J Ind Microbiol Biotechnol (2005) 32: 234–252
DOI 10.1007/s10295-005-0231-4

ORIGINAL PAPER

Mariët J. van der Werf · Renger H. Jellema
Thomas Hankemeier

Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets

Received: 7 July 2004 / Accepted: 10 March 2005 / Published online: 14 May 2005

Abstract Microbial production strains are currently improved using a combination of random and targeted approaches. In the case of a targeted approach, potential bottlenecks, feed-back inhibition, and side-routes are removed, and other processes of interest are targeted by overexpressing or knocking-out the gene(s) of interest. To date, the selection of these targets has been based at its best on expert knowledge, but to a large extent also on 'educated guesses' and 'gut feeling'. Therefore, time and thus money is wasted on targets that later prove to be irrelevant or only result in a very minor improvement. Moreover, in current approaches, biological processes that are not known to be involved in the formation of a specific product are overlooked and it is impossible to rank the relative importance of the different targets postulated. Metabolomics, a technology that involves the non-targeted, holistic analysis of the changes in the complete set of metabolites in the cell in response to environmental or cellular changes, in combination with multivariate data analysis (MVDA) tools like principal component discriminant analysis and partial least squares, allow the replacement of current empirical approaches by a scientific approach towards the selection and ranking of targets. In this review, we describe the technological challenges in setting up the novel metabolomics technology and the principle of MVDA algorithms in analyzing biomolecular data sets. In addition to strain improvement, the combined metabolomics and MVDA approach can also be applied to growth medium optimization, predicting the effect of quality differences of different batches of complex media on productivity, the identification of bioactives in complex mixtures, the characterization of mutant strains, the exploration of the production potential of strains, the assignment of functions to orphan genes, the identification of metabolite-dependent regulatory interactions, and many more microbiological issues.

Keywords Metabolomics · Quenching · Holistic analysis · Biostatistics · Multivariate statistics · Metabolic engineering · Strain improvement · Bioprocess optimization

Introduction

The recently introduced functional genomics technologies are revolutionizing research in the biological sciences. Although these technologies were originally set up to be able to elucidate the role of the many genes of unknown function that were identified in the numerous genome sequencing projects, the true value of these technologies lies in the paradigm shift in methodological approaches in the life sciences that they have initiated: from a reductionistic, one-biomolecule-at-a-time, and hypothesis-driven approach towards a holistic and discovery/question-driven approach. This 'genomics' approach, whether aiming at studying gene function or not, allows one to comprehensively understand and to unbiasedly identify and comprehensively understand biomolecules important for specific biological processes.

Three major functional genomics technologies are recognized that study the keystone biomolecules involved in proper functioning of the cell: mRNAs (transcriptomics, DNA-array technology), proteins (proteomics), and metabolites (metabolomics). The transcriptome, proteome, and metabolome are all context-dependent: they vary in response to different environmental conditions and directly reflect the physiological status of a cell. It is this context-dependency that makes them so valuable in understanding biological functioning.

Metabolomics is the most recent addition to the functional genomics toolbox. It involves the non-targeted, holistic analysis of changes in the complete set of metabolites in the cell (the metabolome) in response to

M. J. Werf (✉) · R. H. Jellema · T. Hankemeier
TNO Quality of Life, P.O. Box 360, Zeist,
3700 AJ, The Netherlands
E-mail: vanderWerf@voeding.tno.nl
Tel.: +31-30-6944071
Fax: +31-30-6944466

environmental or cellular changes. Metabolomics is one step further than metabolic profiling: instead of aiming to obtain an inventory of what metabolites are present in the cell, it aims at quantifying every single metabolite present in the cell (Fig. 1 [85]).

Besides the fact that the holistic 'genomics' concept has only recently evolved, the metabolomics approach is only now technically feasible due to the enormous improvements made in the last decade in two other critical areas of research, i.e. analytical chemistry and bioinformatics. There have been enormous improvements in the ability to separate and detect, specifically and sensitively, large numbers of small molecules. Moreover, fast computing and the implementation of computer algorithms in analytical chemistry and biology makes it now possible to process and interpret large sets of biochemical data generated through a non-biased holistic approach.

Metabolites

Metabolites are low-molecular-weight organic compounds (< 1,000 Da) that are participants in general metabolic reactions or are required for the maintenance, growth, and normal functioning of a cell [4]. This includes essential or nutritionally required compounds that are not synthesized de novo. Macromolecular compounds such as proteins, DNA, structural molecules, and polymeric compounds such as glycogen, but also non-native compounds like xenobiotics are not considered to be metabolites [4].

Metabolites are most generally recognized for their role in cellular metabolism and as carriers of energy and reducing equivalents. However, metabolites also fulfill key roles in regulation [88] and as chemosensors [55]. Moreover, many microorganisms produce secondary metabolites like for instance antibiotics and toxins. Metabolites have also been described as stress protectors for a number of different environmental stresses.

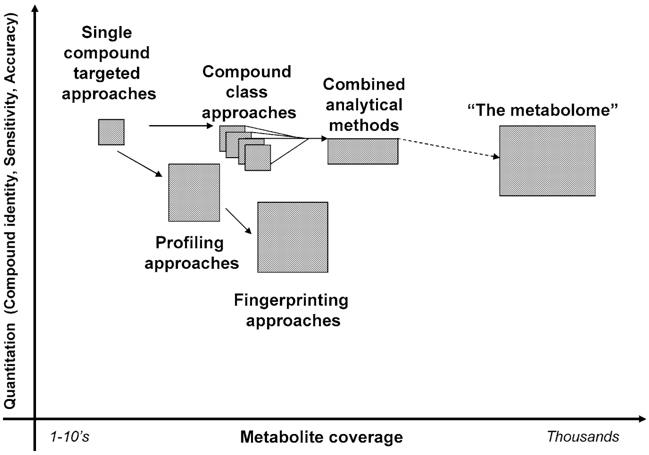

Quantitation (Compound identity, Sensitivity, Accuracy)
Single compound targeted approaches
Compound class approaches
Combined analytical methods
"The metabolome"
Profiling approaches
Fingerprinting approaches
1-10's
Metabolite coverage
Thousands

Fig. 1 Schematic representation indicating the differences between different metabolite analysis approaches (adapted from [85] with permission)

Number of metabolites in the metabolome

The total number of different metabolites that are present in a cell, is yet unknown. In total, almost 20,000 microbial metabolites have been described so far [99]. However, many of these metabolites are secondary metabolites and are only present in relatively few microorganisms. From the recent annotated microbial genome sequences, between 241 and 794 metabolites were deduced to be present in microorganisms [95]. However, it is at this moment impossible to establish the true number of metabolites in the metabolome of a microorganism. This is especially exemplified by the fact that the genome sequencing projects have really taught us how little we still know about cellular functioning: around 40% of the genes present in the microbial genomes are either homologous to genes of unknown function in other organisms or show no homology at all with any of the previously identified genes (i.e. orphan genes [2]). Moreover, an aspect that is largely neglected in genome annotation and subsequent pathway reconstruction studies is the broad substrate-specificity that many of these enzymes have. Schwab [76] stated that the broad substrate-specificity of many enzymes is the major reason why many more metabolites will be present in a metabolome than can be deduced from the genome sequence.

For example, consider the metabolome of *Bacillus subtilis*. The in silico metabolome of this bacterium was deduced from its genome sequence [53] and found to contain 576 different metabolites [95]. In a preliminary experiment, we detected between 300 and 350 different metabolites in *B. subtilis* cells grown in a mineral salts medium, using glucose as the carbon source (unpublished data). From these, 80 are of known identity, which is approximately 25% of the total. This percentage is considerably lower than the 57% of metabolites from the in silico metabolome of *B. subtilis* for which reference compounds can be obtained commercially. Assuming that the ratio between metabolites of known and unknown identity is the same for the complete metabolome of *B. subtilis*, as determined in this experiment, this suggests that in total 1,200–1,400 metabolites are present in the *B. subtilis* metabolome. This is approximately three times more than the 576 deduced from the full genome sequence.

Why metabolomics?

When selecting the functional genomics tool to be used, there are several reasons why metabolomics is the functional genomics technology of choice. First of all, the information that can be derived from the metabolome is very different from that of the genome, transcriptome or proteome, in that each of these levels corresponds to a very different perspective on cellular functioning. The genome can best be understood to represent ‘potential function’, while the transcriptome reflects the ‘functional response’. The proteome and the metabolome together determine the functionality of a cell (Fig. 2).

When going from one biochemical level to the next, information is gained or lost by regulatory events that occur between these levels. For instance, the correlation between the observed changes at the transcriptome and proteome level are poor to moderate ($r^2 = 0.6–0.8$ [21, 24, 33, 39]). Therefore, as the biochemical level of the metabolome is closest to that of the function of a cell (the phenotype; Fig. 2), the study of the metabolome will be the most relevant in order to understand biological functioning. This is especially so as changes in the levels of individual enzymes have, in general, little effect on metabolic fluxes, but do have a significant effect on the concentrations of individual metabolites [30, 75].

Another reason for choosing metabolomics is that once the metabolomics technology platform has been established, it can be applied to any (micro-) organism, in this sense being a truly generic functional genomics platform. This even holds for organisms whose genome has not been sequenced, in this way avoiding large investments in sequencing or constructing microarrays for the microorganism of interest. With metabolomics, an additional advantage of metabolomics is that it can uncover non-genetically modified organisms (GMO) solutions to biological problems. Although increasingly accepted, the use of GMOs is still a problem in foods (e.g. probiotics, fermented foods) or when they need to be released into the environment (e.g. biopesticides, nitrogen-fixing microorganisms).

Notwithstanding the advantages of the metabolomics technology, several challenges remain. Possibly the biggest problem at the moment is that scientists in the life sciences have never been trained to deal with large amounts of data and view these in a holistic manner. On the contrary, they have always been trained to use a completely reductionistic approach [48]. This not only requires the microbiologist to have a ‘holistic’ mind set, but also requires a need to collaborate with scientists from different fields, i.e. especially analytical chemists, statisticians, and informaticians, in order to achieve an optimal approach.

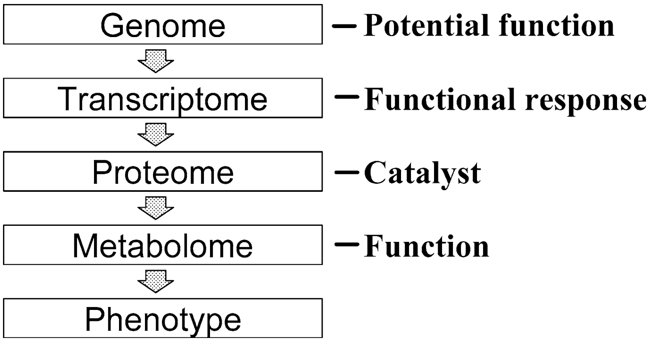


Fig. 2 Biochemical levels in the cell

Furthermore, from a biological point of view, there is a conceptual problem when interpreting metabolomics data. As the relationship between the metabolome and the genome is indirect, how does one decide which of the genes resulting in the formation or degradation of a metabolite identified by metabolomics is the one to knock-out or overexpress?

Also from a more technical point of view, there are still several challenges. Many metabolites, especially signal molecules, are only present in very low concentrations. The sensitivity and dynamic range of analytical instrumentation, when applied in non-target mode, is still not as high as it should be. Another problem is the fact that: (1) there is no commercial software available that allows the automated (pre-) processing of gas chromatography–mass spectrometry (GC-MS) or liquid chromatography–mass spectrometry (LC-MS) files, (2) there is no general commercial reference database available for LC-MS spectra, and (3) the reference database for GC-MS spectra is incomplete. A last challenge relates to the fact that only a limited number of metabolites are commercially available. As many such metabolites are only present in very small concentrations in highly complex mixtures, new analytical chemical approaches need to be developed in order to identify all of these compounds.

Generation of representative biological samples

The key issue in metabolomics is to exploit the information hidden in different metabolome compositions. From a biological point of view, there are several important aspects to be addressed in order to ascertain the collection of representative ‘snapshot’ samples that contain sufficient information/variation.

Experimental design

A first important step when one wants to start generating biological samples is experimental design. In order to improve the information content of the data sets to be generated, thereby improving the accuracy with which relevant parameters can be identified by biostatistics, it is important to use optimal experimental design considerations in advance of the experiments [19]. This starts with a sharp definition of the biological question to be answered. In order to make optimal use of the biostatistical tools, i.e. the statistical ranking of every metabolite measured in relation to the question studied, it is important to move beyond questions like ‘what are the differences?’ to defining exact biological questions that result, preferably, in quantifiable phenotypes, for instance, specific productivity.

Currently, in many transcriptomics and proteomics studies, comparisons are made between results obtained under two different environmental conditions. From the experimental results, in general, the biomolecules showing the largest fold-change are pinpointed as the most relevant genes or proteins for a specific biological process. However, with metabolomics, one can predict beforehand that comparing only two metabolomes does not allow the identification of the most relevant metabolites. For instance, when studying bioproduct improvement, the metabolite that increases the most with increasing productivity is not necessarily the bottleneck. Also, a metabolite strongly decreasing in concentration might be the most relevant for improving product formation, as it might be an inhibitor for one of the enzymes in the biosynthesis pathway. But also, a metabolite that shows only a marginal, but significant, change can be the most important for increased product formation, as it might have a very tight (positive or negative) control over a key biosynthetic enzyme resulting in product formation; and therefore a very small difference in the concentration of such a metabolite may have an enormous effect on product formation. Therefore, it is not the size or the direction of a difference in concentration that identifies a metabolite as being the most relevant for a specific phenotype, but instead, the strength of the correlation of a change in the concentration of a specific metabolite with the variation in the phenotype: i.e. which of the metabolites always correlates with an increased productivity in all the data sets? When setting up an experiment, it is therefore essential to establish the minimal number of metabolomes to be compared that allows a statistically reliable interpretation of the data (see below: ‘Biostatistics’).

Once having established the number of samples one wants to compare, the next issue is to determine how to generate these samples. From pure mathematical considerations, a multifactorial design—that is, one in which different environmental factors are present in many combinations—is likely to be the most informative and thus economical [106]. However, microbiologists in general are not used to applying such a random design. In contrast, today’s microbial production processes are most often established by varying one factor at a time. But which parameter should one change for the best effect? Should one compare different strains producing the same product, should one affect the growth conditions, or should one take samples at different times during growth? Only time will tell which parameters are best varied in order to achieve the largest relevant variations in phenotype.

Reproducible growth of microorganisms

The techniques used for the cultivation of the microorganism may contribute to the overall variation of the experiment. As the biological reproducibility is much less than the analytical reproducibility (for instance in plants, the biological reproducibility is approximately four times lower than the analytical reproducibility [22]) and it is essential to compare many data sets in order to be able to identify metabolites relevant for a biological process of interest (see above: ‘Experimental design’), a prerequisite for metabolomics studies is that microorganisms can be grown in a reproducible manner. Ideally, cells are grown in chemostat cultures, a cultivation technique that gives the possibility to grow microorganisms under constant, carefully controlled conditions [67]. However, industrial fermentations are, in general, fed-batch processes, using complex media. In such complex media, the sequential use of available substrates is likely to occur, which might increase experimental variation, especially when cultures are compared using different sources or batches of such a complex medium component.

Rapid sampling and quenching

In order to obtain representative samples, a sample should be obtained from the culture that is identical to the metabolome of the cells as present when they are harvested. To this end, the metabolic state of the cells as existing in a defined physiological state should be ‘frozen’ (quenched), in order to allow the analysis of the ‘snapshot’ metabolome. Therefore, the (micro-) biological and analytical methods need to be validated, to allow the analysis of samples that are identical to that of the metabolome of the cells when they are harvested, i.e. avoiding biotic or abiotic changes that result in the introduction, conversion, or removal of metabolites/compounds as present in the sample.

Ideally, the time between the cells leaving the culture and their quenching should be zero. Preferably, therefore, rapid sampling techniques should be applied for harvesting the cells [89, 103]. However, especially in view of the rather large number of cells needed for metabolome analysis (see below: ‘Sensitivity’), this might not be possible. In those cases, at least standardized sampling protocol should be used, to ascertain that any changes introduced are similar for all samples taken.

Quenching of metabolism is another critical step when collecting samples. This is especially important, as the turnover of metabolites is even faster than that of mRNAs and proteins. For instance, ATP has a half-life of less than 0.1 s [102]; and therefore the metabolism of the cells should be stopped instantaneously once harvested. In the literature, several quenching methods have been described. These include: (1) rapid filtration through an ultrafiltration membrane followed by immediate freezing of the cells [59, 74], (2), dilution of the cells in perchloric acid [8, 89], (3) dilution of the cells in a methanol solution of −45°C [12, 73], and (4) rapid centrifugation [84]. The cold methanol method seems to be the preferred method for quenching of cells for metabolome studies, as it is a mild method and it allows the concentration of cells (metabolites) by centrifugation.

However, with some microorganisms, for instance *Lactococcus lactis* [41], the application of this method has been report to result in lysis of the cells upon quenching.

Interstitial fluid and compartmentation

After concentrating the cells by centrifugation, the resulting cell pellet consists not only of cellular material (dry weight) and intracellular fluid, but also of fluid that is present between the cells in the pellet (interstitial fluid). This interstitial fluid contains medium components and extracellular metabolites, but after extraction of the (resuspended) cells, a distinction can no longer be made between compounds present intracellularly or extracellularly. For instance, it was established that, for *Saccharomyces cerevisiae*, 50% of the wet weight of the cell pellet was interstitial fluid [94]. Vigorously washing the cells prior to extraction might be a solution to this problem, but disadvantages of this approach include a reduction of the concentrations determined for intracellular metabolites that diffuse freely over the membrane (such as small organic acids) and lysis of the cells during washing. Moreover, in eukaryotic microorganisms, different compartments are present in the cells. As the regulation of transcription, translation, enzyme activities, and other biological processes is only affected by the metabolite concentrations in the direct environment, metabolites should preferably be determined separately in every individual compartment. However, given the relatively large amount of cells needed for full metabolome analyses in view of the sensitivity of the analytical methods and the complexity of the protocols used for separating the different compartments, this is currently not feasible.

Extraction of metabolites from cells

Sample preparation is the step most prone to errors. Sample preparation protocols based on fractionation are not very suitable in metabolomics: the more fractionation steps, the greater the chance that some metabolites will be lost and the less representative the metabolome [97]. Therefore, sample preparation is often minimized for metabolomics studies in order to prevent the loss of individual components. Moreover, sample work-up should be performed under quenched conditions in order to prevent the introduction of changes in the metabolite composition due to residual enzymatic activity present in the samples.

For microbial metabolomics samples, several methods have been described for extracting metabolites from the cells. These include: (1) boiling the cells in an ethanol–buffer solution and subsequent reduction of the volume by evaporation in a rotavapor [29], (2) dilution of the cells in perchloric acid [8, 59], and (3) chloroform extraction at −45°C [12, 73]. Of these methods, chloroform extraction seems to be preferred, as it is a mild method, can be performed under quenched conditions, and does not result in the evaporation of the more volatile metabolites like pyruvate [54].

Analytical chemistry: data acquisition

As said, the key issue in metabolomics is to exploit the information hidden in different metabolome compositions. From an analytical point of view, it is essential that all, or as many as possible, metabolites are being detected with the greatest reliability. To this end, the following issues are important when setting up an analytical metabolomics platform.

Analytical methods

Ideally, the metabolome of the cell is determined selectively in every compartment by (non-invasive) in vivo methods. However, in vivo methods such as NMR and IR are at present not very sensitive and moreover do not separate individual metabolites, therefore requiring the deconvolution of the complex spectra generated. Therefore, most of the holistic analytical metabolomics platforms that are currently being set up rely on invasive techniques. Although the analysis of the first microbial metabolome was achieved by two-dimensional thin layer chromatogtography (TLC) [92], in general more advanced hyphenated techniques like GC-MS, LC-MS, and LC-diode array detection (DAD) are preferred. These methods combine chromatographic procedures for separating metabolites, based upon their physical and chemical properties, coupled with (mass) spectral-based identification of each metabolite.

Hyphenated GC and LC approaches have already been applied for decades in analytical chemistry. However, until recently, the study of metabolites was limited to a handful of compounds at a time which were expected by the researcher to be of particular importance in a given situation—so-called target analysis. In this respect, metabolomics has also introduced a methodological shift in analytical chemistry from target analysis towards holistic analysis.

GC-MS has the advantage of having a high-separation efficiency and providing reproducible retention times, combined with sensitive and selective (electron impact) mass detection. Classically, GC-MS has been applied for the analysis of volatile or medium-polar compounds present in the headspace [70, 101]. However, many metabolites contain polar functional groups that are thermally labile at the temperatures required for their separation or are not volatile at all. In addition, the peak shape of compounds with polar functional groups can be unsatisfactory because of undesirable column interaction, such as irreversible adsorption. Therefore, derivatization of the compounds prior to GC analysis is necessary. In the past decade, derivatization by oximation and subsequent sylilation has proven to be very powerful for the derivatization of alcohol, aldehyde, acid, and amino groups of metabolites, resulting in the analysis of over 200 compounds in a single chromatographic run by GC-MS (Fig. 3 [15, 23, 47]).

In the future, the use of comprehensive GC x GC-MS seems to be very promising [52, 62]. It involves the use of two directly coupled columns, with a cryogenic modulation system at their confluence (cryogenic trap). This allows peaks eluting from the first separation column to be refocused by cryofocusing and subsequently transferred rapidly within a small band by a temperature pulse into a second column with different separation properties. Separation in the second column is generally achieved within a run time of 4–7 s. Especially interesting from a biological perspective is that this approach is expected to also be suitable for the robust and sensitive analysis of individual enantiomers in highly complex biological samples [78].

Notwithstanding the superior resolution, robustness, and dynamic properties of GC, LC approaches are essential for the detection of highly involatile compounds during derivatization or compounds unstable at high temperature, like nucleotides and CoA esters [64], or larger metabolites with a molecular weight above about 500–700 Da. Furthermore, specific detectors can be used, like fluorescence or UV, that are much more sensitive for certain compounds than electron-spray MS generally used for LC-MS. The LC approaches have the additional advantage that derivatization is, in general, not necessary and sample volumes can be only a few microliters or even less. A potentially large problem associated with LC-MS approaches is ion suppression [65], which can jeopardize the determination of the concentration of metabolites. Ion suppression can occur when the response of a compound in the MS detector is suppressed due to the presence of a co-eluting compound that is preferentially ionized. Moreover, as many metabolites are instable compounds at extreme physical conditions, the liquid phase used for the separation of the compounds should be as neutral as possible and the column oven temperature should not be too high. To date only a limited number of holistic LC methods have been reported, using a large number of different approaches [5, 26, 51, 90, 100].

More recently, capillary electrophoresis (CE) approaches were introduced for the holistic analysis of metabolomes [81, 87]. Compared to LC-MS, CE-MS can in some cases be more sensitive and can provide superior separation efficiency; but it is less reproducible with respect to retention time and derivatization, less straightforward to couple to MS detection, and less robust with regards to the presence of salts and peptides.

The CE approaches are a first step towards a future use of laboratory-on-a-chip or micro total analytical systems (μTAS) systems for the analysis of metabolomes [49, 69]. These are methods that have the potential to enable the multiparallel analysis of samples much faster and with higher sensitivity than the currently used methods. However, μTAS systems are currently in their infancy and many different concepts are still being pursued.

Robustness and inertness

Setting up holistic separation methods is, although not yet frequently performed, still not the most challenging job when setting up the analytical platform. The real challenge is to set up an analytical platform that is truly robust and inert so that it allows the quantitative comparison of metabolomes in an automated fashion, the ultimate goal of metabolomics. Therefore, the holistic analytical protocols should be extensively validated with respect to robustness, i.e. variation in retention times and response factors: one has to be absolutely sure that *every* metabolite is detected quantitatively with good precision and accuracy, to be able to detect small differences in concentrations between samples [71]. In

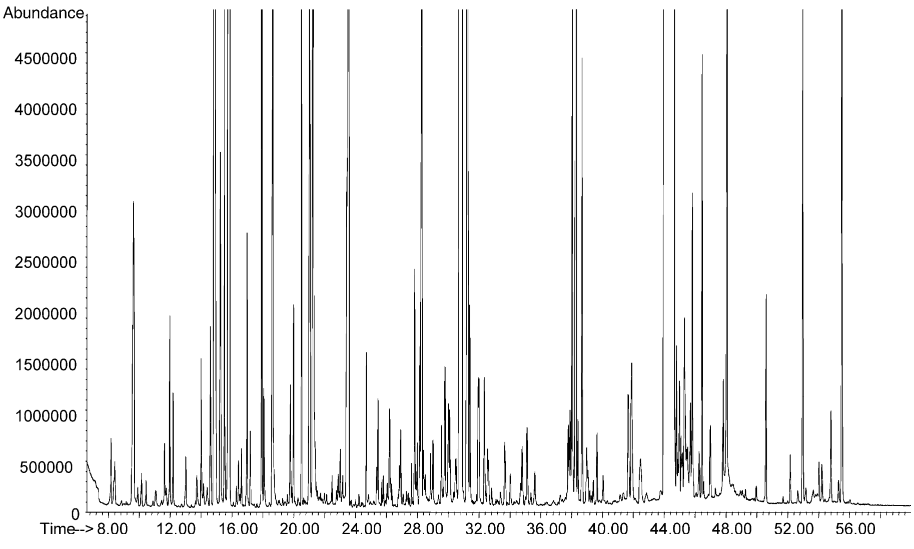


Fig. 3 Full-scan GC-MS chromatogram of a microbial sample derivatized by oximation and subsequent sylilation (unpublished data)

order to control the precision and accuracy of the overall procedure, quality standards should be added at several stages of the sample work-up and analysis procedure to be able to check the variation/efficiency of the different steps. Moreover, one should be absolutely sure that every compound detected is really present in the metabolome and not introduced (either biotically or abiotically) due to changes introduced during analysis.

Sensitivity

As the goal of metabolomics is to analyze *all* metabolites, sensitivity is a highly important aspect. The most commonly used GC-MS instrument for the holistic analysis of biological samples is a quadrupole GC-MS operated in the electron impact (EI) ionization mode. This ionization mode allows the detection of any metabolite eluting from the analytical GC column with comparable response factor in the full-scan acquisition mode rather than selective (and sometimes more sensitive) detection using chemical positive or negative ionization. With GC-EI-quadrupole-MS, typical detection limits are 0.025–2.0 μg metabolite/mL (unpublished data). However, recently much more sensitive instruments have come commercially available, like GC time-of-flight (TOF) MS instruments, that are in general 5–20 times more sensitive than a quadrupole detector. An additional increase in sensitivity can be achieved using comprehensive GC x GC-TOF-MS rather than GC-TOF-MS: due to the smaller peak width obtained in the second separation dimension, an additional five- to ten-fold increase in sensitivity can be gained.

If the screening of a wide range of compounds in full-scan aquisition mode is the aim, ion-trap (Q) LC-MS systems are often used. With a conventional LC set-up, an ion-trap LC-MS system reaches sensitivity comparable to GC-EI-MS using a quadrupole system. Again, more sensitive MS instruments for full-scan screening have recently become available, such as linear ion-trap LC-MS systems, which are approximately 10–20 times more sensitive than conventional ion-trap MS detectors (unpublished data).

However, the overall sensitivity of a method is not only determined by the sensitivity of the analytical instruments available on the market. Samples should also be generated from a large amount of biomass. In addition, the sensitivity can be increased by applying concentration steps like lyophilization and by performing the derivatization reaction in as small a volume as possible.

Quantification

Although when combining metabolomics with multivariate data analysis (MVDA) it is not essential to quantify each metabolite on an absolute scale, the obtained peak areas for the same metabolite in different samples should be comparable with each other on a relative scale. Therefore, the response for the various metabolites should be highly linear. Moreover, to be able to compare the relative concentrations of metabolites analyzed in different samples, the relative standard deviation (RSD) for quantification using these methods should be small. We have obtained RSD values $<10\%$ for both holistic GC-MS and LC-MS methods; and the RSDs were even better (1–4%) for the more stable metabolites (unpublished data). Furthermore, in order to relate these concentrations to the amount of biomass that they were obtained from, it is essential to add internal standards prior to extraction and analysis. In this respect, the dynamic range of analytical methods is important [i.e. the range of concentrations for which (at a high or very low concentration) a change in response as a result of a change in concentration is detected], as is the linear range of the method. This is because the concentrations of the same metabolite in different metabolome samples can vary a great deal, as can the concentrations of different metabolites in the same sample (with a factor of more than 1,000 [97]). The linear range for GC-MS using EI ionization is approximately 10^4–10^5 and that for LC-MS using electrospray ionization is 10^3–10^4.

There are two possible strategies in data analysis: target analysis of a defined number of known (and/or unknown) metabolites or holistic analysis of all the metabolites. With target analysis, the peak areas of a pre-defined list of compounds are determined, and if appropriate response factors are available (via analysis of standards in between or a reference database), quantification is possible. With holistic analysis, a non-biased analysis of all metabolites is carried out without necessarily knowing their identity. The ultimate aim is to generate a peak/metabolite list from every chromatographic run, reporting the concentrations of every metabolite measured by calculating the concentration from the response information of every metabolite as stored in a reference database. For both approaches, a range of quality standards has to be used to control the overall procedure, i.e. extraction, derivatization, and analysis (injection–separation–detection) to correct for possible variation in the response of the detector or injection volume.

Identification of metabolites

For the identification of compounds detected by LC-MS, fractionation of the LC eluent and subsequent nuclear magnetic resonance (NMR) analysis is often applied next to MS/MS experiments. However, as compounds to be identified are often only present in small amounts in very complex mixtures, fractionation and subsequent NMR analysis are not straightforward to apply. Therefore, a combination of LC using MS/MS, MS^n, and high-resolution MS detection to determine the elemental composition is required to identify metabolites. The use of a recently introduced hybrid MS detector such as the linear ion-trap combined with the Fourier transform (FT)-MS detector is a promising system. It should be mentioned that also the use of flow injection into a FT-MS system is reported, but then isomers cannot be separated and therefore not identified and quantified [60, 91].

For the identification of compounds detected by GC-MS, fractionation of peaks is not a straightforward option, although preparative GC is possible. Identification can be achieved using chemical ionization to determine the molecular weight and high-resolution MS detection for the determination of the elemental composition of the molecular ion and characteristic fragments. The MS/MS and MS^n experiments can then be used for further structure elucidation.

Data pre-processing

The data output from most analytical instruments requires significant pre-processing before the differences between the data sets can be analyzed using MVDA tools. In data pre-processing, several aspects should be taken into account and/or corrected:

1. In order to exclude from the biological question irrelevant effects that play an important role in chromatography (such as column changes, temperature differences, differences between columns), the GC data and LC data should be corrected for small shifts in retention time (Fig. 4), drifts in base-line, detector response, and the setting of thresholds with respect to noise.
2. When analyzing complex mixtures like metabolomes, it is common to encounter situations where two or more components elute with a similar retention time [10, 36].
3. Normalization of the data sets by correction for differences in the amount of biomass used to derive the different samples.
4. Data files generated by GC-MS, LC-MS, and LC-DAD are three-dimensional in nature (i.e. chromatographic retention time, spectral information, intensities). The dimension with the spectral information (whether it be MS, fluorescence, or UV spectra) contains highly correlated information. This results in a drastic increase in the likelihood that 'correlations by chance' are identified by the MVDA tools. Therefore, a reduction of dimensions should be achieved, for instance by integrating the different peaks in every mass trace (see Fig. 5).
5. Another critical step before applying MVDA tools is the scaling of the data sets. Scaling approaches are data pre-treatment procedures that allow one to concentrate on the differences between the data sets (e.g. by subtracting the average from all the data), or to ascertain that all variables become equally important (by converting the data into relative responses). Mean scaling and auto-scaling are the most commonly used scaling methods in MVDA.

Currently, the lack of commercially available software for data pre-processing severely hampers the automated (pre-) treatment of the electronic data acquired. Ultimately, deconvolution software is expected to deal with most of these data pre-processing issues. Deconvolution involves the mathematical treatment of analytical data to systematically extract resolved mass spectra for a mixture of components from the raw data. However, the development of deconvolution software, such as AMDIS [34, 82], which allows the holistic, nonbiased analysis of all metabolites whether their identity is known or not, is currently in its infancy.

The ultimate aim is that the combined use of raw data files, data pre-processing tools and a reference

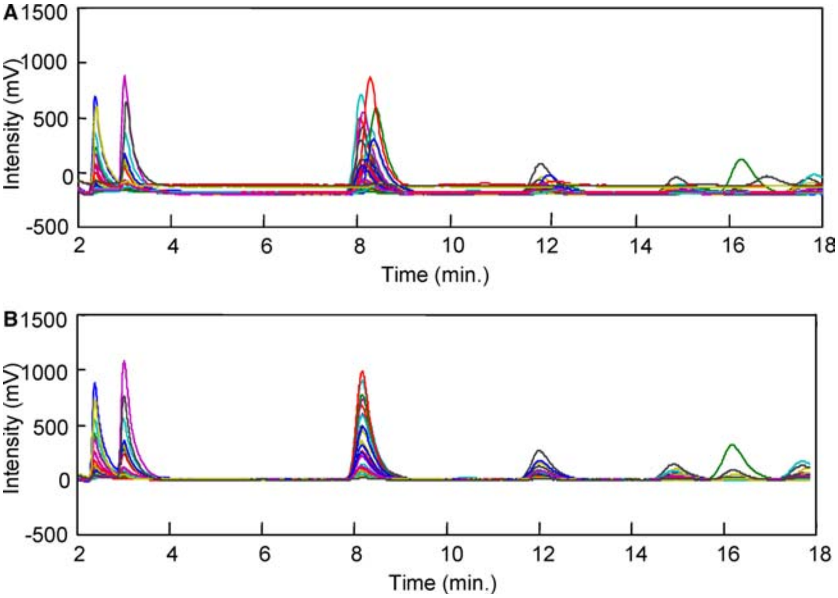


Fig. 4 Effect of time-shift and baseline correction on chromatograms: **a** before correction, **b** after correction

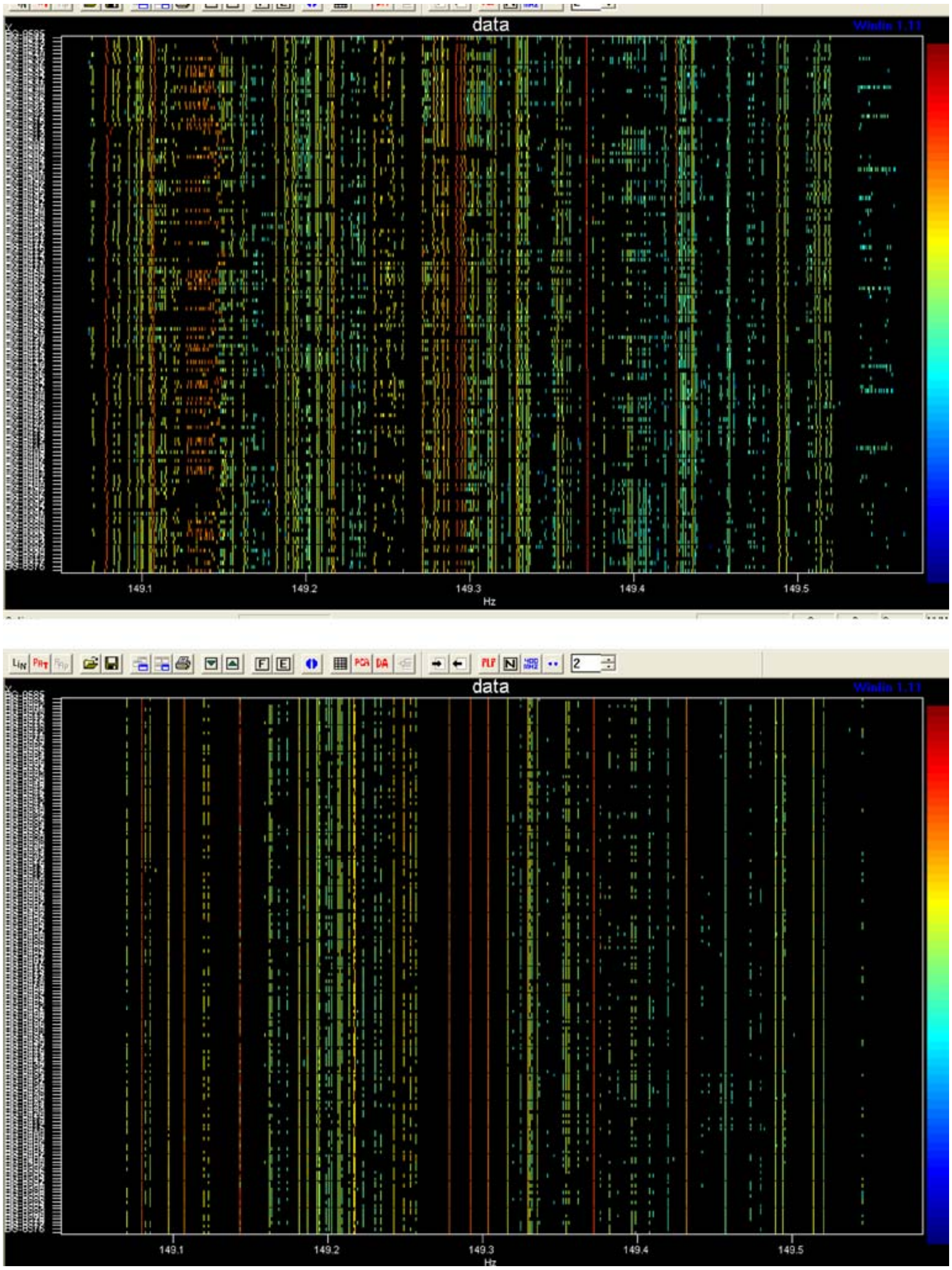


Fig. 5 Plot of a series of samples analyzed by GC-MS before (*upper figure*) and after (*lower figure*) alignment of the retention time and application of an appropriate threshold

database (see above: 'Quantification'), should result in the generation of peak lists that comprise the name of the metabolites, and if metabolites are unknown, unique compound identifiers, quantitative data and quality parameters indicating the reliability with which each peak has been identified and/or quantified. These are the 'clean' data files that are the input for data analysis and biological interpretation, which should also be stored in the data warehouse and should be regularly updated, with information on the identification of compounds that were previously marked as unknown metabolites.

Biostatistics: converting data into information, using MVDA techniques

Central to the metabolomics approach stands the translation of differences in the metabolome compositions into phenotypic differences: the generation of large amounts of data is not the issue, but how to extract information from such large data sets is the crucial topic in metabolomics [98]. This can be achieved by applying unbiased statistical data analysis tools, preferably MVDA ('biostatistics') tools since biological systems are multivariate in nature, i.e. there is an inherent interdependency of the biomolecules [14, 28, 43, 50, 63]. The MVDA (pattern recognition, chemometrics, biometrics) tools are statistical data analysis algorithms that can generate scientific hypotheses by reducing mathematically the many parameters in data sets and visualizing the clustering behavior of parameters.

The MVDA techniques are matrix algebra techniques: i.e. a data set composed of n variables is converted into a vector in n-dimensional hyperspace. For instance, consider a biological sample, A, in which the concentrations of three different metabolites, i.e. x_1, x_2, and x_3, are determined (see Fig. 6). Sample A can be displayed as a vector in a three-dimensional space, in which the axes are the variables, i.e. the metabolites (x_1–x_3). As the concentration of metabolite x_3 is high and the concentrations of x_1 and x_2 are low, the vectors representing sample A point in a direction that is close to the x_3 axis (see Fig. 6). Similarly, two other samples, B and C, which also are high in x_3 and low in x_1 and x_2, can be displayed as vectors in the same three-dimensional space and end up close to the vector describing sample A. Also, sample P, high in x_1 and low in x_2 and x_3, can be displayed in the same three-dimensional space, but its vector will point in a different direction. If only three variables are present, the end-points of these vectors can be represented graphically and it can easily be deduced that samples Q and R contain a high concentration of metabolite x_1 and low concentrations of metabolites x_2 and x_3, as illustrated both by the direction of the vector and by the fact that the end-point of this vector is close to that of sample P, which consists of a similar metabolite mixture (Fig. 6).

Most metabolomes contain over 1,000 different metabolites (variables). Yet, for MVDA it does not matter if a data set contains three or thousands of variables. In general, data sets of n variables are displayed as vectors in an n-dimensional space. However, for the human eye, it is impossible to interpret data sets that are visualized in a multi-hundred or multi-thousand dimensional space. Therefore, it is necessary to project such an n-dimensional space into a two- (or three-) dimensional space.

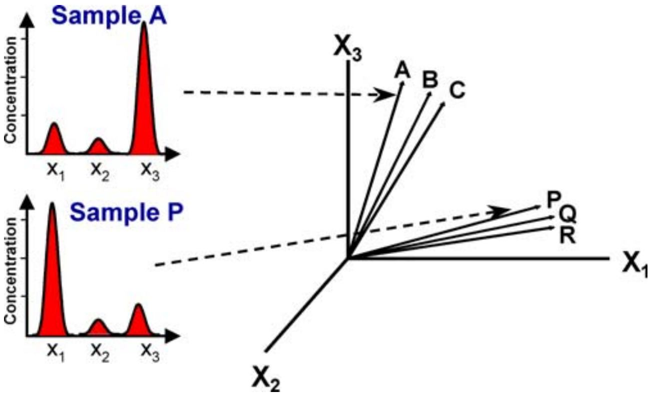


Fig. 6 A graphical representation of the concentrations of three variables (metabolite concentrations: x_1–x_3) in samples A, B, C, P, Q, and R as vectors in a three-dimensional space

It is possible to plot the data sets in two- or three-dimensional plots, usingonly two or three of the original variables (metabolites) of the original data set. However, in the case of a large number of variables, a large number of plots are produced. For this reason, the underlying theme of multivariate analysis is simplification or dimension reduction, using the correlation structure in the data, while retaining as much as possible of the variation present in the data. The idea behind dimension reduction is illustrated in Fig. 7 for principal component analysis (PCA), the most commonly employed MVDA tool. PCA concentrates strongly correlating variables, i.e. variables that vary in a similar way in all data sets, into a new variable. This new variable, a so-called principal component (PC), is a linear combination of the original variables. For instance, x_1–x_3 are the original variables and $0.5{\times}x_1 + 0.6{\times}x_2 + 0.8{\times}x_3$ can be the PC that reflects/displays the correlation between the original variables. The remaining $0.5{\times}x_1 + 0.4{\times}x_2 + 0.2{\times}x_3$ is then either a residue, or can become part of a next PC. In this way, PCA aims at establishing relationships between the m rows (biological samples) and n columns (variables, e.g. metabolites) of a matrix (dimensions $m \times n$). The matrix X is broken down into the product of two smaller matrices and a residual matrix (E_1), as shown in Fig. 7. The two smaller matrices are a row vector and a column vector whose product extracts as large a portion of the variance in the original matrix X as possible. Together, they make up a PC (of the original matrix). The row vector, the $1{\times}n$ matrix in Fig. 7, is known as the loadings matrix and is a common component of all of the metabolomes analyzed. The column vector, the $m{\times}1$ matrix in Fig. 7, is known as the scores matrix and represents the amount of the loadings matrix which is present in each sample. In short, the loadings represent relationships between the variables and the scores those between the samples. Thus, co-varying metabolites can be identified by identifying the variables that are present with a high loading in a PC. Collectively, the loadings and the scores for the first PC represent the first principal component (PC1) and they account for the difference between the variances in X and E_1. This difference is usually expressed as the percentage of the explained variance in X. Subsequently, more PCs that are uncorrelated to each other are derived in a similar manner from the residual matrix (E_1). As each PC is extracted, their value is calculated so as to minimize the value of the residual matrix. This process is continued until a model is obtained in which the number of PCs is considered adequate. The number of relevant PCs established depends on the extent of the correlation between the variables (metabolites). The block diagram depicted in Fig. 7 illustrates this concept for the extraction of the first two components.

If after dimension reduction, a two-dimensional plot is drawn of two PCs (i.e. co-varying metabolites), the similarity of samples can be visualized (see e.g. Fig. 8).

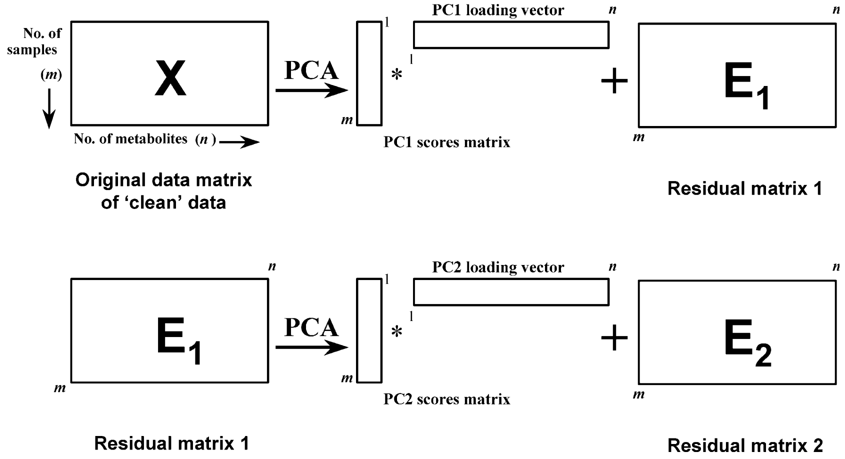


Fig. 7 A block diagram illustrating how the first two principal components are extracted

In this graph, spots (end-points of vectors) represent complete metabolomes. Vectors representing metabolomes that are highly similar will end up close together, while dissimilar metabolomes end up further apart in such a two-dimensional representation. As far fewer plots are necessary to describe a data set with PCs than with the original variables, fewer plots are needed to judge the relations between samples.

When displaying the data in this way, it is possible to identify whether samples are similar or dissimilar, i.e. samples of cells reflecting a similar biological status cluster together (Fig. 8). Moreover, it can easily be judged whether there is significant information in the data sets (metabolomes) that explains the difference between two or more biological states. If such information is present in the data sets (metabolomes), the clusters of samples belonging to the different biological states do not overlap (see e.g. Fig. 8).

After determining that there is information in the data sets that explains the difference between the different biological states, it is also possible to identify the variables (metabolites) that are the most important for these differences. This information can be obtained from a PCA biplot. In a biplot, not only are the end-points of the vectors belonging to the different samples displayed (projected) in a two-dimensional plot spanned by two PCs, but also the original variables are projected in this two-dimensional plot (Fig. 9). From this biplot, the variables that are important for a specific cluster can be identified by selecting those variables (vectors) that point in the direction of the cluster of interest and that after projection result in a long vector, indicating that, in a multidimensional space, this variable vector really points in the direction of the cluster. Also, vectors that point in the opposite direction of the cluster of interest are of importance: these are variables that correlate negatively with the cluster of interest.

Traditionally, MVDA tools are mainly applied as descriptive and/or predictive methods: 'how similar are different data sets' or 'what is the property of an uncharacterized/new data set'. However, much more interestingly, these tools can also be applied as analytical/

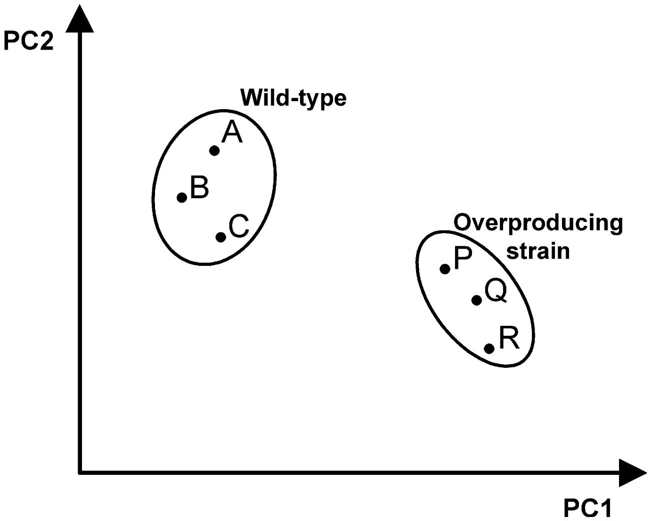


Fig. 8 PCA analysis: visualization of samples A, B, C, P, Q, and R in a two-dimensional space spanned by the first and second PC. Each *dot* in this figure represents the end-point of a vector

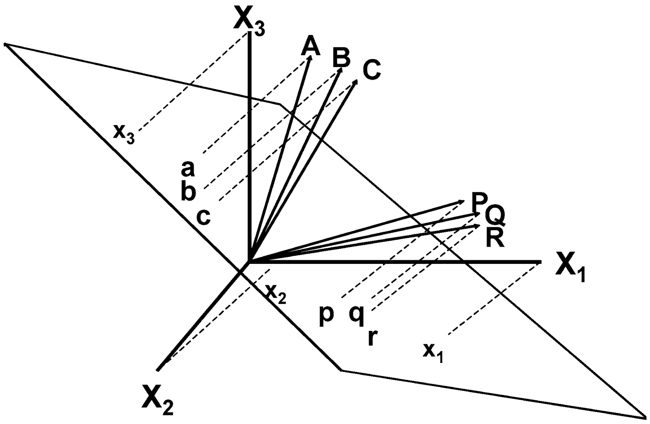


Fig. 9 Biplot: graphical representation of the projection of profiles A, B, C, P, Q, and R and the original variables x_1, x_2, and x_3 onto a plane spanned by two PCs. The *lower case* characters denote the projections

interpretative tools, i.e. to reveal the information embedded in the data: which variables (metabolites) are important for explaining the differences in groups of data [86], which contribute significantly to a specific phenotype, or which co-vary in all the data sets, indicating that they are biologically related and may therefore result in pathway- or metabolite-derived regulatory information. This is key information for further strain or medium improvement programs.

Besides PCA, there are many different MVDA tools that can be divided in two main varieties:

1. Unsupervised methods: these are methods like PCA or hierarchical cluster analysis that visualize relations/patterns in data sets without a priori knowledge.
2. Supervised methods: these include tools like principal component discriminant analysis [37], partial least squares (PLS [27]), or genetic programming [30] that visualize relations/patterns in data sets with a priori knowledge about one or several biological properties of the data sets. For example, information about a specific biological group that a data set belongs to (e.g. wild-type strain vs mutant strain) or a specific quantifiable phenotype that belongs to a specific data set (e.g. specific productivity or yield) is taken into consideration when reducing the dimensions of the data sets.

For identifying the metabolites in data sets that are the most important for a phenotype, the MVDA tool PLS holds especially great promise [42]. PLS is a multivariate regression method that relates the data matrix X to a response or y-variable, like e.g. productivity or yield. As in PCA, in order to reduce the dimension of X, PLS constructs new variables that are linear combinations of the original variables. However, in PCA, the PCs constructed summarize as much of the original information (variation) as possible, irrespective of y-variable information and therefore yields components that are not necessarily predictive/descriptive of the y-variable. In contrast, in PLS, linear combinations of variables are estimated that are highly correlated with respect to the response variable. Thus, the objective criterion for constructing components in PLS is to maximize the *covariance* between the response variable (y) and a linear combination of the original variables (metabolite concentrations; X). Roughly, PLS results in a model (a regression model) that describes a quantifiable variable (phenotype; P) of interest, based on the variables determined (metabolites; x_1, x_2, x_3):

$$[\mathrm{P}] = a_1x_1 + a_2x_2 + a_3x_3 + \cdots$$

By ordering the (relative) statistical importance of the metabolites by virtue of the weight factors (regression factors; a_1, a_2, a_3) estimated by PLS for these metabolites, variables that contribute most to a quantifiable phenotype of interest can be identified. In other words, by applying PLS, targets for metabolic engineering can be identified and ranked, based on their importance in relation to the question under study. By applying PLS, we have successfully identified compounds correlating with productivity (Fig. 10; unpublished data).

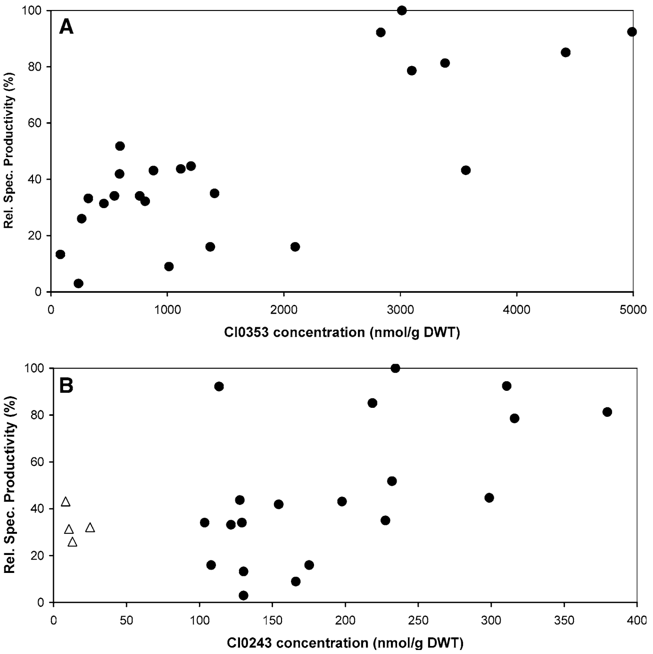


Fig. 10 Correlation between the metabolite concentration in cells and the relative specific productivity of compounds identified by PLS to be important for product formation (unpublished data). **a** Overall correlation and **b** concentration of the relevant compound in the wild type (*filled triangles*) and an overproducing strain obtained by classic mutagenesis (*filled circles*)

For further reading, the reader is referred to [14, 27].

Bioinformatics

Metabolomics, like any other functional genomics technology, generates enormous amounts of data. The generation, storage, analysis, and interpretation of these data requires a bioinformatics infrastructure in order to be able to handle all these data. Such an infrastructure consists of several modules (Fig. 11).

Laboratory information management system

A laboratory information management system (LIMS) is essential to track down how a specific sample was generated (i.e. microorganism source, medium, growth conditions, and so on) and the protocols with which it was collected, extracted, analyzed, and pre-processed [40]. This is essential to be able to judge the quality of data sets when analyzing and interpreting the data (see ‘Biostatistics’ and ‘Biological interpretation’) or when reanalyzing data sets in the future.

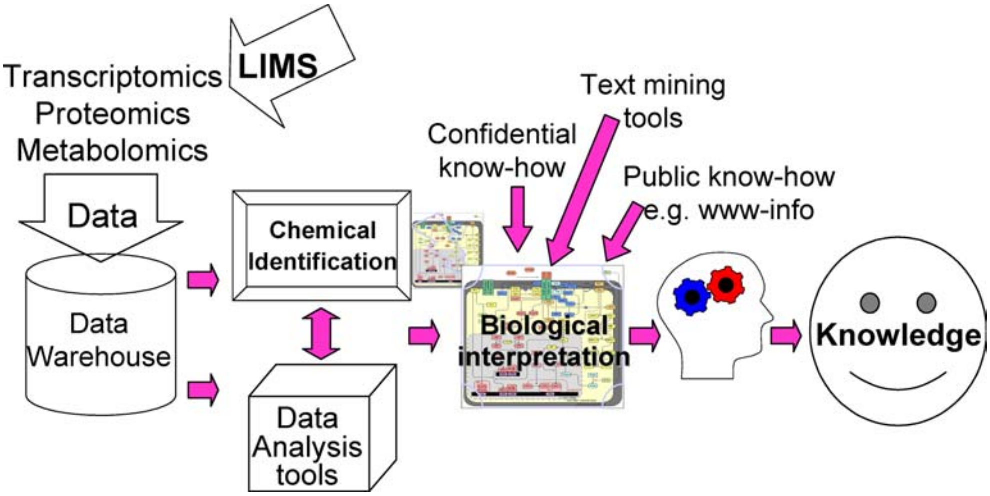


Fig. 11 Schematic presentation of the data warehousing and data interpretation infrastructure for metabolomics (and other functional genomics technologies). *LIMS* Laboratory information management system

Data warehousing

The data files generated by the different analytical methods are very large. For instance, one single LC-MS file is 40 Mb. Not all data in such a file are information-containing data. After data pre-processing (see above), smaller, 'clean' data files can be obtained that still contain all the relevant information. The raw data and processed data need to be stored in a data-warehousing infrastructure in order to allow their future re-analysis and/or re-interpretation [35]. Preferably, such an infrastructure also allows the storage of genomic, transcriptomics, and proteomics data, essential to allow the (future) integration of these data sets into a systems biology approach. Although there is data-warehousing software commercially available, a structure for storing biomolecule data sets is not provided with it, nor is there currently consensus about how to best bring structure in such data. Therefore, at many different places, scientists are independently discovering both what is the best method and which data to store.

Reference database

One of the key challenges in metabolomics will be the identification of the metabolites detected. For instance, of the 576 metabolites of the in silico metabolome of *B. subtilis*, 43% are not commercially available (unpublished data), because they are either inherently instable, or not interesting enough to allow their commercial production. Even more so, the percentage of metabolites for which no reference compounds are available is much larger when analyzing cell extracts of this microorganism. Therefore, an enormous effort is required to identify all these unknowns (see above: 'Identification of metabolites'). Such effort put into the identification of specific metabolites should not be wasted. Retention index, mass spectral, response, and other relevant information, such as synonyms, molecular weight, etc, of the newly identified compounds should be stored in a reference database, to facilitate the identification of the same metabolite in the future.

One such a reference database is commercially available for compounds analyzed by GC-MS (NIST database; http://webbook.nist.gov/chemistry/), but for LC-MS data no such reference database is available. Preferably, however, such a metabolomics reference database should be independent of the analytical methods used, to also be able to correct for compounds analyzed by more than one of the analytical methods that comprise the metabolomics platform.

Biostatistics toolbox

Once 'clean' data sets have been generated, the data should be transferred into information, using data analysis tools. Many MVDA tools are available via commercial software packages such as Matlab (The Mathworks), but data analysis tools developed by oneself, or not commercially available, should also be made accessible via the bioinformatics infrastructure (Fig. 11).

Biological interpretation

After having identified specific [set(s) of] metabolites as being important for the question under study, the issue is to obtain the maximum information about the biological function of these metabolites, preferably as quickly as possible, in the context of the question under study. Although most microbiologists have at their disposal a good knowledge of microbial physiology, one cannot expect that everybody knows everything about the role of every metabolite known in every microorganism in each biological context. Therefore, tools are needed to assist the microbiologist in the interpretation of the potential meaning of the metabolites identified by biostatistics as being relevant for a specific biological question. The next sections describe such tools.

Metabolic pathway diagrams

Visualization of the metabolites identified in metabolic pathways can be an enormous aid in understanding the biological relevance of the identified metabolites in relation to the question under study. Several metabolic pathways are available via the Internet, such as KEGG [44], EcoCyc [46], UM-BBD [17], and WIT [66]. However, these pathways are generally more aimed at being as complete as possible, than as an aid to the biological interpretation of results; and they are, therefore, not always very helpful in the biological interpretation of the results (for instance compare Fig. 12a, b). Preferably, metabolic diagrams should be drawn in such a way that the biological meaning of an identified metabolite becomes immediately apparent, i.e. why is it logical for the cell to increase or decrease the level of a certain metabolite in relation to the question addressed. This involves drawing the diagrams in a way that biologists are familiar with, but essential information in relation to reducing equivalents, energy, and cofactors required or consumed in specific steps should preferably also be visualized in such diagrams. Moreover, pathway diagrams that contain a lot of reactions are generally not very clear and tend to confuse the issue. Therefore, in contrast to what is stated by Li et al. [56], bioinformatics tools that highlight all metabolites that increase or decrease in concentration will only be of limited value for biological interpretation.

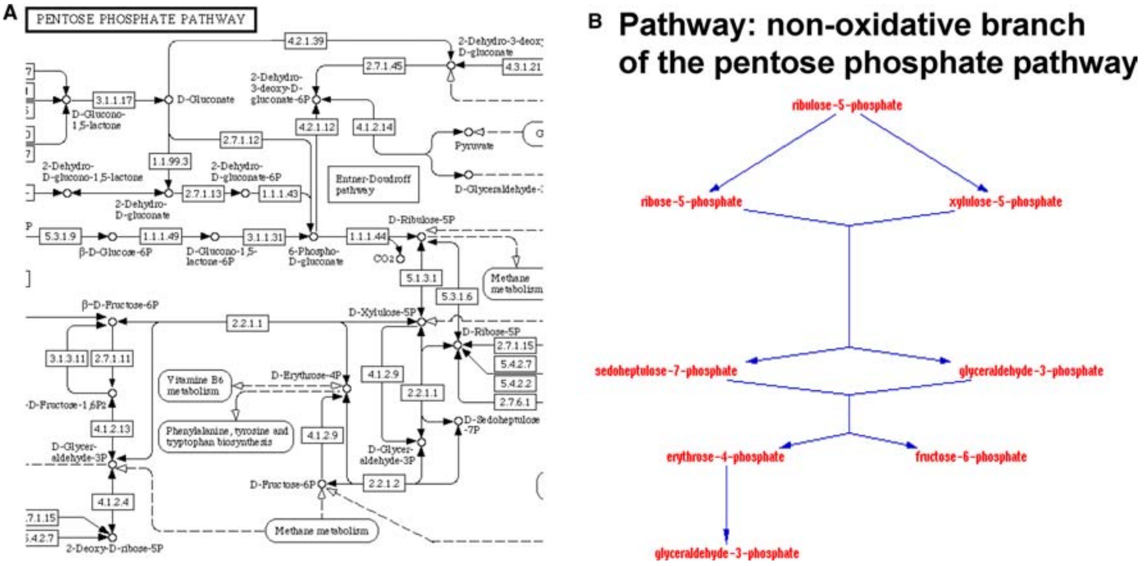
Another risk with the graphical representation of metabolic pathways as presented in databases is that they are incomplete. For instance, *Propionibacterium shermanii* uses a modified citric acid cycle [3] which is not present in the KEGG database. This is especially a problem as metabolic routes are highly variable amongst different microorganisms.

In this respect, the notion about what is a metabolic pathway, especially in view of the fact that they all interact and that there is redundancy, has also to be taken into account [56]. Metabolic pathways usually ignore side-reactions and the interactions of co-substrates. Therefore, the term ‘metabolic neighborhoods’ was introduced and is defined as ‘the set consisting of a central metabolite, all the reactions that include it as a substrate or product, plus all metabolites that take part in those reactions’ [56]. Metabolic neighborhood representations allow local views of all pathways surrounding a specific metabolite of interest, in contrast to most pathway maps present in databases that represent only part of the network [56].

Besides graphical representations of biochemical pathways, metabolic models have been made for several microorganisms [83]. Two types of models can be discriminated. For metabolic control analysis, models consisting of 30–40 reactions are made that describe the pathways between substrate degradation and product formation [11, 105]. More recently, constraints-based models using the full genome sequence of for instance *Escherichia coli* and *S. cerevisiae* have been made [16, 20]. These models consist of all biochemical reactions for which corresponding genes have been identified in the annotated genome and contain over 700 reactions and 400 metabolites. Potentially, these models also allow the feeding of information obtained from metabolome analysis. Unfortunately, however, in many instances different simulation/analysis tools were used, making the exchange and coupling of different biochemical networks impossible and/or very laborious. Therefore, recently, a systems biology markup language was developed that is independent of the software used and that allows the sharing, evaluation, and cooperative development of models [38].

Currently also, tools are being developed that allow the reconstruction of the metabolic network and regulatory interactions, based on the metabolite data

Fig. 12 Representation of the pentose phosphate pathway in: **a** the KEGG database (http://www.genome.ad.jp/kegg/) and **b** BioCyc (http://biocyc.org)



obtained in the experiment [1, 72]. The basis for this approach is that metabolites that correlate (positively or negatively) to each other in all metabolomes are expected to be metabolically or regulatory related to each other, while non-correlated metabolites are expected to be more distantly related.

Databases and scientific literature

In order to summarize/obtain an overview of all reported scientific findings, many databases (generally accessible via the Internet) have appeared and are currently still appearing. One of the largest problems with these databases is their contamination by incorrect information [57]. Even more so, there is a considerable problem in the fact that findings which are identified as hypothetical or putative the first time around seem to become more and more proven, when referred to more distantly. Therefore, it is important that one is able to check the merits of findings reported in such databases, by cross-checking with the experimental data as described in the original publication(s). This is of especial importance as the percentage of information built upon theoretical grounds seems to increase much faster than the percentage of genes/biomolecules whose function has been verified by experimentation.

In this respect, the trend that an increasing number of scientific publications are directly accessible via the Internet as PDF files will be an enormous aid, especially in view of the fact that, inherent to the holistic nature of metabolomics, many targets will pop up whose virtue has to be evaluated. The PDF file-availability of scientific publications via the Internet will both reduce the risk that hypotheses are built upon unreliable information and save a lot of time because they are available online. However, not withstanding the risks of databases, they can be very handy in a first scan about the potential biological role of any selected metabolite.

Text-mining tools

For biological interpretation, not only is the (rapid) availability of the original scientific publication(s) an issue, but also the selection of the most important publications. One would not like to think about the pile of literature one would find when glucose pops up as the most relevant metabolite for a specific biological question. Therefore, text-mining tools that allow context-based searching in e.g. abstracts of scientific manuscripts will be an enormous aid in selecting the most relevant literature only [77].

Metabolite function databases

Currently, several metabolite databases, such as Compound KB [45] and LIGAND [32], are available via the Internet. These databases contain all kinds of information about the chemical properties of such compounds, but so far nothing has been reported in these databases about the biological role these metabolites fulfill in cells. Generally, a link with metabolic pathway database(s) is present, but there is no list of other (organism-specific) roles that have been described for the metabolites (i.e. compatible solute, allosteric affector of a specific enzyme, signaling molecule, and so on).

Microbial metabolomics: state-of-the-art

Metabolomics is currently still in its infancy. So far, it has primarily been used in the biomedical area and recently the first examples for plants were reported [22, 72].

In microbiology, the term metabolomics is not only used to describe comparative comprehensive metabolite studies, but has also been used for dynamic metabolic flux modeling studies [8, 25]. Few examples of microbial metabolomics studies aimed at the non-biased comprehensive study of metabolites have been reported. One of the first papers of what could be described as metabolomics is a study applying three different GC-MS methods for the analysis of fatty acids, amino acids, and carbohydrates in combination with the MVDA tool SIMCA, in order to monitor microbial contamination during the fermentative production of dextran by *Leuconostoc mesenteroides* [18]. The metabolome of *E. coli* has been determined using two-dimensional TLC by Ferenci and co-workers [58, 92, 93]. They relied on the visual inspection of differences and made no effort to identify most of the metabolites. However, using this approach, unexpectedly elevated valine pools were identified under oxidative stress conditions, a metabolite not previously implicated in this biological process [93]. It was speculated that, because of valine's tertiary carbon atom, it could function in 'mopping up' reactive oxygen species. This example illustrates the strength of the metabolomics approach for the identification of metabolites not previously implicated as being important for a specific biological process. Also, a metabolomic study of *S. cerevisiae* was performed, applying NMR and enzymatic detection methods for the detection of a few specific metabolites in combination with MVDA [68]. The differences in concentrations of some of the intracellular metabolites were used to identify the phenotypes of several knock-out mutants.

The groups led by Nishioka and Terabe have in recent years actively published on metabolome studies of *B. subtilis* [6, 7, 61, 80, 81, 87]. Recently also, a study on the metabolome of *Corynebacterium glutamicum* was reported [84]. With both microorganisms, however, due to the relatively poor quality of the biological set-up of the experiments used by these analytical chemists (see above: 'Generation of representative biological samples'), the value of these studies for biology is difficult to judge. The group led by Oliver reported preliminary results relating to a protocol that they are validating for metabolome analysis of *S. cerevisiae* [9]. They applied the protocol set up for the study of a yeast mutant that

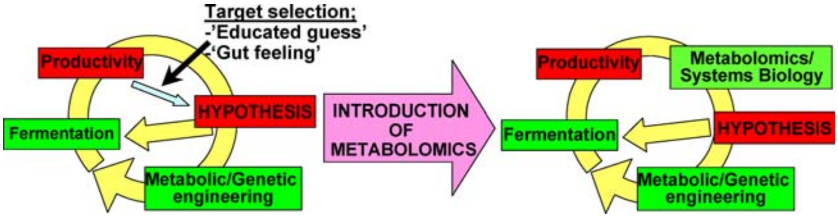


Fig. 13 The metabolomics approach integrated in a bioprocess-optimization cycle

did not show any apparent change in phenotypic characteristics and observed clear differences in the metabolic profiles in comparison with the wild type. The use of metabolic profiling approaches for the authentication of strains [13, 31, 79] and the detection of microbial cross-contamination [18, 104] has been reported before. Also, van Dam et al. [96] reported preliminary metabolome analysis results with *S. cerevisiae*. They reported almost identical concentrations for metabolites in cells obtained from six independently grown steady-state fermentations; and the RSD was only one- to two-fold higher than the RSD obtained from replicate analysis of the same sample.

The impact of metabolomics on industrial microbiology: conclusions and future outlook

In remarkable contrast to the relevance of the metabolome for understanding cellular functioning, the development of technologies that allow the comprehensive analysis of 'snapshot' metabolomes has only just begun. Due to the complexity of this technology, it will likely take a few more years before fully operational and automated metabolomics platforms are available that allow the quantitative analysis and identification of every single metabolite present in a metabolome.

A complicating factor in applying metabolomics, as in any other functional genomics technology, is the unfamiliarity of microbiologists and cell biologists with MVDA tools like PCA and PLS, essential for converting data into information. This is further complicated by the highly inaccessible way in which the relatively few papers in which MVDA tools have been applied in industrial microbiology have been written. Due to this, the huge potential of MVDA tools in cellular biology has yet to become apparent.

Undoubtly, metabolomics will have a strong impact on industrial microbiology in the coming decades. By applying this technology, trial-and-error-based approaches for target selection can be replaced by a scientific way towards not only the selection but also the ranking of targets/leads that are the basis for further improvement of production strains or process conditions [98]. This will not only result in a reduction of research and development time, and thus money, by wasting less time on targets that later prove to be irrelevant or only result in a very minor improvement, but also in a shortening of time-to-market, as greater improvements will be achieved in every cycle of bioprocess optimization (Fig. 13). In addition, new (unexpected) insights will be gained in cellular functioning and the regulation of cellular processes, especially when integrating metabolomics with transcriptomics and/or proteomics into a systems biology approach.

References

1. Arkin A, Shen P, Ross J (1997) A test case of correlation metric construction of a reaction pathway from measurements. Science 277:1275–1279
2. Bailey JE (1999) Lessons from metabolic engineering for functional genomics and drug discovery. Nat Biotechnol 17:616–618
3. Beck S, Schink B (1995) Acetate oxidation through a modified citric acid cycle in *Propionibacterium freudenreichii*. Arch Microbiol 163:182–187
4. Beecher CWW (2003) The human metabolome. In: Harrigan GG, Goodacre R (eds) Metabolic profiling: its role in biomarker discovery and gene function analysis. Kluwer, Boston, pp 311–319
5. Bhattacharya M, Fuhrman L, Ingram A, Nickerson KW, Conway T (1995) Single-run separation and detection of multiple metabolic intermediates by anion-exchange high-performance liquid chromatography and application to cell pool extracts prepared from *Escherichia coli*. Anal Biochem 232:98–106
6. Britz-Mckibbin P, Terabe S (2002) High-sensitivity analyses of metabolites in biological samples by capillary electrophoresis using dynamic pH junction-sweeping. Jpn Chem J Forum 2:397–404
7. Britz-McKibbin P, Nishioka T, Terabe S (2003) Sensitive and high-throughput analysis of purine metabolites by dynamic pH Junction multiplexed capillary electrophoresis: a new tool for metabolomics studies. Anal Sci 19:99–104
8. Bucholz A, Hurlebaus J, Wandrey C, Takors R (2002) Metabolomics: quantification of intracellular metabolite dynamics. Biomol Eng 19:5–15
9. Castrillo JI, Hayes A, Mohammed S, Gaskell SJ, Oliver SG (2003) An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. Phytochemistry 62:929–937
10. Colby BN (1992) Spectral deconvolution for overlapping GC/MS components. J Am Soc Mass Spectrom 3:558–562
11. Dauner M, Sonderegger M, Hochuli M, Szyperski T, Withrich K, Hohmann H-P, Sauer U, Bailey JE (2002) Intracellualr carbon fluxes in riboflavin-producing *Bacillus subtilis* during growth on two-carbon substrate mixtures. Appl Environ Microbiol 68:1760–1771
12. Koning W de, Dam K van (1992) A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. Anal Biochem 204:118–123
13. Nijs M de, Larsen JS, Gams W, Rombouts FM, Wernars K, Thrane Ul, Notermans SHW (1997) Variations in random amplified polymorphic DNA patterns and secondary metabolite profiles within *Fusarium* species from cereals from various parts of the Netherlands. Food Microbiol 14:449Y–457Y

14. Dillon WR, Goldstein M (1984) Multivariate analysis, methods and applications. Wiley, New York
15. Duez P, Kumps A, Mardens Y (1996) GC-MS profiling of urinary organic acids evaluated as a quantitative method. Clin Chem 42:1609–1615
16. Edwards JS, Palsson BO (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97:5528–5533
17. Ellis LBM, Hershberger CD, Wackett LP (1999) The university of minnesota biocatalysis/biodegradation database: specialized metabolism for functional genomics. Nucleic Acids Res 27:373–376
18. Elmroth I, Sundin P, Valeur A, Larsson L, Odham G (1992) Evaluation of chromatographic methods for the detection of bacterial contamination in biotechnical processes. J Microbiol Methods 15:215–228
19. Faller D, Klingmuller U, Timmer J (2003) Simulation methods for optimal experimental design in systems biology. Simulation 79:717–725
20. Famili I, Forster J, Nielsen J, Palsson BO (2003) *Saccharomyces cerevisiae* phentotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proc Natl Acad Sci USA 100:13134–13139
21. Fell DA (2001) Beyond genomics. Trends Genet 17:680–682
22. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000a) Metabolite profiling for plant functional genomics. Nat Biotechnol 18:1157–1161
23. Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000b) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal Chem 72:3573–3580
24. Flutcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999) A sampling of the yeast proteome. Mol Cell Biol 19:7357–7368
25. Förster J, Gombert AK, Nielsen J (2002) A functional genomics approach using metabolomics and *In Silico* pathway analysis. Biotechnol Bioeng 79:703–712
26. Fraser PD, Pinto MES, Holloway DE, Bramley PM (2000) Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. Plant J 24:551–558
27. Geladi P, Kowalski BR (1986) Partial least squares regression: a tutorial. Anal Chim Acta 185:1–17
28. Glassey J, Montague G, Mohan P (2000) Issues in the development of an industrial bioprocess advisory system. Trends Biotechnol 18:136–141
29. Gonzalez B, Francois J, Renaud M (1997) A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol. Yeast 13:1347–1356
30. Goodacre R, Kell DB (2003) Evolutionary computation for the interpretation of metabolomic data. In: Harrigan GG, Goodacre R (eds) Metabolic profiling: its role in biomarker discovery and gene function analysis. Kluwer, Boston, pp 239–256
31. Goodacre R, Rischert DJ, Evans PM, Kell DB (1996) Rapid authentication of animal cell lines using pyrolysis masss spectrometry and auto-associative artificial neural networks. Cytotechnology 21:231–241
32. Goto S, Oluno Y, Hattori M, Nishioka T, Kanehisa M (2002) LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 30:402–404
33. Gygi SP, Rochon Y, Pranza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730
34. Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, Chalmers RA (1999) Deconvoultion gas chromatogrphy/mass spectrometry of urinary organic acids—potential for pattern recognition and automated indentification of metabolic disorders. Rapid Commun Mass Spectrom 13:279–284
35. Hardy N, Fuell H (2003) Databases, data modelling and schemas. In: Harrigan GG, Goodacre R (eds) Metabolic profiling: its role in biomarker discovery and gene function analysis. Kluwer, Boston, pp 277–291
36. Herron NR, Donnelly JR, Sovocool GW (1996) Software-based mass spectral enhancement to remove interferences from spectra of unknowns. J Am Soc Mass Spectrom 7:598–604
37. Hoogerbrugge R, Willig SJ, Kistemaker PG (1983) Discriminant analysis by double stage principal component analysis. Anal Chem 55:1710–1712
38. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531
39. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–934
40. Jenkins H, Hardy N, Beckmann M, et al (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol 22:1601–1606
41. Jensen NBS, Jokumsen KV, Villadsen J (1999) Determination of the phosphorylated sugars of the Embden-Meyerhoff-Parnass pathway in *Lactococcus lactis* using a fast sampling technique and solid phase extraction. Biotechnol Bioeng 63:356–362
42. Johansson D, Lindgren P, Berglund A (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. Bioinformatics 19:467–473
43. Joliffe IT (1986) Principal component analysis. Springer, Berlin Heidelberg New York
44. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The Kegg databases at GenomeNet. Nucleic Acids Res 30:42–46
45. Karp PD (1992) A knowledge base of the chemical compounds of intermediary metabolism. Comput Appl Biosci 8:347–357
46. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrinie-Toole A, Bonavides C, Gama-Castro S (2002) The EcoCyc database. Nucleic Acids Res 30:56–58
47. Katona ZsF, Sass P, Molnar-Perl I (1999) Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography-mass spectrometry. J Chromatogr A 847:91–102
48. Kell DB, Oliver SG (2003) 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
49. Khandurina J, Guttman A (2002) Bioanalysis in microfluidic devices. J Chromatogr A 943:159–183
50. Kohonen T (1995) Self-organizing maps. Springer, Berlin Heidelberg New York
51. Kristal BS, Vigneau-Callahan KE, Matson WR (1998) Simultaneous analysis of the majority of low-molecular-weight, redox-active compounds from mitochondria. Anal Biochem 263:18–25
52. Kueh AJ, Marriott PJ, Wynne PM, Vine JH (2003) Application of comprehensive two-dimensional gas chromatography to drugs analysis in doping control. J Chromatogr A 1000:109–124
53. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G et al (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390:249–266
54. Lange HC, Eman M, van Zuijlen G, Visser D, van Dam JC, Frank J, Teixeira de Mattos MJ, Heijnen JJ (2001) Improved rapid sampling for in vivo kinetics of intracellular metabolites in *Saccharomyces cerevisiae*. Biotechnol Bioeng 75:406–415
55. Lengeler JW (2000) Metabolic networks: a signal-oriented approach to cellular models. Biol Chem 381:911–920

56. Li XJ, Brazhnik Ol, Kamal A, Guo D, Lee C, Hoops S, Mendes P (2003) Databases adnd visualization for metabolomics. In: Harrigan GG, Goodacre R (eds) Metabolic profiling: its role in biomarker discovery and gene function analysis. Kluwer, Boston, pp 293–309
57. Linial M (2003) How incorrect annotations evolve—the case of short ORFs. Trends Biotechnol 21:298–300
58. Liu X, Ng C, Ferenci T (2000) Global adaptations resulting from high population densities in *Escherichia coli* cultures. J Bacteriol 182:4158–4164
59. Lowry OH, Carter J, Ward JB, Glaser L (1971) The effect of carbon and nitrogen sources on the level of metabolic intermediates in *Escherichia coli*. J Biol Chem 246:6511–6521
60. Mano N, Goto J (2003) Biomedical and biological mass spectrometry. Anal Sci 19:3–14
61. Markuszewski MJ, Britz-McKibbin P, Terabe S, Matsuda K, Nishioka T (2003) Determination of pyridine and adenine nucleotide metabolites in *Bacillus subtilis* cell extract by sweeping borate complexation capillary electrophoresis. J Chromatogr A 989:293–301
62. Marriott P, Shellie R, Fergeus J, Ong R, Morrison P (2000) High resolution essential oil analysis by using comprehensive gas chromatographic methodology. Flav Fragr J 15:225–239
63. Martens H, Naes T (1989) Multivariate calibration. Wiley, Chichester
64. McCloskey JA (1990) Constituents of nucleic acids: overview and strategy. Methods Enzymol 193:771–781
65. Nelson MD, Dolan JW (2002) Ion suppression in LC-MS-MS: a case study. LC GC Eur 2002:73–79
66. Overbeek R, Larsen N, Pusch GD, D'Souza M, Slekov E, Kyrpides N, Fonstein M, Maltsev N, Selkov E (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 28:123–125
67. Piper MDW, Daran-Lapujade P, Bro C, Regenberg B, Knudsen S, Nielsen J, Pronk JT (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. J Biol Chem 277:37001–37008
68. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, Dam K van, Oliver SG (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19:45–50
69. Regnier FE, He B, Lin S, Busse J (1999) Chromatography and electrophoresis on chips: critical elements of future integrated, microfluidic analytical systems for life science. Trends Biotechnol 17:101–106
70. Rhodes G, Miller M, McConnel ML, Novotny M (1981) Metabolic abnormalities associated with diabetes mellitus, as investigated by gas chromatography and pattern-recognition analysis of profiles of volatile metabolites. Anal Chem 27:580–585
71. Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J 23:131–142
72. Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13:11–29
73. Ruijter GJG, Visser J (1996) Determination of intermediary metabolites in *Aspergillus niger*. J Microbiol Methods 25:295–302
74. Saez MJ, Lagunas R (1976) Determination of intermediary metabolites in yeast. Critical examination of the effect of sampling conditions and recommendations for obtaining true levels. Mol Cell Biochem 13:73–78
75. Sanford K, Soucaille P, Whited G, Chotani G (2002) Genomics to fluxomics and physiomics—pathway engineering. Curr Opin Microbiol 5:318–322
76. Schwab W (2003) Metabolome diversity: too few genes, too many metabolites? Phytochemistry 62:837–849
77. Shatkay H, Feldman R (2003) Mining the biomedial literature in the genomic era: an overview. J Comput Biol 10:821–855
78. Shellie R, Marriott PJ (2002) Comprehensive two-dimensional gas chromatography with fast enantioseparation. Anal Chem 74:5426–5430
79. Soderstrom B, Frisvad JC (1984) Separation of closely related asymmetrc penicillia by pyrolysis gas chromatography and mycotoxin production. Mycologia 76:408–419
80. Soga T, Ueno Y, Naraoka H, Ohashi Y, Tomita M, Nishioka T (2002) Simultaneous determination of anionic intermediates for *Bacillus subtilis* metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. Anal Chem 74:2233–2239
81. Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M, Nishioka T (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. J Proteome Res 2:488–494
82. Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J Am Soc Mass Spectrom 10:770–781
83. Stephanopoulos GN, Aristidou AA, Nielsen J (1998) Metabolic engineering, principles and methodologies. Academic, San Diego
84. Streikov S, Elstermann M von, Schomburg D (2004) Comprehensive analysis of metabolites in *Corynebacterium glutamicum* by gas chromatography/mass spectrometry. Biol Chem 385:853–861
85. Sweetlove LJ, Last RL, Fernie AR (2003) Predictive metabolic engineering: a goal for systems biology. Plant Physiol 132:420–425
86. Taylor J, King RD, Altmann T, Fiehn O (2002) Application of metabolomics to plant phenotype discrimination using statistics and machine learning. Bioinformatics 18 [Suppl 2]:S241–S248
87. Terabe S, Markuszewksi MJ, Inoue N, Otsuka K, Nishioka T (2001) Capillary electrophoretic techniques toward the metabolome analysis. Pure Appl Chem 73:1563–1572
88. Kuile BH ter, Westerhoff HV (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. FEBS Lett 500:169–171
89. Theobald U, Mailinger W, Reuss M, Rizzi M (1993) In vivo analysis of glucose-induced fast changes in yeast adenine nucleotide pool applying a rapid sampling technique. Anal Biochem 214:31–37
90. Tolstikov VV, 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
91. Tolstikov VV, Lommen A, Nakanishi K, Tanaka N, Fiehn O (2003) Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. Anal Chem 75:6737–6740
92. Tweeddale H, Notley-McRobb L, Ferenci T (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool (Metabolome) analysis. J Bacteriol 180:5109–5116
93. Tweeddale H, Notley-McRobb L, Ferenci T (1999) Assessing the effect of reactive oxygen species on *Escherichia coli* using a metabolome approach. Redox Rep 4:237–241
94. Uribelarrea JL, Pacaud S, Goma G (1985) New method for measuring the cell water content by thermogravimetry. Biotechnol Lett 7:75–80
95. Vaidyanathan S, Goodacre R (2003) Metabolome and proteome profiling for microbial characterization. In: Harrigan GG, Goodacre R (eds) Metabolic profiling: its role in biomarker discovery and gene function analysis. Kluwer, Boston, pp 9–38
96. Dam JC van, Eman MR, Frank J, Lange HC, Dedem GWK, Heijnen SJ (2002) Analysis of glycolytic intermediates in *Saccharomyces cerevisiae* using anion exchange chromatography and electrospray ionization with tandem mass spectrometric detection. Anal Chim Acta 460:209–218

97. Greef J van der, Davidov E, Verheij E, Vogels J, Heijden R van der, Adourian AS, Oresic M, Marple EW, Naylor S (2003) The role of metabolomics in systems biology. In: Harrigan GG, Goodacre R (eds) Metabolic profiling: its role in biomarker discovery and gene function analysis. Kluwer, Boston, pp 171–198
98. Werf MJ van der (2005) Towards replacing closed with open target selection approaches. Trends Biotechnol 23:11–16
99. Vicente MF, Basilio A, Cabello A, Pelaez F (2002) Microbial natural products as a source of antifungals. Clin Microbiol Infect 9:15–32
100. Vogt AM, Ackermann C, Noe T, Jensen D, Kubler W (1998) Simultaneous detection of high energy phosphates and metabolites of glycolysis and the Krebs cycle by HPLC. Biochem Biophys Res Commun 248:527–532
101. Wahl HG, Hoffmann A, Luft D, Liebich HM (1999) Analysis of volatile organic compounds in human urine by headspace gas chromatography-mass spectrometry with a multipurpose sampler. J Chromatogr A 847:117–125
102. Walsh K, Koshland DE (1984) Determination of flux through the branch point of two metabolic cycles. The tricarboxylic acid cycle and the glyoxylate shunt. J Biol Chem 259:9646–9654
103. Weuster-Botz D, Graaf AA de (1996) Reaction engineering methods to study intracellular metabolite concentrations. Adv Biochem Eng 54:75–108
104. Wilkinson SR, Young M, Goodacre R, Morris JG, Farrow JAE, Collins MD (1995) Phenotypic and genotypic differences between certain strains of *Clostridium acetobutylicum*. FEMS Microbiol Lett 125:199–204
105. Wittmann C, Heinzle E (2002) Genealogy profiling through strain improvement by suing metabolic network analysis: metabolic flux genealogy of several generations of lysine-producing corynebacteria. Appl Environ Microbiol 68:5843–5859
106. Yang YH, Speed T (2002) Design issues for cDNA microarray experiments. Nat Genet 3:579–588