptomics), protein translation (proteomics) including post-translational modifications, and the metabolic network (metabolomics), with a view to a 'systems biology' approach of defining the phenotype and bridging the genotype-to-phenotype gap [6].

There is active debate in the research community over the exact definition of the 'metabolome', but it was first defined by Oliver *et al.* [7] as the quantitative complement of all of the low molecular weight molecules present in cells in a particular physiological or developmental state. Another definition states that the metabolome consists 'only of those native small molecules (definable non-polymeric compounds) that are participants in general metabolic reactions and that are required for the maintenance, growth and normal function of a cell' [8].

Although the metabolome is certainly 'complementary' to transcriptomics and proteomics, it might be seen to have special advantages. In particular, it is known from both the theory underlying metabolic control analysis [9,10] and experiment [11] that, although changes in the quantities of individual enzymes might be expected to have little effect on metabolic fluxes, they can and do have significant effects on the concentrations of numerous individual metabolites. In addition, the metabolome is further down the line from gene to function and so reflects more closely the activities of the cell at a functional level. Thus, as the 'downstream' result of gene expression, changes in the metabolome are expected to be amplified relative to changes in the transcriptome and the proteome [12]. As expected, metabolic fluxes (at least as exemplified by glycolysis in trypanosomes) are not regulated by gene expression alone, which provides a further rationale for pursuing metabolomics [13].

In this review we describe the growing field of metabolomics, the needs and means by which metabolome data can be generated, and how this information can be turned into knowledge.

Measuring the metabolome

The ultimate starting point of a metabolomic experiment is to quantify all of the metabolites in a cellular system (i.e. the cell or tissue in a given state at a given point in time). Currently this is impossible, given the lack of simple automated analytical strategies that can effect this in a reproducible and robust way. The main challenges are the chemical complexity and heterogeneity of metabolites, the dynamic range of the measuring technique, the throughput of the measurements, and the extraction protocols. Ideally, metabolomics should be non-biased but, considering the above, at best it can be thought of as 'non-targeted'. Moreover, the paucity of our knowledge with respect to known metabolites is staggering, although perhaps

Corresponding author: Royston Goodacre (r.goodacre@umist.ac.uk).

www.sciencedirect.com 0167-7799/$ - see front matter © 2004 Elsevier Ltd. All rights reserved. doi:10.1016/j.tibtech.2004.03.007

understandable at present; for example, there are estimated to be up to 200 000 different metabolites in the plant kingdom [6] and, even though the numbers might be significantly smaller in individual mammalian systems, the fate of a toxin can lead to a plethora of intermediates and products before it is adequately detoxified [14].

One might ask why bother with metabolomics when transcriptomics and proteomics are currently so popular? Another answer, in addition to the above, is simple and stems from evolution. To measure the amount of, say, a specific fructose 1,6-bisphosphatase from different organisms, one has to know *a priori* the DNA sequences or protein sequences (plus post-translational modifications) from each organism to design suitable complementary oligonucleotides to capture mRNA on a nucleotide array [15] or to effect protein identification via two-dimensional gel electrophoresis and mass spectrometry (MS) [16]. By contrast, the substrate and product of this enzyme, fructose 1,6-bisphosphate and fructose 6-phosphate, have the same basic chemical structure irrespective of the organism from which they are extracted and so, after one has learnt how to quantify these metabolites in parallel and in various sample matrices, a more or less universal approach that spans the species barriers can be adopted.

Alterations in cells, biofluid or even cell media that are induced in response to environmental or developmental stimuli, or to a genetic mutation, result in changes in the quasi steady-state amounts of intermediate pathway metabolites and/or in the end accumulation of terminal metabolites. To capture these changes, the metabolites and their quantities must be monitored both spatially and temporally. Because the metabolic complement is even more dynamic than the proteome, analysis can be envisaged at different levels.

Although it would be ideal to have information on the status of the whole metabolic complement of a cell, there might be instances when it would suffice to derive information on only a portion of the total metabolome. For example, it might be sufficient to monitor selectively only the relevant metabolites that contribute to a specific pathway that is directly associated with function (although this begs the answer in a way that is normally unsupportable because one does not necessarily know *a priori* which pathways to monitor [17]). In some instances, it might be necessary only to monitor changes in the overall or partial metabolic pool structure and to classify samples without determining the quantities of individual metabolites. Box 1 lists some of the common definitions used in metabolomics.

Box 1. Classification of metabolomic approaches

- **Metabolite target analysis**: analysis restricted to metabolites of, for example, a particular enzyme system that would be directly affected by abiotic or biotic perturbation [70].
- **Metabolite profiling**: analysis focused on a group of metabolites, for example, a class of compounds such as carbohydrates, amino acids or those associated with a specific pathway [70].
- **Metabolomics**: comprehensive analysis of the whole metabolome under a given set of conditions [70].
- **Metabolic fingerprinting**: classification of samples on the basis of provenance of either their biological relevance or origin [70].
- **Metabolic profiling**: often used interchangeably with ‘metabolite profiling’; metabolic fingerprinting is commonly used in clinical and pharmaceutical analysis to trace the fate of a drug or metabolite [71].
- **Metabonomics**: measure of the fingerprint of biochemical perturbations caused by disease, drugs and toxins [18,72].

The terminologies are still evolving and there can be overlaps in their definition; however, the above classification highlights the options available for monitoring the metabolome and serves as a very good starting point.

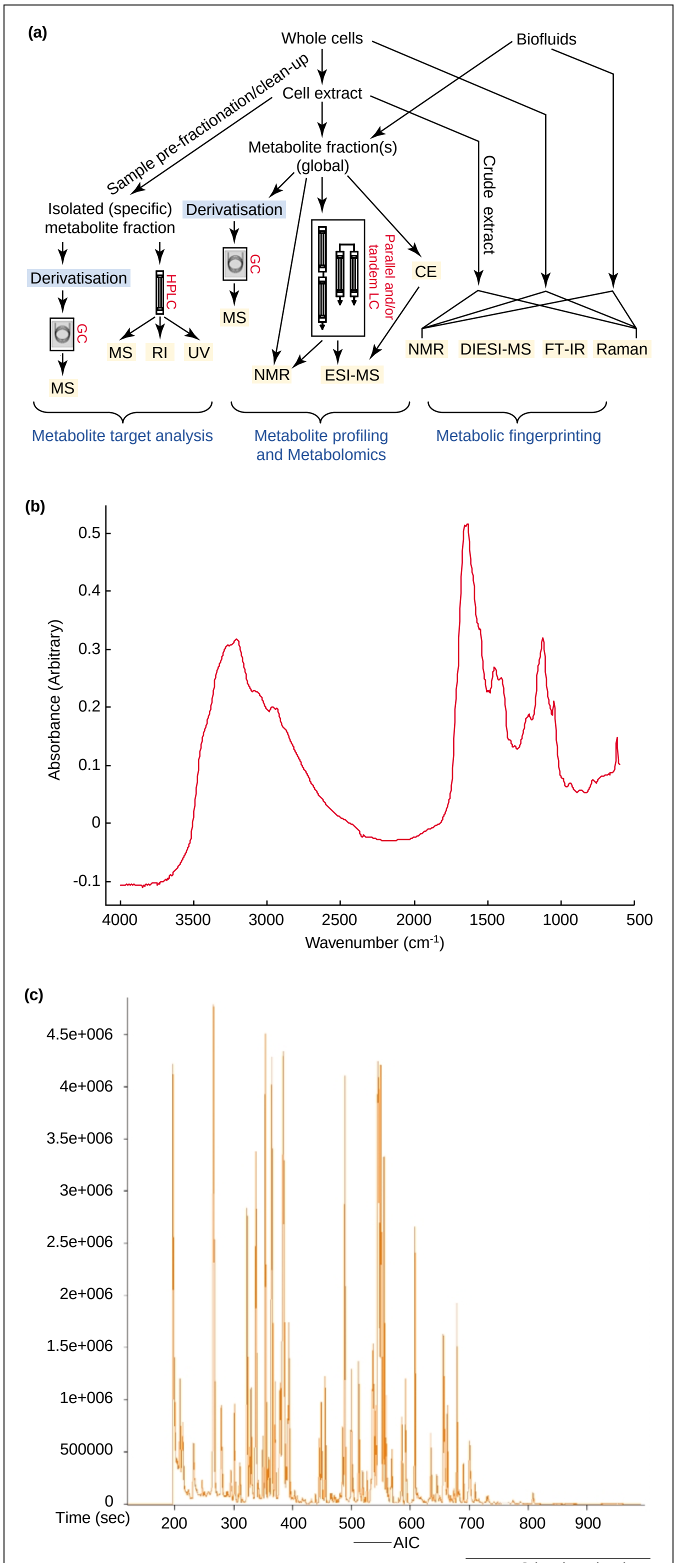
Technology platforms for metabolomics

Metabolites are chemical entities and can be analysed by the standard tools of chemical analysis such as molecular spectroscopy and MS. The resolution, sensitivity and selectivity of these technologies can be enhanced or modified by coupling them to gas chromatograpy (GC) or liquid chromatography (LC) steps. The technologies commonly exploited for different metabolomic strategies are shown in Figure 1. Generally, the technology platform of choice depends on the type of sample to be analysed. Biofluids are perhaps the most easily obtained sample and can be analysed by nuclear magnetic resonance (NMR) with little or no sample preparation [18–20], whereas tissues and cells from animal, plant or microbial systems necessarily require some sample pretreatment.

For metabolite target analysis and metabolite profiling, studies can be geared to monitor specific metabolites by selective analysis and well-developed calibration methods. As suggested in Figure 1, this can be achieved by conventional techniques, such as separation by GC or high performance liquid chromatography (HPLC) coupled to a suitable detection system. A notable recent advance is the use of the so-called ‘hydrophobic interaction chromatography’ (HILIC) method for detecting many polar plant compounds in a MS-friendly manner [21].

NMR spectroscopy has been shown to provide valuable information on metabolites, typically directly from biofluids with little or no sample preparation steps [18,19]. Because NMR is based on the fact that nuclei such as ^{1}H, ^{13}C, ^{31}P can exist at different energy levels in a strong magnetic field because they possess nuclear spin, it can generate valuable structural information. Magic angle spinning NMR can also be used in intact tissues, often giving uniquely powerful insights.

For comprehensive analysis of the metabolome (i.e. metabolomics), it is essential to use strategies that have a wider coverage in terms of the type and number of metabolites analysed. Sample preparation can be elaborate and can involve dividing samples into aliquots before selective enrichment and the analysis of different classes of metabolite in each aliquot. Such pre-fractionation steps and subsequent parallel measurements are required to optimize analyses and to facilitate the detection of even minor changes in a structurally diverse metabolome data set. A combination of several analytical techniques might have to be used for such studies; for example, parallel LC separations can be coupled to MS- and/or NMR-based detection methods. Most extraction procedures reported in the literature so far are less comprehensive and thus biased, because they miss out on some or other metabolites



TRENDS in Biotechnology

Figure 1. Technologies for metabolome analysis. **(a)** General strategies for metabolome analysis. CE, capillary electrophoresis; DIESI, direct-infusion ESI, which can be linked to Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS); NMR, nuclear magnetic resonance; RI, refractive index detection; UV, ultraviolet detection. **(b)** Example of an FT-IR spectrum of a biofluid. In this experiment, 10 μl of rat urine was dried and analysed on a Bruker IFS66 instrument between 400 and 600 cm^{-1}, with 4 cm^{-1} resolution and 256 co-adds (R.G. and G.G.H., unpublished). **(c)** Capillary gas chromatography–time-of-flight–mass spectrometry (GC-TOF-MS) analysis of human serum. In a 15 min run, 722 peaks could be discriminated (W.B.D. and D.B.K., unpublished).

and result in the modelling of only a portion of the metabolome (i.e. metabolite profiles). But even in those investigations, the information content that is obtainable points to the potential of monitoring metabolomes comprehensively.

The current popular method for global metabolite analysis in plants is GC-MS, although this method is limited in the molecular mass of the targets that it can measure and thermolabile ones are necessarily missed. Non-volatile polar metabolites often need to be derivatized to convert them to less-polar, volatile, thermally stable derivatives before they can be separated on a GC column. Although efficient derivatization methods are available, low sample throughput can be a limiting factor in large-scale metabolite profiling because typical metabolite acquisition times are of the order of 10–30 min. Deconvolution is then needed to quantify metabolites that are unresolved by GC. This process is aided by MS, but suitable deconvolution algorithms must be developed. Improved deconvolution algorithms and faster spectral acquisition by time-of-flight (TOF) measurements [22] have, however, resulted in the detection of over 1000 components from plant leaf extracts at a throughput of over 1000 samples per month [23].

Another recent advance is the MSFACTS program developed by Sumner and colleagues [24]. The end result of this procedure is a list of metabolites (either known metabolites found in databases such as KEGG or unique MS profiles associated with a specific GC retention time) and a list of their relative concentrations.

Recent interest in GC-GC-MS is increasing the number of metabolites that can be separated in a single analysis run [25], and methods involving flow injection analysis using direct infusion into electrospray ionization (ESI), coupled to TOF or Fourier transform ion cyclotron resonance (FT-ICR) MS analysis are also becoming popular [26,27]. In particular, FT-ICR-MS is exciting because it is sensitive and, with its high mass resolution $(>10^6)$ coupled to software that can exploit the information in isotope patterns, can produce the empirical formulae for metabolites directly [28].

Metabolic fingerprinting is fast and would be ideally suited for rapid characterizations if prominent changes in the metabolome could be captured in a reproducible manner. Techniques that can handle a large number of samples with minimal sample preparation but are still capable of providing relevant chemical information are well suited for generating rapid fingerprints. In this regard, crude extracts or whole cells can be analysed by MS [29], NMR [11], Fourier transform infrared (FT-IR) or Raman spectroscopies [30]. NMR fingerprinting is currently the method of choice for ‘metabonomics’, but many researchers, aware of its comparatively poor sensitivity, are complementing this approach with MS-based technologies. Although its sensitivity is poor, NMR does provide a uniform detection system of equal sensitivity for all proton-containing molecules; by contrast, MS (particularly in direct infusion) is selectively sensitive, prone to matrix effects (often and erroneously lumped together as ‘ion suppression’) and can be insensitive to some classes of analyte.

A metabolic fingerprinting approach yields data of a similar format to that described above: the first list contains mass-to-charge (*m/z*) ratios, chemical shifts or wavenumbers for MS, NMR or FT-IR, respectively, and the second list contains their relative contribution. Although current technology is far from generating comprehensive information on metabolite pools (metabolomes), there is already sufficient complexity in the information to begin to reconstruct the networks involved [31].

Databases for metabolomics

In a recent study, Lyman and Varian estimated that in 2000 the world produced between 1 and 2 exabytes ($1-2\times10^{18}$ bytes) of 'unique' information per year (http://www.sims.berkeley.edu/how-much-info). This flood of data is roughly 250 megabytes for every man, woman and child on earth! IBM's estimates are that information within the life sciences doubles every 6 months (http://www.bio-itworld.com/champions/janet_perna.html); this data explosion comes from genomic sequencing, the 'omics' and high-throughput screening, as well as the more traditional preclinical and clinical trials.

In his keynote address, *A National Geographic Information System – An Achievable Objective?*, to the Australasian Urban and Regional Information Systems Association in 1990, Henry Nix said

> "Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom."

But can we cope with this torrent of data from metabolomics? It is clear that in the era of postgenomic biology, we shall need good databases, very good data and even better algorithms with which to turn our data into knowledge. The types of database that will be useful for metabolomics are described in Box 2. Curation of these databases is essential if they are to be useful to the wider community. We are all aware of the number of incorrect DNA sequences in databases (e.g. see [32]). It is relatively easy to spot errors in DNA because of its low complexity; however, it will be essential that metabolite profiles are validated and the metadata are complete, so that other researchers can use the same experimental protocol and can compare their profiles against those of others stored in the database.

The issue of precision is acute because, in contrast to the traceable standards that facilitate machine calibration in simple univariate measurements (e.g. blood glucose), there are no simple standards for 'omic measurements, which have hundreds or thousands of variables and for which machine drift can be acute. Freshly made up cocktails of standards, together with the use of advanced transformations [33], might help to solve to this problem. The metadata also need to be captured correctly. A systematic approach already exists for transcriptome data [34] and is being developed for proteomics [35]. Such an approach is also being investigated for metabolomics by Hardy and Fuell [36].

Box 2. Types of database for metabolomics

- Databases storing detailed metabolite profiles, including raw data and detailed metadata (i.e. data about the data) [73].
- Single species-based databases that will store 'relatively' simple metabolite profiles [73].
- Databases storing complex metabolite profile data from many species in many different physiological states [73].
- Databases listing all known metabolites for each biological species. With suitable metadata, these databases could be extended to contain temporal and spatial information.
- Databases such as KEGG [74], compiling established biochemical facts.
- Databases that integrate genome and metabolome data with an ability to model metabolic fluxes [75,76].

A paradigm shift from metabolic pathways to networks and neighbourhoods

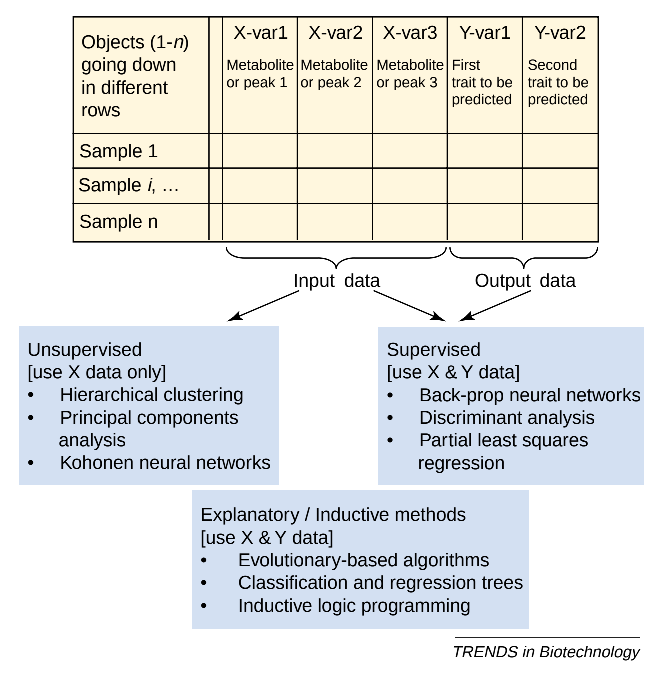
There has been a shift from mental constructs involving metabolic pathways to those based on metabolic networks and neighbourhoods [37,38], and many would argue that the 'Boehringer' metabolic pathways map needs to be updated both radically and conceptually. An excellent example of this is illustrated by the experiments of Willmitzer and colleagues [39] on the carbon sink in potatoes. The aim of these experiments was to increase the amount of starch in the tubers by the 'rational' overproduction of enzymes in the starch synthesis pathway. Rather than producing nice large tubers, however, these experiments decreased the size of the potatoes, suggesting that other pathways, indeed networks, are involved.

This finding is almost universally applicable to other crops, and 'unexpected' effects have been known in metabolic engineering for many years (e.g. see [40,41]). Thus, elucidation and visualization of metabolite neighbourhoods need to be achieved to understand the structural properties of the network [42,43]. This can be done only at the level of the metabolome because fluxes and thus relationships among metabolites through networks cannot be calculated accurately from transcripts or proteins.

Correlation analysis of metabolites is one approach that is being explored to discover novel pathways [44] and hence to infer the metabolic network [45], and this method needs to be linked with good visualization using biochemical network diagrams. Indeed, techniques for network reconstruction that are being developed for transcriptomics [46] are equally applicable to metabolomics and can be overlaid. Such diagrams can be put into context with previously known biochemistry and can also be used to link in transcriptome or proteome data; some excellent software for these applications has been produced by Mendes and colleagues [47].

Hypothesis-generating strategies from metabolome data

Many of the pattern-recognition strategies currently pursued in metabolomics, and indeed in the analyses of all 'omic data, are based on 'unsupervised' techniques [48] (Figure 2), such as hierarchical cluster analysis in which a 'tree-like' dendrogram (as commonly seen in taxonomic



Objects (1-*n*) going down in different rows	X-var1 Metabolite or peak 1	X-var2 Metabolite or peak 2	X-var3 Metabolite or peak 3	Y-var1 First trait to be predicted	Y-var2 Second trait to be predicted
Sample 1					
Sample *i*, …					
Sample n					

Input data

Output data

Unsupervised
[use X data only]
- Hierarchical clustering
- Principal components analysis
- Kohonen neural networks

Supervised
[use X & Y data]
- Back-prop neural networks
- Discriminant analysis
- Partial least squares regression

Explanatory / Inductive methods
[use X & Y data]
- Evolutionary-based algorithms
- Classification and regression trees
- Inductive logic programming

TRENDS in Biotechnology

Figure 2. The chemometric zoo. High-dimensional metabolome data can be analysed in many ways, which can be categorized as unsupervised and supervised learning. When learning is 'unsupervised', the system is shown a set of inputs and then left to cluster the metabolite data into groups. For multivariate analysis this optimization procedure is usually 'simplification' or dimensionality reduction; this means that a large body of metabolite data are summarized by a few parameters with minimal loss of information. After clustering, the ordination plots or dendrograms are then interpreted. When learning is 'supervised', the desired responses (*Y* data or 'traits' or 'classes') associated with each of the inputs (*X* data, or 'metabolome data') are known. The goal is to find a mathematical transformation (model) that will correctly associate all or some of the inputs with the target traits (e.g. whether an animal has been challenged with a drug, the environment that a plant has been grown in or the presence of or susceptibility to disease). In its conventional form, this goal is achieved by minimizing the error between the known target and the model's response (output). In addition, there exist special types of supervised learning that effect explanatory analyses; in other words, the mathematical transformation from input to output data is transparent. Such inductive methods allow one to discover which metabolites (inputs) are key for the separation of the traits to be predicted.

and phylogenetics) is produced. Clustering methods are used to assess, in a multivariate manner, how similar a set of samples are to one another on the basis of their metabolite profiles, although many of the methods used in transcriptomics are poorly reproducible, mathematically unjustified and lack quality metrics for how 'good' the clusters are. Nevertheless, the inclusion of suitable profiles of known provenance (e.g. the profile of the 'knockout' of a gene of known function) means that one can classify unknown samples by their closeness to the known knockouts, a process referred to as 'guilt by association' [49]. However, when several hundred different provenances are analysed, or when disjoint relationship in gene classes arise, this approach is imperfect and alternative strategies must be adopted [50].

Supervised machine learning algorithms [48] are very powerful methods that seek to transform the multivariate data from metabolite profiles into something of biological interest under the guidance of a 'teacher' (Figure 2). The basic idea behind supervised learning is that there are some patterns (e.g. metabolic fingerprints) that have desired responses that are known (e.g. whether an animal has been given a drug or placebo, or has a disease or a susceptibility to it). These two types of data (the representation of the objects and their responses in the system) form pairs that are conventionally called inputs (or *x* data or explanatory variables) and targets (or *y* data). The goal of supervised learning is to find a 'model' or 'mapping' that will correctly associate the inputs with the targets.

Many different algorithms perform supervised learning (see Table 1 for details). One of the most popular types of supervised learning method is based on artificial neural networks (ANNs), which can learn nonlinear as well as linear mappings [51]. But although they are very powerful, the mathematical transformation from metabolite data to the target trait of interest is often largely inaccessible in ANNs [52], and these methods are often perceived as 'black box' approaches to modelling spectra.

It is known from the statistical literature that better (i.e. more robust) predictions can often be obtained when only the most relevant input variables are considered [53]; in other words, 'parsimonious' models tend to generalize better. Thus, the best machine learning techniques not only should give the correct answers, but also should identify a subset of the variables with the maximal explanatory power, thereby providing an interpretable description of what, in biological terms, is the basis for that answer. Such explanatory modelling methods do exist and their salient features are described in Table 1.

The pregenomic era of molecular biology was largely reductionist and qualitative [54], and it relied excessively on a hypothesis-centric view of the world. But not all scientific advances are hypothesis-driven (or hypothesis-dependent) [17]. The iterative process between data gathering and the generation and evaluation of ideas is sometimes referred to as the 'cycle of knowledge' [55] (Figure 3). In the traditional cycle, we have some preconceived notions about the problem domain; experiments are designed to test our hypotheses; and the observations from these experiments are recorded and, by 'deductive' reasoning, considered to be consistent or inconsistent with the hypotheses [56] (Figure 3a). In fact, although this part is normally only implicit, by a process of 'induction' or abduction [57] these observations are synthesized or generalized to refine our accepted wisdom. The cycle then repeats itself until we are happy with the solution to a given problem.

In the early stages of functional genomic programmes, however, we have a situation in which our knowledge is minute: that is, we have no ideas about the role of an orphan open reading frame and there are few if any hypotheses to test [58]. We can, however, design experiments that are based on gene knockouts and controlled overexpression, for example, and observe the effects on the phenotype of the organism. We are then in a position of having collected a great many observations, and the trick is to drive the cycle round via a kind of data-driven or inductive reasoning to generate new hypotheses (Figure 3b).

Evolutionary computing methods [59], classification and regression trees (CART) [60] and inductive logic programming [61] can be considered as 'inductive-reasoning-based' algorithms that are completely data driven and

Table 1. Features of some common supervised learning algorithms

Method	Significant features	Categorical or quantitative [77][a]	Interpretability	Refs
Discriminant analysis	Cluster analysis method; involves projection of test data into cluster space	Categorical	Loadings matrices can give an indication of important inputs	[78]
Partial least squares	Linear regression method	Quantitative	Loadings matrices can give an indication of important inputs	[79]
Discriminant partial least squares	Linear regression method	Categorical	Loadings matrices can give an indication of important inputs	[79]
ANNs	Very popular machine learning methods; can learn nonlinear as well as linear mappings; the main mappings used are multilayer perceptrons and radial basis functions	Both	Mapping from input to output largely opaque; can be improved by pruning or growing ANNs	[80,81]
Rule induction	Based on the growth of a decision tree with predictive segregation of the data; the leaves contain as few different classes as possible; includes CART and fuzzy rule-building expert system	Categorical	Produces uni- or multivariate decision boundaries	[60,82,83]
Inductive logic programming	Uses a specific logic-based language	More categorical than quantitative	Constructs general rules by inductive inference	[61]
Evolutionary computation	Based on concepts of Darwinian selection to generate and to optimize a desired mapping between input and output variables; includes genetic algorithms, genetic programming and genomic computing	Both	Often produces interpretable rules, genetic code and parse trees	[59,84–86]

[a]The output can be either categorical or quantitative. In the former, for example, the metabolome data might have been collected from sera of patients with or without a disease; by contrast, in the latter the output might be considered quantitative such as the level or severity of the disease.

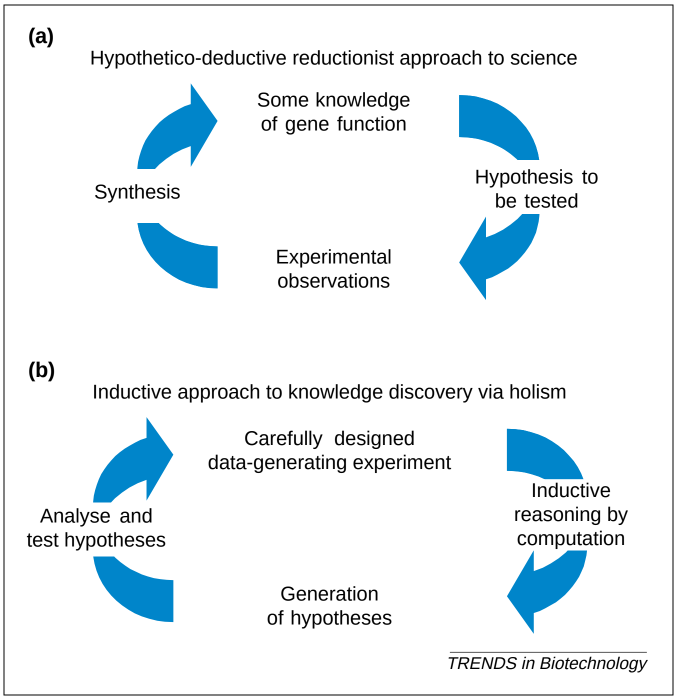

(a)
Hypothetico-deductive reductionist approach to science
Some knowledge of gene function
Hypothesis to be tested
Experimental observations
Synthesis
(b)
Inductive approach to knowledge discovery via holism
Carefully designed data-generating experiment
Inductive reasoning by computation
Generation of hypotheses
Analyse and test hypotheses
TRENDS in Biotechnology

Figure 3. The cycle of knowledge and holism. (a) The traditional cycle of knowledge, in which background knowledge is used to construct a hypothesis to be tested experimentally The experiment produces data that are consistent or otherwise with the hypothesis. In other words, the hypothesis is the starting point. (b) The inductive approach, where there is no real hypothesis and thus the strategy is to generate a hypothesis from the data and not to start with one. This data-driven approach requires computer-based inductive reasoning to turn the data into hypotheses, which can then be tested in the traditional manner. Experimental design is important here, because this determines which experiments are used to populate the search space of possible (useful) experiments. The strategy in which an algorithm chooses which experiments to do is known as 'active learning' [87] and is the strategy of choice. Evolutionary computing methods can be used for active learning [88].

are thus especially appropriate for problems that are data rich, but hypothesis and/or information poor. All of these methods can be used to generate rules and thus hypotheses from suitable examples, and evolutionary computing methods in particular have been used to advantage in metabolomics (e.g. see [62]). Of course, as with any purely inductive method, there are no axioms and so the rules that evolve cannot be proved correct; however, they greatly narrow the search space of possibilities, and by testing them new knowledge will be generated that will lead to an increased understanding of the function of the orphan gene.

From metabolomics to systems biology

> "When a thing was new, people said, 'It is not true'. Later, when the truth became obvious, people said, 'Anyway, it is not important.' And when its importance could not be denied, people said, 'Anyway, it is not new." William James (1842–1920).

'Systems biology' describes a range of techniques, including the 'omics and mathematical modelling, for understanding systems 'as a whole' [63,64], and it is widely recognized that metabolomics will have a major part to play in its development [65]. Emerging trends include the development of suitable mark-up languages for exchanging the models (e.g. see [66]) and a recognition of the close relationship between metabolic engineering [67] and systems biology [68]. To quote Henrik Kacser, one of the architects of metabolic control analysis, "But one thing is clear: to understand the whole, one must study the whole" [69]. The goal of systems biology and metabolomics is to do just that.

Concluding remarks

The field of metabolomics is gaining increasing interest across all disciplines, including functional genomics, integrative and systems biology, pharmacogenomics, and (surrogate) biomarker discovery for drug discovery and therapy monitoring. As more researchers get 'tooled up' for metabolomics, the realization that it is easy to generate floods (or, more accurately, torrents!) of data will become apparent. Thus, in the new postgenomic era of biology, we shall need well-curated databases, very good data with which to populate them, and even better algorithms with which to turn these metabolome data into knowledge.

Acknowledgements

R.G., S.V. and D.B.K. are most grateful to the UK BBSRC, UK EPRSC and UK NERC for supporting our work.

References

1 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512

2 Anderson, S. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290, 457–465

3 Blattner, F.R. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474

4 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

5 The International Human Genome Mapping Consortium (2001) A physical map of the human genome. *Nature* 409, 934–941

6 Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171

7 Oliver, S.G. *et al.* (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 16, 373–378

8 Beecher, C.W.W. (2003) The human metabolome. In *Metabolic Profiling: its Role in Biomarker Discovery and Gene Function Analysis* (Harrigan, G.G. and Goodacre, R., eds), pp. 311–319, Kluwer Academic Publishers

9 Fell, D.A. (1996) *Understanding the Control of Metabolism*, Portland Press

10 Kell, D.B. and Mendes, P. (2000) Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era. In *Technological and Medical Implications of Metabolic Control Analysis* (Cornish-Bowden, A. and Cárdenas, M.L., eds), pp. 3–25, Kluwer Academic Publishers

11 Raamsdonk, L.M. *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 19, 45–50

12 Urbanczyk-Wochniak, E. *et al.* (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* 4, 989–993

13 ter Kuile, B.H. and Westerhoff, H.V. (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* 500, 169–171

14 Wilson, I.D. and Nicholson, J.K. (2003) Topics in xenobiochemistry: do metabolic pathways exist for xenobiotics? The micro-metabolism hypothesis. *Xenobiotica* 33, 887–901

15 Sanders, G.H.W. and Manz, A. (2000) Chip-based microsystems for genomic and proteomic analysis. *Trends Anal. Chem.* 19, 364–378

16 Mann, M. *et al.* (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437–473

17 Kell, D.B. and Oliver, S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays* 26, 99–105

18 Lindon, J.C. *et al.* (2003) So what's the deal with metabonomics? Metabonomics measures the fingerprint of biochemical perturbations caused by disease, drugs, and toxins. *Anal. Chem.* 75, 384A–391A

19 Nicholson, J.K. and Wilson, I.D. (2003) Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* 2, 668–676

20 Brindle, J.T. *et al.* (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics. *Nat. Med.* 8, 1439–1444

21 Tolstikov, V.V. and Fiehn, O. (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.* 301, 298–307

22 Shellie, R. *et al.* (2001) Concepts and preliminary observations on the triple dimensional analysis of complex volatile samples by using GC × GC-TOF MS. *Anal. Chem.* 73, 1336–1344

23 Weckwerth, W. *et al.* (2001) Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS analysis reveals silent metabolic phenotypes. In *Proceedings of the 49th ASMS Conference on Mass Spectrometry* American Society of Mass Spectrometry, Chicago

24 Duran, A.L. *et al.* (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* 19, 2283–2293

25 van Mispelaar, V.G. *et al.* (2003) Quantitative analysis of target components by comprehensive two-dimensional gas chromatography. *J. Chromatogr. A.* 1019, 15–29

26 Ivanova, P.T. *et al.* (2001) Electrospray ionization mass spectrometry analysis of changes in phospholipids in RBL-2H3 mastocytoma cells during degranulation. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7152–7157

27 Allen, J. *et al.* (2003) High-throughput classification of yeast mutants for functional genomics via metabolic footprinting. *Nat. Biotechnol.* 21, 692–696

28 Aharoni, A. *et al.* (2002) Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* 6, 217–234

29 Rashed, M.S. (2001) Clinical applications of tandem mass spectrometry: ten years of diagnosis and screening for inherited metabolic diseases. *J. Chromatogr. B* 758, 27–48

30 Petrich, W. (2001) Mid-infrared and Raman spectroscopy for medical diagnostics. *Appl. Spectrosc. Rev.* 36, 181–237

31 Alm, E. and Arkin, A.P. (2003) Biological networks. *Curr. Opin. Struct. Biol.* 13, 193–202

32 Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet.* 15, 132–133

33 Goodacre, R. and Kell, D.B. (1996) Correction of mass spectral drift using artificial neural networks. *Anal. Chem.* 68, 271–280

34 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365–371

35 Taylor, C.F. *et al.* (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* 21, 247–254

36 Hardy, F. and Fuell, H. (2003) Databases, data modelling and schemas. In *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis* (Harrigan, G.G. and Goodacre, R., eds), Kluwer Academic Publishers

37 Barabási, A-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113

38 Kell, D.B. Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* (in press)

39 Leggewie, G. *et al.* (2003) Overexpression of the sucrose transporter *So*SUT1 in potato results in alterations in leaf carbon partitioning but has little impact on tuber morphology. *Planta* 217, 158–167

40 Westerhoff, H.V. and Kell, D.B. (1987) Matrix method for determining the steps most rate-limiting to metabolic fluxes in biotechnological processes. *Biotechnol. Bioeng.* 30, 101–107

41 Snoep, J.L. *et al.* (1995) Protein burden in *Zymomonas mobilis* – negative flux and growth-control due to overproduction of glycolytic enzymes. *Microbiology* 141, 2329–2337

42 Jeong, H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature* 407, 651–654

43 Wagner, A. and Fell, D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. B.* 268, 1803–1810

44 Weckwerth, W. and Fiehn, O. (2002) Can we discover novel pathways using metabolomic analysis? *Curr. Opin. Biotechnol.* 13, 156–160

45 Steuer, R. *et al.* (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19, 1019–1026

46 Kholodenko, B.N. *et al.* (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12841–12846

47 Jing Li, X. *et al.* (2003) Databases and visualization for metabolomics.

In *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis* (Harrigan, G.G. and Goodacre, R., eds), Kluwer Academic Publishers

48 Hastie, T. *et al.* (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag

49 Altshuler, D. *et al.* (2000) Guilt by association. *Nat. Genet.* 26, 135–137

50 Kell, D.B. and King, R.D. (2000) On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends Biotechnol.* 18, 93–98

51 White, H. (1992) *Artificial Neural Networks: Approximation and Learning Theory*, Blackwell

52 Tickle, A.B. *et al.* (1998) The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Trans. Neural Netw.* 9, 1057–1068

53 Seasholtz, M.B. and Kowalski, B. (1993) The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta* 277, 165–177

54 Maddox, J.S. (1994) Towards more measurement in biology. *Nature* 368, 95

55 Kell, D.B. (2002) Genotype:phenotype mapping: genes as computer programs. *Trends Genet.* 18, 555–559

56 Oldroyd, D. (1986) *The Arch of Knowledge: an Introduction to the History of the Philosophy and Methodology of Science*, Methuen

57 Flach, P.A. and Kakas, A.C. (2000) *Induction and Abduction: Essays on their Relation and Integration*, Kluwer Academic Publishers

58 Brent, R. (2000) Genomic biology. *Cell* 100, 169–183

59 Koza, J.R. *et al.* (2003) *Genetic Programming: Routine Human–Competitive Machine Intelligence*, Kluwer Academic Publishers

60 Breiman, L. *et al.* (1984) *Classification and Regression Trees*, Wadsworth Inc

61 Muggleton, S. (1999) Inductive logic programming: issues, results and the challenge of learning language in logic. *Artif. Intell.* 114, 283–296

62 Kell, D.B. (2002) Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.* 29, 237–241

63 von Bertalanffy, L. (1969) *General System Theory*, George Braziller

64 Kitano, H. (2002) Systems biology: a brief overview. *Science* 295, 1662–1664

65 Weckwerth, W. (2003) Metabolomics in systems biology. *Annu. Rev. Plant Biol.* 54, 669–689

66 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531

67 Kell, D.B. and Westerhoff, H.V. (1986) Metabolic control theory – its role in microbiology and biotechnology. *FEMS Microbiol. Rev.* 39, 305–320

68 Sweetlove, L.J. *et al.* (2003) Predictive metabolic engineering: a goal for systems biology. *Plant Physiol.* 132, 420–425

69 Kacser, H. (1986) On parts and wholes in metabolism. In *The Organization of Cell Metabolism* (Welch, G.R. and Clegg, J.S., eds), pp. 327–337, Plenum Press

70 Fiehn, O. (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genomics* 2, 155–168

71 Harrigan, G.G. and Goodacre, R. (2003) *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, pp. 335, Kluwer Academic Publishers

72 Nicholson, J.K. *et al.* (1999) ‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29, 1181–1189

73 Mendes, P. (2002) Emerging bioinformatics for the metabolome. *Brief. Bioinform.* 3, 134–145

74 Kanehisa, M. *et al.* (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30, 42–46

75 Famili, I. *et al.* (2003) *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13134–13139

76 Förster, J. *et al.* (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 13, 244–253

77 Ellis, D.I. and Goodacre, R. (2002) Rapid and quantitative detection of the microbial spoilage of muscle foods: current status and future trends. *Trends Food Sci. Technol.* 12, 413–423

78 Manly, B.F.J. (1994) *Multivariate Statistical Methods: A Primer*, Chapman & Hall

79 Martens, H. and Næs, T. (1989) *Multivariate Calibration*, John Wiley

80 Rumelhart, D.E. *et al.* (1986) Learning internal representations by error propagation. In *Parallel Distributed Processing* (Volume 1, Foundations) (Rumelhart, D.E. and McClelland, J. eds), MIT Press, Cambridge MA

81 Broomhead, D.S. and Lowe, D. (1988) Multivariable function interpolation and adaptive networks. *Complex Syst.* 2, 321–355

82 Harrington, P.B. (1991) Fuzzy rule-building expert systems: minimal neural networks. *J. Chemometrics* 5, 467–486

83 Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann

84 Bäck, T. *et al.* (1997) *Handbook of Evolutionary Computation*, IOP Publishing/Oxford University Press

85 Reeves, C.R. (2002) *Genetic Algorithms – Principles and Perspectives: a Guide to GA Theory*, Kluwer Academic Publishers

86 Kell, D.B. *et al.* (2001) Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.* 126, 943–951

87 King, R.D. *et al.* (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427, 247–252

88 Vaidyanathan, S. *et al.* (2003) Explanatory optimisation of protein mass spectrometry via genetic search. *Anal. Chem.* 75, 6679–6686