

# Use of metabolomics to discover metabolic patterns associated with human diseases

Oliver Fiehn\* and Joachim Spranger#

\*Max-Planck-Institute of Molecular Plant Physiology, 14424 Potsdam/Golm, Germany

#German Institute of Human Nutrition (DIfE), 14558 Bergholz-Rehbrücke, Germany

key words: metabolomics, diabetes mellitus, mass spectrometry, pattern recognition

## Abstract

Metabolomic techniques aim at detecting unexpected effects comparing stressed/unstressed or mutant/wild type experiments. This chapter asks the question if metabolomics could also go one step further for analyzing changes in metabolic patterns that are associated with the time course of nutritional-dependent human diseases. Such diseases are typically hard to predict, and existing biological markers such as for type 2 diabetes mellitus have limited value for the assessment of individual risks. Many factors may be involved in disease progression such as genetics, nutritional habits, age, or sex, resulting in the need to study large cohorts in order to draw statistically sound conclusions. Due to this inherent biological variability, severe constraints are posed on the validation of the analytical methods used. Using diabetes as an example, the economic and scientific needs for accurate diagnostic tools are discussed with respect to the available analytical and computational approaches for cost-effective high throughput methods.

## Introduction

Functional genomics has been in the focus since recent years for unraveling the structure and function of complex biological mechanisms. The analysis of primary gene products has further been considered as diagnostic and screening tool for disease recognition. Such strategies aim at investigating all gene products simultaneously in order to get a better overview about disease mechanisms and to find suitable therapeutic targets. There are several test cases demonstrating the validity of such approaches. However, there are also severe practical and theoretical constraints known if applying mRNA or protein profiling as universal tool for improved understanding and diagnostics of disease patterns. The paradigm of linear control from gene expression over transcription and translation to metabolic phenotypes has been challenged by multiple experimental observations, such as missing

correlations between mRNA and protein abundances, changes in RNA or protein turnover rates, or the complexity of protein interaction networks. Next, both transcript arrays and proteomics come along with high costs per sample, limiting the number of analyzed biological replicates to a point at which solid statistical evaluations cannot be carried out. It has been estimated that even for isogenic mouse lines, biological variability requires at least 15 replicates per genotype tested. Only in rare cases allow research budgets such high replicate numbers. Researchers try to circumvent the likelihood of false positive findings by setting high thresholds for differential gene expression to 5- or 10-folds. By doing so, the balance bounces back to increased probabilities of false negatives, i.e. genes that were in fact under- or overexpressed at 50% or even at 2- to 3-fold levels. At this point, low cost metabolomic assessments may be considered. Metabolism may be regarded as result of all regulatory steps that respond to the onset or progression of diseases. Small differences in metabolic flux rates may result in amplification of metabolite contents through the pathway network. Although it is disease mechanisms or gene functions are hard to be pinned down by metabolite profiles alone, this amplification effect may be used for screening and diagnostic purposes. Further, use of metabolomics as well as the related metabolic fingerprinting approaches causes costs two to three orders of magnitude lower than transcriptomics or proteomics. This chapter will therefore focus on potential implications of metabolomics as a tool to identify novel metabolic patterns or markers associated with disease status. We will exemplify the potential of this method using the association between specific fats and development of type 2 diabetes as a test case.

### **Type 2 diabetes mellitus has growing impact on human health**

Worldwide, type 2 diabetes mellitus (T2DM) is among the major diseases that is caused by human malnutrition (Warram et al. 1997). High socio-economic burdens arise from T2DM and other nutritional diseases, although disease manifestation could be prevented by timely intervention therapies. Unfortunately, risk assessments in healthy humans are rare, mostly due to lack of adequate assays that had high enough prediction power. Genetic as well as lifestyle factors belong to the major causes of diabetes development (Hu et al. 2001). If diet or alteration in life-style such as increased physical activity is used in intervention therapy, more than half of all T2DM manifestations can be avoided (Tuomilehto et al. 2001, Knowler et al. 2002). Cross-sectional as well as therapeutic studies have demonstrated the efficacy of such therapies to prevent the onset of T2DM in individuals with impaired glucose tolerance.

Therefore, it is crucial to develop accurate tools to predict the risk of individuals to develop T2DM. There are several well-known predictors of T2DM such as biomarkers (cholesterol, triglycerides, LDL and HDL, and lipoproteins) and anthropometric parameters (body-mass index, waist-hip ratio). However, the performance of these risk factors is rather poor. Only 10% of the individuals that later develop T2DM can be identified if these parameters are taken together. Correspondingly, 90% of the people who are classified at risk by the known biomarkers will not develop diabetes (Report of the Expert Committee 1997). Similar results have been found for other metabolism-related diseases, for which biomarkers regularly fail to correctly predict risks causing high numbers of both false positive and false negative results. Therefore, additional diagnostic strategies have utmost clinical importance. It makes perfect sense to use metabolomics for diagnosing metabolic diseases, for which even now metabolites are among the best risk predictors.

The pathogenesis of many complex diseases is considerably influenced by environmental factors as has been shown for cancer, coronary heart disease and T2DM. The prevalence of T2DM is rapidly increasing with the adoption of industrial (western-type) lifestyle.

Accordingly, the lowest numbers of T2DM are found in rural areas where people still follow their traditional lifestyles (Amos et al. 1997). Especially traditional indigenous communities show a large increase in the number of type 2 diabetes manifestations if they change to typical Western diets, e.g., Pima Indians in Arizona, Micronesians in Nauru and Aborigines in Australia (Bennett 1999).

In the year 2025, 300 million adults will suffer from diabetes worldwide, with a prevalence of T2DM of 5.4%. At this time, the majority of diabetes cases will be observed in developing countries, with India and China having more cases than any other country in the world (King et al. 1995). Due to this dramatic raise in type 2 diabetes, these countries will face severe socio-economic consequences, also caused by diabetic complications such as end-stage renal insufficiency and heart disease.

Diet and nutrition are broadly accepted to belong to the major factors in the onset and pathogenesis of type 2 diabetes mellitus. However, specific dietary factors have not been satisfactorily clarified so far. Although it is known by numerous epidemiological studies that a diet rich in vegetables and fruits has significant protective effects (Hu et al. 2001), the exact role of fat and carbohydrates is hotly debated with respect to diabetic risks. Little is known about the potential benefits of other phytochemicals such as phenolic antioxidants. Regularly, a low-fat and high-carbohydrate diet is recommended for prevention of major nutritional diseases such as diabetes, coronary heart attack and other chronic diseases. However, 'fat' and

'carbohydrates' are only weakly defined descriptions of inhomogeneous compound classes. It is getting more and more clear, that different types of fats and carbohydrates have different effects on insulin sensitivity, insulin secretion, and glucose homeostasis. Free fatty acids are long known to be associated with the development of insulin resistance by competing with glucose oxidation (Randle hypothesis, Randle et al. 1965), followed by extensive research in the past decades (Boden 1997). For saturated as well as monounsaturated and polyunsaturated fatty acids (with the exception of n-3 fat) a causal relationship to insulin resistance has been shown in animal models (Storlien et al. 1996). More specifically, high saturated fat intakes are associated with increased risk of glucose intolerance, elevated fasting glucose and a booth in insulin concentrations in epidemiological studies (Hu et al. 2001b). Moreover, increases in fasting insulin levels, lowered insulin sensitivity and higher risk of T2DM manifestation has also been shown to be associated with higher proportions of saturated fatty acids in serum lipids and muscle phospholipids (Vessby et al. 1994, Folsom et al. 1996). In humans, higher proportions of long-chain polyunsaturated fatty acids in skeletal muscle phospholipids have been related to better insulin sensitivity in humans (Borkmann et al. 1993). The role of monounsaturated fatty acids is still not clear, with some reports indicating a high amount of monounsaturated fatty acids to be harmful (Feskens et al. 1995). Correspondingly, insulin sensitivity in healthy individuals was shown to be improved during a three-month dietary intervention study in which monounsaturated fatty acids were substituted. Further, intake and composition of fats both contributed to the beneficial effect: improved insulin sensitivity by substitution of monounsaturated fatty acids was only found for subjects who were consuming less than 37% of energy as fat (Vessby et al. 2001). However, there is still too little data to give accurate recommendations for fat compositions, since also the amount and composition of carbohydrates, polyphenolic antioxidants, micronutrients and physical activity largely effect the pathogenesis of complex diseases such as type 2 diabetes. Accordingly, nutritional research should focus at the contribution and relative importance of different dietary and lifestyle factors simultaneously, in order to overemphasize the role of any single nutrient. It may well turn out that the onset of type 2 diabetes follows more a pattern of dietary components in relation to other factors such as health status, physical activity, age and sex. For such broad overviews of metabolic health status, profiling techniques could be very helpful for screening metabolic changes in response to specific diets.

It is obvious that intervention was highly advantageous and much more successful if it started prior to diabetes manifestation. But how can we exactly predict the risk of healthy persons to

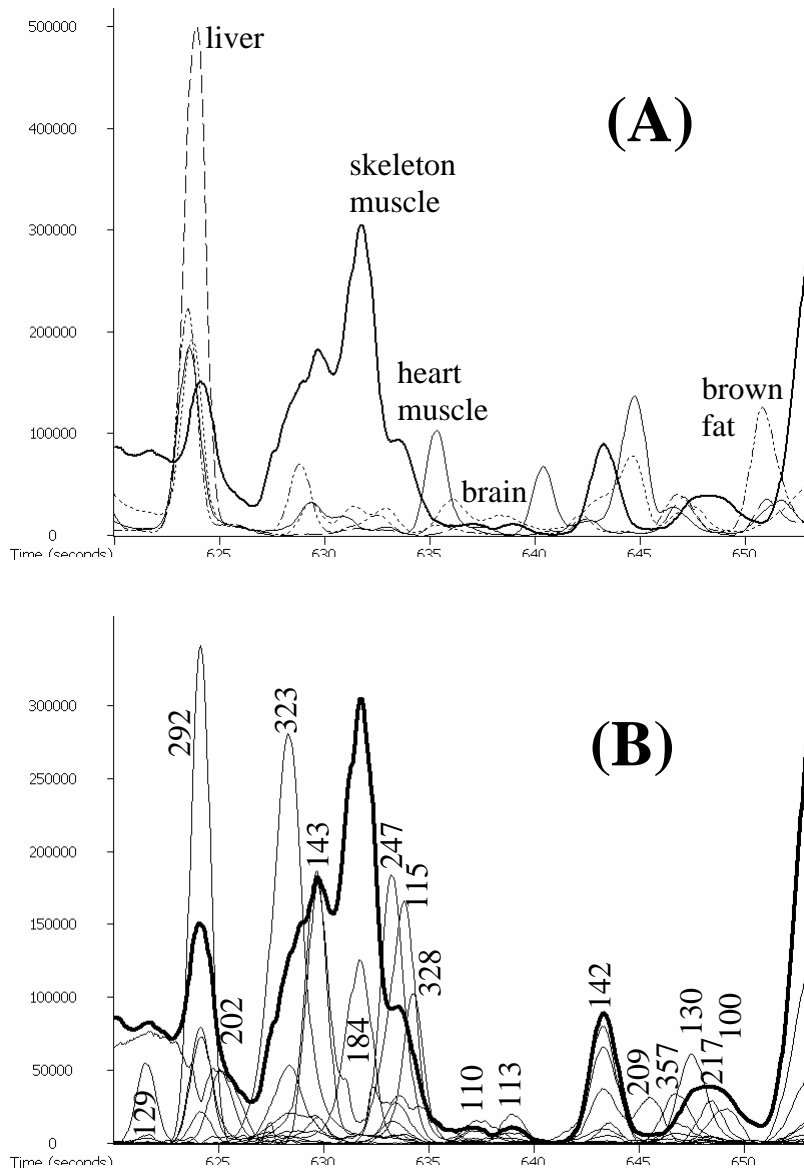


Fig. 1. Comparison of tissue extracts of isogenic mice lines by GC-TOF analysis (unpublished results). **A:** Metabolic profiles of liver, brain, fat, and muscle extracts shown for a 35 s window. Each tissue gave a very specific metabolic profile, indicating the ability to have an unbiased look on metabolism. **B:** For the example of skeleton muscle, metabolites that were only partly resolved by chromatography alone could clearly be distinguished after mass spectral deconvolution and automatic quantitation using optimal ion traces (given in mass-to-charge numbers).

Fig 1

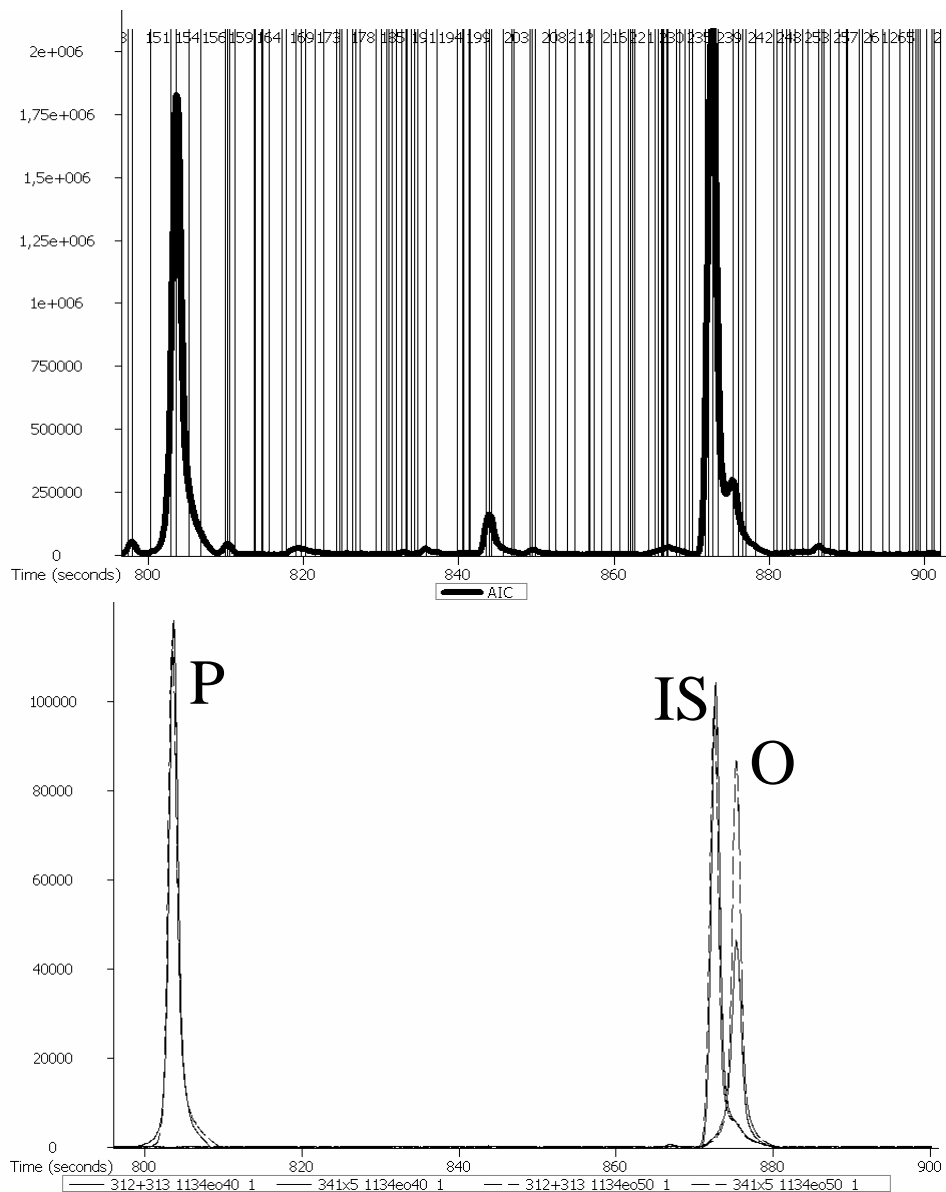


Fig. 2. GC-TOF chromatograms of liver extracts of control and knock out mice lines. Upper panel: automatic peak detection of hundreds of metabolites as indicated by vertical lines, here shown for a 100 s retention time window. Lower panel: Quantification for known metabolites using unique ion traces. Dotted line: knock out mouse line, solid line: control mouse line. P = palmitic acid, O = oleici acidi, IS = internal standard.

Fig 2

develop type 2 diabetes? Screening programs relying on known risk factors result in a high number of individuals falsely identified to be at high risk. Only 2.5% of all subjects screened in a representative population were correctly predicted. Correspondingly, such programs are highly expensive with cost estimations ranging between 4000-8000 US dollars per new case of diabetes identified. If screening programs were more successful and cheaper, high socio-economic burdens caused by type 2 diabetes and its related metabolic diseases could be prevented: for 1997, the American Diabetes Association calculated that diabetes caused 27 billion US dollars in direct medical costs and 32 billion US dollars in lost-productivity costs. Several reports have been published demonstrating that aggressive anti-diabetic therapies in high-risk cohorts can delay or avoid complications. Such pre-manifestation interventions can reduce indirect health-related costs of up to 9000 US dollars per person per year. Rapid and unbiased techniques screening for metabolic changes could thus improve the correct identification of high-risk persons and lower direct and indirect diabetes related costs. Moreover, once metabolic patterns can efficiently characterize diabetes risks, costs and efficacies could be compared to target analyses that are based on just a few metabolite biomarkers. Last, analyzing the metabolome in an unbiased way might also be advantageous to study molecular mechanisms and pathogenesis of type 2 diabetes mellitus.

### **Profiling techniques**

What are the most suitable techniques to unravel the complex association between disease phenotypes with respect to the underlying high biological variation in typical cohorts? RNA microarrays used more and more often to find differences in complex phenotypes by clustering methods. Specific RNA clusters were found associated with progression of breast cancer by van't Veer et al. (2002), enabling risk predictions of individuals for specific disease endpoints. However, transcriptomics and proteomics are still too costly to be used for screening healthy populations. Especially for metabolic diseases, metabolomic analyses might become the preferred tool for diagnostics and prediction of complex diseases. In metabolic diseases as well as phenotypes like obesity, metabolites are key to understanding the role in pathogenesis. Recently, a new lipid has been reported to be anorexigenic which was depending on feeding patterns (Rodriguez et al. 2001). Profiling of metabolic diseases is not a novel technique to clinicians (Tanaka et al. 1980a,b). Since decades, inborn errors are rapidly characterized in screening of newborns as well as other metabolic diseases (Jellum et al. 1988). Despite successful application in acidurias (Kimura et al. 1999, Halket 1999), cervical cancer (Kim et al. 1998), or mitochondrial myopathy (Ning et al. 1996), comprehensive analysis of metabolite patterns is still

not a routine diagnostic tool. Instead, techniques have been diversified in the realm of functional genomics, plant studies and bacterial characterizations. Metabolite analysis can thus broadly be classified into four different directions:

- (a) Classical hypothesis-driven research is still using *metabolite target analysis*. Especially if isotope labeled reference compounds are applied, quantification and identification of biomarkers are unrivalled. However, interpretations are prone to over-simplify research problems since causal effects can hardly be distinguished from simple associations or detection of side-effect correlations.
- (b) This view led to the idea to look for biochemical reactions in a broader way. By *metabolite profiling*, select pathways or compound classes are studied simultaneously by defining dozens to hundreds of target molecules. Analyte identification is usually assured by a combination of compound spectra and chromatographic retention indices, whereas accurate quantification is compromised by the large number of target compounds.
- (c) Both target analysis and metabolite profiling can only find pre-defined compounds. Yet, complexity and flexibility of biochemical networks allow unexpected metabolic responses upon external factors or disease development. Therefore, *metabolomic analysis* tries to detect and quantitate all metabolic signals in truly unbiased ways, irrespective of size, volatility, or other physicochemical parameters. It is thus an extension of metabolite profiling in the sense, that both known and unknown metabolites may be equally important. Metabolomics is challenged by huge differences in molecule structures and abundances, and it poses high demands on software capabilities of the analytical instruments as well as on database structure and implementation. Many of the signals will remain unknown in the first instance, but can be later identified once a signal is demonstrated to be important in biological or clinical studies using relative quantitation.
- (d) If it is aimed at classification of samples rather than comparing detailed biochemical pathways, *metabolic fingerprinting* may be advantageous. Fingerprinting refuses to distinguish all metabolites. Instead, physical spectra of the crude mixtures are acquired without use of chromatography in order to analyze multiple samples in a short time. All spectra are then computed by supervised or unsupervised learning methods in order to

cluster samples according to origin. In some instances, metabolite abundance or spectra characteristics may be sufficient to be evolved directly from the crude spectra.

Metabolite fingerprinting has utilized various physical techniques and was applied to a variety of different research fields. Most often, proton nuclear magnetic resonance is used (<sup>1</sup>H-NMR, (Gavaghan et al. 2000, Raamsdonk et al. 2001), but even fourier transformed infrared spectroscopy (FT-IR, Goodacre et al. 2000) gave sufficient information to distinguish salt stressed tomato fruits from unstressed ones, and to derive information from that spectra to find potential biomarkers. Even more promising might be direct infusion mass spectrometry (FIA-MS). Efficacy of compound ionization and adduct formation in liquid mass spectrometry (and other techniques such as matrix assisted laser desorption/ionization) inevitably depends on matrix effects within the crude mixture itself. Subtle differences in total composition might directly lead to improved or reduced ionization of sample components that therefore give altered intensities in the corresponding mass spectra. Whereas any effort to derive information upon changes in biochemical pathways must remain very doubtful, even if high resolution fourier transform mass spectrometry is used, this amplification of matrix differences may be advantageous for classification purposes. In a recent paper, Vaidyanathan et al. (2001) could show that FIA-MS spectra were sufficient to rapidly detect and discriminate between bacteria strains at competitive costs and high speed. Whereas this application aimed at consumer safety, it can easily be imagined that also other fields of applications could benefit from this approach, for example for screening large cohorts of healthy individuals to get a primary measure about metabolic distances and risk stratifications. Samples of individuals identified at risk could then get analyzed a second time by more time consuming metabolomic techniques in order to strengthen and verify the classification of the first screen, and to pin down changes in particular biochemical pathways. As defined above, metabolomics aims at the simultaneous detection and quantitation of all individual metabolites in a particular biological tissue, as further explained in recent reviews (Fell 2001, Fiehn 2002). To date, metabolomic approaches are exclusively using mass spectrometry to its inherent selectivity, universality, and sensitivity. The largest problem so far is the unambiguous deconvolution of multiple overlapping peaks in multiple samples. Therefore, even the most mature combination of gas chromatography/mass spectrometry (GC/MS) has mostly been used as metabolite profiling technique (Fiehn et al. 2000a). Using improved deconvolution algorithms and faster spectral acquisition by time-of-flight measurements (TOF, Shellie et al. 2001), however, resulted in the detection of more than 1,000 components from 15 mg FW plant leaf extracts and a throughput of more than 1,000 samples per month (Weckwerth et al. 2001). In a test case for over 400 metabolites analyzed by GC/TOF,

average analytical errors were below 10% relative standard deviations even if low abundant compounds were included (which accounted for the largest deviations, Fiehn 2002, unpublished results). With this data, GC/TOF proves to be the 'gold standard' for high throughput metabolomics. However, any GC-based analytical instrument will suffer from its inherent bias against large and thermolabile compounds. Specifically, if it is aimed to elucidate diabetic biomarkers and mechanisms in truly unbiased ways, methods based on liquid chromatography cannot be avoided in order to quantify the amount of neutral and polar lipids, glycosylated or glucuronidated compounds, antioxidants or co-factors. Based on reverse phase chromatography, numerous methods exist to couple LC to MS and tandem MS, especially for unpolar metabolites such as flavonoids or phenolics (Justesen et al. 1998). Only recently was LC/tandem MS extended to highly polar and larger metabolites by applying a variant of normal phases, hydrophilic interaction chromatography (Tolstikov and Fiehn, 2002). However, even most recent software releases of renowned LC/MS manufacturers lack the ability to reliably find and quantitate metabolites in multiple chromatographic runs by mass spectral deconvolution, which is today standard in GC-based methods (Stein 1999). Today, this hampers direct application of LC/MS methods for metabolomic screening studies: it is not enough to detect 1,000 peaks, but the ability to detect and distinguish these 1,000 peaks from novel peaks in multiple runs must be assured with high precision.

As alternative to LC/MS, LC/coulometry detection has been applied to disease recognition (Vigneau-Callahan et al. 2001). Coulometry enables detection of all compounds that have oxidizable or reducible moieties in their molecular structures. Instead of sweeping the applied voltages across redox potential, an array of electrodes with fixed potentials can be used. A 16-channel LC/coulometry array run was sufficient to detect over 1,000 peaks in human plasma, demonstrating that in principle this technology is ripe to be used as complement to mass spectrometric metabolomic approaches (Vigneau-Callahan et al. 2001). However, as the authors admit, peak finding routines were barely adequate for routine operation since only 25% of all peaks were found in more than 5 of 8 comparisons of chromatographic runs from sample pools. No further metabolomic strategy has not been developed so far that aims at recognition of disease progression, effect of therapies or drug target discovery. However, such approaches have been discussed in a few recent conferences, and the number of publications and patents may broaden in the next few years.

By any mass spectrometric, UV, fluorescence, or coulometry-based metabolomic technique, a large proportion of detected signals will not be assigned by exact chemical structure. In part this

is the case because no comprehensive metabolite spectral libraries exist in the public domain. Furthermore, the complexity of plant and animal biochemistry is still underestimated due to decades of hypothesis-driven approaches. However, in all reports published so far, unknown metabolites play an important role in classifying samples and sometimes are in the center of biochemical correlation networks. Therefore, de novo identification of unknown metabolites is a necessity in truly metabolomic approaches. In GC/MS, identification is often hampered by the lack of abundant molecular ions. This may be circumvented by using semi-exact masses and isotope ratios after modified derivatization schemes in order to calculate elemental compositions (Fiehn et al. 2000b). In LC/MS, identification is less error prone and easier to achieve. Starting from chromatographically separated molecular ions, ion trap mass spectrometry can be used to elucidate fragmentation pathways (Drexler et al. 1998) which are aided by quadrupole-time of flight hybrid instruments that deliver exact masses (Blom 2001) as input in mass spectral interpretation software. Whereas in many cases suitable hits can be produced if chemical and biochemical databases are interrogated such as Beilstein or Chemical Abstracts, for many metabolites even this information would not be sufficient. Metabolites have large degrees of freedom not only by the position and sequence of atoms, but also to form isobaric and isomeric enantiomers and diastereomers. This level of detail eventually requires use of one or two-dimensional NMR as has been successfully demonstrated for phytochemicals (Pauli 2000).

### **Method validation for disease recognition**

Metabolomic as well as fingerprinting methods have great potential in health-related research fields to serve for rapid, reliable, sensitive, and cost-effective method to disease risk stratifications. If intended to be applied to large cohorts or as routine clinical instrument, however, the need of method validation cannot be underestimated. Specifically, all steps from sampling human or animal tissues or plasma samples to final data acquisition need to be carefully investigated to reduce additional sources of error that add up to the large variability in mammalian samples that are caused by dietary, genetic, developmental or behavioural factors. This so-called biological variability founds the need to collect data from a large collection of metabolic snapshots, as has been pointed out elaborately in preceding reviews (Kell and Mendes 2000). Consequently, the large number of samples to be analyzed again call for highly robust and automated throughput analytical technologies that include automated peak annotation, data export and database setup prior to statistical evaluations. Such rigid validations must be carried out even in academic laboratories (Krull and Swartz 1999) if results are expected to be reaching beyond low-impact research journals. For disease recognition in screening populations,

ruggedness and repeatability have clear priority over other aspects of validation, i.e. universality, sensitivity, selectivity or comprehensiveness. Ruggedness needs to be tested by as many alterations to existing protocols as possible, since it cannot be assumed that clinical or laboratory co-workers would strictly follow protocol details in all cases. Steps that could lead to gross errors should be pinned down and, eventually, be automated if possible. For example, solvent composition as well as extraction times and temperatures have huge impact on the peak intensities in metabolite profiles. Therefore, validation studies focusing on method robustness must include both small and large variation of protocols, especially if protocols are adopted from other biological disciplines such as plant biology or cell cultures. For example, within-run and between-run repeatability must be tested by comparing aliquots of pooled master samples, including multiple injections and precision tests under varying instrument conditions or variation of derivatization conditions. The methods should be critically assessed if they are able to distinguish between isomers, i.e. of unsaturated fatty acids or monosaccharides. Even if method validation is only aiming at relative quantitation or fingerprinting, sensitivity and selectivity should be assessed (and documented) at least once by spiking internal references, i.e. stable isotope labeled standard compounds. Last, it is of utmost importance of high throughput operations that rigid quality control routines are implemented, including the establishment of quality control charts. Intervention and abort criteria must be developed at all steps of method control, specifically at the level of tolerable and intolerable result output of standard samples and instrumental conditions.

Once these basic steps have been implemented, clinical proof-of-concept studies with known outcome may be carried out, for example diseased-healthy comparisons. If successful, the developed routine operations may then be used for defining biological background variability under a variety of nutritional regimes, in order to be able to accurately distinguish between disease-related metabolite patterns that are associated with pathogenesis or endpoint of diseases like type 2 diabetes mellitus, and metabolic outliers of healthy subjects found at special circumstances. A project like the one outlined above may then generate large databases to apply appropriate tools of multivariate statistics or classification algorithms, aiming at defining certain patterns or single biomarker molecules that are associated with the disease or a certain clinical treatment.

It is not decided, which instrumental technique may be most appropriate for such an approach. It can be assumed that two-dimensional metabolomic techniques have larger problems in instrument quality control, for example for corrections against chromatographic retention time shifts. On the other hand, data variation is also known for metabolic fingerprinting, e.g.

shimming effects in NMR analysis. Moreover, problems associated with instrument inlet parameters (like injection inlets in GC, or electrospray interfaces and skimmers in LC/MS) could be more easily found and eliminated if quality control is based on individual metabolites rather than overall spectra of crude mixtures.

In a test case demonstrating the principle ability to adopt sample preparation protocols derived from plant biology, the authors performed extraction, sample preparation and analysis of different tissues of knock out and control lines of syngenic mice. Each tissue gave a very specific metabolic profile, indicating the ability to have an unbiased look on metabolism (fig.1A). Even for regions in which too many metabolites were observed for clear chromatographic separations, mass spectral deconvolution resulted in automatically purified mass spectra and annotation of so-called unique ions that are best suited for relative quantification. An example is given for the complex mixture of the skeleton muscle tissue extract (fig. 1B). Comparing the metabolite levels from multiple samples is facilitated by this approach. In figure 2, an example is given for the automatic peak finding in liver tissue extracts and the relative quantification by determining peak areas for all peaks using the model ion traces. In this example, the knock out mouse line showed a clear up-regulation of oleic acid after normalization to internal standards, whereas other fatty acids such as palmitic acid remained unaffected. In high throughput applications, all analytical results are then exported to databases for subsequent statistics and data mining.

### **Data mining in metabolic data bases**

Any rigid assessment of disease-related metabolic patterns will need support from well documented and curated data bases. Metabolomics and fingerprinting methods are always prone to data over-fitting, especially if inappropriate classification tools are used, or if success rates and background knowledge can not be adequately compared with test samples. Although the number of variables coming from a single metabolomic experiment may be an order of magnitude smaller than compared to transcript arrays, cost effectiveness allows two order of magnitude more samples to be analysed for a specific experiment. Correspondingly, large piles of clinical and analytical data are expected to enter metabolic data bases. The pure analytical result files must be accompanied by standardized biological, clinical, and method-related background information (meta data) in order to ensure the ability to extract and mine health-related information. The data model must refer to agreed nomenclatures and ontology, and it must allow flexibility in experimental designs as well preserving a rigid and searchable data base structure. It needs to support general concepts such as organ specimen, sexes, nutritional regimes and others,

and it must distinguish raw data from derived data, for example by different types of data normalization. Most importantly in human studies, data safety upon individuals must be guaranteed at all times. Logical correctness and minimization of duplicate entries has to be assured as well as system usability, e.g. by web interfaces for cooperating partners. In this respect, a typical expression data base does not necessarily need to include data mining tools. Rather it should aim at easiness of data downloads and re-formatting into different types of analysis, taking into account differences in dots and commas between European and American partners as well as spreadsheet formats for MS office packages which are most often used by biologists. Such simple problems often hamper existing collaborations and could easily be avoided to enable clinicians or biologists to select data according to their needs and wishes, e.g. by disease type, sex, biological matrix or sampling date.

Next, it might be asked what data mining tool should be used? Although the obvious answer is that there is simply no best algorithm, some guidelines should be followed. First, underlying assumptions have to be clarified, most importantly about the independence of samples, but also about frequency distributions or data structures like canonical or linear data, connected or independent data. It is highly advisable to have statistics expert aboard large-scale health related projects, since experimental designs may support or nullify to draw valid conclusions that reach farther than 'just the next publication'. Researchers should also follow the general rule of statistics, that more replicates and repeats are better than relying on the absolute minimum number of independent samples, specifically, if guided by the limits of analytical precision determined in the method validation part. Next, it is generally accepted that overall results are more trustworthy that are gained by use of different and complementary clustering methods. Any commercial package such as Pirouette, MatLab, SAS, or SPSS could be used, and again, great care should be devoted not to violate the assumptions underlying the different statistical tools. If only classification is asked, supervised learning methods such as discriminant functions analysis, support vector machines, artificial neural networks, decision trees, or evolutionary computing might be most appropriate (Gilbert et al. 2000, Johnson et al. 2000). Several of these techniques should be applied to a certain data set to ensure classification results, taking into consideration that some of the algorithms will allow to follow the path that led to the decision (like evolutionary algorithms) whereas others will only focus on optimal separation (like neural networks). Although it may be appropriate to use cross-validation in classification schemes, it is generally advisable to keep the training data set completely separate from the test data set in order to avoid biased results. With the high number of replicates that can be run in metabolomic or fingerprinting approaches, this recommendation should not pose too huge problems to

practical work. In certain cases, unsupervised learning methods such as hierarchical clustering or principle component analysis may be appropriate, for example if distance measures are to be applied on various classes, and specifically, if the number of expected classes is not known *a priori*. Supervised and unsupervised methods may also be combined, as has been shown for discrimination of silent mutations from wild type strains (Raamsdonk et al. 2001). Once classes or sub-populations have unambiguously been defined, more sophisticated methods may be applied such as comprehensive analysis of co-variance matrices in large metabolic networks (Kose et al. 2001), or time-dependent patterns (Lukashin et al. 2001) that might be most appropriate for analyzing the onset and progression of diseases, and eventual therapeutic success.

## Conclusions

Unbiased detection of unexpected events in biological or clinical studies have made the first steps. A combination of metabolomic and metabolic fingerprinting methods may be appropriate tools if classification and subsequent detailed biochemical analysis is required. For large scale disease-related projects such as diagnosis and prediction of type 2 diabetes mellitus, rigid method validation followed by assessments of 'metabolic baselines' of healthy cohorts is a prerequisite that needs to be complemented by large efforts on the level of experimental designs, data base setup, and appropriate data mining and interpretation schemes. A current bottleneck in metabolomics is observed in the huge diversity and complexity of human and animal metabolomes that goes far beyond classical knowledge found in textbooks or data bases. However, benefits of systematic evaluation of nutritional diseases may also include impact on other disciplines, such as the definition of 'healthy food' and the impact of the composition of putatively beneficial food ingredients such as antioxidants, flavonoids, anthocyanins, carotenoids, or vitamins, specifically in relation to the impact of fats, fibres, and carbohydrates.

## References

- Amos AF, McCarty DJ, Zimmet P. (1997) The rising global burden of diabetes and its complications, estimates and projections to the year 2010. *Diabet Med* 14, 1-85.
- Bennett PH. (1999) Type 2 diabetes among the Pima Indians of Arizona, an epidemic attributable to environmental change? *Nutr Rev* 57, 51-54.
- Blom KF. (2001) Estimating the precision of exact mass measurements on an orthogonal time-of-flight mass spectrometer. *Anal Chem* 73, 715-719.
- Boden G. (1997) Role of fatty acids in the pathogenesis of insulin resistance and NIDDM. *Diabetes* 46, 3-10.
- Borkman M, Storlien LH, Pan DA, Jenkins AB, Chisholm DJ, Campbell LV. (1993) The relation between insulin sensitivity and the fatty-acid composition of skeletal-muscle phospholipids. *N Engl J Med* 328, 238-244.
- Drexler DM, Tiller PR, Wilbert SM, Bramble FQ, Schwartz JC. (1998) Automated identification of isotopically labeled pesticides and metabolites by intelligent 'real time' LC-tandem MS using a bench-top ion trap mass spectrometer. *Rapid Commun Mass Spectrom* 12, 1501-1507.
- Fell DA. (2001) Beyond genomics. *Trends Gen* 17, 680-682.

- Feskens EJ, Virtanen SM, Rasanen L, Tuomilehto J, Stengard J, Pekkanen J, Nissinen A, Kromhout D. (1995) Dietary factors determining diabetes and impaired glucose tolerance A 20-year follow-up of the Finnish and Dutch cohorts of the Seven Countries Study. *Diabetes Care* 18, 1104-1112.
- Fiehn O. (2001) Combining genomics metabolome analysis and biochemical modeling to understand metabolic networks. *Comp Funct Genom* 2, 155-168.
- Fiehn O. (2002) Metabolomics- the link between genotypes and phenotypes. *Plant Mol Biol* 48, 155-171.
- Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. (2000a) Metabolite profiling for plant functional genomics. *Nat Biotechnol* 18, 1157-1161.
- 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.
- Folsom AR, Ma J, McGovern PG, Eckfeldt H. (1996) Relation between plasma phospholipid saturated fatty acids and hyperinsulinemia. *Metabolism* 45, 223-228.
- Gavaghan CL, Holmes E, Lenz E, Wilson ID, Nicholson JK. (2000) An NMR-based metabolomic approach to investigate the biochemical consequences of genetic strain differences, application to the C57BL10J and Alpk,ApfCD mouse. *FEBS Lett* 484, 169-174.
- Gilbert RJ, Rowland JJ, Kell DB. 2000 Genomic computing, explanatory modelling for functional genomics. In, *Proceedings of the genetic and evolutionary computation conference*. Whitley D Goldberg D and Cantú-Paz E (Eds) Morgan Kaufman San Francisco, pp 551-557.
- Goodacre R, Shann B, Gilbert RJ, Timmings EM, McGovern AC, Alsberg BK, Kell DB, Logan NA. (2000) Detection of the dipicolinic acid biomarker in *Bacillus* spores using curie-point pyrolysis mass spectrometry and fourier transform infrared spectroscopy. *Anal Chem* 72, 119-127.
- Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, Chalmers RA. (1999) Deconvolution gas chromatography mass spectrometry of urinary organic acids - Potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun Mass Spectrom* 13, 279-284.
- Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC. (2001) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N Engl J Med* 345,790-797.
- Hu FB, van Dam RM, Liu S. (2001) Diet and risk of Type II diabetes, the role of types of fat and carbohydrate. *Diabetologia* 44, 805-817.
- Jellum E, Kvittingen EA, Stokke O. (1988) Mass spectrometry in diagnosis of metabolic disorders. *Biomed Environ Mass Spectrom* 16, 57-62
- Johnson HE, Gilbert RJ, Winson MK, Goodacre R, Smith AR, Rowland JJ, Hall MA, Kell DB. (2000) Explanatory analysis of the metabolome using genetic programming of simple interpretable rules. *Genet Program Evol Mach* 1, 243-258.
- Justesen K, Knuthsen P, Leth T. (1998) Quantitative analysis of flavonols, flavone and flavanones in fruits, vegetables and beverages by high-performance liquid chromatography with photo-diode array and mass spectrometric detection. *J Chromatogr A* 799, 101-110.
- Kell DB, Mendes P. (2000) Snapshots of systems. In, *Technological and medical implications of metabolic control analysis*. Eds., Cornish-Bowden AJ and Cárdenas ML (Kluwer Academic Publishers) pp 3-25.
- Kim K-R, Park H-G, Paik M-J, Ryu H-S, Oh KS, Myung S-W, Liebich HM. (1998) Gas chromatographic profiling of urinary organic acids from uterine myoma patients and cervical cancer patients. *J Chromatogr B* 712, 11-22.
- Kimura H, Yamamoto T, Seiji Y. (1999) Automated metabolic profiling and interpretation of GC/MS data for organic aciduria screening, a personal computer-based system. *Tohoku J Exp Med* 188, 317-344.
- King H, Aubert RE, Herman WH. (1998) Global burden of diabetes, 1995-2025, prevalence, numerical estimates, and projections. *Diabetes Care* 21, 1414-1431.
- Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM. (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 346, 393-403.
- Kose F, Weckwerth W, Linke T, Fiehn O. (2001) Visualising plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics* 17, 1198-1208.
- Krull IS, Swartz M. (1999) Analytical method development and validation for the academic researcher. *Anal Lett* 32, 1067-1080.
- Lukashin AV, Fuchs R. (2001) Analysis of temporal gene expression profiles, clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* 17, 405-414.
- Ning C, Kuhara T, Inoue Y, Zhang CH, Matsumoto M, Shinka T, Furumoto T, Yokota K, Matsumoto I. (1996) Gas chromatographic mass spectrometric metabolic profiling of patients with fatal infantile mitochondrial myopathy with de Toni-Fanconi-Debre syndrome. *Acta Paed Japon* 38, 661-666.
- Pauli GF. (2000) Higher order and substituent chemical shift effects in the proton NMR of glycosides. *J Nat Prod* 63, 834-838.

- Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol* 19, 45-50.
- Randle PJ, Garland PB, Newsholme EA, Hales CN. (1965) The glucose fatty acid cycle in obesity and maturity onset diabetes mellitus. *Ann N Y Acad Sci* 131, 324-333.
- Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus *Diabetes Care*. (1997); 20,1183-1197.
- de Fonseca FR, Navarro M, Gomez R, Escuredo L, Nava F, Fu J, Murillo-Rodriguez E, Giuffrida A, LoVerme J, Gaetani S, Kathuria S, Gall C, Piomelli D. (2001) An anorexic lipid mediator regulated by feeding *Nature* 414, 209-212.
- Shellie R, Marriot P, Morrison P. (2001) Concepts and preliminary observations on the triple dimensional analysis of complex volatile samples by using GC x GC – TOF MS. *Anal Chem* 73, 1336-1344.
- 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.
- Storlien LH, Baur LA, Kriketos AD, Pan DA, Cooney GJ, Jenkins AB, Calvert GD, Campbell LV. (1996) Dietary fats and insulin action. *Diabetologia* 39, 621-631.
- Tanaka K, Hine DG, West-Dull A, Lynn TB. (1980a) Gas-chromatographic method of analysis of urinary organic acids I Retention indices of 155 metabolically important compounds. *Clin. Chem.* 26, 1839-1846.
- Tanaka K, West-Dull A, Hine DG, Lynn TB, Lowe T. (1980b) Gas-chromatographic method of analysis of urinary organic acids II Description of the procedure and its application to diagnosis of patients with organic acidurias. *Clin Chem* 26, 1847-1853.
- 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.
- Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, Keinänen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M, Salminen V, Uusitupa M, Aunola S, Cepaitis Z, Moltchanov V, Hakumaki M, Mannelin M, Martikkala V, Sundvall J. (2001) Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 344,1343-1350.
- Vaidyanathan S, Rowland JJ, Kell DB, Goodacre R. (2001) Discrimination of aerobic endospore-forming bacteria via electrospray-ionization mass spectrometry of whole cell suspensions. *Anal Chem* 73, 4134-4144.
- van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-536.
- Vessby B, Aro A, Skarfors E, Berglund L, Salminen I, Lithell H. (1994) The risk to develop NIDDM is related to the fatty acid composition of the serum cholesterol esters. *Diabetes* 43, 1353-1357.
- Vessby B, Tengblad S, Lithell H. (1994) Insulin sensitivity is related to the fatty acid composition of serum lipids and skeletal muscle phospholipids in 70-year-old men. *Diabetologia* 37, 1044-1050.
- Vessby B, Unsitupa M, Hermansen K, Riccardi G, Rivellese AA, Tapsell LC, Nalsen C, Berglund L, Louheranta A, Rasmussen BM, Calvert GD, Maffetone A, Pedersen E, Gustafsson IB, Storlien LH (2001) Substituting dietary saturated for monounsaturated fat impairs insulin sensitivity in healthy men and women, The KANWU Study. *Diabetologia* 44, 312-319.
- Vigneau-Callahan KE, Shestopalov AI, Milbury PE, Matson WR, Kristal BS . (2001) Characterization of diet-dependent metabolic serotypes: Analytical and biological variability issues in rats. *J Nutr* 131, 924-932
- Warram JH, Kopczynski J, Janka HU, Krolewski AS. (1997) Epidemiology of non-insulin-dependent diabetes mellitus and its macrovascular complications. A basis for the development of cost- effective programs. *Endocrinol Metab Clin North Am* 26, 165-88.
- Weckwerth W, Tolstikov VV, Fiehn O. (2001) Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS analysis reveals silent metabolic phenotypes. *Proc 49th ASMS Conf Mass Spectrom All Top*, 1-2.