DRUG DISCOVERY TODAY TARGETS

Metabolomics: from pattern recognition to biological interpretation

Wolfram Weckwerth and Katja Morgenthal

Metabolomics is a technology that aims to identify and quantify the metabolome – the dynamic set of all small molecules present in an organism or a biological sample. In this sense, the technique is distinct from metabolic profiling, which looks for target compounds and their biochemical transformation. The combination of both approaches is an emerging technique for the characterization of biological samples and for drug treatment. Metabolomics has proven to be very rapid and superior to any other post-genomics technology for pattern-recognition analyses of biological samples. Changing steady state concentrations and fluctuations of metabolites that occur within milliseconds are a result of biochemical processes such as signalling cascades: metabolomic techniques are instrumental in measuring these changes rapidly and sensitively. Metabolite data can be complemented by protein, transcript and external (environmental) data, thereby leading to the identification of multiple physiological biomarkers embedded in correlative molecular networks that are not approachable with targeted studies.

Metabolomics is gaining increasing interest in drug discovery and disease diagnostics and treatment [1–3]. The concept was recently introduced as the global analysis of all metabolites in a sample (metabolomics) and the analysis of metabolic responses to drugs or diseases (metabonomics) [4–8]. It is an extension of existing methods that look at target compounds and their biochemical transformation. Novel developments in mass detectors and techniques like LC–MS, GC–MS and CE–MS (see glossary) support the approach of measuring the whole metabolome, rather than only target compounds. The most important difference from former protocols is the idea of revealing metabolite dynamics in an unbiased manner, according to the general definition of 'omics technologies' [9,10]. The aim here is to study gene function and fundamental metabolic principles in biochemical networks, for example, the investigation of a disease as a multifactorial phenomenon. Metabolomic technology enables the rapid, accurate and precise analysis of metabolites and subsequent pattern recognition of biological samples, identification of biomarkers and the direct analysis of the biochemical consequences of mutations and drug treatment. In contrast to other post-genomic technologies, such as proteomics and transcriptomics, metabolomics has, within a few years, developed into a routine process with respect to sample throughput and robustness. Analytical techniques, such as GC–MS, LC–MS and NMR, are capable of detecting hundreds of individual chemical structures, and their scope ranges from metabolite fingerprinting to an extensive metabolite profiling. This comprehensiveness reveals a very high number of unidentified chemical structures. Thus, in the near future, the structural identification of potential metabolite candidates associated with diseases will be a major task for biological interpretation [11].

Wolfram Weckwerth*
Katja Morgenthal
Max Planck Institute of Molecular Plant Physiology,
14424 Potsdam,
Germany
*e-mail:
weckwerth@mpimp-golm.mpg.de

1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03609-X

GLOSSARY

Analytical techniques

LC–MS
Liquid chromatography, commonly high performance liquid chromatography (HPLC) on any kind of stationary phase (e.g. reversed phase, normal phase or ion exchange), coupled with mass spectrometry. Analytes are separated by their chemical properties (e.g. hydrophobicity, hydrophilicity or charge).

GC–MS
Gas chromatography, commonly applied to coated capillary columns, coupled to mass spectrometry. Analytes are separated by their boiling point and their interaction with the liquid layer covering the capillary in the gas phase.

CE–MS
Capillary electrophoresis coupled to mass spectrometry. Electrically charged analytes are separated by their mobility in a capillary filled with an electrolyte under the influence of an electric field.

Mass analyzers

Single quadrupole
A mass-selective ion filter that consists of four electronically operated metal rods.

Triple quadrupole
Three quadrupole devices coupled in a linear array. In the single reaction monitoring mode (SRM), selected 'parent ions' pass through the first quadrupole and enter a second quadrupole device used as a collision cell to generate product ions. The third quadrupole serves as a selective mass filter for the determination of the resulting 'daughter ions'.

Iontrap
Collects and stores ions by forcing them into stable orbits and subsequently releases them mass-selectively. Collected and stored 'parent ions' can also be fragmented and thus 'daughter ions' can be analyzed.

Time-of-flight (TOF)
Ions are simultaneously accelerated resulting in the same kinetic energy for any given ion. Along an evacuated flight tube with a fixed distance ions are separated by their mass to charge ratio and velocity, respectively.

Statistical analysis

Principal components analysis (PCA)
An unsupervised method without *a priori* information about sample groups or replicates for the visualization of multidimensional data by dimensionality reduction. The method generates a new set of linear combinations of the original variables – metabolites or spectral data – called principal components (PCs). PCs are required to be orthogonal to each other and form a new coordinate system for separation of the samples. The greatest variance of the dataset is explained by the first principal components, in an ideal case 95% variance for the first three components.

Independent component analysis (ICA)
A technique comparable to PCA except that the new components do not have to be orthogonal to each other, and thus can be better optimized for statistical independence.

The application of biostatistics and novel mathematical frameworks will have a strong role in the extraction of biologically meaningful information from such huge datasets. In multivariate data analysis, problems arise from small sample numbers in contrast with the high number of measured metabolites, and the resulting high dimensionality of the data matrix. Novel algorithms and statistical analysis have to be improved and established for test cases and learning sets.

After a general introduction to metabolomic technology in diagnostics, pattern recognition and biomarker discovery, this review will focus on novel developments in structural elucidation, biostatistics and the biological interpretation of metabolite profiles as a fingerprint of metabolic networks. A stochastic model of metabolic networks is introduced that has lead to a novel understanding of co-regulation in biochemical networks. In this context, the complementation of metabolomic data with protein and transcript data is described. The review will also discuss whether integrative measurements are able to deliver causality for understanding the complexity of regulatory network structures.

Metabolomic techniques

In view of the chemical and physical diversity of small biological molecules, the challenge remains of developing protocols to gather the whole 'metabolome'. No single technique is suitable for the analysis of different types of molecules, which is why a mixture of techniques has to be used [9]. In proteomics and transcriptomics, problems arise from the sheer number of dynamically fluctuating transcripts and proteins, as well as posttranscriptional and posttranslational regulation and modification. In the field of metabolomics, the general estimations of the size and the dynamic range of a species-specific metabolome are at a preliminary stage. We do not know how many metabolites and derivatives of known metabolites are expected in mammals, plants and bacteria – in other words, we are looking at the tip of the iceberg without knowing what is below.

Metabolic fingerprinting and metabonomics with high sample throughput but decreased dynamic range and the deconvolution of individual components achieve a global view of the *in vivo* dynamics of metabolic networks: there are excellent reviews covering this topic, including NMR, direct infusion mass spectrometry, and/or infrared spectroscopy [4,12–16]. GC–MS and LC–MS technology achieve a lower sample throughput but provide unassailable identification and quantitation of individual compounds in a complex samples.

Major steps forward in these technologies have made it possible to match specific demands with specific instruments and novel developments in the performance of mass analyzers (Table 1). However, it is important to note that each type of technology exhibits a bias towards certain compound classes, mostly due to ionization techniques, chromatography and detector capabilities. GC–MS has evolved as an imperative technology for metabolomics due to its comprehensiveness and sensitivity [8,17–27]. The coupling of GC to time-of-flight (TOF) mass analyzers is an emerging technology. High scan rates provide accurate peak deconvolution of complex samples [28–32]. GC–TOF–MS capabilities provide an improvement over conventional GC–MS analysis in the analysis of ultracomplex samples, which is particularly important for the

TABLE 1

Standard techniques for non-targeted and targeted metabolite analysis

	Sensitivity	Throughput	Comprehensiveness
NMR	Low	Low–high	Low–high
IR	Low	High	Low
LC–NMR	Low	Low	High
LC–MS	Medium	High	High
GC–MS	High	High	High
CE–MS	High	Medium	High
LC–EC–MS	High	High	High
LC–UV	Medium–high	High	Very low

metabolomics approach [20,21,25]. Ultracomplex samples contain hundreds of coeluting compounds that vary in abundance by several orders of magnitude. Thus, accurate mass spectral deconvolution and a broad linear dynamic range represent indispensable prerequisites for high quality spectra and peak shapes. Modern GC–TOF–MS applications and incorporated mass spectral deconvolution algorithms fulfil these requirements.

Other promising technologies are CE–MS [33–37], particularly for the analysis of polar and thermolabile compounds, and electrochemical detection in parallel with LC–MS for analysis of redox-active compounds [38–40]. The coupling of LC–MS is the most established technique used for targeted identification and quantitation of specific metabolites in complex mixtures [41] [42–44]. By contrast, for the non-targeted analysis of all compounds in a complex sample, novel deconvolution algorithms have to be implemented [45–48], taking into account the differences in data acquisition capacities of specific mass analyzers, for example quadrupole TOF instruments and ion traps [45,47,49,50].

Structural elucidation of unknown metabolites and metabolite mass spectra libraries

The structural elucidation of small molecules is a challenging task. Typically, the compound of interest has to be purified to homogeneity using several chromatographic steps before being analyzed by NMR and MS [11,51–54]. Moreover, hyphenated LC–NMR–MS techniques are increasingly used for the elucidation of molecular structures [52,55]. Biomass is a limiting factor for metabolites of low abundance in a biological sample [51] but these metabolites can be easily and reproducibly detected and quantified in GC–MS and LC–MS analysis by assigning m/z values. Thus, there is a huge discrepancy between the amount of sample needed for metabolomic analysis that is able to detect unidentified peaks of very low abundance in a chromatogram, and the amount of sample required for structural elucidation of unknown compounds. At present, lists of unidentified chemical structures from metabolomic analyses are exponentially growing. Presumably, comprehensive species-wide structural elucidation and generation of metabolite libraries is only possible as a combined effort between many research groups, recalling the exerted efforts of whole-organism genome sequencing projects and proteomics projects like the Human Proteome Organisation (www.hupo.org). The challenge for metabolomics is not only to elucidate unknown chemical structures but also to put meta-information, such as sample origin, tissue and experimental conditions, into an accessible format [see also standards for proteomics [56] and microarrays [57], or the Plantontology website (www.plantontology.org) that try to define standards for such experimental data and meta-information]. One of the most comprehensive databases providing chemical and structural properties of compounds is the US National Institute of Standards and Technology database (www.nist.gov/srd/nist1a.htm) [30,31,58,59]. Several attempts to standardize mass spectrometry data can be found in the literature, and Table 2 shows links for open access databases and services [60,61]. Unfortunately, different communities follow different computational standards, thereby limiting access to data. This is also a general problem with proteomics data [62,63]. In our opinion, only user-friendly and biology-oriented databases will become generally accepted. Linking information available in databases (e.g. gene annotation) with experimental data (e.g. raw MS data, experimental design and meta-information) will be a crucial feature. It would be desirable to create a database combining peak and spectral deconvolution of raw files from various instruments, mass spectral libraries, library search algorithms, annotation of structures, genes and functions and multivariate data analyses provided with normalization and transformation protocols (Figure 1). Because there are many software and internet packages available that perform parts of the work (Table 2), these databases might develop into open access tools like the US National Center for Biotechnology Information (www.ncbi.nlm.nih.gov).

Interpretation of metabolomic data

Most current metabolite profiling approaches rely on the measurement of steady-state levels, therefore, they do not necessarily reflect the flux of metabolites through a network of pathways. However, this flux is important for revealing alterations in enzymatic reactions, which in-turn affect regulatory processes and putative drug targets in metabolism. Flux profiling is possible in artificial biological systems that are associated with the loss of the natural habitat, for example, cell cultures, microorganism cultures or tissue samples [64–69]. However, The ultimate clues on *in vivo* dynamics are only revealed in intact systems in their natural environment. The question is whether steady state measurements can be exploited to reveal flux alterations and *in vivo* regulation of enzymatic reactions. Early work by Arkin and Ross [70–74] demonstrated the need to introduce stochastic models for the interpretation of metabolic networks. We developed an analogous model system that interprets the metabolite correlations observed in metabolomic datasets after multivariate data mining

TABLE 2

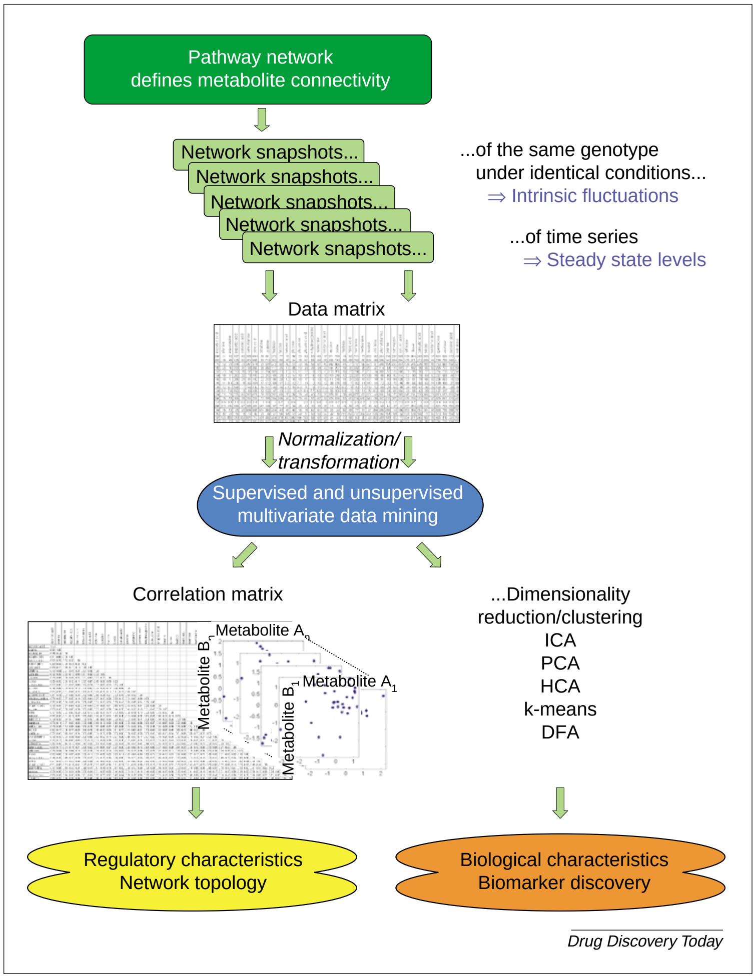
Selection of commercial and non-commercial web-sites of databases, mass spectral libraries, peak deconvolution and multivariate data analysis

Name	Publisher	URL	Application
Automated Mass Spectral Deconvolution and Identification System (AMDIS)	US National Institute of Standards and Technology	http://chemdata.nist.gov/mass-spc/amdis	GC mass spectral library
AnalyzerPro	MatrixAnalyzer Spectralworks	www.spectralworks.com	High throughput LC–MS and GC–MS data processing engine
Component Detection Algorithm (CODA)	ACD Labs	www.acdlabs.com	Prediction of mass spectral, fragmentation for LC and GC, Simulation of LC and GC chromatograms
MassFrontier	Thermo Electron	www.highchem.com	Management, interpretation, and evaluation of mass spectra
MetAlign	Plant Research International PRI	www.metalign.nl	Analysis, alignment and comparison of GC–MS or LC–MS datasets
MetaGeneAlyse	Max Planck Institute of Plant Physiology	http://metagenealyse.mpimp-golm.mpg.de	Multivariate analysis of gene expression and metabolite data
CSBDB	Max Planck Institute of Plant Physiology	http://csbdb.mpimp-golm.mpg.de	Access to public mass spectra libraries and metabolite profiling experiments
ArMet	University of Wales, Aberystwyth	www.armet.org	Framework for the description of plant metabolomics experiments
MeT-RO	UK Centre for Plant and Microbial Metabolomic Analysis	www.metabolomics.bbsrc.ac.uk/MeT-RO.htm	Primary and secondary metabolite profiling approach

and connects these correlations to the underlying enzymatic pathway structure [9,75–78]. These correlations can reveal alterations in enzymatic activity and in the differential analysis of various metabolic states [25,78–80]. This is possible using multiple metabolomic datasets originating from different experimental conditions and/or biological tissues. Intrinsic biological fluctuation of independent samples results in altered patterns of observed correlations. The observation of correlated metabolites offers the chance to investigate whole metabolite network dynamics based on correlation network topologies (Figure 1) [9,25,80]. Changes in the network topology point to regulatory hubs in the biochemical network (Figure 1) because the correlation matrix of metabolite pairs is a fingerprint of the enzymatic and regulatory reaction network. It is further possible to compare the measured correlation network with the proposed underlying reaction network and the corresponding numerically resolved correlation network [9]. Here it becomes evident that correlations cannot be predicted only on the basis of pathway connectivity. Most pairs of metabolites neighboured in the reaction network show a low correlation, whereas other metabolite pairs that are far apart from each other in the reaction network exhibit a strong correlation [9,76,77,79,80]. Primarily regulatory properties, especially the modulation of enzyme activity, serve as a source of changes in the topology of the correlation network [9,25,77]. However, the interrelation of an enzymatic reaction network and the resulting correlation matrix is the basis of the connectivity between measured metabolomic networks and the underlying biochemical regulatory *in vivo* network [9,76]. Any alteration in the reaction network (e.g. disease versus control sample, inhibition of enzyme activity, genetic suppression or enhancement of a reaction or addition of new pathways) will result in different correlation matrices [9,25,78]. Because we are looking for differences in correlations, these phenomena can be analyzed using multivariate statistical analysis. Furthermore, the approach can be extended to systems analysis looking for co-regulation of metabolites, proteins, transcripts or any other external or internal parameter measurable in the system to reveal a holistic picture of metabolism.

Functional studies of networking systems

Molecular genetics has enabled studies of gene and protein function using expression profiling, knockout mutants, antisense expression, promotor-controlled expression, RNA interference and targeted mutations [81–83]. These techniques hold great promise for functional genomics, which is attempting to describe the function and interaction of genes based on high-throughput techniques, and systems biology, which explores the interactions between various parts of a biological system (e.g. metabolic pathways, organelles and cells), to design an overall picture of metabolism. An emerging technology with significant relevance for drug discovery is the systematic screening for small molecules as protein affectors to manipulate and investigate protein function and activity [84,85]. Known as chemical genetics or chemical biology, this technology is used for target discovery using purified proteins combined with computer-supported structural simulation [86,87] and physiology-based screening



Drug Discovery Today

FIGURE 1

Multivariate data analysis using metabolomics data. The active pathway network is representative of the biological sample. The metabolite measurements take snapshots of the system and the data are collected in multidimensional data matrices. Multivariate statistical analysis enables pattern recognition and biomarker identification using, for instance, unsupervised statistical analysis such as PCA or ICA. Metabolite correlation analysis enables comparison between differential network structures and the identification of regulatory hubs within these networks. The whole process can be extended to integrative data matrices consisting of metabolites, proteins, transcripts and environmental data (for details see text and [9,21,25,76–78]). Abbreviations: ICA, independent components analysis; PCA, principal components analysis; HCA, hierarchical cluster analysis; DFA, discriminatory function analysis.

approaches, for instance, in cell cultures. The approach is not only useful for identifying novel drugs acting on specific targets but also for identifying networking in different biological processes. Complex biological systems with divergent and/or redundant signalling pathways might require multicomponent interventions in different perturbation points to modulate a disease pathway [84,85,88]. The assessment of these systems using metabolomics and multilevel data integration (proteins, transcripts and external parameters) opens the way to a novel definition of a disease or a living system as a molecular circuitry of interrelated actions of metabolites, proteins and gene networks [9,21,89–91]. However, before data reveal their interrelation, extended statistical and mathematical concepts are required for the integrative analysis of multifactorial phenomena. Normalization and transformation procedures are of great importance, as is the detection of significant correlations of the components, which is based on clustering, dimensionality reduction, mutual information and machine learning [8,14,21,25,92–94]. The disproportion of only a few experiments contrasting with thousands of variables (metabolites, proteins, transcripts and external parameters) is a challenge for the statistical analysis of these data. However, most of the data mining tools described below are closely related – based on covariance and/or correlations within a data matrix – and therefore have the potential for comparison of results originating from different procedures (Figure 1).

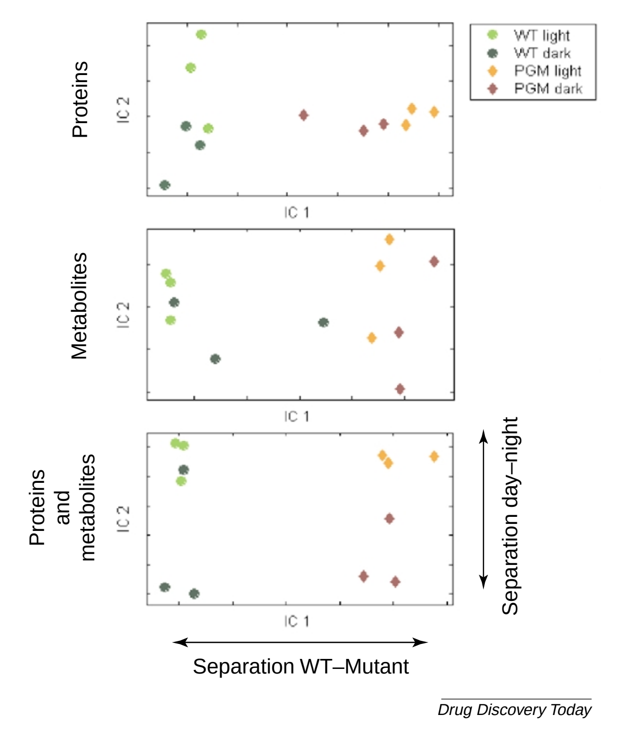
Multilevel-data integration and improvement of pattern recognition and biomarker identification

Causal, or ‘orthogonal’, connectivity of a networking system is revealed with the integration of orthogonal data, such as kinetics, for metabolites, proteins, transcripts or environmental data [9,21,78,95–98]. In an excellent review by Searls [99], data integration was identified as being a bottleneck for future research in drug discovery. The increasing recognition of the complexity of multifactorial diseases demands these kinds of approaches.

Recently, we combined multivariate metabolomic and proteomic data and time-series measurements to reveal protein–metabolite correlations [21,78]. Different methods of multivariate statistical analysis can be explored for the interpretation of these data. In a typical chemometric analysis by Raamsdonk *et al.* [100], NMR spectra of complex yeast extracts were normalized and analyzed by unsupervised principal components analysis (PCA). Owing to an indecisive separation of different samples, the PCA was followed by a supervised discriminant function analysis using *a priori* information based on spectral replicates.

The discrimination of the samples enables the identification of novel components. These components are interpretable as inherent biological characteristics [101,102]. However, the use of *a priori* information bears the risk that ‘noise’ in the measurements – generated by NMR or GC–MS, for instance – is overestimated, especially in the case of a low number of samples or replicates. Therefore, to circumvent these problems and to identify as many meaningful components as possible in a dataset, it is advantageous to seek methods that improve unsupervised analysis [103].

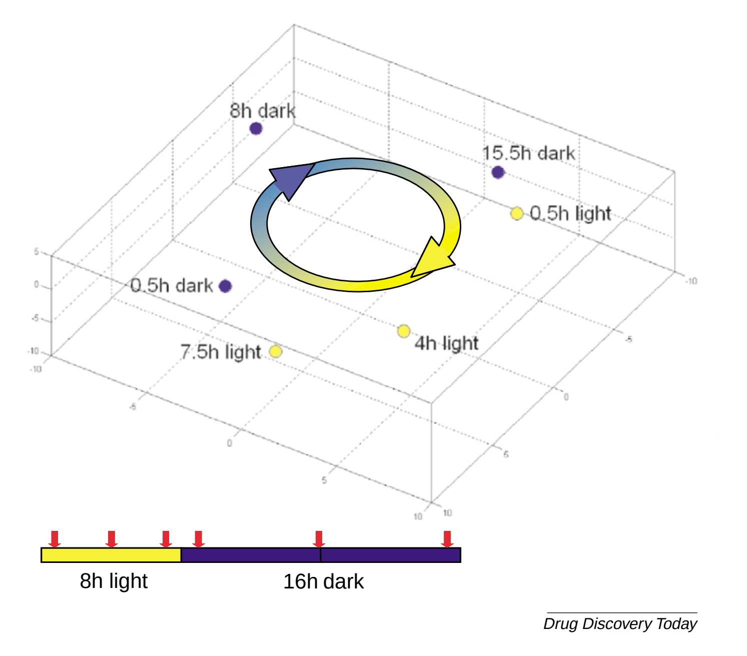
Using an integrated metabolite–protein data matrix, we exploited the improvements of independent components analysis (ICA) [103,104] compared with PCA. Interestingly, the metabolite and the protein data matrices alone showed a different grouping of the samples, but the combination of the data (in other words the correlative metabolite–protein data matrix) improved the separation of the samples (Figure 2) [78]. More importantly, it was possible to unambiguously assign the independent components to the different biological characteristics: mutant and wild-type plants were separated on the first independent



Drug Discovery Today

FIGURE 2

Pattern recognition using independent components analysis. Wild-type (WT) *Arabidopsis thaliana* plants were compared with a starchless *Arabidopsis* phosphoglucomutase mutant throughout a diurnal rhythm [78]. Each dot in the graph represents the average of a triplicate analysis of leaf metabolites and proteins. The integrated metabolite–protein data matrix shows an improved separation of the samples. The components are assignable to the separation of the wildtype and the mutant plant: independent component 1 (IC1) and the diurnal rhythm (IC2).



Drug Discovery Today

FIGURE 3

Principal components analysis based on metabolite–protein data of plant day–night rhythm. The samples were taken throughout an 8 h light/16 h dark rhythm. Visible is the diurnal trajectory (arrow). The transitions from dark to light (15.5 h dark → 0.5 h light) and from light to dark (7.5 h light → 0.5 h dark) are close together according to their sampling time.

component, and the diurnal rhythm (the differences between day and night samples) was found on the second independent component (Figure 2). Biomarkers that are responsible for these different biological characteristics can easily be classified because of the optimized separation using ICA and an integrated metabolite–protein dataset [78]. Evidently, this kind of analysis depends strongly on the comprehensiveness and accuracy of the profiling method, in this case metabolite and protein detection. Assuming that the techniques will improve, more proteins and metabolites can be identified and accurately quantified, the integrated analysis will have great promise.

Molecular coherence in biological systems

Statistical analyses of multivariate datasets enable the visualization of biological and molecular coherence. This coherence is based on the inherent correlative behaviour of metabolites, proteins and transcripts in response to environmental conditions. Correlation networks represent fingerprints of biochemical interactions, such as the regulation of enzyme activity, protein associations and the interplay of anabolism and catabolism. The details of the fingerprint form a coherent perception of network structures, clustering of metabolites or grouping of samples (Figure 1). The challenge is to interpret the conformation as a whole and not as a sum of single details, provided that the analytical accuracy for data generation is high enough. A disease is then described as the correlative action of many molecular factors: a multifactorial phenomenon. Intriguingly, 'coherent perception' is a technical term in psychology used in, for instance, the recognition of human faces. The presentation of shapes is often used as an example to explore multidimensional data where single variables add up to a holistic perception [105–107]. The observation of the overall shape, rather than of single constituting details, gives an immediate understanding. The question remains if this concept is transferrable to the perception of coherence in multivariate datasets constituted of molecular interaction. Examples might be the visualization of the diurnal rhythm of a plant based on metabolite–protein data shown in figure 3 [78] or the developmental stages of a fish embryo described in a study by Viant *et al.* [102].

Conclusion

Metabolomics has developed into a proteomics and transcriptomics-complementing technology within only a few years [9]. In combination with techniques for functional analysis of genes, it is hoped that a holistic picture of metabolism can be formed. This is the first time that integrative data matrices comprise transcript, protein and metabolite data, thus enabling us to look at interacting component networks. In combination with mathematical models of metabolism and statistical assessment of these data, it is foreseeable that we will reach a higher level of biological understanding. These techniques will lead to the identification of physiological and clinical biomarkers that are not approachable with targeted studies [1]. Novel models of the structures of metabolic networks and the application of multivariate statistical analysis have demonstrated that biological and biochemical interpretation of high dimensionality data matrices is possible [9,25,76,78,79,102]. However, at the present time, caution is advisable about the accuracy and robustness of the technologies, especially in combination with statistical evaluation of the data. Normalization and transformation procedures might have a strong impact on clustering the data and therefore influence any biological interpretation. The challenge is to improve the accuracy of the methods with respect to matrix effects, dynamic range and the sheer number of compounds in real-world samples.

Acknowledgments

We are grateful to Matthias Scholz and Joachim Selbig for fruitful discussions. We apologize to all colleagues whose work was not cited or insufficiently cited in this review.

References

1 Frank, R. and Hargreaves, R. (2003) Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov.* 2, 566–580

2 Matsumoto, I. and Kuhara, T. (1996) A new chemical diagnostic method for inborn errors of metabolism by mass spectrometry - Rapid, practical, and simultaneous urinary metabolites analysis. *Mass Spectrom. Rev.* 15, 43–57

3 Horning, E.C. and Horning, M.G. (1971) Metabolic Profiles - Gas-Phase Methods for Analysis of Metabolites. *Clin. Chem.* 17, 802–809

4 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

5 Trethewey, R.N. *et al.* (1999) Metabolic profiling: a Rosetta Stone for genomics? *Curr. Opin. Plant Biol.* 2, 83–85

6 Tweeddale, H. *et al.* (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("Metabolome") analysis. *J. Bacteriol.* 180, 5109–5116

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

8 Fiehn, O. *et al.* (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18, 1157–1161

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

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

11 Nassar, A.E.F. and Talaat, R.E. (2004) Strategies for dealing with metabolite elucidation in drug discovery and development. *Drug Discov. Today* 9, 317–327

12 Nicholson, J.K. *et al.* (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.* 1, 153–161

13 Castrillo, J.O. and Oliver, S.G. (2004) Yeast as a touchstone in post-genomic research: Strategies for integrative analysis in functional genomics. *J. Biochem. Mol. Biol.* 37, 93–106

14 Goodacre, R. *et al.* (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22, 245–252

15 Kell, D.B. (2004) Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* 7, 296–307

16 Dunn, W.B. and Ellis, D.I. (2005) Metabolomics: Current analytical platforms and methodologies. *TrAC Trends Anal. Chem.* 24, 285–294

17 Sauter, H. *et al.* (1988) Metabolic Profiling of Plants - a New Diagnostic-Technique. *Abstracts of Papers of the American Chemical Society* 195, 129

18 Roessner, U. *et al.* (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* 23, 131–142

19 Roessner, U. *et al.* (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13, 11–29

20 Weckwerth, W. *et al.* (2001) Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS analysis reveals silent metabolic phenotypes. *Proceedings of the 49th ASMS Conference on Mass spectrometry and Allied Topics, American Society of Mass Spectrometry, Chicago*, 1-2

21 Weckwerth, W. *et al.* (2004) Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 4, 78–83

22 Wagner, C. *et al.* (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* 62, 887–900

23 Broeckling, C.D. *et al.* (2005) Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.* 56, 323–336

24 Webb, J.W. *et al.* (1986) Metabolic Profiling of Corn Plants Using HPLCc and GC/MS. *Abstracts of Papers of the American Chemical Society* 191, 70

25 Weckwerth, W. *et al.* (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7809–7814

26 Jonsson, P. *et al.* (2004) A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Anal. Chem.* 76, 1738–1745

27 Fiehn, O. *et al.* (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* 72, 3573–3580

28 Watson, J.T. *et al.* (1990) Renaissance of Gas-Chromatography Time-of-Flight Mass-Spectrometry - Meeting the Challenge of Capillary Columns with a Beam Deflection Instrument and Time Array Detection. *J. Chromatogr.* 518, 283–295

29 Veriotti, T. and Sacks, R. (2001) High-speed GC and GC/time-of-flight MS of lemon and lime oil samples. *Anal. Chem.* 73, 4395–4402

30 Stein, S.E. and Scott, D.R. (1994) Optimization and Testing of Mass-Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* 5, 859–866

31 Stein, S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.* 10, 770–781

32 Tong, C.S. and Cheng, K.C. (1999) Mass spectral search method using the neural network approach. *Chemom. Intell. Lab. Syst.* 49, 135–150

33 Schmitt-Kopplin, P. and Frommberger, M. (2003) Capillary electrophoresis-mass spectrometry: 15 years of developments and applications. *Electrophoresis* 24, 3837–3867

34 Soga, T. *et al.* (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* 2, 488–494

35 Soga, T. *et al.* (2002) Pressure-assisted capillary electrophoresis electrospray ionization mass spectrometry for analysis of multivalent anions. *Anal. Chem.* 74, 6224–6229

36 Soga, T. *et al.* (2002) Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem.* 74, 2233–2239

37 Sato, S. *et al.* (2004) Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant J.* 40, 151–163

38 Gamache, P.H. *et al.* (2004) Metabolomic applications of electrochemistry/mass spectrometry. *J. Am. Soc. Mass Spectrom.* 15, 1717–1726

39 Kristal, B.S. *et al.* (1998) Simultaneous analysis of the majority of low-molecular-weight, redox-active compounds from mitochondria. *Anal. Biochem.* 263, 18–25

40 Kaddurah-Daouk, K. *et al.* (2004) Bioanalytical advances for metabolomics and metabolic profiling. *Pharmagenomics* January 2004, 46–52

41 Josephs, J.L. and Sanders, M. (2004) Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadruople mass spectrometers. *Rapid Commun. Mass Spectrom.* 18, 743–759

42 Huhman, D.V. and Sumner, L.W. (2002) Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry* 59, 347–360

43 Shockcor, J.P. *et al.* (2003) LC-MS/MS approach to 'metabonomics' - What can it do for drug discovery/development? *Drug Metab. Rev.* 35 (suppl. 1), 1–1

44 Yang, L. *et al.* (2002) Investigation of an enhanced resolution triple quadrupole mass spectrometer for high-throughput liquid chromatography/tandem mass spectrometry assays. *Rapid Commun. Mass Spectrom.* 16, 2060–2066

45 Tolstikov, V.V. *et al.* (2003) Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.* 75, 6737–6740

46 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

47 von Roepenack-Lahaye, E. *et al.* (2004) Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* 134, 548–559

48 Dear, G.J. *et al.* (1999) The rapid identification of drug metabolites using capillary liquid chromatography coupled to an ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.* 13, 456–463

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

50 Kenney, B. and Shockcor, J.P. (2003) Metabonomic studies. *Pharmagenomics.* November/December 2003, 56-63

51 Watt, A.P. *et al.* (2003) Metabolite identification in drug discovery. *Curr. Opin. Drug Discov. Devel.* 6, 57–65

52 Corcoran, O. and Spraul, M. (2003) LC-NMR-MS in drug discovery. *Drug Discov. Today* 8, 624–631

53 Edlund, P.O. and Baranczewski, P. (2004) Identification of BVT.2938 metabolites by LC/MS and LC/MS/MS after *in vitro* incubations with liver microsomes and hepatocytes. *J. Pharm. Biomed. Anal.* 34, 1079–1090

54 Liu, D.Q. and Hop, C. (2005) Strategies for characterization of drug metabolites using liquid chromatography-tandem mass spectrometry in conjunction with chemical derivatization and on-line H/D exchange approaches. *J. Pharm. Biomed. Anal.* 37, 1–18

55 Wolfender, J.L. *et al.* (2003) Liquid chromatography with ultraviolet absorbance-mass spectrometric detection and with nuclear magnetic resonance spectroscopy: a powerful combination for the on-line structural investigation of plant metabolites. *J. Chromatogr. A.* 1000, 437–455

56 Orchard, S. *et al.* (2004) Advances in the development of common interchange standards for proteomic data. *Proteomics* 4, 2363–2365

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

58 Baumann, C. *et al.* (2000) A library of atmospheric pressure ionization daughter ion mass spectra based on wideband excitation in an ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.* 14, 349–356

59 Stein, S.E. (1994) Estimating Probabilities of Correct Identification from Results of Mass-Spectral Library Searches. *J. Am. Soc. Mass Spectrom.* 5, 316–323

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

61 Kopka, J. *et al.* (2005) GMD@CSB.DB: the Golm metabolome database. *Bioinformatics* 21, 1635–1638

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

63 Pedrioli, P.G. *et al.* (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* 22, 1459–1466

64 Roessner-Tunali, U. *et al.* (2004) Kinetics of labelling of organic and amino acids in potato tubers by gas chromatography-mass spectrometry following incubation in C-13 labelled isotopes. *Plant J.* 39, 668–679

65 Fischer, E. and Sauer, U. (2003) Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.* 270, 880–891

66 Wahl, A. *et al.* (2004) Serial C-13-based flux analysis of an L-phenylalanine-producing E-coli strain using the sensor reactor. *Biotechnol. Prog.* 20, 706–714

67 Roscher, A. *et al.* (2000) Strategies for metabolic flux analysis in plants using isotope labelling. *J. Biotechnol.* 77, 81–102

68 Bonarius, H.P.J. *et al.* (2001) Metabolic-flux analysis of continuously cultured hybridoma cells using (CO2)-C-13 mass spectrometry in combination with C- 13-lactate nuclear magnetic resonance spectroscopy and metabolite balancing. *Biotechnol. Bioeng.* 74, 528–538

69 Wittmann, C. and Heinzle, E. (2001) Application of MALDI-TOF MS to lysine-producing Corynebacterium glutamicum - A novel approach for metabolic flux analysis. *Eur. J. Biochem.* 268, 2441–2455

70 Arkin, A. *et al.* (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* 149, 1633–1648

71 Rao, C.V. *et al.* (2002) Control, exploitation and tolerance of intracellular noise. *Nature* 420, 231–237

72 Arkin, A. *et al.* (1997) A test case of correlation metric construction of a reaction pathway from measurements. *Science* 277, 1275–1279

73 Vance, W. *et al.* (2002) Determination of causal connectivities of species in reaction networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5816–5821

74 Samoilov, M. *et al.* (2001) On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos* 11, 108–114

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

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

77 Morgenthal, K. *et al.* Metabolomic networks in plants: transitions from pattern recognition to biological interpretation. *Biosystems* (in press)

78 Morgenthal, K. *et al.* (2005) Correlative GC-TOF-MS based metabolite profiling and LC-MS based protein profiling reveal time-related systemic regulation of metabolite-protein networks and improve pattern recognition for multiple biomarker selection. *Metabolomics* 1, 109–121

79 Camacho, D. *et al.* (2005) The origin of correlations in metabolomics data. *Metabolomics* 1, 53–63

80 Weckwerth, W. and Steuer, R. (2005) Metabolomic networks from a systems perspective: from experiment to biological interpretation. In: *Metabolome Analysis: Strategies for Systems biology* (Vaidyanathan *et al.*, eds), pp. 265–289, Springer

81 Albertsen, H. (2000) Genetic profiting and microarray technology. *Journal of Clinical Ligand Assay* 23, 283–292

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

83 Eisenberg, D. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826

84 Stockwell, B.R. (2000) Chemical genetics: Ligand-based discovery of gene function. *Nat. Rev. Genet.* 1, 116–125

85 Stockwell, B.R. (2004) Exploring biology with small organic molecules. *Nature* 432, 846–854

86 Koch, M.A. *et al.* (2003) Protein structure similarity as guiding principle for combinatorial library design. *Biol. Chem.* 384, 1265–1272

87 Austin, C.P. (2003) The completed human genome: implications for chemical biology. *Curr. Opin. Chem. Biol.* 7, 511–515

88 Keith, C.T. *et al.* (2005) Multicomponent therapeutics for networked systems. *Nat. Rev. Drug Discov.* 4, 71–78

89 van der Greef, J. *et al.* (2004) The role of analytical sciences medical systems biology. *Curr. Opin. Chem. Biol.* 8, 559–565

90 Fernie, A.R. *et al.* (2004) Innovation - Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5, 763–769

91 Sharom, J.R. *et al.* (2004) From large networks to small molecules. *Curr. Opin. Chem. Biol.* 8, 81–90

92 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

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

94 Steuer, R. *et al.* (2002) The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* 18(*Suppl. 2*), S231–S240

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

96 Aharoni, A. *et al.* (2002) Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *OMICS* 6, 217–234

97 Hirai, M.Y. *et al.* (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10205–10210

98 Clish, C.B. *et al.* (2004) Integrative biological analysis of the APOE*3-Leiden transgenic mouse. *OMICS* 8, 3–13

99 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58

100 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

101 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

102 Viant, M.R. (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem. Biophys. Res. Commun.* 310, 943–948

103 Scholz, M. *et al.* (2004) Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* 20, 2447–2454

104 Diamantaras, K. and Kung, S. (1996) Principal Component Neural Networks, *Wiley*

105 Chernoff, H. (1973) Use of Faces to Represent Points in K-Dimensional Space Graphically. *J. Am. Stat. Assoc.* 68, 361–368

106 Kleiner, B. and Hartigan, J.A. (1981) Representing Points in Many Dimensions by Trees and Castles. *J. Am. Stat. Assoc.* 76, 260–269

107 Flury, B. and Riedwyl, H. (1981) Graphical Representation of Multivariate Data by Means of Asymmetrical Faces. *J. Am. Stat. Assoc.* 76, 757–76